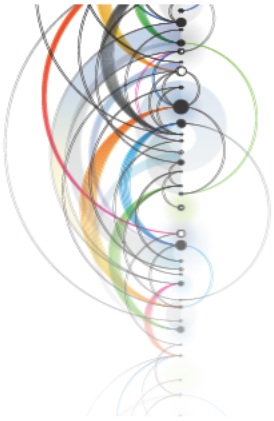
RESEARCH & PRACTICE IN ASSESSMENT



Abstract

Examinee effort can impact the validity of scores on higher education assessments. Many studies of examinee effort have briefly noted gender differences, but gender differences in test-taking effort have not been a primary focus of research. This review of the literature brings together gender-related findings regarding three measures of examinee motivation: attendance at the assigned testing session, time spent on each test item, and self-reported effort. Evidence from the literature is summarized, with some new results presented. Generally, female examinees exert more effort, with differences mostly at very low levels of effort—the levels at which effort is most likely to impact test scores. Examinee effort is positively correlated with conscientiousness and agreeableness, and negatively correlated with work-avoidance. The gender differences in these constructs may account for some of the gender differences in test-taking effort. Limitations and implications for higher education assessment practice are discussed.

AUTHORS

Christine E. DeMars, Ph.D.
James Madison University

Bozhidar M. Bashkov, M.A.
James Madison University

Alan B. Socha, M.A.
James Madison University

The Role of Gender in Test-Taking Motivation under Low-Stakes Conditions

Test-taking motivation is important to many university assessment efforts, because higher education assessments often have low or no consequences for individual students but high consequences for the university. Test-taking motivation has been extensively studied. Examinees score higher when the test has some stakes for them, such as a grade in a course (Sundre, 1999; Sundre & Kitsantas, 2004; Terry, Mills, & Sollosy, 2008; Wolf & Smith, 1995; Wolf, Smith, & DiPaulo, 1996), course placement, promotion, graduation (DeMars, 2000), admissions (Cole & Osterlind, 2008), hiring decisions (Rothe, 1947), or simply knowing that faculty and employers will see the scores (Liu, Bridgeman, & Adler, 2012). Following the terminology of Wise (2009, p. 154), the phrase *low-stakes* will be used here to describe tests with no personal stakes for examinees, regardless of the stakes for institutions or instructors.

When the test has no personal stakes, examinees who report higher effort tend to score somewhat higher (Cole, Bergin, & Whittaker, 2008; Eklöf, 2007; Schiel, 1996; Sundre & Kitsantas, 2004; Wolf & Smith, 1995). As a result, examinees' levels of proficiency will likely be underestimated when examinees do not give their best effort to a low-stakes test. Specifically, lack of examinee motivation can impact the reliability and validity (e.g., increase construct-irrelevant variance) of the inferences one can make from test scores. This includes inferences about gender differences in test scores.

CORRESPONDENCE

Email
demarsce@jmu.edu

Although gender differences have seldom been the primary focus of motivation studies, many studies have briefly noted gender differences in test-taking motivation among university students on low-stakes tests. The purpose of this integrative review is to bring together a variety of evidence to illustrate ways in which these differences are revealed. When available, new data are described after each section to add to the existing evidence from the published literature. Finally, potential explanations of gender differences in test-taking motivation are examined, as well as implications for higher education assessment practice.

Lack of examinee motivation can impact the reliability and validity (e.g., increase construct-irrelevant variance) of the inferences one can make from test scores.

Absence at Test Administration

Students who are extremely unmotivated may simply not show up at the assigned test administration. Swerdzewski, Harmes, and Finney (2009) studied a group of students who failed to attend an assigned testing session but later attended a make-up session. Although scores on the test had no consequences, students could not enroll for the following semester until the assessment requirement was completed; most students eventually complied and came to one of the make-up sessions. Those who attended the regular session were labeled Attenders and those who attended the make-up session were labeled Avoiders.

Male students were less likely to attend the regular session. Aggregating data from tables and text included in Swerdzewski et al. (2009), and assuming that those who provided complete data at the make-up session were representative of the total group of Avoiders and that those in the comparison group were representative of the total group of Attenders, about 30% of male students, compared to 22% of female students, failed to attend the regular testing session.

Although some students likely missed the regular testing session due to reasons other than willful noncompliance, three pieces of evidence suggest that a large portion of the non-attendance was related to motivation. First, Avoiders scored much lower than the Attenders on a fine arts test and a science test (-0.74 and -0.77 standard deviation units, respectively). Second, only 12.7% of Avoiders tried on at least 90% of the items, compared to 47.2% of Attenders. Finally, self-reported effort was 0.42 standard deviation units higher for Attenders.

It might be argued that the Avoiders skipped the assessment and then performed poorly and exerted little effort in the make-up session simply because they had low levels of knowledge. The Avoiders did have somewhat lower average grades (2.80 compared to 3.02), but this difference does not seem large enough to explain their difference in test performance. Further, SAT scores were equivalent for the two groups (1165 compared to 1163).

Overall, it appears that male students are less likely than female students to exert even the minimal effort to show up for an assigned testing session. Not attending the testing session may represent extremely low levels of test-taking motivation.

Rapid Guessing

Another indicator of very low motivation is responding to test items without taking the time to read the question. On a partly-speeded test, rapid guessing may occur toward the end of the test, if examinees run out of time. When time limits are ample or nonexistent, extremely rapid responding is more likely to indicate that the examinee put no effort into selecting an answer. Schnipke (1995) coined the term *solution behavior* to describe responses in which the examinee attempted to choose the correct answer and *rapid guessing behavior* to describe responses in which the examinee simply rapidly chose a response. Wise and Kong (2005) proposed the *response time effort* (RTE) index, the percent of items on which an examinee engaged in solution behavior.

