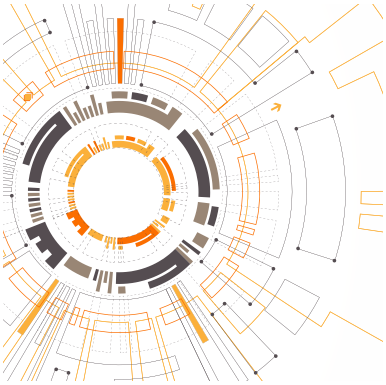

Abstract

Priming incoming and second-year college students with questions about effort prior to completing low-stakes assessments has increased test-taking effort. We extended this research by randomly assigning college seniors to one of three priming conditions prior to completing low-stakes assessments: answering three questions about intended effort that infused positive self-identity, answering three questions about intended effort that incorporated the university's creed, or answering no priming questions. The self-identity questions resulted in higher self-reported effort than the control condition, the university creed questions resulted in higher testing time than the control condition, and neither priming condition increased test performance. However, Pell Grant eligibility moderated the priming effect on effort. Priming resulted in self-reported effort for Pell eligible students being the same or higher than noneligible students. Likewise, ethnicity moderated the priming effect on test scores. White students scored higher than underrepresented students in the control condition, but this difference disappeared with priming.

AUTHORS

Sara J. Finney, Ph.D.
James Madison University

Stuart A. Miller, M.A.
Auburn University

Kendall M. McGoey, Ph.D.
Auburn University

Increasing Expended Effort on Low-Stakes Accountability Tests via Priming: Effectiveness with Graduating University Students

Test scores reported for accountability mandates are often gathered in low-stakes contexts (Cole & Osterlind, 2008; Mathers et al., 2018; Roohr et al., 2016; Smith & Smith, 2004; Wise & DeMars, 2010). For example, in higher education, outcomes assessments typically yield group-level data (incoming student achievement versus senior-level student achievement; students who have versus have not experienced educational programming). Accreditors use this group-level data to evaluate institutional impact on student learning (Liu, 2017). Federal funding is contingent on accreditation (Council for Higher Education Accreditation, 2022). Yet, performance on these assessments may have no personal consequences for students completing them.

CORRESPONDENCE
Email
finneysj@jmu.edu

Student Effort on Low-Stakes Tests

The validity of the interpretation of outcomes assessment data collected in high- versus low-stakes contexts may not be equal because of the discrepancy in expended effort (Simzar et al., 2015; Sundre & Kitsantas, 2004; Sungur, 2007; Wise & Smith, 2016). Students in a high-stakes testing context (e.g., certification testing) have something personal to gain or lose as a result of their performance. Thus, students tend to put forth enough effort to demonstrate their ability. In contrast, students assessed in a low-stakes context with no personal consequence of test performance may not expend the effort necessary to display their ability. Given the difference in expended effort across high- and low-stakes assessment contexts, a potential simple solution is to make institutional accountability tests high-stakes

in nature. However, institutional accountability assessments are often not tied to a single course but rather to a set of courses or experiences. In turn, performance on these assessments cannot inform course grades. Moreover, assessments used for institutional accountability or improvement may not have the psychometric properties necessary for individual-level high-stakes decisions, such as graduation. Thus, many institutional accountability assessments must retain their low-stakes nature to students.

In turn, validity concerns associated with expended effort must be acknowledged when reporting and interpreting test scores gathered in low-stakes contexts (Wise, 2020). A meta-analysis found a positive relation between students' expended effort and test scores in both K-12 and higher education contexts (Silm et al., 2020). The average correlation between self-reported expended effort and test performance was $r = 0.33$ and the average correlation between response time (a behavioral measure of expended effort) and performance was $r = 0.72$, indicating these two measures are representing different forms of effort. Further, the educational level of the student was a significant moderator of the effect sizes, with the average relation between effort and test performance being stronger for university students than for K-12 students.

Students assessed in a low-stakes context with no personal consequence of test performance may not expend the effort necessary to display their ability.

Moreover, on average, effort is generally lower for students later in their educational experience. This difference in effort was found in K-12 (5th graders compared to 8th graders) and college contexts (incoming college students compared to second-year college students) when expended effort was measured using self-report measures and response time (Rios & Guo, 2020; Soland, 2018; Thelk et al., 2009). In turn, there is greater need to increase effort for more advanced students. If more advanced college students expend less effort than incoming students on accountability tests, value-added estimates will be biased downward (Finney et al., 2016).

Given concern about the accuracy of inferences from low-stakes assessments, the purpose of our study was to examine the effectiveness of a short priming intervention to increase test-taking effort from graduating university students. Our study built upon previous studies of this intervention by examining the effectiveness of the intervention with senior-level students and adapting the intervention to prime students' connection with their institution. Before detailing our results, we review studies of preemptive strategies to combat low test-taking effort.

Strategies to Increase Effort in Low-Stakes Testing Contexts

Four strategies have been examined to increase students' effort and improve test score accuracy (Wise & DeMars, 2005): (a) offering external incentives; (b) increasing test relevance; (c) modifying assessment design; and (d) promising feedback on performance. A meta-analysis evaluated the effectiveness of these strategies (Rios, 2021). Offering feedback ($d = -0.01$) or modifying the assessment design (e.g., remove mentally-taxing items; $d = 0.13$) had smaller impact on expended effort compared to offering external incentives (e.g., money; $d = 0.36$) and increasing the relevance of assessments for students ($d = 0.21$).

There are practical issues with providing external incentives and increasing test relevance. The financial burden of incentives may be prohibitive. Moreover, incentives for performance may be perceived as inappropriate by stakeholders (e.g., board of visitors, faculty; Wise & DeMars, 2005). With respect to increasing the test relevance (students' subjective value of assessments), modifying instructions to highlight the importance of results for institutional reputation and personal benefit increased students' effort (Hawthorne et al., 2015; Liu et al., 2015). However, modifying instructions may be prohibited, especially if messaging is not truthful (e.g., deceptively telling students that employers see test scores; Finney et al., 2018).

Because we could not justify the cost of external incentives or increase test relevance, we turned to a newly examined strategy to increase effort: the question-behavior effect (QBE). Asking people questions about their behavior toward a target action (e.g., volunteering) increases their likelihood of performing the behavior (Levav & Fitzsimons, 2006; Wilding et al., 2016). The QBE has been supported for behaviors such as increasing voting, helping, and exercising (Spangenberg et al., 2016; Wood et al., 2016).

