DISSERTATION

CHEATING ON ONLINE ASSESSMENT TESTS: PREVELANCE AND IMPACT ON VALIDITY

Submitted by

Thomas M. Cavanagh

Department of Psychology

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 2014

Doctoral Committee:

Advisor: Kurt Kraiger

Alyssa Gibbons
Kim Henry
Travis Maynard

ABSTRACT

CHEATING ON ONLINE ASSESSMENT TESTS: PREVELANCE AND IMPACT ON VALIDITY

Online tests are a relatively efficient way to assess large numbers of job candidates and are becoming increasingly popular with organizations. Due to their unproctored nature, however, online selection tests provide the potential for candidates to cheat, which may undermine the validity of these tests for selecting qualified candidates. The purpose of this study was to test the appropriateness of utility theory as a framework for understanding decision-making in regard to cheating on an online cognitive ability test (CAT) by manipulating the probability of passing the test with cheating, the probability of being caught cheating, and the value of being caught cheating in two samples: 518 adults recruited through Amazon mTurk, and 384 undergraduate students. The probability of being caught cheating significantly affected performance on the CAT for the mTurk sample, but not for the student sample, and significantly moderated the relationship between CAT score during session one and CAT score during session two for the student sample. Neither the probability of being caught cheating nor the value of being caught cheating significantly affected CAT performance or validity in either sample. Findings regarding the prevalence and effectiveness of cheating are discussed.

ACKNOWLEDGMENTS

It might be a cliché, but no one makes it through graduate school alone. During my time at Colorado State University, I have been blessed with a loving and supporting community. I would like to thank Kurt Kraiger for being such a supportive and nurturing advisor. He gave me a lot of rope, but never quite enough with which to hang myself. I would not be where I am today without his help and guidance. I would also like to thank Dr. Kim Henry, whose dedication to teaching and to her students is an inspiration. She is a role model that I will aspire to live up to for the rest of my life, both personally and professionally. Thanks to Dr. Zinta Byrne, for her honest and caring feedback, her unflagging efforts to polish my rough edges, and all the opportunities she gave me to succeed. Thanks to my true cohos, Janet Peters, Christy Smith, and Erica Solove. I would have been lost without their help and your friendship. Thanks to Christa Kiersch, Stefanie Putter, Natalie Wolfson, and Paige Lancaster for showing us the way. Thanks to Cameron Kiersch, Jay Peters, Ben Panico, Will Lancaster, and Kyle Sandell for bringing a bit of testosterone, tomfoolery, and wanton destruction into my graduate school years. I have no doubt I would have spent my free time drinking boxed wine and painting toenails if it were not for their influence. Finally, thanks to Casey Onder for always believing in me.

TABLE OF CONTENTS

INTRODUCTION

Unproctored Internet testing (UIT) is a relatively new selection procedure whereby selection tests are administered to job candidates via the Internet and without the presence of a human proctor (Tippins, 2009; Tippins et al., 2006; Lievens & Burke, 2011). The evolving use of technology in selection procedures was noted as early as 2003 (Chapman & Webster, 2003). By 2008, 100% of Fortune 500 companies employed some sort of online application procedure (Younger, 2008), and two-thirds of all employers used some sort of Internet testing as part of their application procedures (Fallaw, Solomonson, & McClelland, 2009). Recent research published by practitioners shows continued interest in the topic (e.g., Hense, 2009; Gibby, 2009; Reynolds, 2009).

Benefits of UIT

For organizations, there are many perceived benefits to UIT that make it attractive as a selection procedure, including reduced costs, an increased applicant pool, and consistency in administration and scoring of selection tests.

Reduced Costs

UIT reduces the costs of test administration and screening time for job applicants (Gibby et al., 2009). Organizations do not need to hire and train proctors, or send them to testing locations; testing equipment does not need to be purchased, distributed, or maintained; and compared to traditional testing programs, it is cheaper and easier to update and adjust Internet delivered selection tests (Tippins, 2009). Internet testing is also "scalable," which means that organizations can drastically increase the number of candidates who complete the selection test without an accompanying increase in administration costs. The cost of maintaining an Internet

test is roughly the same regardless of how many candidates actually complete the test (Naglieri et al., 2004).

There are multiple empirical and case studies that demonstrate these savings. Bank of America, for example, replaced telephone screening with more efficient and objective unproctored Internet screening for candidates for a customer service position, and were rewarded with a statistically significant 68% drop in mean time spent screening candidates (from 23 minutes to eight minutes), without any drop in predictive validity (Hense, 2009). Proctor and Gamble instituted an unproctored Internet cognitive ability test, and, in one year, reduced the number of supervised paper-and-pencil tests administered in Japan by 10,000 (Gibby, 2009).

Increased Applicant Pool

After time and cost savings, one of the most lauded benefits of UIT is its potential to increase the size of the applicant pool (Chapman & Webster, 2003; Naglieri et al., 2004; Tippins, 2006). Because UIT can be completed anytime, it makes selection tests available to individuals who might not be able to attend a proctored testing administration during normal business hours (Tippins, 2009). Because UIT can be completed anywhere, it opens up selection tests to individuals from geographically diverse regions, including applicants from rural areas, international applicants, and, importantly, candidates with physical disabilities that might make it difficult to travel to a proctored testing location (Chapman & Webster, 2003; Naglieri et al., 2004; Reynolds, 2009; Tippins, 2009b). Furthermore, if the applicant pool substantially increases, but the number of candidates selected remains constant, then, as long as more highly qualified candidates can be identified, the utility of the test increases (Tippins, 2009b).

Consistency in Test Administration/ Scoring

Reduced costs would be a dubious benefit of UIT if it came at the price of the quality of test administration. In addition to costs savings, though, UIT is extremely useful for ensuring consistency in both the administration and scoring of selection tests (Naglieri et al., 2004; Tippins, 2009). UIT, for example, can be used to precisely standardize instructions provided to test-takers and enforcement of time limits (Reynold et al., 2009). Technology utilized in UIT can be used to score tests objectively, accurately, and almost immediately (Naglieri et al., 2004; Tippins, 2009). UIT can also eliminate inconsistencies in test administration and scoring that arise from test administrator biases in regard to candidate characteristics such as race, weight, or age (Chapman & Webster, 2003).

Summary

By considering the benefits of reduced costs, increased applicant pool, and consistency in test administration and scoring, it is not difficult to understand why UIT is increasingly used by organizations as a selection procedure. These benefits, however, must be weighed against the drawbacks of UIT, which will be discussed next.

Drawbacks of UIT

UIT has many benefits, but they come at a price. Potential drawbacks of UIT include problems with technology, compromised test security, lack of environmental standardization, and cheating.

Problems with Technology

In order to complete UIT, candidates need Internet access and an electronic device capable of connecting to the Internet. These requirements can lead to problems with Internet connectivity and computer processing speed (Tippins, 2006; Tippins, 2009). This in turn can

cause candidate frustration and reduce completion rates (Hense, 2009), and can compromise the standardization of administration, which can in turn compromise test validity (Potosky & Bobko, 2004). For example, if candidates need to watch a video or examine a figure in order to answer a test item, but the video or figure fails to load for some candidates, they will be unlikely to answer that item correctly, regardless of their underlying ability. If this happens on several items during the test administration, candidates might give up without completing the test.

Compromised Test Security

Another major drawback of UIT is that administering tests online might severely threaten test security. When test content is placed online, candidates can copy it in order to study it themselves or share it with other candidates (Lievens & Burke, 2011; Naglieri et al., 2004). In the case of proprietary tests, competitors might be able to copy the test material in order to use in their own product design or marketing (Chapman & Webster, 2003). This proliferation of test materials potentially undermines the validity of the selection instrument, especially in the case of cognitive ability tests (Lievens & Burke, 2011). Candidates who have access to test items prior to test administration can find the correct answers to these items using forbidden outside resources, allowing them to achieve a score on the test that is not truly reflective of the underlying knowledge, skill, or ability the test was designed to measure.

Lack of Environmental Standardization

Though UIT helps ensure standardization in instructions across testing situations, it does nothing to ensure the environment in which candidates complete the test is standardized (Naglieri et al., 2004). Good testing practices require that candidates complete the test under conditions that facilitate their best possible performance on the test (Tippins, 2009). With UIT, however, candidates are allowed to take the test wherever they please, which means that the test-

taking environment could be full of noise, people, or other distractions, in addition to the technological idiosyncrasies listed above (e.g., Internet connection speed, processing speed of the electronic device used to complete the test) (Potosky & Bobko, 2004; Reynolds et al., 2009; Tippins, 2009b). These environmental conditions can affect candidate performance, which in turn can affect the reliability and validity of a candidate's test score (Beaty, 2011; Lievens & Burke, 2011; Naglieri et al., 2004).

Discrimination

Though one of the benefits of UIT is that it enlarges the applicant pool by making the test more convenient for candidates to complete, there is currently a debate about for which candidates UIT is more convenient. Because UIT requires an electronic device with Internet access, UIT may unfairly deny employment opportunities to candidates who do not have such access. These include candidates of lower socioeconomic status, older candidates, minorities, and candidates outside of the United States (Naglieri et al., 2004; Nye, 2008; Reynolds et al., 2009). Even when candidates within these groups can access the test, their lack of familiarity with modern technology may negatively influence perceptions of the organization and test performance (Naglieri, 2004; Nye, 2008). Although outside the scope of this investigation, this phenomenon raises important questions about UIT and its potential for adverse impact (Chapman & Webster, 2003). Minorities in the categories listed above might perform poorly on UIT because they are unfamiliar or uncomfortable with the technology upon which those tests are administered (e.g., difficulty navigating the testing interface), and not because of any deficit in the underlying ability the test is designed to measure. Qualified minority candidates could be unfairly denied job opportunities because of a characteristic (e.g., familiarity with the technology upon which the test is administered) that is unrelated to job performance.

Cheating

By far the greatest concern discussed in the literature on UIT is the impact of cheating on test scores (for example Beaty et al., 2011; Lievens & Burke, 2011; Naglieri et al., 2004; and for a thorough discussion, see Tippins et al., 2006). The next section of this manuscript will cover that topic in detail.

Summary

UIT offers many benefits to organizations, but comes with accompanying challenges and drawbacks, including problems with technology, compromised test security, lack of environmental standardization, and cheating. Cheating is perhaps the greatest concern of both practitioners and researchers, and will be discussed in detail below.

Cheating

Cheating is routinely noted as one of the biggest concerns associated with UIT (Tippins, 2009a; Tippins, 2009b; Tippins et al., 2006). Although often discussed, cheating is rarely explicitly defined, at least within the organizational literature.

Tippins (2006) defined cheating as any strategy, "by which people attempt to 'game' or compromise the testing situation for their personal advantage (or the advantage of others), resulting in test scores that do not accurately reflect an individual's standing on whatever the test is measuring," (p. 206). Lievens and Burke (2011) defined it as, "obtaining a score through prohibited materials, others' help or others impersonating applicants so that applicants' scores do not reflect their standing on the construct" (pp. 817-818). The educational literature can be of some help here, in which cheating is defined as the use of prohibited materials or assistance to undermine the assessment process (Garavalia et al. 2007). Drawing upon each of these definitions, in this manuscript, I define cheating as the purposeful use of prohibited materials or

assistance to undermine the validity of the assessment. Cheating is problematic precisely because it compromises test validity. Since selection tests are designed to identify highly qualified candidates for job positions; cheating undermines their ability to do so, meaning that the candidates who score well are not necessarily the best candidates for the job (Lievens & Burke, 2011).

Methods of Cheating

Methods of cheating are limited only by test-takers' imagination. Several of the more common methods, including the use of outside resources, taking advantage of compromised test security, and using a test surrogate, are discussed below.

Use of Outside Resources

UIT is, by definition, unproctored, meaning that there is no proctor present to guarantee that the candidate completes the test without the use of prohibited materials (Lievens, 2002). Candidates completing UIT have Internet access, which means that they have access to a nearly limitless knowledge through browser search engines; this access may provide them answers to test questions. Outside resources could also include reference books or even a knowledgeable friend (Lievens & Burke, 2011). The use of these resources may very well alter a candidate's score so that it no longer accurately reflects the candidate's standing on whatever construct is being tested (Lievens, & Burke).

Compromised Test Security

Compromised test security refers specifically to a situation in which test-takers have access to test questions prior to completing the test (Drasgow, Nye, Guo, & Tay, 2009; Lievens & Burke, 2011; Naglieri et al., 2004; Tippins, 2009). As with access to outside resources, having access to test questions before officially taking the test could undermine test validity (Lievens &

Burke, 2011) by allowing candidates to become more comfortable with test content, and to memorize answers in advance of the test (Tippins, 2009). Note too that endorsing a test with compromised test security could be an ethical violation. The APA *Ethical Principles of Psychologists and Code of Conduct* (American Psychological Association, 2002), section 9.11 specifically requires that psychologists make reasonable efforts to maintain the integrity and security of test materials.

Use of Test Surrogate

A test surrogate refers to someone other than the identified candidate completing the selection instrument (Tippins, 2006). As several authors have noted, it is nearly impossible to verify candidate identity during UIT (Beaty, 2011; Lievens & Burke, 2011; Naglieri, 2004; Tippins, 2006; Tippins, 2009). The consequences of this method are obvious: though the test-taker may have performed extremely well on the selection instrument, it is difficult to verify that the person being hired and the test-taker are the same individual, or that the score on the UIT reflects the candidate's knowledge or ability to perform well on the job. In this case, the validity of the test might be extremely low for those job candidates who used surrogates, because the test score is not reflective of their individual abilities.

Summary

Candidates can cheat in several ways, including using outside materials, taking advantage of compromised test security, and using a test surrogate. Cheating is useful for candidates because it potentially allows them to exaggerate their qualifications for a specific job. Cheating is problematic for organizations for the exact same reason: if candidates cheat, their scores are not reflective of their qualifications for employment, meaning that the candidates selected are not

necessarily the most qualified for the job, wasting organizational resources and negatively impacting organizational effectiveness.