RTE provides an unobtrusive way to collect motivation data for each item, and thus does not rely on examinee judgments of their own motivation (Kong, Wise, & Bhole, 2007; Wise & Kong, 2005). RTE is based on the notion that examinees who are not motivated will exhibit rapid guessing behavior; that is, they will rapidly respond to items without taking the time to read or fully consider the items. With this approach, response times are a proxy for motivation. Thus, rapid guessing can decrease test score validity (Wise & DeMars, 2010). In part, this is because the correctness of answers resulting from rapid guessing behavior will be at or near chance levels, as the answers are essentially random (Wise & DeMars, 2010; Wise & Kong, 2005). RTE scores have high internal consistency, are correlated with other measures of test-taking effort, and are uncorrelated with external measures of proficiency (Wise & DeMars, 2010; Wise & Kong, 2005). Because RTE is based on response times, this method is only feasible with computer-based tests where the software permits collection of response times. RTE scores have several uses: (a) to indicate levels of examinee effort, (b) to provide information on the dynamics of examinee motivation, and (c) to supply data for motivation filtering (Wise & Kong, 2005).

To compute RTE, examinee item responses are first classified as either exhibiting rapid guessing behavior (i.e., when examinees appear to supply an answer without considering the item) or solution-based behavior (i.e., when examinees attempt to find the best answer for the item; DeMars, 2007; Kong et al., 2007; Wise, 2009; Wise & Cotton, 2009). Thus, a time threshold defining response times that are too short for an examinee to have a chance to read and consider the item must be set for each item (DeMars & Wise, 2010; Swerdzewski, Harmes, & Finney, 2011). Several methods exist for setting the time threshold (see DeMars, 2007; Kong et al., 2007; Swerdzewski et al., 2011).

An index, item solution behavior (SB_{ij}), is then assigned a value of 0 or 1 based on whether an examinee's response time is below (i.e., rapid guessing behavior) or above (i.e., solution-based behavior) the threshold (DeMars & Wise, 2010; Kong et al., 2007; Wise & Kong, 2005). Because the thresholds are based on the minimum amount of time needed to read and consider the items, this index will only identify responses which the researchers are reasonably certain are noneffortful (Kong et al., 2007). The proportion of items on which the examinee exhibited solution behavior is the examinee's RTE score (DeMars, 2007; DeMars & Wise, 2010; Kong et al., 2007; Swerdzewski et al., 2011; Wise, 2009; Wise & Cotton, 2009; Wise & DeMars, 2010; Wise & Kong, 2005).

RTE values can be used in motivation filtering (Swerdzewski et al., 2011; Wise & Cotton, 2009; Wise & Kong, 2005). In fact, motivation filtering using RTE scores has been found to be favorable compared to using self-reported measures (Wise & Kong, 2005). Motivation filtering involves removing data from unmotivated examinees. Doing so should result in higher mean test scores, lower test score standard deviations, and higher correlations between test scores and external measures (i.e., convergent validity evidence) of ability when examinee effort is not related to actual proficiency (DeMars, 2007; Wise & Cotton, 2009; Wise & Kong, 2005; Wise & DeMars, 2010). In this case, motivation filtering reduces construct-irrelevant variance (Wise & Cotton, 2009). If, however, examinee effort were related to actual proficiency, data would be filtered from the lower part of the proficiency distribution, which would artificially inflate the mean of the remaining scores (Wise & DeMars, 2010). Thus, external measures of proficiency should be used to examine whether examinee effort (i.e., RTE) is related to proficiency prior to motivation filtering. For example, Wise and Kong (2005) showed that filtering students based on RTE on a university assessment made no difference in the average SAT score before and after filtering.

Importantly, gender differences can be misestimated if examinee motivation is not taken into account.

Importantly, gender differences can be misestimated if examinee motivation is not taken into account. For example, some studies found that female students exhibit more solution-based behavior than male students (Wise & Cotton, 2009; Wise & DeMars, 2010). One study found that, when motivation differences were ignored, female students showed sizeable gains between two time periods, whereas male students showed virtually no gains (Wise & DeMars, 2010). However, when examinees with the lowest RTE were removed from the data, both male and female students showed clear gains. Thus, without taking motivation into account, observed differences in mean score changes may misrepresent the actual difference in mean changes by the degree to which there are differences in rapid guessing behavior between the groups. In a study of middle school and high school students (Wise, Kingsbury, Thomason, & Kong, 2004), only 27 out of 2,382 students had RTE scores less than .90, but 23 of these 27 students were boys. Freund and Rock (1992) studied a behavior conceptually related to rapid guessing: pattern-marking (random marking of responses or systematic strings such as ABCDABCD). On the National Assessment of Educational Progress (NAEP), pattern-marking was more common among male adolescents than female adolescents, and the gender gap was greater among high school seniors than among 8th graders.

However, not all studies have found gender differences in RTE. On a test of scientific and quantitative reasoning administered under low-stakes conditions, gender was only very slightly correlated with RTE (Wise, Pastor, & Kong, 2009).

Empirical Study

RTE data were available from a science test administered to a random sample of university students and four business tests administered to students majoring in business. The science test was administered to a random sample of university students with 45-70

Far more men than women exhibited rapid guessing behavior on over half of the items.

cumulative credit hours during the spring 2009, spring 2010, and spring 2012 semesters. The test is used to directly measure objectives of the General Education program. It is low-stakes for students. The business tests were used to assess objectives from core courses taken in the first two years of the college of business curriculum. On the business tests, students who did not complete the tests had points deducted from their class grades, but the points earned did not depend on how well they scored on the test. Additionally, this group of students was required to take nine 30-item tests, spread over four weeks, outside of class time (see DeMars, 2007, for more information on this series of tests). This testing burden likely made the tests even less motivating. Only tests administered during the last week were included because tests administered in the last week tended to invoke far more rapid guessing than tests administered in the first week.