In the initial study examining the QBE in a low-stakes assessment context (Finney & McFadden, 2023), incoming first-year college students completing low-stakes institutional accountability tests were randomly assigned to one of three question conditions: no questions (No Question Condition); answering five intended effort questions (e.g., "I will engage in good effort throughout the test"; Intended Effort Condition); and answering five intended effort questions with reference to a positive self-identity (e.g., "As a conscientious test-taker, I will engage in good effort throughout the test"; Self-Identity Condition). Students then completed two unproctored cognitive assessments (information literacy test and oral communication test). Finney and McFadden (2023) hypothesized they would find higher effort for both question conditions than the No Question Condition and that the Self-Identity Condition would elicit higher effort than the Intended Effort Condition. Moreover, based on previous research (Rios et al., 2014; Soland, 2018), they hypothesized that approximately 15% of incoming college students in the No Question Condition would be flagged to be filtered from the dataset due to low effort and a lower percentage would be filtered in both question conditions. Filtering non-effortful responses (i.e., motivation filtering) is a practical approach to produce more valid interpretations of assessment data. That is, low effort results in construct-irrelevant variance in test scores; scores do not only reflect ability but also motivation. Construct-irrelevant variance can be addressed by motivation-filtering. Two criteria have been used to filter scores provided by students who expend low effort: 1) self-identifying being unmotivated via self-report measures (Rios et al., 2014; Swerdzewski et al., 2011); and 2) rapidly responding to items as measured by response time (Wise & Kong, 2005). Finney and McFadden (2023) used both methods.

As expected, students primed with either set of questions exhibited higher self-reported effort, lower proportions of rapid responding to items, and a lower percentage of data filtered from the dataset due to low effort. When conducting motivation filtering based on self-reported effort, 15.3% of students in the No Question Condition were filtered from the dataset due to low effort. The percentage filtered was reduced to 11.8% and 10.3% for students in the Intended Effort and Self-Identity Conditions, respectively. These findings were replicated using response time, where 15.6% of incoming students in the No Question Condition were filtered due to low response time, but only 11.9% were removed in the question conditions. Because including a positive self-identity increased the QBE in the context of voting behavior (Bryan et al., 2011), Finney and McFadden (2023) hypothesized that the Self-Identity Condition (e.g., "conscientious test-taker") would result in greater expended effort in a testing context. Contrary to their hypothesis, there was no significant difference in effort across the two question conditions.

To extend the study of the QBE in a testing context, Finney et al. (2024) examined the QBE with second-year college students and explored if gender moderated the effect. These more advanced college students were randomly assigned to the same three conditions, then two cognitive assessments were administered, and self-reported effort and response time were collected. There was no effect of QBE condition for students identifying as male for either measure of effort. Recall, when examining incoming college students, Finney and McFadden (2023) found no difference in self-reported effort or response-time effort when questions included or excluded positive self-identity. Both question conditions prompted more effort than the No Question Condition for incoming students. Likewise, for second-year female students, Finney et al. (2024) found no difference in self-reported effort when questions included or excluded positive self-identity and both question conditions resulted in more effort than the No Question Condition. However, response times for female students were significantly higher in the Self-Identity Condition compared to the No Question Condition, with no difference in response time between the No Question Condition and the Intended Effort Condition. The moderating effect of gender and inclusion of positive self-identity wording needs further study.

Because we could not justify the cost of external incentives or increase test relevance, we turned to a newly examined strategy to increase effort: the question-behavior effect (QBE).

In a third study of the QBE intervention with incoming college students, McFadden and Finney (2025) examined whether administering a second "dose" of questions could combat the decrease in examinee effort later in a testing session. In the previous studies (Finney & McFadden, 2023; Finney et al., 2024), five priming questions were administered at the start of the testing session but never again once testing was underway. McFadden and Finney (2025) randomly assigned incoming students to one of three question conditions prior to completing two low-stakes tests: answering three questions about intended effort directly before the first test in a session; answering three questions about intended effort directly before each test in

a session; and answering no priming questions. Administering a second dose of questions directly before the second test in a session significantly increased response-time effort and self-reported effort for the more difficult test. Moreover, the effects were found when reducing the priming questions from five to three questions. McFadden and Finney (2025) did not examine the utility of questions with self-identity wording or possible differential effects across gender.

In a fourth study of the QBE with first-year college students, Finney and Pastor (2025) examined if initially non-compliant students (those who completed testing after the testing deadline) would respond to the priming intervention to the same extent as compliant students (those who tested on time). Moreover, they examined if the priming effect would be similar when reducing the questions from three to one question. They randomly assigned first-year students to one of five priming conditions prior to completing a low-stakes test: answering one or three questions about intended effort, answering one or three questions about intended effort that infused positive self-identity, or answering no priming questions. Priming conditions were crossed by testing compliance status (students who tested on time versus late). Compliance status did not moderate the priming effect for self-reported effort; no questions resulted in significantly and practically lower self-reported effort than both three-question conditions. Compliance status moderated the priming effect for response time effort, with the three self-identity questions being effective for both compliant and non-compliant students. Thus, they demonstrated that priming with three questions is a quick and effective strategy to increase test-taking effort for first-year students, including those not initially compliant with testing requests.

Purpose of the Current Study

The current study extends the previous study of the QBE intervention in six important ways, aligned with the following six research questions.

1) Does the QBE emerge for effort and test scores with graduating senior students?

The QBE has been examined for first-year college students (Finney & McFadden, 2023; Finney & Pastor, 2025; McFadden & Finney, 2025) and second-year college students (Finney et al., 2024). We examined if the QBE intervention would be effective with university students who were completing assessments two to four months prior to college graduation. The graduating seniors were not exposed to the QBE intervention in prior low-stakes testing contexts. Given the limited study of the differential effectiveness of test-taking effort interventions across academic year of the student (O'Neil et al., 1995), we explored (rather than hypothesized) the effectiveness of the QBE intervention for these senior college students. It was unclear 1) if the intervention would have similar effectiveness as was found for incoming college students, 2) if the effectiveness would be differential across gender as was found with second-year students, or 3) if the effectiveness would be nil for students at this late point in college. Moreover, we further examined the utility of three rather than five priming questions.

2) Do differences in the self-identities primed in the questions impact the size of the QBE?

Similar to Finney and McFadden (2023), we framed priming questions in terms of general positive self-identities (e.g., "motivated student") and examined if this wording resulted in higher effort than no questions. However, we also created what may be more relevant self-identity priming questions. This study is the first to examine self-identity prompts that reflected a university's creed, which may be more pertinent to college students. Specifically, we explored whether there was a difference in effort if the self-identity primed in questions was generally positive (e.g., "conscientious test-taker) or specific to the positive characteristics of students noted in the university creed (e.g., "as someone who believes in hard work").

Administering a second dose of questions directly before the second test in a session significantly increased response-time effort and self-reported effort for the more difficult test.