Prevalence of Cheating on UIT

There is currently little to no empirical evidence on the extent of cheating, the conditions that encourage or discourage cheating, and/or the impact of cheating on the validity and effectiveness of UIT (Tippins, 2009b). Despite several calls for such research (e.g., Naglieri et al., 2004; Tippins, 2006), there have been only a handful of studies on the topic, all of which are limited methodologically (Arthur, 2009; Beaty, 2011; Nye, 2008). To confuse the topic even more, these studies have often found contradictory results. Some researchers have found no differences on proctored vs. unproctored tests, others have found higher scores in the unproctored condition, and still others have found higher scores in the proctored condition (Do, 2009). Many of these studies exist only as conference presentations, and are thus not widely distributed, and may also be suspect methodologically. Three representative studies, their findings, and limitations, are discussed below.

Arthur (2009)

A within-subjects design study (N=296) was conducted in which test takers were considered to have experienced high-stakes testing and low-stakes retesting. Specifically, participants first completed an unproctored cognitive ability test administered over the Internet as job applicants (time one; high stakes), then as research participants (time two; low stakes). The test was speeded to make cheating more difficult. Candidates were applying for a variety of positions across numerous industries.

Candidates who scored more than one standard error of the measurement lower on time two vs. time one were considered likely cheaters. Using this metric, 7.77% of participants were

considered likely cheaters. It is important to note that this could be because of actual cheating, or because at time two individuals were less motivated to perform well on the test (e.g., they had already received jobs). Because it is impossible to distinguish between the two possible causes of score differences, 7.77% represents the upper limit for cheating, with the actual level falling anywhere between 0 and 7.77%.

Limitations. The study included no control group, the researchers did not manipulate aspects of the testing situation to make cheating more or less attractive, and only those candidates who were hired were retested.

Do (2005)

Study 1: Undergraduate sample. Using an undergraduate sample, the researcher compared performance of proctored (n=252) and unproctored conditions (n=163) on cognitive ability tests. To motivate the students, high performing test-takers were entered into a lottery for a $100 prize.

Despite no significant differences in self-reported SAT or ACT scores, participants in the proctored condition performed significantly better than those in the unproctored condition, though the effect size was small (d=-.25).

Study 2: Field sample. In a second study using a field sample, the researcher analyzed data from 12,620 job incumbents and applicants for an entry-level management position with a retail organization. The proctored condition included 3,116 individuals and the unproctored condition included 9,504.

For the cognitive portion of the test, participants in the unproctored condition scored slightly higher (than those in the proctored condition. Probably owing to the large sample size, this small difference was significant, though the effect size was quite small (d=-.09).

Limitations. In this study, the only scores compared were mean scores on proctored vs. unproctored conditions. Many variables aside from cheating could have affected test performance. Furthermore, the researcher was only able to detect effective cheating, and was not able to estimate prevalence of cheating or identify possible cheaters. There was also no analysis of how cheating effected the validity of the test.

Nye (2008)

Eight-hundred and fifty-six European job applicants seeking positions as customer service agents for a large international call center in the UK were administered two parallel forms of an online speeded attention to detail test. The first time they completed the test they did so in an unproctored environment, at a time and place of their own choosing. The second time they completed the test, they did so in a proctored setting at the company's staffing agency.

After controlling for regression to the mean, participants in the proctored condition performed significantly better than those in the unproctored condition, although the effect size was small (d=.29). Though, there was no evidence of cheating at the group level, the researchers analyzed changes in individual scores to detect likely cheaters, who were defined as individuals whose scores changed by more than 1.96 standard deviations between testing conditions. Of the 856 applicants in the dataset, only four met this criterion. The researchers therefore concluded that cheating was almost nonexistent in this study.

Limitations. Because there was no control condition that took the test twice under proctored conditions, to the researchers estimated and statistically corrected for practice effects and regression to the mean. It is impossible to know how accurate the estimate and statistical corrections were. Furthermore, the researcher did not manipulate any aspects of the testing situation that could possibly affect the likelihood of cheating.

Summary

Though the studies above give us some estimate of the prevalence of cheating, they lack the rigorous methodological control available in lab studies. Without a control group for comparison, it is difficult to parse out score changes due to the unreliability of measures, practice effects, changes in motivation, and regression to the mean from those due to cheating. Furthermore, differences in-group means only reflect effective cheaters (i.e., those individuals who were able to use cheating to effectively alter their score), not those individuals who cheated ineffectively (i.e., without changing their score, or even inadvertently lowering their score).

Detecting Cheating

Detecting possible cheaters is a difficult tasking, and proving cheating occurred is often an impossible one (Haney & Clark, 2007; Tippins, 2009b). Several methods, however, have been developed to estimate the prevalence of cheating and will be adopted into this study. They are discussed below.

Detecting cheating in UIT is important for two reasons. The first is to estimate the prevalence of cheating at a group level in order to determine the utility of UIT as a selection procedure and estimate the impact of cheating on the validity of the selection system. The second is to identify individual cheaters in order to eliminate them from the applicant pool or have them complete the selection test under different conditions in order to obtain more accurate scores.

The first objective is relatively easy to achieve using statistical analyses (Guo & Drasgow, 2010; Haney & Clark, 2007). For example, if two groups take the same selection test under conditions that manipulate the ease of cheating (e.g., proctored vs. unproctored conditions), and the mean score of the group in which cheating is relatively easy is significantly higher than that of the group in which cheating is relatively difficult, it is likely that cheating was

more prevalent in the condition in which cheating was relatively easy (Arthur, Glaze, Villado, & Taylor, 2009; Beaty, Fallon, Shepherd, & Barrett, 2002; Haney & Clark, 2007). Likewise, if the same group takes a selection test in an unproctored condition, and then in a proctored condition, we would expect the mean score to change very little (assuming high reliability), or even to improve upon the second administration (due to practice effects) (Arthur, 2009; Nye, 2008). If we see the opposite pattern (the mean score in the proctored condition drops significantly), it is likely that cheating occurred in the unproctored condition, and this led to inflation in the mean score. If appropriate criteria are available, statistical analyses can also be used to estimate the validity of the test (e.g., the correlation between the test and a specific criterion, the factor structure of the test, etc.) under different testing conditions. In this case, organizations might not only estimate whether cheating occurs and leads to score inflation, but how cheating affects the validity of the test (e.g., Beaty et al., 2011).

Compared to detecting cheating at the group level, detecting individual cheaters in testing scenarios is considered to be much more difficult, if not impossible (Haney & Clark, 2007; Tippins, 2009b). Statistical procedures can often suggest that cheating occurred within a group, but cannot identify individual cheaters. For example, if the mean score of an unproctored condition is significantly higher than that of a proctored condition, it suggests that cheating occurred in the unproctored condition. However, there is no way to determine with certainty which individuals' scores are accurate, and which individuals' scores are artificially inflated through cheating. Even in a test-retest situation, significant differences in individuals' scores from time one to time two might be reflective of cheating, or they might be reflective of practice effects, changes in motivation or personal well-being, or other sources of measurement error (Arthur, 2009; Haney & Clark, 2007). Because accusing specific individuals of cheating is

associated with potentially life changing consequences, these accusations can make organizations vulnerable to civil or even criminal claims, not to mention the various ethical implications of doing so (Haney & Clark, 2007; Tippins, 2009b). Organizations, therefore, should be extremely confident that an individual cheated before making such an accusation. This type of confidence can rarely be achieved through statistical analysis alone.

The current study will use methods designed to detect cheating at both the group and individual level. These methods are described below.

Group Mean Differences Between Proctored and Unproctored Conditions

Assuming cheating raises test scores, then comparing the mean score between two different testing conditions should be an effective way to detect the presence of cheating, with the mean score of the condition with more cheaters being significantly higher than the mean score of the condition with fewer cheaters (Beaty, 2011). If we suspect that cheating is more common in unproctored settings, then we can compare a proctored and unproctored condition, and, if cheating really is more common in the unproctored condition, we would expect to see a significantly higher mean score for that condition (see, for example, Arthur, 2009; Beaty, 2011; Do, 2005). This technique, however, cannot be used to identify cheating at the individual level (Haney & Clark, 2007; Tippins, 2009b).

In the current study, there will be several conditions in which the participants take the test in proctored and unproctored situations, allowing for group mean comparisons across those conditions to estimate the impact of cheating at the group level.

Within Person Differences in Test Performance

Another way of detecting cheating is to look at within person differences in test performance across testing conditions (Haney & Clark, 2007). If candidates score significantly

higher in a condition that potentially made cheating easier (e.g., an unproctored condition) than in a condition that potentially made cheating more difficult (e.g., a proctored condition), it is possible that these candidates cheated (Naglieri et al., 2004; Hense, 2009; Lievens & Burke, 2011). Though this method allows for the identification of individuals who likely cheated, it is problematic in that researchers must decide how much higher the individual must score in the condition that potentially made cheating easier in order to be labeled a likely cheater. This difference has traditionally been set at three standard deviations (Hartshorne & May, 1928, as quoted in Haney & Clark, 2007), though others have used a difference of 1.96 standard deviations (see above, Nye, 2008), The second problem with this method is that scores might change on the second testing for reasons that have nothing to do with cheating, such as practice effects, regression to the mean, or changes in motivation (Nye, 2008).

In the current study, a number of participants will take a cognitive ability test twice in a proctored situation (proctored/ proctored), whereas others will take the test once in an unproctored situation, and again in a proctored situation (unproctored/ proctored). This will allow the researchers to analyze not only if there are significant within-person score changes amongst those participants who took the unproctored test followed by the proctored test, but also how those score changes compare to score changes amongst participants who took the proctored test twice. This will help eliminate statistical explanations for any score changes.

Similar Incorrect Answers

When, despite such a method being prohibited, candidates have the opportunity to work with others on the test, similar wrong answers might be indicative of individuals who collaborated on the test. This method was famously used by Jacob and Levitt (2002), in part, to identify unusual answer strings on standardized tests that indicated that public school teachers

had been changing answers on students' tests to artificially inflate their grades. This method has been used in educational settings for quite a long time (e.g., Bird, 1927, as cited in Haney & Clark, 2007), and is based on the idea that students should share incorrect answers at a no more than chance level (Haney & Clark, 2007). It is important to note that, because some incorrect answers are more likely to be chosen than others, baselines for the likelihood of choosing a certain incorrect answer should be empirically determined, and not based on theoretical distributions (Haney & Clark, 2007).

In the current study, items with no correct answer will be included on the cognitive ability test, and, for half of the participants, access to a prohibited answer key will also be provided. The answer key will provide "correct" answers to the unsolvable items. "Correct" answers to these items (at a higher level than chance) will be considered evidence of possible cheating.

Self-Report

One way to discover if a test-taker has cheated is simply to ask. In the current study, participants will be guaranteed anonymity, assured that their answer will have no negative consequences, and then asked if they cheated on the UIT, and, if so, how.

The Decision to Cheat

One question that has been severely under-researched, at least in the organizational literature, is why people cheat, and what conditions facilitate or prevent cheating (Tippins, 2006; Tippins, 2009b). In Tippins (2006), Fritz Drasgow is quoted as saying, "I think the most pressing need is to understand the psychology underlying cheating by job applicants. With a good model, practitioners could confidently decide when UIT could be effectively utilized and when cheating would be so likely that test scores were meaningless," (p. 218). Despite that article being written

over seven years ago, however, the research on UIT within the organizational literature has remained atheoretical.

Though cheating on UIT is a relatively new topic, cheating has been studied for decades within the educational literature (for examples see Haney & Clark, 2007). Rettinger (2007) proposed that cheating is a decision, and that we can understand cheating behavior as a decision making process. Based on this perspective, Rettinger then described how utility theory, a well-established judgment and decision-making theory, can be used to explain test-takers' decisions to cheat.

Utility Theory

Utility theory is a judgment and decision making theory that can be used to explain how individuals make decisions when they are unsure of outcomes (Rettinger, 2007; for a thorough review of judgment and decision making processes, see Weber & Johnson, 2009). In these situations, individuals compile a list of possible decisions (e.g., to cheat or not to cheat), and a list of possible outcomes (e.g., pass the test, fail the test, get caught cheating). Individuals then assign a subjective value to each possible outcome (e.g., how much is it worth to do well on this test? What are the consequences of being caught cheating?), and then estimate the probability of each outcome occurring (e.g., there's a 10% chance I'll be caught cheating). They then multiply the subjective value of an outcome by the probability of its occurrence, and this product is referred to as the "expected utility" of the outcome. Individuals then sum the expected utilities for each possible decision, yielding the "expected value," and choose the decision with the highest expected value.

An example might be useful to clarify the application of utility theory to cheating. John, who lives in Los Angeles, is trying to choose between driving to Mammoth and driving to Tahoe

to go skiing for the weekend, with the hope of getting some fresh powder. He checks the weather report and sees there is a 90% of 3" of snow in Mammoth, and a 30% of 12" of snow in Tahoe. He begins the decision making process by mentally listing his options, which in this case we will restrict to driving to Mammoth or driving to Tahoe. He then compiles a list of possible outcomes for each decision. If he drives to Mammoth, it might snow 3", or it might not snow at all. If he drives to Tahoe, it might snow 12", or it might not snow at all. He then assigns a subjective value to each possible outcome, which, for the sake of this example, we will express in dollars. Snow in neither Mammoth nor Tahoe are worth equally little to him, say $0; 3" of snow in Mammoth is worth $20; and 12" of snow in Tahoe is worth four times as much, $80. Next, he estimates the probability of each outcome: the probability of no snow in Mammoth equals 10%; the probability of 3" of snow in Mammoth equals 90%. The probability of no snow in Tahoe equals 70%; the probability of 12" of snow in Tahoe equals 30%. John next multiplies the subjective value of each outcome by the probability of its occurrence, yielding the expected utility of that outcome, and then adds them together for each decision, yielding the expected value. For Mammoth, this is $0 X .10 plus $20 X .90, which equals 1.8. For Tahoe, this is $0 X .7 plus $80 X .3, which equals 2.4. Because the expected value of driving to Tahoe is higher, John decides to drive to Tahoe.