Table 1 shows the mean gender difference in RTE. Negative differences indicate lower RTE for men. The degree of the gender gap varied, but men had somewhat lower average RTE on every test. Although RTE was consistently lower for men, what is not evident from Table 1 is that the gender gap was particularly large at the low end of the RTE distribution—far more men than women exhibited rapid guessing behavior on over half of the items. For illustration, the RTE distribution is plotted in Figure 1 for Business Test Q, the test with the greatest gender difference in RTE. Although a minority of men were at the extreme low end, there were far more men than women in this extreme group. The main graph does not include examinees with RTE = 1, because the percentages in this group were much higher than the percentages with any other value of RTE. Instead, these values are shown on a bar chart inlay; although the majority of both male and female examinees had RTE = 1, more women than men were in this extremely high group. The same pattern persisted in Business Test R, as shown in Figure 2, even though the mean gender difference in RTE was smaller for this test. Overall, the gender difference in RTE was small on all tests, but it was most noticeable at low values of RTE. This matters because it is the students exhibiting extremely low effort who are likely to score much lower than they are capable of scoring.

Table 1
Gender Differences in RTE

	N		
Test	Men	Women	Gender Difference in RTE
Science	260	446	-0.01
Business Test Q	215	178	-0.10
Business Test R	214	178	-0.04
Business Test S	208	207	-0.04
Business Test T	208	205	-0.06

Self-Reported Test-Taking Effort

Although many studies do not separate the results by gender, most studies that provide scores by gender tend to show slightly higher levels of self-reported effort for female examinees.

Not attending a required test administration or rapid guessing during the test captures only the lowest levels of test-taking motivation. Self-report scales, on the other hand, may be able to capture a wider range of motivation. In some form, these scales include questions asking the examinees how hard they tried on the test, often for the purpose of studying relationships between motivation and test performance. Hoyt (2001) found that in a sample of college students taking a low-stakes General Education test, 22% reported giving little or no effort to the mathematics subtest, 8% reported giving little or no effort to the English subtest, and 15% reported little or no effort on the critical thinking subtest. Similarly, Schiel (1996), using a larger sample of over 20,000 college and university students, found the percent reporting little or no effort varied from 4-28%, depending on the subtest.

Although many studies do not separate the results by gender, most studies that provide scores by gender tend to show slightly higher levels of self-reported effort for female examinees. Wise et al. (2009) administered a measure of assessment citizenship to university students participating in mandatory low-stakes assessment. Assessment citizenship was a concept modeled on the idea of academic citizenship; students high on this trait would agree that they

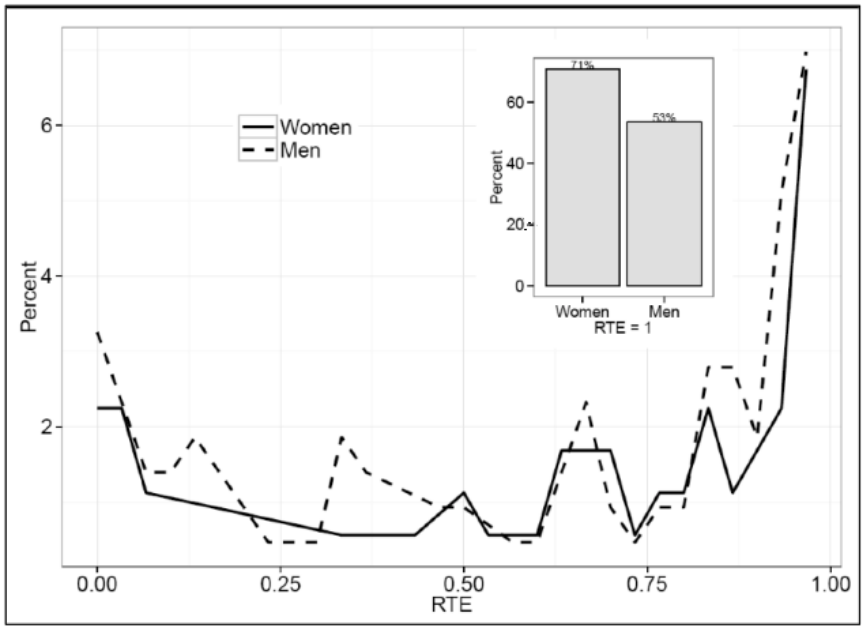


Figure 1. Distribution of RTE on Business Test Q, by gender.

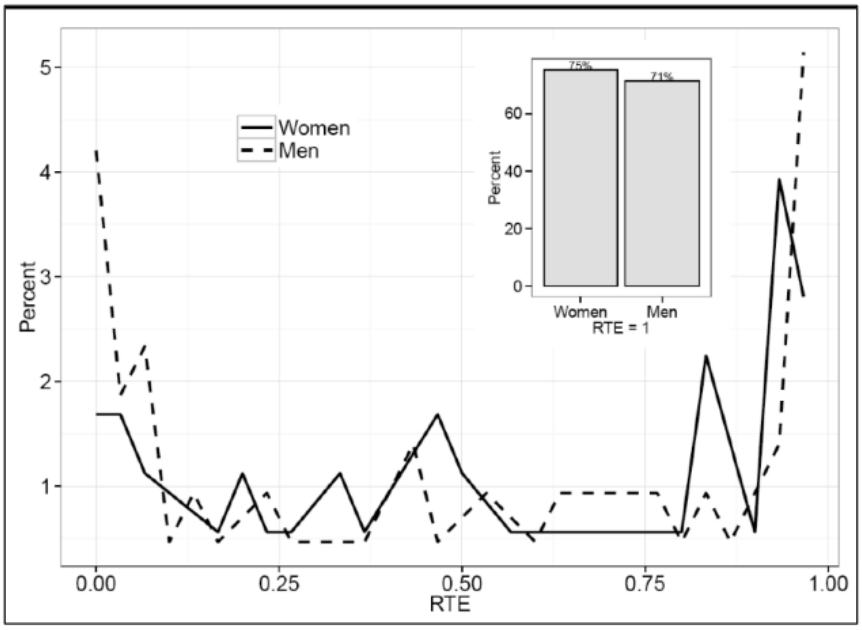


Figure 2. Distribution of RTE on Business Test R, by gender.

had a responsibility, as members of the university community, to comply with requests for participation and exert reasonable effort so that the university could collect valid data. They found a gender difference of 0.22 standard deviation units, with female students reporting more cooperativeness. Cole et al. (2008) administered four General Education tests and asked students to report effort for each test. Gender differences in effort, with negative values indicating greater effort reported by women, ranged from -0.41 standard deviations in English to 0.18 standard deviations in social studies, with intermediate standardized differences of -0.22 in math and -0.02 in science. It seems that, as with studies of RTE, the gender differences in effort vary with the subject area.