3) Is the QBE moderated by student characteristics?

Finney et al. (2024) found the QBE for students identifying as female, but not male. Differential QBE associated with gender identity needs further examination. Moreover, other student characteristics (ethnicity, transfer status, first-generation status, Pell Grant eligibility) may moderate the QBE. Thus, we explored differential effects for different student sub-populations.

4) Could an increase in student motivation through priming decrease the amount of data removed, thus providing a cost benefit to an institution having to outsource its testing needs?

If priming students increases effort, a lesser amount of invalid data due to disengagement would need to be removed from the dataset. Given the commercial outcome measures administered at the university (i.e., at cost per student), any data removed due to low effort is wasteful. Priming students to give good effort may result in a cost benefit by increasing the amount of useable data.

5) Are results consistent when effort is operationalized using time and self-reported effort?

Finney and McFadden (2023) found the QBE to be similar when effort was operationalized using response-time effort and self-reported effort. Response-time effort was computed using the number of items for which the student did not rapidly respond. Finney et al. (2024) used the total time spent completing a test. They found different QBE results when operationalizing effort as total test time versus self-reported effort. For total test time from female students, the Self-Identity Condition resulted in more effort than the No Question Condition, whereas for self-reported effort, both the Self-Identity Condition and the Intended Effort Condition resulted in more effort than the No Question Condition. We employed total test time to operationalize effort, in addition to self-reported effort. Given that time-based measures of effort and self-reported effort have a low correlation and different nomological nets (Akhtar & Firdiyanti, 2023), we were prepared that the QBE may emerge for one measure of effort but not the other. We examined both effort measures since assessment practitioners may only be able to gather one measure.

6) Are test scores improved due to priming effort?

Priming questions that reflected a university's creed may be more pertinent to college students.

If priming enhances effort on the test, then it could result in increased test performance. Previous studies of the QBE (e.g., Finney & McFadden, 2023; Finney & Pastor, 2025; McFadden & Finney, 2025) examined its effects when gathering outcomes assessment data at pre-test or baseline. These incoming students had not received instruction in the test domain. Thus, even though effort was increased, there was no expectation that test scores would be increased. Increased effort would not translate to increased test scores if students do not have ability in the domain (Rios, 2021). However, for the current study of senior students, it was of great interest to evaluate if performance levels were impacted by priming and if this impact on test scores was differential across student characteristics (e.g., Pell Grant eligibility).

Methods

Procedures

In the current study, graduating seniors from a large R1 public southern university were enrolled in a zero-credit graduation course and asked to complete an assessment that would evaluate an institutional-level learning outcome focused on students' intercultural competence. To evaluate this learning outcome, the University used the HEIghten Intercultural Competency and Diversity (ICD) assessment, created by ETS© and administered by Territorium.

All graduating students enroll in a common graduation course where they must complete four assignments to receive their diploma after degrees are confirmed. Once the course is live for student participation, a hold is placed on the student's account until all four items are complete. The hold only impacts diploma delivery and does not impact a student's ability to graduate from the University. The hold does not impact financial aid disbursement, nor does it impact a student's ability to add or drop courses.

At the beginning of each semester, each enrolled student in the graduation course receives an e-mail informing them of said enrollment, the hold that has been applied to their student account, and the steps to take to have the hold removed by the end of the semester. All students are encouraged by the notification e-mail to navigate to the Canvas LMS platform where the graduation course assignments are located and to complete each assignment within

the allotted time frame. The first assignment a student must complete within the course is a one-hour assessment, which is a part of the University's efforts to continuously improve its general education curriculum. Each year the University measures student learning associated with one of nine learning outcomes that are directly aligned to the core curriculum.

Although all students complete the assessment for graduation purposes, they had the choice of releasing their scores for the research study. That is, after the assessment was complete, students were provided a research statement with information about the study and IRB approval of the study. Once students confirmed that they read the research statement, they were then provided with the opportunity to opt in or out of the study.

The graduation course launched on January 11, 2023, and students were asked to complete this one-hour assessment by March 16, 2023. Beyond directions on how to access the assessment, students were only encouraged to "give their best effort" on the assessment.

Participants

A total of 3,457 graduating seniors were enrolled in the Spring 2023 graduation course. Of those enrolled, 3,413 were marked as either completing the ICD assessment or completing a waiver for a prior semester test completion. For the purposes of this study, 3,311 students completed the ICD assessment. Detailed instructions within the Canvas course directed students to this assessment based on the last digit of a student's ID number. Students were then randomly assigned to one of the three priming conditions: 1) Control Condition with no priming questions before completing the assessment; 2) Positive Self-Identity Condition with priming questions created by Finney and McFadden (2023) answered before completing the assessment; and 3) University Creed Condition with priming questions infused with the university's creed language answered before the assessment (see Appendix for priming questions).

Of importance, although 3,311 students completed the assessment, only students who gave consent to use their data for this study were analyzed. Therefore, after removing students who (1) did not give consent for the study, (2) had duplicate or missing data, or (3) took over the allotted one-hour time provided via the Territorium site, the final sample size was N = 2,204 (Control Condition = 656; Positive Self-Identity Condition = 852; University Creed Condition = 696). This sample of 2,204 students was used to evaluate the impact of priming on self-reported effort, response time on the test, and test performance.

The 2,204 students who completed the ICD assessment and provided consent represented the demographics of the University. The majority of students self-identified as female (54.6%) and White (85.9%). Other ethnicities included Black or African American (2.2%), American Indian or Alaska Native (0.2%), Asian (2.2%), Hispanics of any race (3.3%), Native Hawaiian or Other Pacific Islander (0.04%), Two or More Races (2.2%), Nonresident Alien (3.8%), and Race and Ethnicity Unknown (0.1%). The sample was predominately non-transfer students (84.4%), non-first-generation/continuing students (88.0%), and non-Pell eligible students (88.9%). Importantly, because student characteristic information was gathered through the University's system and not reported at the time of the assessment, not all 2,204 students had data for every demographic characteristic. Thus, our analyses that evaluate the following moderators are based on different sample sizes: gender (n = 2,135), ethnicity (n = 2,052), transfer status (n = 2,074), first-generation status (n = 2,004), and Pell eligibility (n = 2,135).

All graduating students enroll in a common graduation course where they must complete four assignments to receive their diploma after degrees are confirmed.