We can apply the same logic to a candidate completing a UIT. After John's great weekend skiing powder at Tahoe, he decides to apply for a position at a national corporation that uses UIT as part of its selection procedure. For simplicity's sake, we'll limit John to two options: not cheating on the test, or cheating on the test. If he decides not to cheat on the test, he will either pass the test (get hired or move on to the next stage of the selection process) or fail the test (neither getting hired nor moving on to the next stage of the selection process). If he decides to

cheat, there are three possible outcomes: he will pass the test, he will fail the test, or he will get caught cheating. John then assigns a subjective value to each of these outcomes. Whether he decides to cheat or not, passing the test is probably very valuable to him. For the sake of the example, let us say it is worth $10,000 (the increase in salary over his current job). Whether he cheats or not, failing the test is worth very little to him, say $0. Being caught cheating might prevent him from applying to the same company in the future, so it actually represents a cost to John, say -$1,000. John then assigns a likelihood to each possible outcome; that is, he estimates his probability of passing and failing the test with and without cheating, and the probability of being caught cheating. He estimates that his probability of passing the test without cheating is .7, and failing the test without cheating is .3. He then estimates his probability of passing the test with cheating is .8, failing the test with cheating .1, and being caught cheating .1. John next multiplies the subjective value of each outcome by the probability of its occurrence, yielding the expected utility of that outcome, and adds them together for each decision, yielding the expected value. For the decision not to cheat, this is $0 X .3 plus $10,000 X .70, which equals 7,000. For the decision to cheat, this would be $0 X .1 plus $10,000 X .8 plus -$1,000 X .1, which equals 7,900. Because the expected value of cheating is higher, John decides to cheat on the test.

Utility theory supports several hypotheses concerning cheating behavior, listed below:

Hypothesis 1a: Group mean scores on a cognitive ability test will be significantly higher in conditions in which the subjective evaluation of the probability of passing the test with cheating is high, compared to conditions in which it is low.

Hypothesis 1b: A significantly greater proportion of participants will self-report cheating behavior in conditions in which the subjective evaluation of the probability of passing the test with cheating is high, compared to conditions in which it is low.

Hypothesis 1c: Participants in conditions in which the subjective evaluation of the probability of passing the test with cheating is high will answer significantly more of the fake cognitive ability items (CAT) items correctly on the first version of the cognitive ability test, compared to participants in conditions in which it is low.

Hypothesis 1d: A significantly greater proportion of participants will score high enough on the CAT during session one to be excused from the vigilance task in conditions in which the subjective evaluation of the probability of passing the test with cheating is high, compared to conditions in which it is low.

Hypothesis 2a: Group mean scores on a cognitive ability test will be significantly higher in conditions in which the subjective evaluation of the probability of being caught cheating is low, compared to conditions in which it is high.

Hypothesis 2b: A significantly greater proportion of participants will self-report cheating behavior in conditions in which the subjective evaluation of the probability of being caught cheating is low, compared to conditions in which it is high.

Hypothesis 2c: Participants in conditions in which the subjective evaluation of the probability of being caught cheating is low will answer significantly more of the fake CAT items correctly on the first version of the cognitive ability test, compared to participants in conditions in which it is high.

Hypothesis 2d: A significantly greater proportion of participants will score high enough on the CAT during session one to be excused from the vigilance task in conditions in which the subjective evaluation of the probability of being caught cheating is low, compared to conditions in which it is high.

Hypothesis 3a: Group mean scores on a cognitive ability test will be significantly higher in conditions in which the subjective value of being caught cheating is low, compared to conditions in which it is high.

Hypothesis 3b: A significantly greater proportion of participants will self-report cheating behavior in conditions in which the subjective value of being caught cheating is low, compared to conditions in which it is high.

Hypothesis 3c: Participants in conditions in which the subjective value of being caught cheating is low will answer significantly more of the fake CAT items correctly on the first version of the cognitive ability test, compared to participants in conditions in which it is high.

Hypothesis 3d: A significantly greater proportion of participants will score high enough on the CAT during session one to be excused from the vigilance task in conditions in which the subjective value of being caught cheating is low, compared to conditions in which it is high.

Hypothesis 4a: The average CAT score for self-reported cheaters will be significantly higher than the average CAT score for participants who do not self-report cheating behavior.

Hypothesis 4b: A significantly greater proportion of self-reported cheaters will "pass" the CAT (i.e., score high enough to be excused from the vigilance task).

Hypothesis 4c: Self-reported cheaters will answer significantly more of the fake CAT items "correctly" than participants who did not self-report cheating behavior.

Hypothesis 5a: There will be a significant, positive relationship between CAT performance during session 1 and CAT performance during session 2.

Hypothesis 5b: The relationship between CAT performance during session 1 and CAT performance during session 2 will be moderated by experimental condition, such that the relationship will be stronger in conditions that discourage cheating (i.e., when the probability of passing the test with cheating is low, when the probability of being caught cheating is high, and when the value of being caught cheating is high) and weaker in conditions that encourage cheating (i.e., when the probability of passing the test with cheating is high, when the probability of being caught cheating is low, and when the value of being caught cheating is low).

Hypothesis 5c: The relationship between CAT score during session 1 and CAT score during session 2 will be moderated by self-reported cheating behavior, such that the relationship will be weaker for those who self-reported cheating, and stronger for those who did not.

Hypothesis 6a: There will be a significant, positive relationship between CAT score during session one and self-reported SAT score.

Hypothesis 6b: The relationship between CAT score during session one and self-reported SAT score will be moderated by experimental condition, such that the relationship will be stronger in conditions that discourage cheating (i.e., when the probability of passing the test with cheating is low, when the probability of being caught cheating is high, and when the value of being caught cheating is high) and weaker in conditions that encourage cheating (i.e., when the probability of passing the test with cheating is low, when the probability of being caught cheating is low, and when the value of being caught cheating is low).

Hypothesis 6c: The relationship between CAT score during session 1 and self-reported SAT score will be moderated by self-reported cheating behavior, such that the relationship will be weaker for those who self-reported cheating, and stronger for those who did not.

Hypothesis 7a: There will be a significant, positive relationship between CAT score during session 1 and self-reported ACT score.

Hypothesis 7b: The relationship between CAT score during session 1 and self-reported ACT score will be moderated by experimental condition, such that the relationship will be stronger in conditions that discourage cheating (i.e., when the probability of passing the test with cheating is low, when the probability of being caught cheating is high, and when the value of being caught cheating is high) and weaker in conditions that encourage cheating (i.e., when the probability of passing the test with cheating is high, when the probability of being caught cheating is low, and when the value of being caught cheating is low).

Hypothesis 7c: The relationship between CAT score during session 1 and self-reported ACT score will be moderated by self-reported cheating behavior, such that the relationship will be weaker for those who self-reported cheating, and stronger for those who did not.

Limitations of Utility Theory

A major drawback of utility theory is that it assumes humans are perfectly rational and perfectly accurate computational machines, when, actually, irrational information (such as emotion) often plays a major role in human decision making processes (Weber & Johnson, 2009). Despite this limitation, however, utility theory has been demonstrated as a relatively

accurate heuristic to anticipate and explain people's decisions (Weber & Johnson, 2009). Rational decision-making, via utility theory is thus the focus of the current investigation, not emotion. Emotion will be measured, however, and, if necessary, statistically controlled.

Conclusion

UIT represents an area of organizational psychology where practice is far outpacing research. There is a split between practitioners who embrace the benefits of UIT, and researchers who are wary of its drawbacks (Tippins, 2006). Both groups agree that new technologies provide a tremendous opportunity for testing and selection, but that this opportunity comes with a corresponding need for the ethical and professional use of these technologies, and a need for our science to better understand their impact (Naglieri et al., 2004).

In many ways, the biggest problem for UIT is a lack of empirical data. Beaty (2011), for example, noted that, " There are literally no published studies, as far as we know, that present data showing what happens to the predictive validity of a test when it is taken offsite, via the Internet, and administered to job applicants," (pp. 1-2). Similarly, Tippins (2009b), lamented, "There is little if anything in the literature that indicates the extent of cheating on employment tests," (p. 69). Without this empirical data, and a theory to guide research and practice, both scientists and practitioners lack the knowledge and a framework to know how and when to best utilize UIT (Tippins, 2006).

Contributions of the Current Study

The current study attempts to fill some of the gaps in the UIT literature by proposing and testing a model that explains how candidates decide to cheat on UIT, estimating the prevalence of cheating under various testing conditions, estimating the impact of cheating on test validity, and testing methods for detecting cheating at both the individual and group levels. Though high-

quality field research has been conducted on this topic, these studies lacked the precise control available in lab experiments, another contribution of the current study.

Proposing and Testing a Model that Explains How Candidates Decide to Cheat on UIT

One of the greatest limitations of the literature on UIT is that it lacks a cohesive model to guide research and application (Tippins, 2006). Little is understood about the psychology of the testing process, including why candidates decide to cheat, or not to cheat (Beaty, 2011; Tippins, 2006). The current study will address that limitation by proposing and testing utility theory as a model to understand candidates' cheating decisions.

Estimate the Prevalence of Cheating under Various Testing Conditions

The potential for cheating behavior represents one of the greatest challenges to the full-scale implementation of UIT. It is assumed that cheating is widespread, and that cheating on unproctored tests is more common than on proctored tests (Tippins, 2009b; Tippins, 2006), however, there is virtually no empirical data on the prevalence of cheating on UIT, or what conditions encourage or discourage cheating (Beaty, 2011; Tippins, 2009b; Tippins, 2006). The current study will address this limitation by estimating the prevalence of cheating under various proctored and unproctored conditions.

Estimating the Impact of Cheating on Validity

Cheating is a concern in UIT primarily because it might affect test validity and hence the utility of decisions made with test scores (Tippins, 2006; Tippins, 2009; Beaty, 2011; Lievens & Burke, 2011). Despite this concern, there is very little research on the impact of cheating on test validity in UIT (Beaty, 2011). The current study will not only investigate the prevalence of cheating, but also the impact of cheating on validity.

Test Methods for Detecting Cheating at Both Group and Individual Level.

Several methods exist for detecting cheating at both the individual and group levels (Haney & Clark, 2007). The current study will provide estimates of the effectiveness of these methods, taking advantage of controlled laboratory conditions that allow the researcher to be more confident in the claims that cheating has occurred (e.g., the use of an answer sheet with the answers to impossible questions and self-report confessions of cheating).

Benefits of a Lab Study

Though providing invaluable information about UIT under real-use conditions, field studies are limited in study design and sample size and lack control measures (Beaty, 2011). In field studies, it is impossible to determine whether or not score changes reflect cheating, or whether they reflect changes in motivation, practice effects, or statistical artifacts such as the unreliability of the measure or regression to the mean (Nye, 2008). The current study will address these challenges by investigating cheating using rigorous scientific methodology, including the presence of a control condition, manipulation of pertinent variables, and measurement of possible confounding variables.

METHOD

Participants and Procedure

Participants. Sample 1 consisted of 384 college-aged students recruited through the participant pool at Colorado State University. These were students enrolled in PSY100 who must complete six research credit hours as part of their final grades. Of 174 participants reporting gender, 124 (71.3%) identified as female.

Sample 2 consisted of 518 workers recruited through Amazon's mTurk. mTurk participants were all recruited from the United States. 337 (65.1%) identified as female, and the average age was 37.2 years old (*sd*=12.6). The directions that mTurk workers were given can be found in Appendix A.

Materials

Cognitive ability test. The Scholastic Achievement Test (SAT) is a highly valid test of cognitive ability (Frey & Detterman, 2004; Sackett & Borneman, 2008). The SAT correlates highly with college GPA (Sackett & Borneman, 2008), as well as general cognitive ability (Frey & Detterman, 2004). General cognitive ability, in turn, is a strong predictor of job performance across jobs (Schmidt & Hunter, 2004).

Previously administered versions of the SAT are publically available. For this study, these versions were collected and combined into two similar, 30-minute test versions. These versions were then uploaded to the survey site Qualtrics (www.qualtrics.com). Participants were presented with the first version of the test during the first session of the study, and the second version of the test during the second session of the study.

Answer key. In half the conditions, participants were shown the correct answer to each question after they had answered it. The testing interface included a back button, allowing participants to go back and change their answers once they had been shown the correct answer, facilitating cheating.

Fake questions. Several altered questions were added to the cognitive ability test. These questions were similar in appearance to the other questions, but none or all of the answer choices were correct. The answer key, however, had a single "correct" answers to these questions.

Self-report SAT/ ACT score. It is likely that participants who are confident that they can pass the test without cheating (i.e., higher ability participants) will be less likely to cheat. Self-reported SAT/ ACT score has been shown to be relatively accurate compared to actual SAT/ ACT score (Mayer, 2006), and SAT/ ACT scores themselves are a valid and reliable proxy measure for general intelligence (Frey & Detterman, 2004; Sackett, 2008). Therefore, participants were asked to provide this information, which in turn were used as criteria against which to compare the validity of CAT scores as estimates of cognitive ability across the various experimental conditions, and between self-reported cheaters and non-self-reported cheaters.

Vigilance task. A version of a psychomotor vigilance task was used in this study. In this task, participants were asked to stare at a black background upon which, every 30-45 seconds, a small, red dot appeared (Dinges & Powell, 1985). When participants saw the red dot, they were required to press a button on their keyboard to acknowledge it.

This task was specifically designed to be monotonous and unpleasant to perform for long periods of time. There is evidence supporting performance on this task as a valid and reliable measure of vigilance (Loh, Lamond, Dorrian, Roach & Dawson, 2004; Wilkinson & Houghton, 1982), however performance on this task was not an outcome of importance for this study;

rather, the threat of having to complete this monotonous, unpleasant task served to increase the value of performing well on the cognitive ability test. Participants were told that if they performed well enough on the cognitive ability test, they would be excused from the psychomotor vigilance task. If they performed poorly, however, they would have to complete the unpleasant task for 35 minutes.

Suspicion of ulterior motive for the study survey. A survey was designed to detect participant suspicion of an ulterior motive for the study (as compared to the vigilance task cover story).

Self-report survey of cheating behavior. Towards the end of the study, once participants were debriefed as to the true purpose of the experiment and assured that their answers were completely confidential, and that they would not be punished for their behavior, participants were asked several questions about their cheating behavior during the study.

Procedure

The procedure was the same for both samples unless otherwise noted.