Similar results have been reported for secondary students taking low-stakes tests. Eklöf (2007) found that test-taking motivation was about 0.33 standard deviation units higher for girls than for boys among Swedish 14-15 year-olds taking the TIMSS (Trends in International

Although most men, like most women, report reasonable effort, the disproportionate gender ratio in the low range could bias estimates of gender differences in learning.

Math and Science Study) test. O'Neil, Abedi, Miyoshi, and Mastergeorge (2005) studied self-reported effort among 12th graders on released TIMSS items at low-achieving schools. Among students tested under the typical low-stakes instructions, self-reported effort was 0.21 standard deviation units lower for male students. Across many countries, 15-year-olds taking PISA (Programme for International Student Assessment) self-reported their test-taking effort as well as how hard they would have tried if the test counted toward their class grades. Butler and Adams (2007) used the difference between these values as a measure of relative effort. Girls reported slightly higher relative effort than boys. Karmos and Karmos (1984) administered a survey asking middle school students about their attitudes on standardized tests, specifically referring to a test they had recently taken. Three of the items related to effort on the test. Girls reported higher effort, with effect sizes ranging from 0.43 to 0.50 standard deviation units. Brown and Walberg (1993) found no gender differences in self-reported effort on a standardized achievement test, but they studied younger students (grades 3-8).

Empirical Study

To further examine the relationship between gender and self-reported test-taking effort, data were collected from 3,903 women and 2,345 men participating in a university assessment day in spring 2011 and 2012. To motivate students, the university's use of the results was emphasized, but the scores had no impact on student grades or other individual consequences. After completion of the 2.5 hour testing session, students reported their effort using the Student Opinion Scale (SOS; Sundre, 1997; Sundre & Moore, 2002). The scale contains five items pertaining to the student's effort during the assessments, each rated on a 5-point scale from *Strongly Disagree* to *Strongly Agree*. In previous literature, responses to this scale have shown that the item parameters are invariant across gender (Thelk, Sundre, Horst, & Finney, 2009), so comparisons of male and female examinees are reasonable.

In our data, there was very little difference in mean scores on the SOS; the mean for men was 3.62 (*SD* = 0.86) and the mean for women was 3.70 (*SD* = 0.75; Cohen's *d* = -0.10). However, male examinees' effort was more variable (variance ratio = 1.34). Figure 3 shows the distribution of effort. More men reported levels at or below the scale midpoint. More women reported levels between 3.4 and 4.8. Students at the very low end of the effort range may be the ones who could sabotage the test results. Hoyt (2001) and Schiel (1996) each found that the score gap on several tests was smallest between moderate effort and best effort; scores increased most between no effort and little effort, and again between little effort and moderate effort. In Figure 3, clearly there are more men in the problematic range. Although most men, like most women, report reasonable effort, the disproportionate gender ratio in the low range could bias estimates of gender differences in learning.

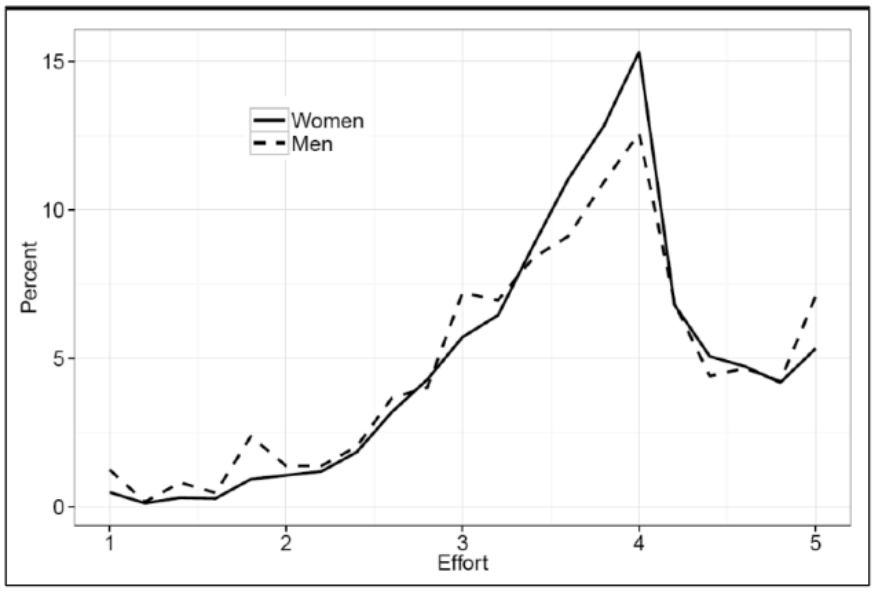


Figure 3. Distribution of self-reported effort, by gender.

As in the literature cited above, in our data self-reported effort was moderately correlated with test performance (correlations ranged from r = .22 to r = .34), but not with SAT scores (r = .05 with SAT verbal and r = .06 with SAT math). The lack of correlation between effort and SAT scores suggests that low test-taking effort yielded low test scores and not the other way around. Thus, the test scores of the subgroup of students who reported very low effort may not be representative of their knowledge. There appeared to be more male than female examinees in the very low end of the effort distribution, which may distort gender differences in test scores.