Measures

Analyze & Act Test Performance

The HEIghten Intercultural Competency and Diversity (ICD) assessment is a 74-item measure. The ICD assessment produces two scale scores: Analyze & Act scores, which range from 150 to 180 based on 40 cognitive (right or wrong) items and Approach scores, which range from 90 to 150 based on 34 noncognitive (Likert-type) items. The Approach scores reflect students' view of themselves, which is dispositional in nature and not used as a measure of learning at the university. For the Analyze & Act dimension, there are three proficiency levels: "Developing" (scores between 150 to 157), "Proficient" (scores between 158 to 174), and "Advanced" (scores between 175 to 180). Additionally, six Analyze & Act subscale scores

can be computed: Self-Awareness, Cultural Knowledge Application, Suspending Judgment/ Perspective Taking, Social Monitoring, Emotion Regulation, and Behavior Regulation. Subscale scores range from 1 to 10, with higher scores representing higher ability.

We only employed the total Analyze & Act score for our study. Specifically, students' average Analyze & Act performance serves as one of the three measures (in addition to self-report effort and response time) used to assess the impact of the QBE. This performance score was calculated and provided by Territorium.

Students with scores ≤ 15 on the Effort subscale were deemed unmotivated and identified as responses to be filtered in the dataset.

Self-Reported Effort

At the end of the ICD assessment, the Student Opinion Survey (SOS; Pastor et al., 2023; Thelk et al., 2009), a 10-item measure consisting of two subscales (Effort and Importance) was completed. The Effort subscale measures test-taker's reported effort put forth on a test ("I engaged in good effort throughout this test"), whereas the Importance subscale measures the degree to which students perceived the test to be important (e.g., "Doing well on this test is important to me"). Students responded to the SOS items using a 5-point Likert-type scale (1= Strongly Disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, and 5 = Strongly Agree).

We assessed the impact of the three priming conditions on the Effort subscale scores themselves and when the effort scores were used for motivation filtering. Based on Swerdzewski et al.'s (2011) suggestion, students with scores ≤ 15 on the Effort subscale were deemed unmotivated and identified as responses to be filtered in the dataset.

Response Time

Territorium provided students with one hour to complete the ICD assessment. Each students' time spent on the assessment was converted to minutes for the purpose of evaluating the effectiveness of the priming questions.

Results

Prior to conducting analyses, the three dependent variables (self-reported effort, response time, and test performance) were assessed for normality. All three variables were approximately normally distributed with skewness and kurtosis less than |2|: effort displayed -0.99 skew and 0.90 kurtosis; response time displayed -0.82 skew and 0.99 kurtosis; and test performance displayed -1.33 skew and 1.64 kurtosis.

Effect of Priming Condition across the Three Outcomes

To examine the effect of priming condition on the three outcomes of interest (self-reported effort, response time, and test performance), we conducted separate one-way ANOVAs for each outcome. Levene's test of equality of variance across groups was not significant for self-reported effort (p =.61), response time (p =.22) or test performance (p = .53), indicating the assumption of homogeneity of variance was not violated. As expected, there were significant effects of priming on self-reported effort [F (2, 2201) = 5.0, p = .007)] and response time [F (2, 2201) = 3.6, p = .027)]. For self-reported effort, the Positive Self-identity priming questions resulted in higher effort than the No Question condition (d = 0.16). For response time, the University Creed priming questions resulted in students spending more time on the test than the No Question conditions (d = 0.14). However, the effect on test performance was not significant [F (2, 2201) = 2.5, p = .085). Table 1 provides descriptive statistics and effect size information.

Group Differences by QBE Condition and Student Characteristics

To be diligent, we conducted several factorial ANOVAs to evaluate if priming condition was moderated by any student characteristic for each outcome of interest. In other words, we assessed the significance of the interaction between several student characteristics (gender, ethnicity, transfer status, first-generation status, and Pell eligibility) and the priming condition with respect to self-reported effort, response time, and performance.

First, gender, transfer status, and first-generation status did not interact with priming condition for any of the three outcome variables. Thus, general statements about the impact of priming on these outcomes can be made across these student groups.

Table 1
Descriptive statistics across priming conditions for the three outcomes of interest

Condition	*N*	Self-reported Effort				Response Time			Test Performance	
		Mean	SD	*d*	% to be filtered	Mean	SD	*d*	Mean	SD
No Questions	656	4.30_{a}	.62	-	5.34%	38.35_{a}	7.44	-	172.24_{a}	6.22
Positive Self-Identity	852	4.40_{b}	.60	.16	3.87%	38.68_{ab}	7.83	.04	171.95_{a}	6.24
University Creed	696	4.35_{ab}	.58	.08	3.30%	39.45_{b}	8.05	.14	171.49_{a}	6.56

Note. Effort scores can range from 1 to 5, with higher scores reflecting higher levels of self-reported effort. Response time was reported in minutes. Within a column, means with different subscripts are statistically significantly different. *d* = Cohen's *d* effect size when comparing the No Questions condition to each question condition. Self-reported effort "% to be filtered" is the percentage of students in that condition whose scores were flagged for removal due to low self-reported effort (at or below a summed score of 15 across five effort items as suggested by Swerdewski et al. (2011)). Mean test performance in each condition is based on all students without filtering low-effort students.

However, when examining self-reported effort, Pell eligibility interacted with priming condition. Prior to estimating and interpreting this effect, the assumption of homogeneity of variance was assessed and supported (Levene's test, $p = .20$). The interaction indicated that the question-behavior effect was stronger for students who were Pell eligible [interaction effect: F (2, 2129) = 5.90, $p = .003$]. When we did not prime students, Pell eligible students put forth significantly less effort (4.15) than those who were not Pell eligible (4.33). However, when we primed students with University Creed questions, the Pell eligible students put forth an equal (not significantly different) amount of effort (4.38) as students who were not Pell eligible (4.35). Notably, when we primed students with positive self-identity questions, the Pell eligible students put forth significantly more effort (4.54) than students who were not Pell eligible (4.39). As shown in Table 2, it appears that we have the opportunity to enhance expended effort from Pell eligible students via priming questions.

When we primed students with positive self-identity questions, the Pell eligible students put forth significantly more effort (4.54) than students who were not Pell eligible (4.39).

Table 2
Interaction of Pell Grant eligibility and priming condition on self-reported effort

Condition	Pell Eligible				Not Pell Eligible			
	N	Mean	SD	*d*	*N*	Mean	SD	*d*
No Questions	70	4.15_{a}	.65	-	566	4.33_{a}	.65	-
Positive Self-Identity	106	4.54_{b}	.50	.69	713	4.39_{a}	.60	.10
University Creed	60	4.38_{b}	.66	.35	620	4.35_{a}	.58	.03

Note. Effort scores can range from 1 to 5, with higher scores reflecting higher self-reported effort. Within columns, means with different subscripts are statistically significantly different. *d* = Cohen's *d* effect size when comparing the No Questions condition to each question condition.