Session 1: Online survey. Participants were told they were participating in a study designed to test performance on an online vigilance task. The majority of participants were provided a link to the study, and allowed to take it at a time and place of their own choosing, before a certain deadline. After following the link, they were asked to complete an online consent form. A smaller subset of participants was asked to sign up for a convenient time to take the test in a proctored computer lab on campus. They were required to verify their identification, and then assigned a computer on which to complete the study.

Next, participants were told that the study was designed to test performance on an online vigilance task. The vigilance task was described, and then participants were asked to perform the

task for eight minutes in order to familiarize themselves with it. The true purpose of this practice session was to show participants how boring the vigilance task was.

Participants were then told that the researchers were only interested in individuals within a certain cognitive ability range. Thus, participants were asked to complete a cognitive ability test before they competed the vigilance task. Participants were told that those who performed above a certain minimum score on the CAT were excused from having to perform the vigilance task, but those who failed to achieve the minimum score had to perform the vigilance task for an additional 35 minutes. Participants' cumulative scores were noted at the top of the page for each question, as well as the minimum score needed to be excused from the vigilance task. This essentially served as an added encouragement for participants to cheat; if participants knew they were not going to pass without cheating, they would theoretically be more likely to decide to cheat. As noted above, participants in half the conditions were shown the correct answer to the questions and given the opportunity to go back and change them.

Participants were then warned against cheating, and told of the procedures in place to detect cheaters, and the punishment that cheaters would face if caught.

After completing the CAT, participants who scored above the minimum were excused from the rest of the study; those who scored below it were required to complete the vigilance task. All participants were then told they were randomly selected to retake the test in a proctored computer lab on campus, and asked to sign-up for a convenient time and date to do so.

Session 2: Proctored exam. All participants from sample one were asked to come into a proctored computer lab in order to complete the second part of the study (participants from sample 2 only completed the first session of the study). Participants were assigned a computer

and given a link to the second part of the study. They were reminded of the instructions for the CAT, and then administered the alternate version of the CAT.

After completing the CAT, participants were asked several questions concerning whether they suspected the study was about something other than a vigilance task, and when they began to suspect this. The true purpose of the study (i.e., to investigate cheating on unproctored, online tests) was revealed to participants, and they were asked whether or not they cheated on the exam, and, if so, what method they used.

Participants were then thoroughly debriefed and thanked for their time. They were reminded that the study was ongoing, and asked not to share their research experience with any other students.

Manipulations

According to utility theory, participants' decisions to cheat or not should be based on sum of the values of outcomes of not cheating (i.e., passing/ failing the test) or cheating (passing/ failing/ getting caught), as well as the probabilities of each of those outcomes. Thus, in this experiment, the value of passing/ failing the test, the probability of passing/ failing the test without cheating, and probability of being caught cheating were manipulated.

Subjective value of passing/ failing the test. The value of passing/ failing the test for participants in sample one was the same across conditions. Participants were told that if they achieved a minimum score they would be excused from the rest of the experiment, but if they failed, they would need to perform the vigilance task for 35 minutes. Time is a valuable commodity for most people, and saving time is often cited as one motivation for student cheating (e.g., Rettinger, 2007).

Subjective evaluation of the likelihood of passing/failing the test without cheating. Participants' subjective evaluation of the probability of passing/ failing the test was not directly manipulated. It is likely, however, that high ability students would consider themselves more likely to pass the test, and low ability students would consider themselves more likely to fail. Therefore, self-reported SAT/ ACT score was measured.

Subjective evaluation of the probability of passing the test with cheating. Participants' subjective evaluation of the probability of passing with cheating was manipulated by providing access to an answer key in the "high probability of passing the test with cheating," condition, which essentially ensured a 100% chance of passing the test by cheating (i.e., using the answer key). Participants in the "low probability of passing the test with cheating" condition were not provided with an answer key. It is important to note that it is impossible to determine the exact probability of passing-by-cheating for students in the "low probability of passing the test with cheating," condition, who could potentially use other methods of cheating (e.g., using a test surrogate, searching the Internet for answers, etc.). In any case, it seems reasonable to assume that the probability of passing by cheating for these participants is substantially lower than for participants provided the answer key. It is also worth noting that it would require more effort to cheat, which, all else being equal, means that these participants would be less likely to seek out other cheating strategies.

Subjective evaluation of the probability of being caught cheating. For sample one, participants' subjective evaluation of the probability of being caught cheating was manipulated in two ways.

First, participants in the "proctored" condition were required to take the test in a proctored computer lab on campus. The presence of proctors increased the probability that they would be caught cheating, as compared to taking the test in an unproctored environment.

Second, all participants were told that a certain percentage of participants would be randomly selected to take another version of the test in a proctored environment to validate their scores and detect cheaters. There were two "probability of being caught cheating," conditions: high and low. In the "high probability of being caught cheating" condition, participants were told that 90% of participants would be randomly selected to take another version of the test in a proctored environment. In the "low probability of being caught cheating," condition participants were told that 10% of participants would be randomly selected to take another version of the test in a proctored environment.

This deception also provided the researchers with an excuse to ask all participants to attend the second experimental session in a proctored computer lab on campus.

For sample two, the probability of getting caught cheating was manipulated by telling participants that a certain percentage of responses (again, 10% or 90%) would be analyzed for suspicious cheating activity.

Subjective value of being caught cheating. For sample one, participants' subjective evaluation of the value of being caught cheating was directly manipulated by stating a punishment for being caught cheating. In the "high subjective value of being caught cheating" condition, participants were told that, if they were caught cheating, they would forfeit all of their research credits for the semester and not be allowed to complete any more. Research credits represent a large portion of participants' PSY100 grades, so this punishment was designed to be particularly severe. Participants in the "low subjective value of being caught cheating condition"

were told that, if they were caught cheating, they would be asked to retake the CAT; this was designed to be a relatively lenient punishment.

For sample two, participants' subjective evaluation of the value of being caught cheating was directly manipulated by stating that, if caught cheating, they would not be paid ("high value of being caught cheating" condition), or that they would have to retake the cognitive ability test ("low value of being caught cheating" condition).

RESULTS

Two samples were collected for analysis in this study. Descriptive statistics for each sample are reported below.

Student Sample Descriptives

384 individuals participated in the study from the student sample. Of those reporting gender, 124 (71.3%) identified as female, whereas 50 (28.7%) identified as male. The average self-reported SAT score was 66.4% (*sd*=0.1; *n*=91), and the average self-reported ACT score was 25.1 (*sd*=3.5; *n*=318). Participants achieved an average score of 21 (out of 35 possible; *sd*=1.1, *n*=384) on the cognitive ability test (CAT) during session one (*KR20*=.853), and 20.4 (*sd*=5.8, *n*=244) during session two (*KR20*=.820). 62 participants (16.1%) performed well enough on the CAT during session one to be excused from the vigilance task. See Tables 1 and 2 for descriptive statistics and frequencies, respectively. See Table 3 for a correlation matrix of key variables.

Of 238 student participants with data, only 10 (4.2%) suspected the study was about cheating. 37 participants of 237 with data self-reported cheating behavior (15.6%). These results suggest that the study protocol was successful at obscuring the true purpose of the study and providing opportunities for participants to cheat if they so desired in the student sample (see Table 2).

mTurk Descriptives

518 individuals in the United States between the ages of 18 and 76 (*m*=37.3, *sd*=12.6) participated in the online mTurk survey. 387 (65.1%) of these individuals identified as female. The average self-reported SAT score was 75.5%, (*sd*=0.2, *n*=197) and the average self-reported ACT score was 26.8 (*sd*=5.8, *n*=154) (SAT was converted into a percentage to account for

scoring changes that occurred in 2005). ACT had unacceptable levels of skew (*skew*=-1.7, *se*=0.2) and kurtosis (*kurtosis*=4.7, *se*=0.4). Standardized ACT scores were computed, six outliers (participants with standardized scores with an absolute value greater than three) were identified, and their scores were removed from the sample. This resolved the skew and kurtosis problems, resulting in a mean ACT score of 27.3 (*sd*=4.7, *n*=151). Participants achieved an average score of 23.9 (out of 35 possible; *sd*=5.6) on the cognitive ability test (CAT), which showed adequate reliability (*KR20*=.842, 35 items). Seven participants (1.4%) performed well enough on the CAT to be excused from the vigilance task. See Tables 4 and 5 for descriptive statistics and frequencies, respectively, and Table 6 for a correlation matrix of key variables.

Deception was an integral part of this study. To test whether the deception was successful, participants were asked prior to debriefing whether they suspected the study was about something other than cognitive ability and vigilance (which was used as a cover story). Of 518 participants, only five (1.0%) suspected the study was about cheating. Another important aspect of the study is whether or not participants really would cheat, and/or admit it; 56 participants (10.8%) self-reported cheating behavior. These results suggest that the study protocol was successful at obscuring the true purpose of the study and providing opportunities for participants to cheat if they so desired (see Table 5).

Sample comparisons

Several t-tests and chi-squared tests were used to compare the two samples on key variables. First, I compared the samples on mean ACT score, mean SAT score, mean CAT score, and the number of fake items answered correctly. One of the assumptions associated with t-tests is that the dependent variable is normally distributed. To test this assumption, I requested histograms of each of the above variables. All of the variables, with the exception of the number

of fake items answered correctly, approximated a normal distribution, supporting this assumption, so all variables were compared using t-tests, except for number of fake items answered correctly, which was analyzed using a Mann-Whitney U test. Homogeneity of variance results are reported with each variable.

First I compared the samples on ACT mean scores. Levene's test of homogeneity of variance was significant for this variable, $F(1, 467)=15.383$, $p<.001$, so equal variances were not assumed. The t-test revealed that the mTurk sample self-reported significantly higher ACT scores (n=151, m=27.3, sd=4.7) compared to the student sample (n=318, m=25.1, sd=3.5), $t(233.676)=5.593$, $p<.001$, d=0.531 (see Table 7).

Next I compared the samples on SAT mean scores. Levene's test of homogeneity of variance was only marginally significant for this variable, $F(1, 286)=3.396$, p=.066, so the t-test proceeded as normal. The t-test revealed that the mTurk sample self-reported significantly higher SAT scores (n=197, m=75.5%, sd=0.2) compared to the student sample (n=91, m=66.4%, sd=.1), $t(286)=4.315$, $p<.001$, d=0.577 (see Table 7).

Next I compared the samples on CAT mean scores. Levene's test of homogeneity of variance was significant for this variable, $F(1, 900)=16.535$, p<.001, so equal variances were not assumed. The t-test revealed that the mTurk sample performed significantly higher on the CAT (n=518, m=23.8, sd=5.6) compared to the student sample (n=384, m=21, sd=6.5), $t(749.224)=6.770$, $p<.001$, d=0.478 (see Table 7).

Next I compared the samples on the number of fake items answered correctly using a Mann-Whitney U test. The mTurk sample (n=518, *mean rank*=326.625) answered significantly fewer fake items correctly than the student sample (n=384, *mean rank*=619.951), U=34771.000, $p<.001$, Wendt's r=.650 (see Table 8).

I then compared the samples on dichotomous variables using a series of chi-square tests. I first compared the number of participants in each sample who, prior to debriefing, suspected the true purpose of the study was to investigate cheating behavior. In total, prior to debriefing, 15 of 680 participants (2.2%) suspected the true purpose of the study was to investigate cheating behavior. Prior to debriefing, five of 518 participants (1%) in the mTurk sample suspected the true purpose of the study was to investigate cheating behavior, compared to 10 of 162 participants (6.1%) in the student sample. This difference was statistically significant, $X^2(1)=15.514$, $p<.001$, $\phi=.151$ (see Table 9), more participants from the student sample than the mTurk sample suspected the true purpose of the study.

Next I compared the number of participants in each sample who performed well enough on the CAT to be excused from the vigilance task. In total 69 of 902 participants (7.7%) performed well enough on the CAT to be excused from the vigilance task. Seven of 518 participants (1.4%) in the mTurk sample performed well enough on the CAT to be excused from the vigilance task, compared to 62 of 384 participants (16.2%) in the student sample. This difference was statistically significant, $X^2(1)=68.324$, $p<.001$, $\phi=.275$ (see Table 10).

Finally, I compared the number of participants in each sample who self-reported cheating behavior. In total 86 of 680 participants (12.7%) self-reported cheating behavior. 56 of 518 participants (10.9%) in the mTurk sample self-reported cheating behavior, compared to 30 of 162 participants (18.5%) in the student sample. This difference was statistically significant, $X^2(1)=6.636$, $p=.010$, $\phi=.099$ (see Table 11).

Because the two samples came from different populations, and significantly differed on all key variables, they were analyzed separately.

Student Sample

Hypotheses 1a, 2a, and 3a: Mean difference in CAT score depending on experimental condition. I used a 2x2x2 ANCOVA to test hypotheses 1a, 2a, and 3a (i.e., mean differences in CAT score depending on experimental condition) in the student sample. ANCOVA requires that several assumptions be met, which were tested prior to running the analyses. To test for normality of errors, I requested descriptive statistics, as well as Kolmogorov-Smirnov and Shapiro-Wilk statistics, of the residuals for each group in the model. None of the skew or kurtosis statistics for any of the groups were larger than twice the standard error term, and none of the Kolmogorov-Smirnov or Shapiro-Wilk statistics were significant, suggesting that the residuals were normally distributed. Likewise, normality plots supported the assumption of normality of errors. A histogram suggested that the dependent variable (i.e., CAT score) was normally distributed. Self-reported ACT score was significantly related to CAT score, and thus was retained as a covariate; SATper was not. CAT score was plotted against self-reported ACT score, and results suggested that a linear relationship did exist, supporting the assumption of linearity of regression. I checked for an interaction between each of the independent variables and standardized test scores to test the assumption of homogeneity of regression; this interaction was not significant, supporting the assumption. Finally, I assessed homogeneity of variance by requesting Levene's test of equality of error variances. Results indicated a significant difference in the error variance of the dependent variable across groups: $F(7, 285)=4.856$, $p<.001$. ANCOVA, however, is robust to violations of this assumption, especially when the group sizes are approximately even (as they are in this case), and when the dependent variable is normally distributed within each group, which histograms of CAT within each group revealed to be the case; thus, I continued with the ANCOVA.