Possible Explanations of Gender Differences in Test-Taking Motivation

Given conscientiousness and agreeableness are somewhat related to effort and women score higher on these attributes, it could be that the gender difference in test-taking motivation is due in part to gender differences in personality.

Two questions of interest to both researchers and practitioners might be *why* some examinees are more willing to engage in effort on low-stakes tests and why this tendency relates to gender. Some would attribute differences in examinee motivation to individual differences or personality traits. Specifically, one might expect students who are more agreeable or conscientious to also be more compliant with requests to cooperate in test-taking. Indeed, previous research has found small and not always consistent gender differences in conscientiousness, with women typically reporting being more conscientious (Feingold, 1994) and dutiful (Costa, Terracciano, & McCrae, 2001) than men. Gender differences in agreeableness have been more prominent and consistent, with women scoring higher on agreeableness than men (Costa et al., 2001; Feingold, 1994). Further, Marrs and Sigler (2012) compared study strategies for male and female college students in efforts to explain the lower academic performance of male students. They found that female students tended to employ a "deep approach" to learning, which involved engaging in the material at a deeper level, whereas male students tended to utilize a "surface approach," which involved tasks requiring minimal effort (e.g., memorization). Marrs and Sigler also found that female students were much more academically motivated than male students (d = .44). This is not surprising, given several studies have shown that work-avoidance is negatively related to motivation and achievement (Meece, Blumenfeld, & Hoyle, 1988; Meece & Jones, 1996). Given these findings, it seems reasonable to believe that male students are also more work-avoidant, in addition to being less conscientious and less agreeable than female students. These gender differences in personality may well be the key to explaining, at least in part, the gender differences in test-taking motivation.

Empirical Study

As part of campus-wide assessment for accountability purposes in spring 2011 and 2012, students completed a battery of tests which included measures of conscientiousness, agreeableness, and work-avoidance in addition to the Student Opinion Scale. Both conscientiousness and agreeableness are subscales of the Big Five Inventory (John & Srivastava, 1999). Work-avoidance was measured via a subscale of the Achievement Goal Questionnaire (Finney, Pieper, & Barron, 2004).

Table 2
Gender Differences in Personality Traits

	Women			Men			
Trait	Mean	*SD*	*N*	Mean	*SD*	*N*	Cohen's *d*
Conscientiousness	34.03	5.47	1866	32.08	5.53	1114	.36
Agreeableness	36.21	5.35	1861	33.88	5.51	1118	.43
Work-Avoidance	11.13	4.77	3892	12.72	5.41	2333	-.32

As expected, both conscientiousness and agreeableness had about the same small positive relationship with test-taking effort (r = .22 and r = .19, respectively). Although small, both of these correlations are in the expected direction. Thus, they further support the meta-analytic findings in the literature (Costa et al., 2001; Feingold, 1994). In addition, the average conscientiousness and agreeableness scores for men and women are quite different (Table 2), with women scoring higher on both of these measures. The complete distributions of these traits are graphed in Figures 4 and 5. Unlike RTE and self-reported effort, where the gender

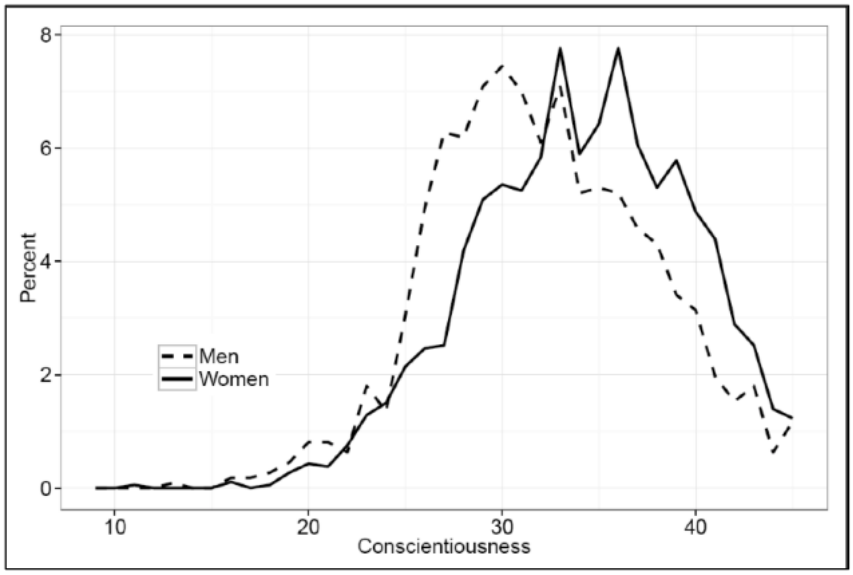


Figure 4. Distribution of Conscientiousness by gender.

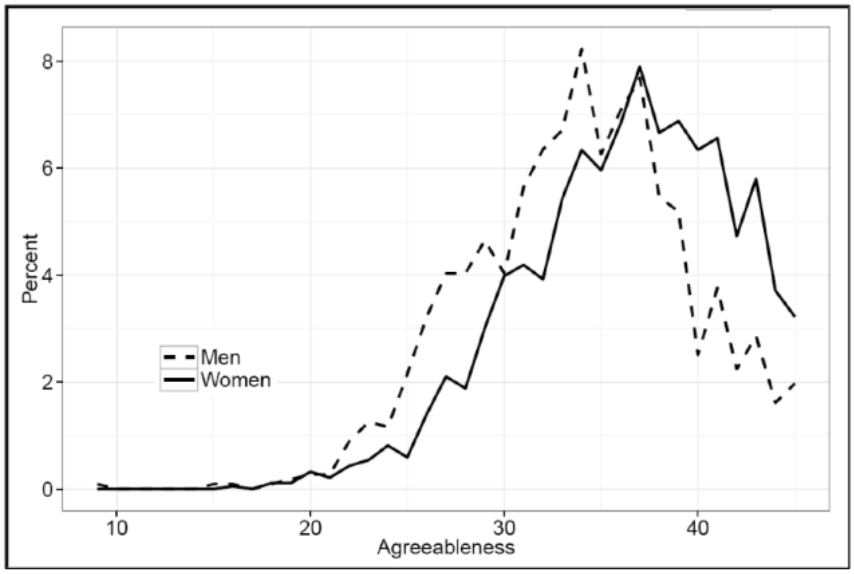


Figure 5. Distribution of Agreeablesness by gender.

differences were limited to extreme scores, these traits show fairly sizeable differences in the means. Given conscientiousness and agreeableness are somewhat related to effort and women score higher on these attributes, it could be that the gender difference in test-taking motivation is due in part to gender differences in personality. This logic is further supported by the relationship between work-avoidance and effort.