Further, ethnicity moderated the effect of priming on test performance. Prior to estimating and interpreting this effect, the assumption of homogeneity of variance was assessed and supported (Levene's test, $p = .08$). The interaction effect [F (2, 2046) = 3.23, $p = .040$)] uncovered a positive effect for underrepresented students. Specifically, in the No Question condition (typical testing condition), underrepresented students scored significantly lower (170.49) than White students (172.91). However, when primed with either set of questions, there was no difference in test scores across the two student groups (see Table 3). It appears that priming students can

Table 3
Interaction of ethnicity and priming condition on test performance

Condition	White				Underrepresented				
	N	Mean	SD	d_W	*N*	Mean	SD	d_U	d_E
No Questions	551	172.91	5.74	-	65	170.49	7.10	-	.41
Positive Self-Identity	701	172.46	5.74	.08	92	172.33	5.41	.29	.02
University Creed	583	172.05	6.20	.14	60	172.05	5.10	.25	.00

Note. d_E = Cohen's *d* effect size comparing performance for the two ethnic groups within each priming condition (i.e., White students are .41 SDs higher on test performance than underrepresented students in the No Questions condition, which is statistically significant, but the two ethnic groups are not significantly different on test performance in the two priming conditions, *d* = .02 and .00). *dw* = Cohen's *d* effect size comparing the No Questions condition to each question condition for White students. *du* = Cohen's *d* effect size comparing the No Questions condition to each question condition for underrepresented students.

increase test scores for underrepresented students, erasing performance differences across the underrepresented and majority ethnic groups.

Each test costs the university $8 per student. When filtering based on self-reported effort, we found the following number of students would have been removed from the three groups for low effort. For the Control Group, 35 students out of 656 would have been removed, which fortunately is a low amount of filtering. The goal with priming is to reduce this number even though it is already low (which again is a positive outcome for this testing program). For the Positive Self-Identity Group, 33 out of 852 students would have been removed. For the University Creed Group, 23 out of 696 students would have been removed. To put this into a cost perspective, the Control Group removed $280 worth of test scores, the Positive Self-Identity Group removed $264 worth of test scores, and the University Creed removed $184 worth of test scores. Based on the low removal rates, we determined that for the purposes of our study, any cost benefits based around priming for low effort would only be warranted for an institution or office that had financial hardships or had much more disengagement on assessments. For institutions or offices that do not have financial constraints or have limited disengagement (as we did), priming for low effort is likely not going to be effective to impact costs.

Discussion

Priming students with questions about their intended effort prior to completing low-stakes assessments has been found to significantly increase self-reported effort and response time (e.g., Finney & McFadden, 2023). Although this strategy has been effective for students early in their college career, it had not been examined with graduating university students nor had possible moderating student characteristics been adequately examined (e.g., Finney et al., 2024). Thus, we randomly assigned college seniors to one of three priming question conditions prior to completing a low-stakes assessment for institutional accountability: answering three questions about intended effort that infused positive self-identity, answering three questions about intended effort that incorporated the University's Creed, or answering no priming questions (control).

It appears that priming students can increase test scores for underrepresented students, erasing performance differences across the underrepresented and majority ethnic groups.

When simply examining the impact of the priming conditions, one would infer that priming seniors with self-identity questions resulted in higher self-reported effort than the control condition, priming with the University creed resulted in higher time on the test than the control condition, and neither priming condition increased test performance over no priming. No student characteristics moderated the priming effect on response time. Moreover, we found that gender, transfer status, and first-generation status did not moderate the priming effect on the three outcomes. However, Pell Grant eligibility did moderate the impact of priming on self-reported effort with priming increasing effort from eligible students who were lower in effort

than non-eligible students in the control condition. In short, priming resulted in self-reported effort for eligible students being equal or higher than non-eligible students.

Likewise, the effect of priming on test scores was moderated by ethnicity. White students scored higher on the test than underrepresented students in the non-priming control condition (typical testing condition). However, this difference in test performance disappeared with priming. That is, underrepresented students had test scores not significantly different than White students if they were primed with either set of questions, which was an unexpected but positive outcome of priming. In turn, without priming, when disaggregating data and reporting test performance differences by student group, inferences would have been made about White students performing better than underrepresented students. Likely, numerous discussions would have occurred postulating why underrepresented students performed 0.41 SDs lower than White students (e.g., opportunity to learn). However, priming with either question type would lead to very different conclusions about student performance differences (or lack thereof). Clearly, further study of the effect of priming on test performance is needed.

Recall, when examining effective motivation interventions, a meta-analysis (Rios, 2021) found the largest effect on effort ($d = 0.36$) involved paying for performance (may not be acceptable or financially feasible at an institution) and the next largest effect ($d = 0.21$) involved changing the relevance of the test (easier than incentives, but still difficult in many contexts and requires more testing time than priming). When collapsing across student characteristics, the priming intervention had smaller effect sizes ($d = 0.16$ for self-identity primes on self-reported effort and $d = 0.14$ for creed primes on response time) than the more costly interventions. However, the priming effect in this testing context was larger than prior QBE studies with prosocial behaviors, which is encouraging (Wilding et al., 2016: blood donation Hedges' $g = 0.06$, voting Hedges' $g = 0.06$). Importantly, when examining the priming effect by Pell eligibility status and ethnicity, the priming effects are much larger for some student populations. Specifically, for Pell eligible students, self-reported effort for self-identity primes ($d = 0.69$) and creed primes ($d = 0.35$) was over a third of a standard deviation higher than no primes. These effect sizes are similar to or exceed the effects associated with providing incentives and changing the test relevance, interventions that cost more money and time. Likewise, for underrepresented students, self-identity primes ($d = 0.29$) and creed primes ($d = 0.25$) resulted in test performance being at or over a quarter standard deviation higher than no primes and these effect sizes align with or exceed the effects of increasing test relevance ($d = 0.27$) and external incentives ($d = 0.21$) on test performance (Rios, 2021).

Priming students with questions about their intended effort prior to completing low-stakes assessments has been found to significantly increase self-reported effort and response time.