The ANCOVA model as a whole explained a significant amount of variance in CAT score: $F(8, 284)=11.344$, $p<.001$ (see Table 12). The factor representing the subjective evaluation of the probability of passing the test with cheating was significant: $F(1, 284)=11.303$, $p<.001$, $\eta^2=.030$. The estimated marginal mean for conditions in which the subjective evaluation of the probability of passing the test with cheating was high was 20.480 (*se*=0.462; see Table 13), compared to 22.788 conditions in which it was low (*se*=.504). So although the factor was significant, it was significant in the wrong direction, and no evidence was found to support hypothesis 1a, that participants would perform better on the CAT in conditions in which the probability of passing the test with cheating was high, compared to conditions in which it was low.

The factor representing the subjective evaluation of the probability of being caught cheating was not significant: $F(1, 284)=.0.519$, $p=.472$, $\eta^2=.001$. Thus, I failed to reject the null hypothesis for hypothesis 2a, indicating that (controlling for ACT) group mean scores on the CAT did not significantly differ in conditions in which the subjective evaluation of the probability of being caught cheating was low (*estimated marginal mean*=21.9, *se*=0.5), compared to conditions in which it was high (*estimated marginal mean*=21.4, *se*=.5; see Table 13).

Finally, the factor representing the subjective value of being caught cheating was not significant: $F(1, 284)=.006$, $p=.937$, $\eta^2<.001$. Thus, I failed to reject the null hypothesis for hypothesis 3a, indicating that (controlling for ACT) group mean scores on the CAT did not significantly differ in conditions in which the subjective value of being caught cheating was low (*estimated marginal mean*=21.7, *se*=0.5), compared to conditions in which it was high (*estimated marginal mean*=21.6, *se*=0.5; see Table 13).

The interaction between the probability of passing the test with cheating and probability of being caught cheating was non-significant: $F(1, 284)=0.336$, $p=.563$, $\eta^2=.001$, as was the interaction between the probability of passing the test with cheating and the value of being caught cheating: $F(1, 284)=0.318$, $p=.574$, $\eta^2=.001$, and the interaction between the probability of being caught cheating and the value of being caught cheating: $F(1, 284)=0.188$, $p=.665$, $\eta^2<.001$. Finally, the three-way interaction between the probability of passing the test with cheating, the probability of being caught cheating, and the value of being caught cheating, was non-significant: $F(1, 284)=0.647$, $p=.422$, $\eta^2=.002$.

Results of this analysis suggest that, although, the subjective evaluation of the probability of being caught cheating and the subjective value of being caught cheating did not significantly affect CAT score at the group level, the subjective evaluation of the probability of passing the test with cheating did, although in the wrong direction. Thus, no evidence was found to support hypotheses 1a, 2a or 3a in the student sample.

mTurk Sample

Hypotheses 1a, 2a, and 3a: Mean difference in CAT score depending on experimental condition. I used a 2x2x2 ANCOVA to test hypotheses 1a, 2a, and 3a (i.e., mean differences in CAT score depending on experimental condition) for the mTurk sample. Analysis of the residuals revealed the presence of six outliers that resulted in unacceptable levels of skew and kurtosis, as well as significant Kolmogorov-Smirnov and Shapiro-Wilk statistics, of the residuals for several of the experimental conditions. After the removal of these outliers, none of the skew or kurtosis statistics for any of the groups were larger than twice the standard error term, and none of the Kolmogorov-Smirnov or Shapiro-Wilk statistics were significant, suggesting that the residuals were normally distributed.

Likewise, normality plots supported the assumption of normality of errors. A histogram suggested that the dependent variable (i.e., CAT score) was normally distributed. ACT was significantly related to CAT score, and thus was retained as a covariate; SAT score was not, and thus was dropped. CAT score was plotted against self-reported ACT scores, and results suggested that a linear relationship did exist, supporting the assumption of linearity of regression. I checked for an interaction between each of the independent variables and self-reported ACT score to test the assumption of homogeneity of regression; this interaction was not significant, supporting the assumption. Finally, I assessed homogeneity of variance by requesting Levene's test of equality of error variances. Results indicated no significant difference in the error variance of the dependent variable across groups: $F(7,137)=1.872$, $p=.079$. Because these assumptions were met, I proceeded with the ANCOVA analysis.

The ANCOVA model as a whole explain a significant amount of variance in CAT score: $F(8, 136)=19.841$, $p<.001$ (see Table 14). The factor representing the subjective evaluation of the probability of passing the test with cheating was significant: $F(1, 136)=19.639$, $p<.001$, $\eta^2=.073$. Thus, the data supported hypothesis 1a, indicating that (controlling for ACT) group mean scores on the CAT significantly differed in conditions in which the subjective evaluation of the probability of passing the test with cheating was high (*estimated marginal mean*=26.2, *se*=0.5; see Table 15), compared to conditions in which it was low (*estimated marginal mean*=23.5, *se*=0.4).

The factor representing the subjective evaluation of the probability of being caught cheating was not significant: $F(1, 136)=.652$, $p=.421$, $\eta^2=.002$. Thus, I failed to reject the null hypothesis for hypothesis 2a, indicating that (controlling for ACT) group mean scores on the CAT did not significantly differ in conditions in which the subjective evaluation of the

probability of being caught cheating was low (*estimated marginal mean*=24.6, *se*=0.4), compared to conditions in which it was high (*estimated marginal mean*=25.1, *se*=.4; see Table 15).

Finally, the factor representing the subjective value of being caught cheating was not significant: $F(1, 136)=1.679$, $p=.197$, $\eta^2=.006$. Thus, I failed to reject the null hypothesis for hypothesis 3a, indicating that (controlling for ACT) group mean scores on the CAT did not significantly differ in conditions in which the subjective value of being caught cheating was low (*estimated marginal mean*=25.3, *se*=0.4), compared to conditions in which it was high (*estimated marginal mean*=24.5, *se*=0.4; see Table 15).

The interaction between the probability of passing the test with cheating and probability of being caught cheating was non-significant: $F(1,136)=0.001$, $p=.982$, $\eta^2<.001$, as was the interaction between the probability of passing the test with cheating and the value of being caught cheating: $F(1, 136)=0.723$, $p=.397$, $\eta^2=.003$. The interaction between the probability of being caught cheating and the value of being caught cheating, however, was significant: $F(1, 136)=4.157$, $p=.043$, $\eta^2=.015$. To probe the interaction, six pairwise comparisons were made comparing each of the four possible combinations of high/ low probability of being caught cheating and high/ low value of being caught cheating, using a Bonferroni correction for experimentwise error rate. Results revealed that the only significant difference was between conditions low in both probability of being caught cheating and value of being caught cheating and conditions low in probability of being caught cheating but high in value of being caught cheating (*mean difference*=2.973, *se*=0.861, *p*=.004, *d*=3.453). See Table 16 for group means and standard errors, and Table 17 for results of all of the comparisons. Finally, the three-way interaction between the probability of passing the test with cheating, the probability of being

caught cheating, and the value of being caught cheating, was non-significant: $F(1, 136)=0.164$, $p=.686$, $\eta^2=.001$.

Results of this analysis suggest that, although, the subjective evaluation of the probability of being caught cheating and the subjective value of being caught cheating did not significantly affect CAT score at the group level, the subjective evaluation of the probability of passing the test with cheating did, providing support for hypothesis 1a, but not for hypotheses 2a or 3a in the mTurk sample.

Student Sample

Hypotheses 1b, 2b, and 3b: Proportion of self-reported cheaters by experimental condition. In the student sample, I tested hypotheses 1b, 2b, and 3b using two different sets of statistical analyses. First, a chi-square test of independence was use to compare the proportion of self-reported cheaters across all eight experimental conditions. In total, 31 of 185 (16.8%) of participants self-reported cheating. The results of this analysis revealed no significant difference in self-reported cheating across the eight conditions, $X^2(8)=10.834$, $p=.211$, $\phi_c=.185$ (see Table 18).

Next, I split the data by the three experimental manipulations (i.e., high/ low probability of passing the test with cheating; high/ low probability of being caught cheating; high/ low value of being caught cheating) and compared the proportion of participants within each manipulation who self-reported cheating behavior. Alpha level was again corrected for experimentwise error rate by dividing alpha by the number of tests, yielding a corrected alpha of .017 (i.e., .05/3).

In conditions in which the subjective evaluation of the probability of passing the test with cheating was low, 11 of 91 participants (12.1%) self-reported cheating behavior, compared to 20 of 94 participants (21.3%) in conditions in which it was high. This difference was non-

significant, $X^2(1)=2.799$, $p=.094$, $\phi=.123$, providing no evidence in support of hypothesis 1b: participants were not significantly more likely to self-report cheating behavior in conditions in which the probability of passing the test with cheating was high, compared to those conditions in which the probability of passing the test with cheating was low (see Table 19)

In conditions in which the subjective evaluation of the probability of being caught cheating was low, 9 of 84 participants (10.7%) self-reported cheating behavior, compared to 22 of 101 (21.8%) in conditions in which it was high. This difference was significant at the .05 level, $X^2(1)=4.027$, $p=.045$, $\phi=.148$, but not at the more stringent .017 level. Thus, I was unable to reject the null hypothesis, and no evidence was found in support of hypothesis 2b, that participants were more likely to self-report cheating behavior in conditions in which the probability of being caught cheating was low, compared to conditions in which the probability of being caught cheating was high (see Table 20).

In conditions in which the subjective value of being caught cheating was low, 16 of 92 participants (17.4%) self-reported cheating behavior, compared to 15 of 93 (16.1%) in conditions in which it was high. This difference was not significant, $X^2(1)=0.053$, $p=.818$, $\phi=-.017$. Thus, I were unable to reject the null hypothesis, and no evidence was found in support of hypothesis 3b, that participants were more likely to self-report cheating behavior in conditions in which the subjective value of being caught cheating was low, compared to conditions in which it was high (see Table 21).

From these results, it seems that none of the manipulated factors had an impact on self-reported cheating behavior in this sample, and no evidence was found to support hypotheses 1b, 2b, or 3b.

mTurk Sample

Hypotheses 1b, 2b, and 3b: Proportion of self-reported cheaters by experimental condition. In the mTurk sample, I tested hypotheses 1b, 2b, and 3b using two different sets of statistical analyses. First, a chi-square test of independence was use to compare the proportion of self-reported cheaters across all eight experimental conditions. In total, 56 of 518 (10.8%) of participants self-reported cheating. The results of this analysis revealed no significant difference in self-reported cheating across the eight conditions, $X^2(7)=2.833$, $p=.900$, $\phi_c=.074$ (see Table 22).

Next, I split the data by the three experimental manipulations (i.e., high/ low probability of passing the test with cheating; high/ low probability of being caught cheating; high/ low value of being caught cheating) and compared the proportion of participants within each manipulation who self-reported cheating behavior. Alpha level was corrected for experimentwise error rate by dividing alpha by the number of tests, yielding a corrected alpha of .017 (i.e., .05/3).

In conditions in which the subjective evaluation of the probability of passing the test with cheating was low, 29 of 286 participants (10.1%) self-reported cheating behavior, compared to 27 of 232 participants (11.6%) in conditions in which it was high. This difference was non-significant, $X^2(1)=0.298$, $p=.585$, $\phi=.024$, providing no evidence in support of hypothesis 1b: participants were not significantly more likely to self-report cheating behavior in conditions in which the probability of passing the test with cheating was high, compared to those conditions in which the probability of passing the test with cheating was low (see Table 23).

In conditions in which the subjective evaluation of the probability of being caught cheating was low, 27 of 256 participants (10.6%) self-reported cheating behavior, compared to 29 of 262 (11.1%) in conditions in which it was high. This difference was not significant,

$X^2(1)=0.037$, $p=.848$, $\phi=.008$ Thus, I was unable to reject the null hypothesis, and no evidence was found in support of hypothesis 2b, that participants were more likely to self-report cheating behavior in conditions in which the probability of being caught cheating was low, compared to conditions in which the probability of being caught cheating was high (see Table 24).

In conditions in which the subjective value of being caught cheating was low, 27 of 255 participants (10.6%) self-reported cheating behavior, compared to 29 of 263 (11.0%) in conditions in which it was high. This difference was not significant, $X^2(1)=0.026$, $p=.872$, $\phi=.007$. Thus, I was unable to reject the null hypothesis, and no evidence was found in support of hypothesis 3b, that participants were more likely to self-report cheating behavior in conditions in which the subjective value of being caught cheating was low, compared to conditions in which it was high (see Table 25).

From these results, it seems that none of the manipulated factors had an impact on self-reported cheating behavior in this sample, and no evidence was found to support hypotheses 1b, 2b, or 3b.

Student Sample

Hypothesis 1c, 2c, and 3c: The number of fake items that participants answer correctly will differ depending on experimental condition. In the student sample, the variable representing the number of fake items that participants answered correctly had neither unacceptable skew (*skew*=0.346, *se*=0.125) nor unacceptable kurtosis (*kurtosis*=-0.367, *se*=0.248). I attempted to use parametric tests with this variable, but the residuals of these models were invariably skewed (as with the mTurk sample- see below), and showed patterns with the predicted values that violated assumptions of independence of errors and normality. Consequently, for all analyses involving the number of fake items that participants answered

correctly, nonparametric tests were used. These tests are essentially equivalent to standard analyses, without the assumption of normality for the dependent variable.

I first used a Kruskal-Wallis test to compare the number of fake items answered correctly across the eight experimental conditions. Results revealed no significant differences across the groups, $H(7)=6.424$, $p=.491$ (see Table 26).

Next, I compared the number of fake items answered correctly within each experimental manipulation individually using Mann-Whitney U tests. Alpha level was corrected for experimentwise error rate by dividing alpha by the number of tests, yielding a corrected alpha of .017 (i.e., .05/3).

There was no significant difference in number of fake items answered correctly between conditions in which the probability of passing the test with cheating was low ($n=195$) compared to conditions in which it was high ($n=189$), $U=16703$, $p=.100$, Wendt's $r=.094$ (see Table 27); nor between conditions in which the probability of being caught cheating was low ($n=191$) compared to conditions in which it was high ($n=193$), U=18424.5, $p=.995$, Wendt's $r=.000$ (see Table 28); nor between conditions in which the value of being caught cheating was low ($n=196$) compared to conditions in which it was high ($n=188$), $U=18415.500$, $p=.994$, Wendt's $r=.000$ (see Table 29).