As expected, work-avoidance was negatively related to test-taking effort (r = -.23), indicating that the higher one's work-avoidance, the less effort one would likely expend on a battery of low-stakes tests. Moreover, women scored lower on this measure than men, which was not surprising based on previous research (Meece & Jones, 1996; Steinmayr & Spinath, 2008). Figure 6 shows the complete distribution of work-avoidance for men and women. Again, the negative relationship between work-avoidance and test-taking effort coupled with a noticeable gender difference in work-avoidance scores further supports the logic that personality traits may be a promising source in attempts to explain the gender gap in test-taking motivation.

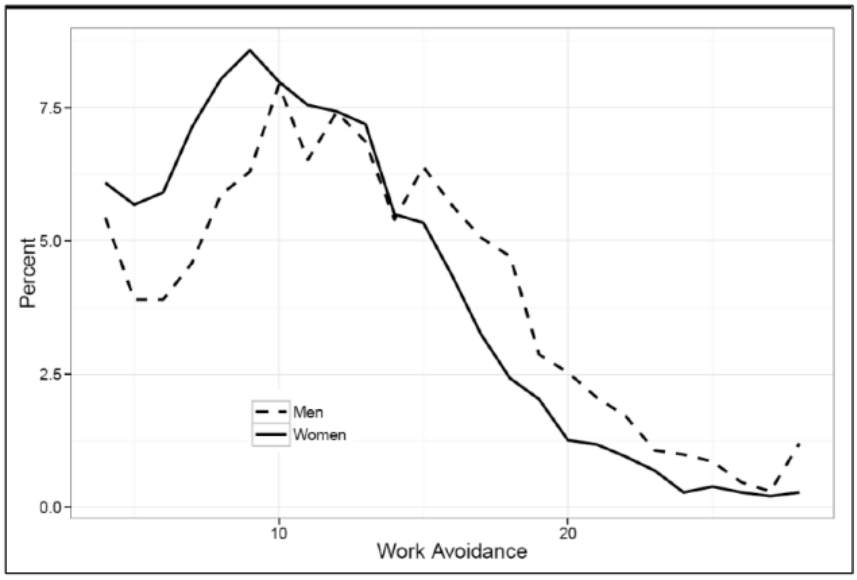


Figure 6. Distribution of Work-Avoidance by gender.

Discussion

Test-taking motivation has been the focus of considerable research in higher education assessment efforts. Previous research has linked high test-taking effort to better performance on specific tests, but not to external measures of proficiency. Thus, test-taking motivation merits close examination, in order to ensure the inferences based on low-stakes assessments are valid. In the current paper, we focused on the role of gender in test-taking motivation—an area that has received indirect attention in research but is equally important in making accurate inferences based on test scores. The purpose of the paper was to draw upon multiple sources of evidence in the existing literature reporting small but consistent gender differences in test-taking motivation and to compare these findings against our own data to investigate the phenomenon more fully. Specifically, we explored gender differences across three different indicators of test-taking motivation documented in the literature: test session attendance, rapid guessing, and self-reported test-taking effort. Where possible, we also included results from our own data, which further supported the trend of lower test-taking motivation among men than women. Finally, based on previous findings, we explored the gender gap in test-taking effort in the context of several personality traits, which we considered a possible pathway to understanding why women tend to expend more effort on low-stakes assessments.

Studies reporting examinee motivation by gender have consistently found higher self-reported effort for females than males.

We first reviewed research on the most basic level of test-taking motivation under low-stakes settings—showing up at an assigned testing session. Absence at an assigned test administration essentially indicates extremely low levels of motivation. Although only one known study has provided this type of evidence of test-taking motivation, the study revealed two compelling findings: (a) male avoiders disproportionately outnumbered their female counterparts (i.e., many more males than females failed to attend their assigned testing session); and (b) failure to attend the assigned testing session was largely related to low motivation. Thus, at the minimum level of test-taking motivation needed to show up to a testing session, males appeared to be less motivated than females. This result could be due to gender differences in personality, which we discuss later. Alternatively, it could be due to other, unmeasured variables.

Second, we examined rapid guessing via RTE, an unobtrusive indicator of test-taking motivation based on response times collected when computerized tests are administered. Specifically, an RTE score indicates the proportion of test items on which an examinee spent a minimally-adequate amount of time to read and consider the response options based on a preset time threshold for each item. Used as a proxy for motivation, RTE scores are especially useful in flagging rapid responses. Given effort is not related to proficiency in general, filtering out data from extremely unmotivated examinees (i.e., rapid responders) can reduce the construct-irrelevant variance in test scores, and thus boost the validity of inferences one

wishes to draw from responses that are now at least minimally effortful. With respect to gender, both prior research and our analyses showed that men tended to engage in rapid guessing more frequently than women. Moreover, in our data samples the gender gap was especially evident in the lower end of the distribution. These slight but fairly consistent findings across samples further support the idea that gender does indeed have a role in test-taking motivation in low-stakes conditions, with women being more motivated than men. As such, this difference should be taken into account when comparing test scores between men and women, provided item response times are available.