One may ask if these effect sizes ($d = 0.69, 0.35, 0.29, 0.25$) are non-negligible in an absolute sense; they are. Recently proposed benchmarks for effect sizes from causal studies of education interventions on student achievement in Pre-K to 12 are the following: less than 0.05 is "small", 0.05 to less than 0.20 is "medium", and 0.20 or greater is "large" (Kraft, 2020). These benchmarks were based on 1,942 effect sizes from 747 randomized control trials evaluating education interventions with standardized test outcomes. Although our study does not focus on Pre-K to 12 populations or educational interventions (e.g., interventions to increase reading or math), it does evaluate a motivational intervention using standardized test outcomes with a population of students. Moreover, echoing methodologists who focus on effect sizes (e.g., Kraft, 2020), cost matters when evaluating effect sizes for policy decisions. Effect sizes should be considered relative to the costs of implementing the intervention, strategy, or program. Administrators may want to observe a 0.50 effect size for important outcomes to justify implementing an expensive program. In our context of priming students, we would argue that smaller effects ($d = 0.15$) support implementing this low-cost, quick intervention, as it is a desired step toward more trustworthy data and in turn more valid interpretations and decisions. That is, any increase (big or small) in students' effort to show their true ability is desired to make valid inferences from test scores and in turn institutional decisions about curriculum and programming. Gathering test data that does not represent the construct of interest due to student disengagement can result in decisions that are wrong, inefficient, and that cause harm. Fortunately, priming not only resulted in non-negligible effect sizes, but it is also at no or low cost to implement.

Implications for Assessment Practitioners

Effect sizes should be considered relative to the costs of implementing the intervention.

Students approach testing with different perceptions and prior testing experiences, which impact their engagement and in turn their test scores and inferences from the scores. We need to try to understand how these personal characteristics interact with testing conditions, and how testing conditions can be altered to produce more valid inferences from test scores. Some of these alterations may be quite minor, such as priming students to expend the necessary effort to show their ability. We are not suggesting that a priming intervention (or any motivation intervention) can address systemic inequities in educational measurement (see instead Forzani et al., 2024; Randall et al., 2022, 2024; Russell, 2023; Sireci, 2020). Rather, if priming reduces or eliminates group differences in effort and performance (as found for Pell Grant eligibility and ethnicity), priming may offer more valid inferences about differences in test engagement or performance across groups. We believe this is particularly important given prior research showing incorrect inferences from test scores directly impact college students. Randall et al. (2024, p. 2) provide a powerful example regarding college-level placement testing: "Often based on placement test scores, Black (66%) and Hispanic (53%) students are placed in remedial courses (see Gilman, 2019; Nastal, 2019; Ngo & Melguizo, 2020) more frequently than White (36%) students (Chen, 2016). While some may argue that students need additional support for math and literacy, it has been shown that too often placement tests misplace students (Scott-Clayton et al., 2014). For too many students, misplacement becomes an academic death sentence (Klausman & Lynch, 2022)." Assessments used for student placement (e.g., math, world language) or for remediation programming (e.g., early alert assessments) may be perceived as low stakes to students, but decisions based on them can result in stigmatization. If priming students to give effort on low-stakes assessments results in more accurate test scores and thus more appropriate decisions for even a handful of students, we believe it is well-worth the few minutes to prime them. With respect to the institution, if priming results in more accurate interpretations about the ability levels of various student groups, then institutions may avoid unnecessary programming that (mis)targets specific student populations and the deficit narratives that often accompany group differences in performance.

Although not necessarily a purpose for the current study, priming questions were reduced from five in previous studies to only three. It was promising that priming for good effort was still effective with a reduction in questions, even with this graduating student sample. Thus, we recommend assessment practitioners use one to two minutes to ask students to answer three priming questions prior to engaging in low-stakes assessments. Priming can be done solely at the beginning of the testing session; however, priming the students before each test has been shown to be promising (McFadden & Finney, 2025).

We recognize that our method for priming students (whether primes are administered via computer prior to the test or via a test cover sheet if tests are administered paper/pencil) is a simple and quick strategy, yet unknown barriers could exist that prevent a desired outcome. Should institutions encounter any logistical barriers to priming, those barriers should be shared out to the assessment community. In our opinion, those couple of minutes used for priming are worth increasing the quality of inferences from the test scores reported for accountability and improvement purposes (i.e., low cost but high benefit). More specifically, if institutions are gathering measures of motivation (self-reported, response time) and motivation is quite high, then allocating a couple minutes for this intervention may not be worthwhile. However, if motivation is low or variable across students (as is common for low-stakes tests), priming to increase effort may have great benefit with very limited "cost".

Moreover, if institutions are not collecting motivation data, but instead assuming motivation is high, we warn against this dangerous assumption. If institutions are not collecting motivation data because it takes time to gather self-reported effort or a timing mechanism is not available, we strongly suggest these institutions err on the side of caution, assume that at least some students are disengaged, and employ a strategy to increase engagement during low-stakes testing. Here we offer a very quick and cheap strategy to enhance students' test-taking effort.

Given that time-based measures of effort and self-report effort have a low correlation and different nomological nets (Akhtar & Firdiyanti, 2023), we were not surprised by the

different priming effect across these two operationalizations of effort. While we feel comfortable recommending the general strategy of priming to increase effort on low-stakes institutional accountability tests, we feel it premature to advise on self-identity versus creed questions given their differential effects on the two effort measures. If pressed, we would suggest the following. Based on this single study, either priming condition will positively impact test scores for underrepresented senior students, which is encouraging but needs to be replicated. If an institution operationalizes effort via self-report measures, we recommend using the self-identity priming questions as they had the largest effect. If an institution operationalizes effort via response time, we recommend the creed-infused questions. Again, these suggestions are tentative and future studies are needed to examine the stability of these effects, in addition to other considerations discussed below.

Limitations of the Current Study and Call for Future Research

Given we found the effect of priming was moderated by certain student characteristics, we hope future studies can better capture complete information on student characteristics. Notably, we recognize the extreme crudeness of our classification of White and underrepresented students. Our sample was representative of the University's demographics. Thus, we were limited in the comparison that could be made. Future research on the QBE in accountability testing contexts should be conducted with more diverse student populations. We perceive this study as the beginning of this line of research. Moreover, future studies should examine the generalizability of the priming effect across different types of institutions (very large, very small, community colleges).

Additionally, previous studies have explored the self-identity questions with incoming first-year students (Finney & McFadden, 2023; Finney & Pastor, 2025). Given our results comparing the self-identity questions and creed questions with graduating seniors, we encourage studies that explore if first-year students are impacted by creed-infused questions. It may be that a university's creed resonates more or less with incoming students depending on the university's culture (Miller & Finney, 2024). More generally, the creed-infused questions need more examination given this study was the first. Creed phrasing may be effective at institutions where the creed is ever-present and important to students but may have no impact at institutions where the creed is not highlighted or not perceived as important by students. Thus, we encourage future studies that evaluate creed phrasing to always include a control group (no priming), but also non-creed-infused priming questions. These types of designs will allow for a better understanding of what type of phrasing in the primes is most impactful and for who. Regarding prime phrasing more generally, we strongly encourage others to evaluate different phrasing of the primes, beyond self-identity or creed-infused wording. We believe that suggestions from current students may be particularly helpful to design the most impactful primes. These are research questions worth pursuing.