The number of fake items that participants answered correctly did not significantly differ by experimental condition, providing no support for hypotheses 1c, 2c, or 3c in the student sample.

mTurk Sample

Hypothesis 1c, 2c, and 3c: The number of fake items that participants answer correctly will differ depending on experimental condition. In the mTurk sample, the number of fake items that participants answered correctly was extremely skewed (*skew*=1.792, *se*=0.107) and had high kurtosis (*kurtosis*=4.901, *se*=0.214). Because the items were fake, and impossible to answer correctly except by chance or through accessing the answer key, it makes sense that most participants would answer most of these items incorrectly, which was the case (*m*=0.8, *sd*=0.9), leading to the problems with skew and kurtosis. Several transformations of the original variable were attempted, none of which, however, were able to approximate a normal distribution. Consequently, for all analyses involving the number of fake items that participants answered correctly, nonparametric tests were used. These tests are essentially equivalent to standard analyses, without the assumption of normality for the dependent variable.

I first used a Kruskal-Wallis test to compare the number of fake items answered correctly across the eight experimental conditions. Results revealed no significant differences across the groups, $H(7)=6.955$, $p=.434$ (see Table 30).

Next, I compared the number of fake items answered correctly within each experimental manipulation individually using Mann-Whitney U tests. Alpha level was corrected for experimentwise error rate by dividing alpha by the number of tests, yielding a corrected alpha of .017 (i.e., .05/3).

There was no significant difference in number of fake items answered correctly between conditions in which the probability of passing the test with cheating was low (*n*=286) compared to conditions in which it was high (*n*=232), *U*=30530, *p*=.089, Wendt's *r*=.080 (see Table 31); nor between conditions in which the probability of being caught cheating was low (*n*=256)

compared to conditions in which it was high (n=252), U=33501, p=.982, Wendt's r=.001 (see Table 32); nor between conditions in which the value of being caught cheating was low (n=255) compared to conditions in which it was high (n=263), U=32829, p=.653, Wendt's r=.021 (see Table 33).

The number of fake items that participants answered correctly did not significantly differ by experimental condition, providing no support for hypotheses 1c, 2c, or 3c in the student sample.

Student Sample

Hypotheses 1d, 2d, and 3d: Proportion of participants who performed well enough on the CAT to be excused from the vigilance task by experimental condition. In the student sample, I tested hypotheses 1d, 2d, and 3d using two different sets of statistical analyses. First, a chi-square test of independence was use to compare the proportion of self-reported cheaters across all eight experimental conditions. In total, 62 of 384 (16.2%) of participants performed well enough on the CAT to be excused from the vigilance task. The results of this analysis revealed no significant difference in self-reported cheating across the eight conditions, $X^2(8)$=7.480, p=.486, ϕ_c=.074 (see Table 34).

Next, I split the data by the three experimental manipulations (i.e., high/ low probability of passing the test with cheating; high/ low probability of being caught cheating; high/ low value of being caught cheating) and compared the proportion of participants within each manipulation who performed well enough on the CAT to be excused from the vigilance task. Alpha level was corrected for experimentwise error rate by dividing alpha by the number of tests, yielding a corrected alpha of .017 (i.e., .05/3).

In conditions in which the subjective evaluation of the probability of passing the test with cheating was low, 25 of 195 participants (12.8%) performed well enough on the CAT to be excused from the vigilance task, compared to 37 of 189 participants (19.6%) in conditions in which it was high. This difference was non-significant, $X^2(1)=3.236$, $p=.072$, $\phi=.092$, providing no evidence in support of hypothesis 1d: participants were not significantly more likely to performed well enough on the CAT to be excused from the vigilance task in conditions in which the probability of passing the test with cheating was high, compared to those conditions in which the probability of passing the test with cheating was low (see Table 35).

In conditions in which the subjective evaluation of the probability of being caught cheating was low, 32 of 191 participants (16.8%) performed well enough on the CAT to be excused from the vigilance task, compared to 30 of 193 (15.5%) in conditions in which it was high. This difference was not significant, $X^2(1)=0.104$, $p=.747$, $\phi=-.016$. Thus, I was unable to reject the null hypothesis, and no evidence was found in support of hypothesis 2d, that participants were more likely to performed well enough on the CAT to be excused from the vigilance task in conditions in which the probability of being caught cheating was low, compared to conditions in which the probability of being caught cheating was high (see Table 36).

In conditions in which the subjective value of being caught cheating was low, 34 of 196 participants (17.4%) performed well enough on the CAT to be excused from the vigilance task, compared to 28 of 188 (14.9%) in conditions in which it was high. This difference was not significant, $X^2(1)=0.427$, $p=.514$, $\phi=-.033$. Thus, I was unable to reject the null hypothesis, and no evidence was found in support of hypothesis 3d, that participants were more likely to performed well enough on the CAT to be excused from the vigilance task in conditions in which

the subjective value of being caught cheating was low, compared to conditions in which it was high (see Table 37).

From these results, it seems that none of the manipulated factors had an impact on whether participants performed well enough on the CAT to be excused from the vigilance task in this sample, and no evidence was found to support hypotheses 1d, 2d, or 3d in the student sample.

mTurk Sample

Hypotheses 1d, 2d, and 3d: Proportion of participants in each condition who performed well on the CAT to be excused from the vigilance task. In the mTurk sample, I tested hypotheses 1d, 2d, and 3d using two different sets of statistical analyses. First, I attempted to use a chi-square test of independence to compare the proportion of self-reported cheaters across all eight experimental conditions. However, for each of the experimental conditions, less than five individuals were predicted to occupy each cell, violating one of the key assumptions for chi-square. Thus, this analysis was not run and I moved on to the next set of analyses.

I split the data by the three experimental manipulations (i.e., high/ low probability of passing the test with cheating; high/ low probability of being caught cheating; high/ low value of being caught cheating) and compared the proportion of participants within each manipulation who self-reported cheating behavior. Alpha level was corrected again for experimentwise error rate by dividing alpha by the number of tests, yielding a corrected alpha of .017 (i.e., .05/3).

Because of the small number of predicted members in each cell, I used a Fisher's exact test in lieu of a chi-square test of independence to test hypotheses 1d, 2d, and 3d (i.e., to compare the proportion of participants in each condition who performed well enough on the CAT to be excused from the vigilance task).

In total, only seven of 518 participants (1.4%) performed well enough on the CAT to be excused from the vigilance portion of the study. In conditions in which the subjective evaluation of the probability of passing the test with cheating was low, one of 286 participants (0.4%) performed well enough on the CAT to be excused from the vigilance task, compared to six of 232 participants (2.6%) in conditions in which it was high. This difference was significant at the .05 level, but not at the more stringent .017 level, with Fisher's exact test yielding a significance level of p=.049, OR=0.210. Thus, I found no evidence in support of hypothesis 1d, that participants were more likely to perform well enough on the CAT to be excused from the vigilance task in conditions in which the probability of passing the test with cheating was high, compared to those conditions in which the probability of passing the test with cheating was low (see Table 38).

In conditions in which the subjective evaluation of the probability of being caught cheating was low, four of 256 participants (1.6%) performed well enough on the CAT to be excused from the vigilance task, compared to three of 262 (1.2%) in conditions in which it was high. This difference yielded a *Fisher's exact* significance of p=.722, OR=1.297. Thus, I were unable to reject the null hypothesis, and no evidence was found in support of hypothesis 2d, that participants were more likely to perform well enough on the CAT to be excused from the vigilance task in conditions in which the probability of being caught cheating was low, compared to conditions in which the probability of being caught cheating was high (see Table 39).

In conditions in which the subjective value of being caught cheating was low, three of 255 participants (1.2%) performed well enough on the CAT to be excused from the vigilance task, compared to four of 263 (1.5%) in conditions in which it was high. This difference was not significant, yielding a *Fisher's exact* significance of p=1.000, OR=0.730. Thus, I was unable to

reject the null hypothesis, and no evidence was found in support of hypothesis 3d, that participants were more likely to perform well enough on the CAT to be excused from the vigilance task in conditions in which the subjective value of being caught cheating was low, compared to conditions in which it was high (see Table 40).

From these results, I concluded that none of the factors manipulated in this study influenced the probability of participants performing well enough on the CAT to be excused from the vigilance task in the mTurk sample.

Student Sample

Hypothesis 4a: Self-reported cheaters will perform significantly better on the CAT compared to participants who did not self-report cheating behavior. In the student sample, a t-test was used to test hypothesis 4a (i.e., participants who self-reported cheating behavior will perform significantly better on the CAT than those who did not self-report cheating behavior). To test the assumption that the dependent variable is normally distributed, I requested a histogram of participants' performance on the CAT; the histogram approximated a normal distribution, supporting this assumption. Levene's test for equality of variances was non-significant, $F(1, 183)=0.393$, $p=.531$, meaning that the variances did not significantly differ between the two groups.

The t-test revealed that although participants' who self-reported cheating behavior performed better on the CAT ($n=31$, $m=23.1$, $sd=6.6$) compared to those who did not self-report cheating behavior ($n=154$, $m=20.9$, $sd=6.1$), this difference was not significant, $t(183)=-1.808$, $p=.072$, $d=0.346$ (see Table 41). Thus, no evidence was found to support hypothesis 4a, that participants who self-reported cheating behavior performed significantly better on the CAT than those who did not self-report cheating behavior in the student sample.

mTurk Sample

Hypothesis 4a: Self-reported cheaters will perform significantly better on the CAT compared to participants who did not self-report cheating behavior. In the mTurk sample, a t-test was used to test hypothesis 4a (i.e., participants who self-reported cheating behavior will perform significantly better on the CAT than those who did not self-report cheating behavior). The histogram of participants' performance on the CAT approximated a normal distribution, and Levene's test for equality of variances was non-significant, $F(1,516)=1.675$, $p=.196$, meaning that the variances did not significantly differ between the two groups.

The t-test revealed that although participants' who self-reported cheating behavior performed better on the CAT ($n=56$, $m=24.5$, $sd=6.2$) compared to those who did not self-report cheating behavior ($n=462$, $m=23.8$, $sd=5.5$), this difference was not significant, $t(516)=-0.876$, $p=.382$, $d=0.119$ (see Table 42). Thus, no evidence was found to support hypothesis 4a, that participants who self-reported cheating behavior performed significantly better on the CAT than those who did not self-report cheating behavior in the mTurk sample.

Student Sample

Hypothesis 4b: Proportion of self-reported cheaters who perform well enough on the CAT to be excused from the vigilance task, compared to participants who did not self-report cheating behavior. Hypothesis 4b stated that a greater proportion of individuals who self-reported cheating would perform well enough on the CAT to be excused from the vigilance task. I used a chi-square test of independence to test this hypothesis in the student sample.

In total, 29 of 185 (16.2%) participants performed well enough on the CAT to be excused from the vigilance task. Of participants who self-reported cheating behavior, nine of 31 (14.9%) performed well enough on the CAT to be excused from the vigilance task, compared to 20 of 154

(17.4%) of those who did not self-report cheating behavior. This difference was statistically significant, $X^2(1)=5.026$, $p=.025$, $\phi=.165$ (see Table 43), however in the opposite direction as hypothesized.

mTurk Sample

Hypothesis 4b: Proportion of self-reported cheaters who perform well enough on the CAT to be excused from the vigilance task, compared to participants who did not self-report cheating behavior. Hypothesis 4b stated that a greater proportion of individuals who self-reported cheating would perform well enough on the CAT to be excused from the vigilance task. Because of the small number of predicted members in each cell in the mTurk sample, I used a Fisher's exact test in lieu of a chi-square test of independence.

In total, seven of 518 (1.4%) participants performed well enough on the CAT to be excused from the vigilance task. Of participants who self-reported cheating behavior, four of 56 (7.1%) performed well enough on the CAT to be excused form the vigilance task, compared to three of 462 (6.5%) of those who did not self-report cheating behavior. This difference was statistically significant, yielding a *Fisher's exact* significance of $p=.003$, $OR=6.620$ (see Table 44), providing support for hypothesis 4b: participants who self-reported cheating behavior were significantly more likely to perform well enough on the CAT to be excused from the vigilance task, compared to those who did not self-report cheating behavior in the mTurk sample.

Student Sample

Hypothesis 4c: Self-reported cheaters will answer a greater number of fake items correctly, compared to participants who did not self-report cheating behavior. Due to the previously discussed issues with the variable representing the number of fake items answered correctly in the student sample, a Mann-Whitney U analysis was used to test hypothesis 4c, that

self-reported cheaters would answer a greater number of fake items correctly than participants who did not self-report cheating behavior. Results of the analysis suggested that participants who self-reported cheating (n=31) did not answer significantly more of the fake items correctly than participants who did not self-report cheating (n=154), U=1880, p=.053, Wendt's r=.212 (see Table 45). Thus, I found no evidence in support of hypothesis 4c in the student sample.

mTurk Sample

Hypothesis 4c: Self-reported cheaters will answer a greater number of fake items correctly, compared to participants who did not self-report cheating behavior. Due to the previously discussed issues with the variable representing the number of fake items answered correctly in the mTurk sample, a Mann-Whitney U analysis was used to test hypothesis 4c, that self-reported cheaters would answer a greater number of fake items correctly than participants who did not self-report cheating behavior. Results of the analysis suggested that participants who self-reported cheating (n=31) did not answer significantly more of the fake items correctly than participants who did not self-report cheating (n=154), U=1880, p=.053, Wendt's r=.055 (see Table 46). Thus, I found no evidence in support of hypothesis 4c in the mTurk sample.

Student Sample

Hypotheses 5a, 5b, and 5c: CAT score during session one will be significantly, positively related to CAT score during session 2; this relationship will be moderated by experimental condition and by self-reported cheating behavior. I used linear regression to test hypotheses 5a, (i.e., that participants' CAT score during session one would be significantly related to participants' CAT score during session two) and 5b (i.e., that experimental condition would moderate the relationship between participants' CAT scores during session one and session two).