Next, we examined what is by far the most widely used indicator of test-taking motivation: self-report measures. Unlike the other two methods, self-report measures typically capture a wider range of examinee motivation, and thus the relationship between scores on such measures and test performance has been widely studied. Studies reporting examinee motivation by gender have consistently found higher self-reported effort for females than males. The data we analyzed also supported this trend. Specifically, we found a very small mean difference in self-reported effort (females scoring higher); however, upon examination of the distribution of effort scores, we discovered a much larger gender gap at the low end of the distribution across multiple tests, indicating men tended to report lower effort than women below the scale midpoint. Again, this gender difference in examinee motivation may appear trivial at the mean level, but it could severely bias the examination of gender differences in test scores; thus, it should be considered when making such comparisons.

Is possible to observe sizeable gender differences in performance on low-stakes assessments partly or fully due to gender differences in test-taking motivation.

Finally, the gender gap in test-taking motivation was examined in the context of personality differences in efforts to provide one plausible explanation of why such a gap in motivation exists. Both prior research and our empirical results indicated that women score higher on conscientiousness and agreeableness and lower on work-avoidance than men do. Furthermore, our analyses showed a positive relationship of effort with conscientiousness and agreeableness, and a negative relationship between effort and work-avoidance. All of these relationships were of modest magnitude but in the expected direction, based on theory and previous findings in the literature. As such, we believe these personality traits provide at least a partial explanation of why women tend to expend more test-taking effort on low-stakes assessments.

Implications for Practice

The array of findings based on prior research and new empirical data presented here clearly indicate a small but consistent gender difference in test-taking motivation under low-stakes conditions. Across a variety of measures of examinee motivation women appear to expend higher levels of effort than men. Although the size of this gender gap appears to vary across age groups and subject areas, it certainly has an impact on test scores. As demonstrated in one study (Wise & DeMars, 2010), gender differences in motivation could almost completely account for gender differences in test scores. Thus, under low-stakes testing conditions, it is of utmost importance to examine not only motivation but also the effect of gender, especially when there is interest in comparing test scores by gender. Assessment practitioners could control for effort by filtering noneffortful responses through RTE screening, when response time data are available, or by collecting other measures of test-taking motivation (e.g., self-report measures), which could then be used as covariates in the analyses.

In addition to applying statistical methods to control for effort in low-stakes conditions, researchers have proposed several approaches to enhance test-taking motivation directly. Such methods include increasing the stakes of the test (e.g., requiring a passing score or including the score in a course grade), conveying to students the importance of assessment for curricular improvement, providing valuable feedback to students regarding their performance, offering monetary incentives, or utilizing a computer-based testing environment which prompts students to expend more effort when they engage in rapid guessing behavior (Wise, 2009). For paper-and-pencil test administrations, researchers have discovered that proctors overseeing the test sessions have a significant impact on the engagement and motivation of examinees, and as a result, on their test scores (Lau, Swerdzewski, Jones, Anderson, & Markle, 2009). Any and all of these methods could be applied in practice in higher education assessment under low-stakes conditions to improve the validity of inferences drawn from test scores.

These are but a few examples of how the findings from the literature and the new empirical evidence presented here could benefit practitioners of higher education assessment. Perhaps the most important take-home message for practice is to be aware that it is possible to observe sizeable gender differences in performance on low-stakes assessments partly or fully due to gender differences in test-taking motivation. Fortunately, there are various methods to empirically investigate this possibility and control for a motivation effect moderated by gender. We presented three such methods, as well as recommendations for ways to increase test-taking motivation in efforts to combat threats to validity of score comparisons and overall test score interpretation. We encourage future research to explore the extent to which these and other motivational enhancement efforts developed in recent years are effective in narrowing the gender gap in test-taking motivation under low-stakes conditions and reducing the construct-irrelevant variance introduced by low levels of effort.

Gender is often used as a proxy variable in research for the very reason that men and women do differ on a wide range of variables that may be difficult to obtain compared to simply recording students' gender.

Limitations and Future Directions

While we identified numerous sources of evidence suggesting a consistent pattern of low test-taking motivation among men compared to women, as well as likely explanations for this pattern, our investigation was limited in several ways. First, we were unable to conduct a thorough meta-analysis of the examinee motivation literature as it pertains to gender differences simply because results are rarely broken down by gender in most published research. It is our hope that once higher education and assessment professionals become aware of the small but consistent gender differences in test-taking motivation under low-stakes conditions, more evidence will be cumulated and this phenomenon will be investigated and understood more fully.

Furthermore, we were able to explore only three personality traits that could allude to the gender gap in test-taking effort. Other important variables certainly exist that could account for gender differences. In fact, gender is often used as a proxy variable in research for the very reason that men and women do differ on a wide range of variables that may be difficult to obtain compared to simply recording students' gender (Bashkov & Finney, 2013). However, the three personality variables discussed in this study appeared essential to understanding at least in part why men and women expended different amounts of effort on the assessments. Future research should explore these and other personality traits further, in order to reach a better understanding of the role of gender in test-taking motivation under low-stakes conditions.

References

Bashkov, B. M., & Finney, S. J. (2013). Applying longitudinal mean and covariance structures (LMACS) analysis to assess construct stability over two time points: An example using psychological entitlement. *Measurement and Evaluation in Counseling and Development, 46*, 289-314. doi: 10.1177/0748175613497038

Brown, S. M., & Walberg, H. J. (1993). Motivational effects on test scores of elementary students. *Journal of Educational Research, 86*, 133-136. doi: 10.1080/00220671.1993.9941151

Butler, J., & Adams, R. J. (2007). The impact of differential investment of student effort on the outcomes of international studies. *Journal of Applied Measurement, 8*, 279-304.