We also call for longitudinal studies of the priming effect. At some institutions, students complete the low-stakes assessments at multiple time points so value-added or growth estimates can be computed and reported to accreditors. Recall, when more advanced college students expend less effort on tests than incoming students, value-added estimates are biased downward (Finney et al., 2016). Future studies should evaluate if value-added estimates are less attenuated for students who were primed at both time points compared to students who were not primed.

We encourage future studies that evaluate creed phrasing to always include a control group (no priming), but also non-creed-infused priming questions.

Finally, future studies could examine if utilizing a simple "yes" or "no" response option rather than a Likert response scale impacts the QBE. Basic research in the domain of the QBE suggests "yes" or "no" responses may be effective (e.g., Spangenberg et al., 2016). If asking a few "yes" or "no" questions prior to low-stakes institutional testing produces significantly and practically higher effort and reduces differences in test performance across student subpopulations, this strategy would be an attractive option to increase the validity of inferences from assessment scores.

References

Akhtar, H. & Firdiyanti, R. (2023). Test-taking motivation and performance: Do self-report and time-based measures of effort reflect the same aspects of test-taking motivation? *Learning and Individual Differences*, *106*, 102323. https://doi.org/10.1016/j.lindif.2023.102323

Bryan, C. J., Walton, G. M., Rogers, T. & Dweck, C.S. (2011). Motivating voter turnout by invoking the self. *PNAS*, *108*(31), 12653-12656. http://doi.org/10.1073/pnas.1103343108

Chen, X. (2016). *Remedial coursetaking at U.S. public 2- and 4-year institutions: Scope, experiences, and outcomes* (NCES 2016-405). U.S. Department of Education. National Center for Education Statistics. https://nces.ed.gov/pubs2016/2016405.pdf

Cole, J. S., & Osterlind, S. J. (2008). Investigating differences between low- and high-stakes test performance on a general education exam. *The Journal of General Education*, *57*(2), 119–130. https://doi.org/10.1353/jge.0.0018

Council for Higher Education Accreditation (2022, March 3). *Accreditation & recognition*. https://www.chea.org/about-accreditation

Finney, S. J. & McFadden, M. E. (2023). Examining the question-behavior effect in low-stakes testing contexts: A cheap strategy to increase examinee effort. *Educational Assessment*, *28*(4), 211-228. https://doi.org/10.1080/10627197.2023.2222588

Finney, S. J., Myers, A. J., & Mathers, C. E. (2018). Test instructions do not moderate the indirect effect of perceived test importance on test performance in low-stakes testing contexts. *International Journal of Testing*, *18*(4), 297–322. https://doi.org/10.1080/15305058.2017.1396466

Finney, S. J., & Pastor, D. A. (2025). Priming non-compliant students to expend test-taking effort: How many primes are needed? *Journal of Experimental Education*. Advance online publication. https://doi.org/10.1080/00220973.2025.2459392

Finney, S. J., Schaefer, K. E., & McFadden, M. E. (2024). Priming examinees to give good effort: Differential utility across gender identity. *Journal of Experimental Education*. Advance online publication. https://doi.org/10.1080/00220973.2024.2310678

Finney, S. J., Sundre, D. L., Swain, M. S., & Williams, L. M. (2016). The validity of value-added estimates from low-stakes testing contexts: The impact of change in test-taking motivation and test consequences. *Educational Assessment*, *21*(1), 60–87. https://doi.org/10.1080/10627197.2015.1127753

Forzani, E., Corrigan, J., Slomp, D., & Randall, J. (2024) Prioritizing equitable social outcomes with and for diverse readers: A conceptual framework for the development and use of justice-based reading assessment. *Educational Psychologist*, *59*(4), 291-314. https://doi.org/10.1080/00461520.2024.2418400

Gilman, H. (2019). Are we whom we claim to be? A case study of language policy in community college writing placement practices. *Journal of Writing Assessment*, *12*(1). https://escholarship.org/uc/item/3kh925t2

Hawthorne, K. A., Bol, L., Pribesh, S., & Suh, Y. (2015). Effects of motivational prompts on motivation, effort, and performance on a low-stakes standardized test. *Research & Practice in Assessment*, *10*, 30–39. https://www.rpajournal.com/dev/wp-content/uploads/2015/06/A3.pdf

Klausman, J., & Lynch, S. (2022). From ACCUPLACER to informed self-placement at Whatcom Community College: Equitable placement as evolving practice. In J. Nastal, M. Poe, & C. Toth (Eds.), *Writing placement in two-year colleges: The pursuit of equity in postsecondary education* (pp. 59–83). The WAC Clearinghouse, University Press of Colorado. DOI: 10.37514/PRA-B.2022.1565.2.02

Kraft, M. A. (2020). Interpreting effect sizes for education interventions. *Educational Researcher*, *49*(4), 241 – 253. https://doi.org/10.3102/0013189X20912798

Levav, J., & Fitzsimons, G. J. (2006). When questions change behavior: The role of ease of representation. *Psychological Science*, *17*(3), 207-213. https://pubmed.ncbi.nlm.nih.gov/16507060/

Liu, O. L. (2017). Ten years after the spellings commission: From accountability to internal improvement. *Educational Measurement: Issues and Practice*, *36*(2), 34–41. https://doi.org/10.1111/emip.12139

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

Mathers, C. E., Finney, S. J., & Hathcoat, J. (2018). Student learning in higher education: A longitudinal analysis and faculty discussion. *Assessment and Evaluation in Higher Education*, *43*, 1211-1227. DOI: 10.1080/02602938.2018.1443202

McFadden, M. E., & Finney, S. J. (2025). Investigating the impact of multiple priming questions on examinee effort during low-stakes testing. *International Journal of Testing*, *25*(1), 109–133. https://doi.org/10.1080/15305058.2024.2414425

Miller, S.A. & Finney, S.J. (2024). Enhancing student effort for improved institutional accountability data: The impact of motivation priming interventions. *Assessment Update*, *36*(6), 8-9. https://doi.org/10.1002/au.30421

Nastal, J. (2019). Beyond tradition: Writing placement, fairness, and success at a two-year college. *Journal of Writing Assessment*, *12*(1). http://journalofwritingassessment.org/article.php?article=136

Ngo, F., & Melguizo, T. (2020). The equity cost of inter-sector math misalignment: Racial and gender disparities in community college Student outcomes. *The Journal of Higher Education*, *92*(3), 410–434. https://doi.org/10.1080/00221546.2020.1811570

O'Neil, H., Sugrue, B., & Baker, E. L. (1995) Effects of motivational interventions on the National Assessment of Educational Progress mathematics performance. *Educational Assessment*, *3*(2), 135-157. https://doi.org/10.1207/s15326977ea03022