To test hypotheses 5a, I used simple linear regression (SLR). One of the assumptions associated with SLR is that a linear relationship exists between Y and X. A scatter-plot confirmed that a linear relationship appeared to be an appropriate fit for the data. Another assumption associated with SLR is that the residuals of the model are normally distributed. After the model was run, the residuals were saved and plotted on a histogram; the histogram approximated a normal distribution, confirming this assumption. Another assumption associated with SLR is homoscedasticity. To check this assumption, residuals were plotted against the predictor variable (i.e., CAT score during session one). The plot revealed no detectable pattern, confirming this assumption. Because all of the assumptions were met, I proceeded with the SLR analysis.

CAT score during session one explained a significant amount of variance in CAT score during session two, R=.719, R^2=.517, F(1, 192)=205.480, p<.001. Participants' CAT during session one significantly predicted participants' CAT score during session two (b=.655, se=0.046, p<.001, β=0.719). Thus, I was able to reject the null hypothesis, providing evidence for hypothesis 5a (see Table 47).

To test hypothesis 5b, I used multiple linear regression (MLR). One of the assumptions of MLR is that the residuals are normally distributed; to test this assumption, I requested a histogram and p-plot of the residuals, both of which provided evidence in support of the assumption of normality. Another assumption of MLR is homogeneity of variance; to test this assumption, the predicted values were plotted against the unstandardized residuals. The scatter plot revealed no pattern, supporting the assumption of homogeneity of variance. Because all of the assumptions were met, I proceeded with the MLR analysis.

To test for main effects of experimental condition on participants' CAT score during session two, dummy coded variables representing each of the four experimental factors (i.e., the subjective evaluation of the probability of passing the test with cheating, the subjective evaluation of the probability of being caught cheating, the subjective value of being caught cheating, proctored/ unproctored) were added to the model along with participants' CAT score during session one. A "1" on the dummy coded variables represents the higher version of the variable (e.g., high subjective evaluation of the probability of passing the test with cheating) whereas a "0" on the dummy coded variables represents the lower version of the variable (e.g., a low subjective evaluation of the probability of passing the test with cheating). This model was also significant, R=.726, R^2=.514, F(5, 188)=41.802, p<.001, although the additional amount of the variance in CAT score during session two explained by the model was not, ΔR^2=.010, ΔF(4, 188)=0.943, p=.440 (see Table 48).

In this model, participants' CAT score during session one continued to significantly predict participants' CAT score during session two, b=0.655, se=0.046, t(188)=14.124, p<.001, β=0.719. The dummy coded variable representing the subjective evaluation of the probability of passing the test with cheating was non-significant, b=0.863, se=0.594, t(188)=1.452, p=.148, β=0.076; participants' mean scores on the CAT during session two did not significantly differ as a function of the subjective evaluation of the probability of passing the test with cheating. The dummy coded variable representing the subjective evaluation of the probability of being caught cheating was non-significant, b=-0.677, se=0.599, t(188)=-1.130, p=.260, β=-0.060; participants' mean scores on the CAT during session two did not significantly differ as a function of the subjective evaluation of the probability of being caught cheating. The dummy coded variable representing the subjective value of being caught cheating was non-significant, b=-0.493,

se=0.588, *t*(188)=-0.839, *p*=.403, *β*=-0.044; participants' mean scores on the CAT during session two did not significantly differ as a function of the subjective value of being caught cheating. The dummy coded variable representing the proctored/ unproctored conditions was non-significant, *b*=-0.531, *se*=1.297, *t*(188)=-0.410, *p*=.683, *β*=-0.023; participants' mean scores on the CAT during session two did not significantly differ as a function of the proctored/ unproctored conditions.

Next, interaction terms were created by multiplying each of the four experimental factors by participants' CAT score during session one, and then these terms were added to the model. These interaction terms represent the moderating effect of each experimental factor on the relationship between participants' CAT score during session one and participants' CAT score during session two. This model was significant, *R*=.737, R^2=.543, *F*(9, 184)=24.316, *p*<.001, but the additional amount of variance explained by the interaction terms was not, ΔR^2=.017, ΔF(4, 184)=1.691, *p*=.154 (see Table 49). The interaction term representing the moderating effect of subjective evaluation of the probability of passing the test with cheating on the relationship between CAT score during session one and CAT score during session two was non-significant, *b*=-0.132, *se*=0.095, *t*(184)=-1.393, *p*=.165, *β*=-0.262; the relationship between participants' CAT score during session one and participants' CAT scores during session two was not significantly moderated by the subjective evaluation of the probability of passing the test with cheating.

The interaction term representing the moderating effect of the subjective evaluation of the probability of being caught cheating on the relationship between CAT score during session one and CAT score during session two was non-significant, *b*=-0.130, *se*=0.096, *t*(184)=-1.356, *p*=.177, *β*=-0.258; the relationship between participants' CAT scores during session one and

participants' CAT scores during session two was not significantly moderated by the subjective evaluation of the probability of being caught cheating.

The interaction term representing the moderating effect of the subjective value of being caught cheating on the relationship between CAT score during session one and CAT score during session two was non-significant, b=-0.130, se=0.094, t(188)=-1.378, p=.170, β=-0.260; the relationship between participants' CAT scores during session one and participants' CAT scores during session two was not significantly moderated by the subjective value of being caught cheating.

The interaction term representing the moderating effect of the proctored/ unproctored conditions on the relationship between CAT score during session one and CAT score during session two was non-significant, b=0.031, se=0.230, t(184)=0.137, p=.892, β=0.028; the relationship between participants' CAT scores during session one and participants' CAT scores during session two was not significantly stronger in the proctored condition, as compared to the unproctored condition.

Results revealed that condition did not moderate the relationship between CAT score during session one and CAT score during session two, providing no support for hypothesis 5b.

To test hypothesis 5c (i.e., that the relationship between CAT score during session one and CAT score during session two will be moderated by self-reported cheating behavior), I added a dummy coded variable representing self-reported cheating (0=participant did not self-report cheating; 1=participant did self-report cheating) to the model regressing CAT score during session two on CAT score during session one (i.e., hypothesis 5a). The model including the dummy coded cheating term was significant, R=.724, R^2=.524, F(2, 175)=96.512, p<.001, explaining an additional 1.3% of the variance in CAT score during session two, ΔR^2=.130, ΔF(1,

175)=4.906, p=.028 (see Table 50). Self-reported cheating behavior was a significant predictor of CAT score during session two, b=-1.734, se=0.783, t(175)=-2.215, p=.028, β=-0.117; participants who self-reported cheating on the CAT during session one scored, on average, 1.734 points lower on the CAT during session two than participants who did not self-report cheating on the CAT during session one.

Next, an interaction term was created by multiplying CAT score during session one by the dummy coded variable representing self-reported cheating behavior, and this term was added to the model. This interaction term represented the moderating effect of self-reported cheating behavior on the relationship between CAT score during session one and CAT score during session two. This model was significant, R=.733, R^2=.538, F(3,174)=67.422, p<.001, explaining an additional 1.3% of the variance in CAT score during session two, ΔR^2=.013, ΔF(1, 174)=4.919, p=.028 (see Table 51). The term representing the interaction between CAT score during session one and self-reported cheating behavior significantly moderated the relationship between CAT score during session one and CAT score during session two, b=-0.262, se=0.118, t(174)=-2.218, p=.028, β=-0.426, such that the relationship was weaker for participants who self-reported cheating, compared to those who did not.

The results of the above analyses provide evidence in support of hypothesis 5c: self-reported cheating behavior significantly moderated the relationship between CAT score during session one and CAT score during session two.

Student Sample

Hypotheses 6a, 6b, and 6c: CAT score during session one will be significantly, positively related to self-reported SAT score; this relationship will be moderated by experimental condition and by self-reported cheating behavior. In the student sample, I used linear regression to test hypotheses 6a, (i.e., that participants' CAT score during session one would be significantly related to participants' self-reported SAT score) and 6b (i.e., that experimental condition would moderate the relationship between participants' CAT scores during session one and self-reported ACT score).

To test hypotheses 6a, I used simple linear regression (SLR). A scatter-plot confirmed that a linear relationship appeared to be an appropriate fit for the data. After the model was run, the residuals were saved and plotted on a histogram; the histogram approximated a normal distribution. Residuals were plotted against the predictor variable (i.e., CAT score during session one), and the plot revealed no detectable pattern. Because all of the assumptions were met, I proceeded with the SLR analysis.

CAT score during session one did not explain a significant amount of variance in self-reported SAT score, R=.189, R^2=.036, $F(1, 89)$=3.293, p=.073. Participants' CAT during session one did not significantly predict participants' self-reported SAT score, b=0.004, se=0.002, t(89)=1.815, p=.073, β=0.189. Thus, I was unable to reject the null hypothesis, providing no evidence in support of hypothesis 6a in the student sample (see Table 52).

To test hypothesis 6b, I used multiple linear regression (MLR). I requested a histogram and p-plot of the residuals, both of which provided evidence in support of the assumption of normality. The predicted values were plotted against the unstandardized residuals, and the scatter

plot revealed no pattern, supporting the assumption of homogeneity of variance. Because all of the assumptions were met, I proceeded with the MLR analysis.

To test for main effects of experimental condition on participants' self-reported SAT score, dummy coded variables representing each of the four experimental factors (i.e., the subjective evaluation of the probability of passing the test with cheating, the subjective evaluation of the probability of being caught cheating, the subjective value of being caught cheating, proctored/ unproctored) were added to the model along with participants' CAT score during session one. A "1" on the dummy coded variables represents the higher version of the variable (e.g., high subjective evaluation of the probability of passing the test with cheating) whereas a "0" on the dummy coded variables represents the lower version of the variable (e.g., a low subjective evaluation of the probability of passing the test with cheating). This model was also non-significant, R=.217, R^2=.047, F(5, 85)=0.842, p=.524, nor was the additional amount of the variance in self-reported SAT score explained by the model, ΔR^2=.012, ΔF(4, 85)=0.257, p=.905 (see Table 53).

In this model, participants' CAT score during session one again failed to significantly predict participants' self-reported SAT score, b=0.004, se=0.002, t(85)=1.777, p=.079, β=0.194. The dummy coded variable representing the subjective evaluation of the probability of passing the test with cheating was non-significant, b=-0.003, se=0.030, t(85)=-0.097, p=.923, β=-0.011; participants' mean self-reported SAT score did not significantly differ as a function of the subjective evaluation of the probability of passing the test with cheating. The dummy coded variable representing the subjective evaluation of the probability of being caught cheating was non-significant, b=0.010, se=0.030, t(85)=0.331, p=.742, β=0.036; participants' mean scores on the self-reported ACT score did not significantly differ as a function of the subjective evaluation

of the probability of being caught cheating. The dummy coded variable representing the subjective value of being caught cheating was non-significant, b=0.002, se=0.030, t(85)=0.083, p=.934, β=0.009; participants' mean scores on the self-reported SAT score did not significantly differ as a function of the subjective value of being caught cheating. The dummy coded variable representing the proctored/ unproctored conditions was non-significant, b=0.072, se=0.076, t(85)=0.946, p=.347, β=0.109; participants' mean scores on the self-reported SAT score did not significantly differ as a function of the proctored/ unproctored conditions.

Next, interaction terms were created by multiplying each of the four experimental factors by participants' CAT score during session one, and then these terms were added to the model. These interaction terms represented the moderating effect of each experimental factor on the relationship between participants' CAT score during session one and participants' self-reported SAT score. This model was also non-significant, R=.268, R^2=.072, F(9, 81)=0.694, p=.713, as was the additional amount of variance explained by the interaction terms, ΔR^2=.024, ΔF(4, 81)=0.532, p=.713 (see Table 54). The interaction term representing the moderating effect of subjective evaluation of the probability of passing the test with cheating on the relationship between CAT score during session one and self-reported SAT score was non-significant, b=-0.001, se=0.005, t(81)=-0.110, p=.913, β=-0.044; the relationship between participants' CAT score during session one and participants' self-reported SAT score was not significantly moderated by the subjective evaluation of the probability of passing the test with cheating.

The interaction term representing the moderating effect of the subjective evaluation of the probability of being caught cheating on the relationship between CAT score during session one and self-reported SAT score was non-significant, b=-0.004, se=0.005, t(81)=-0.870, p=.387, β=-0.348; the relationship between participants' CAT scores during session one and participants'

self-reported ACT score was not significantly moderated by the subjective evaluation of the probability of being caught cheating.

The interaction term representing the moderating effect of the subjective value of being caught cheating on the relationship between CAT score during session one and self-reported SAT score was non-significant, b=0.002, se=0.005, t(81)=0.519, p=.605, β=0.218; the relationship between participants' CAT scores during session one and participants' self-reported SAT scores was not significantly moderated by the subjective value of being caught cheating.

The interaction term representing the moderating effect of the proctored/ unproctored conditions on the relationship between CAT score during session one and self-reported SAT score was non-significant, b=0.014, se=0.016, t(81)=0.853, p=.396, β=0.422; the relationship between participants' CAT scores during session one and participants' self-reported ACT score was not significantly stronger in the proctored condition, as compared to the unproctored condition.

Results revealed that condition did not moderate the relationship between CAT score during session one and self-reported SAT score, providing no support for hypothesis 6b in the student sample.

To test hypothesis 6c (i.e., that the relationship between CAT score during session one and self-reported SAT score will be moderated by self-reported cheating behavior), I added a dummy coded variable representing self-reported cheating (0=participant did not self-report cheating; 1=participant did self-report cheating) to the model regressing self-reported SAT score on CAT score during session one (i.e., hypothesis 6a). The model including the dummy coded cheating term was non-significant, R=.298, R^2=.089, F(2, 50)=2.442, p=0.097, as was the additional variance explained by the model, ΔR^2=.033, ΔF(1, 50)=1.812, p=.184 (see Table 55).

Self-reported cheating behavior was not a significant predictor of self-reported SAT score, b=-0.072, se=0.054, t(50)=-1.346, p=.184, β=-0.185.