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*, 609–624. doi: 10.1016/j.cedpsych.2007.10.002

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

Costa, P. T. Jr., Terracciano, A., & McCrae, R. (2001). Gender differences in personality traits across cultures: Robust and surprising findings. *Journal of Personality and Social Psychology, 81*, 322-331. doi: 10.1037/0022-3514.81.2.322

DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education, 13*, 55-77. doi: 10.1207/s15324818ame1301_3

DeMars, C. E. (2007). Changes in rapid-guessing behavior over a series of assessments. *Educational Assessment, 12*, 23-45. doi: 10.1080/10627190709336946

DeMars, C. E., & Wise, S. L. (2010). Can differential rapid-guessing behavior lead to differential item functioning? *International Journal of Testing, 10*, 207-229. doi: 10.1080/15305058.2010.496347

Eklöf, H. (2007). Test-taking motivation and mathematics performance in TIMSS 2003. *International Journal of Testing, 7*, 311–326. doi: 10.1080/15305050701438074

Feingold, A. (1994). Gender differences in personality: A meta-analysis. *Psychological Bulletin, 116*, 429-456. doi: 10.1037/a0022247

Finney, S. J., Pieper, S. L., & Barron, K. E. (2004). Examining the psychometric properties of the Achievement Goal Questionnaire in a general academic context. *Educational and Psychological Measurement, 64*, 365-382. doi: 10.1177/0013164403258465

Freund, D. S., & Rock, D. A. (1992, April). *A preliminary investigation of pattern-marking in 1990 NAEP data*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA. (ERIC Document Reproduction Service No. ED 347189)

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*, 71-85.

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality theory and research* (Vol. 2, pp. 102–138). New York, NY: Guilford Press.

Karmos, A. H., & Karmos, J. S. (1984). Attitudes toward standardized achievement tests and their relation to achievement test performance. *Measurement and Evaluation in Counseling and Development, 17*, 56-66.

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

Lau, A. R., Swerdzewski, P. J., Jones, A. T., Anderson, A. D., & Markle, R. E. (2009). Proctors matter: Strategies for increasing examinee effort on general education program assessments. *Journal of General Education, 58*, 196-217. doi: 10.1353/jge.0.0045

Liu, O. L., Bridgeman, B., & Adler, R. M. (2012). Measuring learning outcomes in higher education: Motivation matters. *Educational Researcher, 41*, 352-362.

Marrs, H., & Sigler, E. A. (2012). Male academic performance in college: The possible role of study strategies. *Psychology of Men & Masculinity, 13*, 227-241. doi: 10.1037/a0022247

Meece, J. L., Blumenfeld, P. C., & Hoyle, R. H. (1988). Students' goal orientations and cognitive engagement in classroom activities. *Journal of Educational Psychology, 80*, 514-523. doi: 10.1037/0022-0663.80.4.514

Meece, J. L., & Jones, M. G. (1996). Gender differences in motivation and strategy use in science: Are girls rote learners? *Journal of Research in Science Teaching, 33*, 393-406. doi: 10.1002/(SICI)1098-2736(199604)33:4<393::AID-TEA3>3.0.CO;2-N

O'Neil, H. F., Abedi, J., Miyoshi, J., & Mastergeorge, A. (2005). Monetary incentives for low-stakes tests. *Educational Assessment, 10*, 185-208. doi: 10.1207/s15326977ea1003_3

Rothe, H. F. (1947). Distribution of test scores in industrial employees and applicants. *Journal of Applied Psychology, 31*, 480–483.

Schiel, J. (1996). *Student effort and performance on a measure of postsecondary educational development* (ACT Rep. No. 96-9). Iowa City, IA: American College Testing Program.

Schnipke, D. L. (1995, April). *Assessing speededness in computer-based tests using item response times*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco. (ERIC Document Reproduction Service No. ED383742)

Steinmayr, R., & Spinath, B. (2008). Sex differences in school achievement: What are the roles of personality and achievement motivation? *European Journal of Personality, 22*, 185-209. doi: 10.1002/per.676

Sundre, D. L. (1997, April). *Differential examinee motivation and validity: A dangerous combination*. Paper presented at the annual meeting of the American Educational Research Association. Chicago, IL.

Sundre, D.L. (1999, April). *Does examinee motivation moderate the relationship between test consequences and test performance?* Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec, Canada. (ERIC Document Reproduction Service No. ED432588)

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. doi: 10.1016/S0361-476X(02)00063-2

Sundre, D. L., & Moore, D. L. (2002). The Student Opinion Scale: A measure of examinee motivation. *Assessment Update, 14*, 8-9.

Swerdzewski, P. J., Harmes, J. C., & Finney, S. J. (2009). Skipping the test: Using empirical evidence to inform policy related to students who avoid taking low-stakes assessments in college. *Journal of General Education, 58*, 167-195.

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*, 162-188. doi: 10.1080/08957347.2011.555217

Terry, N., Mills, L., & Sollosy, M. (2008). Student grade motivation as a determinant of performance on the Business Major Field ETS Exam. *Journal of College Teaching & Learning, 5*, 27-32.

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*, 129-151. doi: 10.1353/jge.0.0047

Wise, S. L. (2009). Strategies for managing the problem of unmotivated examinees in low-stakes testing programs. The *Journal of General Education, 58*, 152-166. doi: 10.1353/jge.0.0042

Wise, S. L., & Cotton, M. R. (2009). Test-taking effort and score validity: The influence of student conceptions of assessment. In D. M. McInerney, G. T. L. Brown, G. Arief, & D. Liem (Eds.), *Student perspectives on assessment: What students can tell us about assessment for learning* (pp. 187-205). Charlotte, NC: Information Age Publishing.

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

Wise, S. L., Kingsbury, G. G., Thomason, J., & Kong, X. (2004, April). *An investigation of motivation filtering in a statewide achievement testing program*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego.

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. doi: 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*, 185-205. doi: 10.1080/08957340902754650

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

Wolf, L. F., Smith, J. K., & DiPaulo, T. (1996, April). *The effects of test specific motivation and anxiety on test performance*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.