Pastor, D. A., Patterson, C. R., & Finney, S. J. (2023). Development and internal validity of the Student Opinion Scale: A measure of test-taking motivation. *Journal of Psychoeducational Assessment*, *41*(2), 209-225. https://doi.org/10.1177/07342829221140957

Randall, J., Poe, M., Oliveri, M. E., & Slomp, D. (2024) Justice-oriented, antiracist validation: Continuing to disrupt White supremacy in assessment practices. *Educational Assessment*, *29*(1), 1-20. https://doi.org/10.1080/10627197.2023.2285047

Randall, J., Slomp, D., Poe, M., & Oliveri, M. E. (2022) Disrupting White supremacy in assessment: Toward a justice-oriented, antiracist validity framework. *Educational Assessment*, *27*(2), 170-178. https://doi.org/10.1080/10627197.2022.2042682

Russell, M. (2023) Shifting educational measurement from an agent of systemic racism to an anti-racist endeavor. *Applied Measurement in Education*, *36*(3), 216-241. https://doi.org/10.1080/08957347.2023.2217555

Rios, J. (2021). Improving test-taking effort in low-stakes group-based educational testing: A meta-analysis of interventions. *Applied Measurement in Education*, *34*(2), 85-106. https://doi.org/10.1080/08957347.2021.1890741

Rios, J. A., Liu, O. L., & Bridgeman, B. (2014). Identifying low-effort examinees on student learning outcomes assessment: A comparison of two approaches. *New Directions for Institutional Research*, *2014*(161), 69–82. https://doi.org/10.1002/ir.20068

Rios, J. A., & Guo, H. (2020). Can culture be a salient predictor of test-taking engagement? An analysis of differential non effortful responding on an international college-level assessment of critical thinking. *Applied Measurement in Education*, *33*(4), 263–279. https://doi.org/10.1080/08957347.2020.1789141

Roohr, K. C., Liu, H., & Liu, O. L. (2016). Investigating student learning gains in college: A longitudinal study. Studies in *Higher Education*, *42*(12), 2284–2300. https://doi.org/10.1080/03075079.2016.1143925

Scott-Clayton, J., Crosta, P. M., & Belfield, C. R. (2014). Improving the targeting of treatment: Evidence from college remediation. *Educational Evaluation and Policy Analysis*, *36*(3), 371–393. https://doi.org/10.3102/0162373713517935

Silm, G., Pedaste, M., & Täht, K. (2020). The relationship between performance and test-testing effort when measured with self-report or time-based instruments: A meta-analytic review. *Educational Research Review*, *31*, 100335. https://doi.org/10.1016/j.edurev.2020.100335.

Simzar, R. M., Martinez, M., Rutherford, T., Domina, T., & Conley, A. M. (2015). Raising the stakes: How students' motivation for mathematics associates with high-and low-stakes test achievement. *Learning and Individual Differences*, *39*, 49–63. https://doi.org/10.1016/j.lindif.2015.03.002

Sireci, S. (2020). Standardization and UNDERSTANDardization in educational assessment. *Educational Measurement: Issues and Practice*, *39*(3), 100–105. https://doi.org/10.1111/emip.12377

Smith, L. F., & Smith, J. K. (2004). The influence of test consequences on national examinations. *North American Journal of Psychology*, *6*(1), 13–25.

Soland, J. (2018). The achievement gap or the engagement gap? Investigating the sensitivity of gaps estimates to test motivation. *Applied Measurement in Education*, *31*(4), 312–323. https://doi.org/10.1080/08957347.2018.1495213

Spangenberg, E. R., Kareklas, I., Devezer, B., & Sprott, D. E. (2016). A meta-analytic synthesis of the question-behavior effect. *Journal of Consumer Psychology*, *26*(3), 441-458. http://doi.org/10.1016/j.jcps.2015.12.004

Sundre, D. L., & Kitsantas, A. (2004). An exploration of the psychology of the examinee: Can examinee self-regulation and test-taking motivation predict consequential and non-consequential test performance? *Contemporary Educational Psychology*, *29*, 6–26. https://doi.org/10.1016/S0361-476X(02)00063-2

Sungur, S. (2007). Contribution of motivational beliefs and metacognition to students' performance under consequential and nonconsequential test conditions. *Educational Research and Evaluation*, *13*, 127–142. https://doi.org/10.1080/13803610701234898

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., Sundre, D. L., Horst, J. S., & Finney, S. J. (2009). Motivation matters: Using the Student Opinion Scale (SOS) to make valid inferences about student performance. *Journal of General Education*, *58*, 131–151. https://doi.org/10.1353/jge.0.0047

Wilding, S., Conner, M., Sandberg, T., Prestwich, A., Lawton, R., Wood, C., Miles, E., Godin, G., & Sheeran, P. (2016). The question-behaviour effect: A theoretical and methodological review and meta-analysis. *European Review of Social Psychology*, *27*(1), 196-230. http://doi.org/10.1080/10463283.2016.1245940

Wise, S. L. (2020). Six insights regarding test-taking disengagement. *Educational Research and Evaluation*, *26*(5-6), 328–338. https://doi.org/10.1080/13803611.2021.1963942

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 non-effort 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., & Smith, L. F. (2016). The validity of assessment when students don't give good effort. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of human and social conditions in assessment* (pp. 204–220). New York, NY: Routledge.

Wood, C., Conner, M. T., Miles, E., Sandberg, T., Taylor, N. J., Godin, G., & Sheeran, P. (2016). The impact of asking intention or self-prediction questions on subsequent behavior: A meta-analysis. *Personality and Social Psychology Review*, *20*, 245-268. https://doi.org/10.1177/1088868315592334

Appendix

Priming Questions with Positive Self-Identity

Please think about the test you are about to complete. Mark the answer that best represents how you feel about each of the statements below.

A =Strongly Disagree B =Disagree C =Neither Agree nor Disagree D =Agree E =Strongly Agree

1. As a conscientious test-taker, I will engage in good effort throughout the test.

2. I, a motivated student, will give my best effort on this test.

3. As a hardworking student, I will persist to completion of the test.

Priming Questions with Creed-Infused Self-Identity

Please think about the test you are about to complete. Mark the answer that best represents how you feel about each of the statements below.

A =Strongly Disagree B =Disagree C =Neither Agree nor Disagree D =Agree E =Strongly Agree

1. As a someone who believes in education, which gives me knowledge to work wisely, I will engage in good effort throughout the test.

2. As someone who believes that this is a practical world and that I can count only on what I earn, I will give my best effort on this test.

3. As someone who believes in hard work, I will persist to the completion of the test.