Next, an interaction term was created by multiplying CAT score during session one by the dummy coded variable representing self-reported cheating behavior, and this term was added to the model. This interaction term represented the moderating effect of self-reported cheating behavior on the relationship between CAT score during session one and self-reported SAT score. This model was non-significant, R=.312, R^2=.098, F(3, 49)=0.464, p=.499, as was the additional amount of the variance in self-reported SAT score explained by the model, ΔR^2=.009, ΔF(1, 49)=.464, p=.499 (see Table 56). The term representing the interaction between CAT score during session one and self-reported cheating behavior did not significantly moderate the relationship between CAT score during session one and self-reported SAT score, b=-0.006, se=0.008, t(49)=-0.681, p=.499, β=0.008.

The results of the above analyses provide no evidence in support of hypothesis 6c: self-reported cheating behavior did not significantly moderate the relationship between CAT score during session one and self-reported SAT score in the student sample.

mTurk Sample

Hypotheses 6a, 6b, and 6c: CAT score will be significantly, positively related to self-reported SAT score; this relationship will be moderated by experimental condition and by self-reported cheating behavior. In the mTurk sample, I used linear regression to test hypotheses 6a, (i.e., that participants' CAT score would be significantly related to self-reported SAT score) and 6b (i.e., that experimental condition would moderate the relationship between participants' CAT scores and self-reported SAT score).

To test hypotheses 6a, I again used SLR. A scatter-plot confirmed that a linear relationship that a linear relationship existed between Y and X. The plot the residuals of the CAT scores on self-reported cheating behavior revealed no detectable pattern, confirming the assumption of homoscedasticity. Because all of the assumptions were met, I proceeded with the SLR analysis.

Studentized deleted residuals were calculated for all cases and used to detect outliers. Cases with a studentized deleted residual with an absolute value greater than two were removed from the analyses. This led to the removal of 17 cases.

CAT score explained a significant amount of variance in self-reported SAT, R=.269, R^2=.073, $F(1,178)$=13.925, p<.001. Participants' CAT score significantly predicted self-reported SAT (b=0.007, se=0.002, p<.001, β=0.269). Thus, I was able reject the null hypothesis, providing support for hypothesis 6a in the mTurk sample (see Table 57).

To test hypothesis 6b, I used multiple linear regression (MLR). I requested a histogram and p-plot of the residuals, both of which provided evidence in support of the assumption of normality. The predicted values were plotted against the unstandardized residuals, and this scatter plot revealed no pattern, supporting the assumption of homogeneity of variance. None of the part and partial plots revealed a detectable pattern. Because all of the assumptions were met, I proceeded with the MLR analysis.

To test for main effects of experimental condition on self-reported SAT, dummy coded variables representing each of the three experimental factors (i.e., the subjective evaluation of the probability of passing the test with cheating, the subjective evaluation of the probability of being caught cheating, the subjective value of being caught cheating) were added to the model along with participants' CAT score. A "1" on the dummy coded variables represents the higher version

of the variable (e.g., high subjective evaluation of the probability of passing the test with cheating) whereas a "0" on the dummy coded variables represents the lower version of the variable (e.g., a low subjective evaluation of the probability of passing the test with cheating). This model was also significant, R=.331, R^2=.110, F(4, 175)=5.393, p<.001, however the additional amount of variance explained by the experimental conditions was not, ΔR^2=.037, ΔF(3, 175)=2.436, p=.066 (see Table 58).

In this model, participants' CAT score again significantly predicted self-reported SAT, b=0.008, se=0.002, t(175)=4.115, p<.001, β=0.298. The dummy coded variable representing the subjective evaluation of the probability of passing the test with cheating was also significant, b=-0.040, se=0.018, t(175)=-2.234, p=.027, β=-0.162; participants in the condition in which the subjective likelihood of passing the test with cheating was high self-reported significantly lower SAT scores. The dummy coded variable representing the subjective evaluation of the probability of being caught cheating was non-significant, b=-0.020, se=0.018, t(175)=-1.133, p=.259, β=-0.082; self-reported SAT did not significantly differ as a function of the subjective evaluation of the probability of being caught cheating. The dummy coded variable representing the subjective value of being caught cheating was non-significant, b=0.019, se=0.018, t(175)=1.041, p=.299, β=0.075; self-reported SAT did not significantly differ as a function of the subjective value of being caught cheating.

Next, interaction terms were created by multiplying each of the three experimental factors by participants' CAT score, and then these terms were added to the model. These interaction terms represent the moderating effect of each experimental factor on the relationship between participants' CAT scores and self-reported SAT. This model was significant, R=.342, R^2=.117, F(7, 172)=3.244, p=.003, however the additional amount of variance explained by the

interaction terms was not, ΔR^2=.007, ΔF(3, 172)=0.447, p=.719 (see Table 59). The interaction term representing the moderating effect of subjective evaluation of the probability of passing the test with cheating on the relationship between CAT score and self-reported SAT was non-significant, b=0.003, se=0.004, t(172)=0.729, p=.467, β=0.319; the relationship between participants' CAT score and self-reported SAT was not significantly moderated by the subjective evaluation of the probability of passing the test with cheating.

The interaction term representing the moderating effect of the subjective evaluation of the probability of being caught cheating on the relationship between CAT score and self-reported SAT was non-significant, b=-0.001, se=0.004, t(172)=-0.286, p=.775, β=-0.121; the relationship between participants' CAT scores and self-reported SAT was not significantly moderated by the subjective evaluation of the probability of being caught cheating.

The interaction term representing the moderating effect of the subjective value of being caught cheating on the relationship between CAT score and self-reported SAT was non-significant, b=-0.003, se=.004, t(172)=-0.844, p=.400, β=-0.337; the relationship between participants' CAT scores and self-reported SAT was not significantly moderated by the subjective value of being caught cheating.

The results suggest that experimental condition did not moderate the relationship between CAT score and self-reported SAT score, providing no evidence in support of hypothesis 6b in the mTurk sample.

To test hypothesis 6c (i.e., that the relationship between CAT score and self-reported SAT will be moderated by self-reported cheating behavior), I added a dummy coded variable representing self-reported cheating (0=participant did not self-report cheating; 1=participant did self-report cheating) to the model regressing self-reported SAT on CAT score (i.e., hypothesis

6a). Studentized deleted residuals were calculated for all cases and used to detect outliers. Cases with a studentized deleted residual with an absolute value greater than two were removed from the analyses. This led to the removal of 15 cases.

The model including the dummy coded cheating term was significant, R=.281, R^2=.079, F(2, 179)=7.651, p=.001, explaining an additional 2.1% of the variance in self-reported SAT, ΔR^2=.021, ΔF(1, 179)=4.151 p=.043 (see Table 60). Self-reported cheating behavior was a significant predictor of self-reported SAT, b=-0.064, se=0.031, t(179)=-2.038, p=.043, β=-0.146; participants who self-reported cheating on the CAT during session reported scoring, on average,-0.064 points lower on the SAT than participants who did not self-report cheating on the CAT .

Next, an interaction term was created by multiplying CAT score by the dummy coded variable representing self-reported cheating behavior, and this term was added to the model. This interaction term represented the moderating effect of self-reported cheating behavior on the relationship between CAT score and self-reported SAT. This model was significant, R=.299, R^2=.090, F(3, 178)=5.838, p=.001, but the additional amount of variance explained by the interaction terms was not, ΔR^2=.011, ΔF(1, 178)=2.117, p=.147 (see Table 61). The term representing the interaction between CAT score and self-reported cheating behavior did not significantly predict self-reported SAT, b=-0.009, se=0.006, t(178)=-1.455, p=.147, β=-0.541.

The results of the above analyses provide no evidence in support of hypothesis 6c, that self-reported cheating behavior significantly moderated the relationship between CAT score and self-reported SAT in the mTurk sample.

Student Sample

Hypotheses 7a, 7b, and 7c: CAT score during session one will be significantly, positively related to self-reported ACT score; this relationship will be moderated by experimental condition and by self-reported cheating behavior. In the student sample, I used linear regression to test hypotheses 7a, (i.e., that participants' CAT score during session one would be significantly related to participants' self-reported ACT score) and 7b (i.e., that experimental condition would moderate the relationship between participants' CAT scores during session one and self-reported ACT score).

To test hypotheses 7a, I used simple linear regression (SLR). A scatter-plot confirmed that a linear relationship appeared to be an appropriate fit for the data. After the model was run, the residuals were saved and plotted on a histogram; the histogram approximated a normal distribution. Another assumption associated with SLR is homoscedasticity. Residuals were plotted against the predictor variable (i.e., CAT score during session one). The plot revealed no detectable pattern. Because all of the assumptions were met, I proceeded with the SLR analysis.

CAT score during session one explained a significant amount of variance in self-reported ACT score, R=.456, R^2=.208, $F(1, 316)$=83.082, p<.001. Participants' CAT during session one significantly predicted participants' self-reported ACT score (b=0.244, se=0.027, p<.001, β=0.456). Thus, I was able to reject the null hypothesis, providing evidence for hypothesis 7a in the student sample (see Table 62).

To test hypothesis 7b, I used multiple linear regression (MLR). I requested a histogram and p-plot of the residuals, both of which provided evidence in support of the assumption of normality. Predicted values were plotted against the unstandardized residuals. The scatter plot

revealed no pattern, supporting the assumption of homogeneity of variance. Because all of the assumptions were met, I proceeded with the MLR analysis.

To test for main effects of experimental condition on participants' Self-reported ACT score, dummy coded variables representing each of the four experimental factors (i.e., the subjective evaluation of the probability of passing the test with cheating, the subjective evaluation of the probability of being caught cheating, the subjective value of being caught cheating, proctored/ unproctored) were added to the model along with participants' CAT score during session one. A "1" on the dummy coded variables represents the higher version of the variable (e.g., high subjective evaluation of the probability of passing the test with cheating) whereas a "0" on the dummy coded variables represents the lower version of the variable (e.g., a low subjective evaluation of the probability of passing the test with cheating). This model was also significant, R=.495, R^2=.245, $F(5, 312)$=20.293, p<.001, as was the additional amount of the variance in self-reported ACT score explained by the model, ΔR^2=.037, $\Delta F(4, 312)$=3.847, p=.005 (see Table 63).

In this model, participants' CAT score during session one continued to significantly predict participants' self-reported ACT score, b=0.248, se=0.026, $t(312)$=9.370, p<.001, β=0.464. The dummy coded variable representing the subjective evaluation of the probability of passing the test with cheating was also significant, b=1.236, se=0.360, $t(312)$=3.439, p<.001, β=0.176; participants in the condition in which the probability of passing the test with cheating reported scoring, on average, 1.236 points higher on the ACT. The dummy coded variable representing the subjective evaluation of the probability of being caught cheating was non-significant, b=-0.323, se=0.358, $t(312)$=-0.900, p=.369, β=-0.046; participants' mean scores on self-reported ACT did not significantly differ as a function of the subjective evaluation of the

probability of being caught cheating. The dummy coded variable representing the subjective value of being caught cheating was non-significant, b=-0.355, se=0.359, t(312)=-0.990, p=.323, β=-0.051; participants' mean scores on self-reported ACT did not significantly differ as a function of the subjective value of being caught cheating. The dummy coded variable representing the proctored/ unproctored conditions was non-significant, b=0.931, se=0.806, t(312)=1.156, p=.249, β=0.063; participants' mean scores on self-reported ACT did not significantly differ as a function of the proctored/ unproctored conditions.

Next, interaction terms were created by multiplying each of the four experimental factors by participants' CAT score during session one, and then these terms were added to the model. These interaction terms represent the moderating effect of each experimental factor on the relationship between participants' CAT score during session one and participants' self-reported ACT score. This model was significant, R=.511, R^2=.262, F(9, 308)=12.126, p<.001, but the additional amount of variance explained by the interaction terms was not, ΔR^2=.016, ΔF(4, 308)=1.692, p=.152 (see Table 64). The interaction term representing the moderating effect of subjective evaluation of the probability of passing the test with cheating on the relationship between CAT score during session one and self-reported ACT score was significant, b=-0.133, se=0.056, t(308)=-2.381, p=.018, β=-0.439; the relationship between participants' CAT score during session one and participants' self-reported ACT was significantly moderated by the subjective evaluation of the probability of passing the test with cheating, such that the relationship was weaker for participants in conditions in which the subjective probability of passing the test with cheating was high.

The interaction term representing the moderating effect of the subjective evaluation of the probability of being caught cheating on the relationship between CAT score during session one

and self-reported ACT score was non-significant, b=0.035, se=0.054, t(308)=.650, p=.516, β=0.114; the relationship between participants' CAT scores during session one and participants' self-reported ACT scores was not significantly moderated by the subjective evaluation of the probability of being caught cheating.

The interaction term representing the moderating effect of the subjective value of being caught cheating on the relationship between CAT score during session one and self-reported ACT score was non-significant, b=-0.052, se=0.055, t(308)=-0.939, p=.348, β=-0.171; the relationship between participants' CAT scores during session one and participants' self-reported ACT score was not significantly moderated by the subjective value of being caught cheating.

The interaction term representing the moderating effect of the proctored/ unproctored conditions on the relationship between CAT score during session one and self-reported ACT score was non-significant, b=-0.078, se=0.142, t(308)=-0.551, p=.582, β=-0.119 the relationship between participants' CAT scores during session one and participants' self-reported ACT score was not significantly stronger in the proctored condition, as compared to the unproctored condition.

Results revealed that condition did not moderate the relationship between CAT score during session one and self-reported ACT score, except for conditions in which the probability of passing the test with cheating was high, providing partial support for hypothesis 7b in the student sample.

To test hypothesis 7c (i.e., that the relationship between CAT score during session one and Self-reported ACT score will be moderated by self-reported cheating behavior), I added a dummy coded variable representing self-reported cheating (0=participant did not self-report cheating; 1=participant did self-report cheating) to the model regressing self-reported ACT score

on CAT score during session one (i.e., hypothesis 7a). The model including the dummy coded cheating term was significant, R=.553, R^2=.305, F(2, 151)=33.187, p<.001, explaining an additional 7.5% of the variance in self-reported ACT score, ΔR^2=.075, ΔF(1, 151)=16.310, p<.001 (see Table 65). Self-reported cheating behavior was a significant predictor of self-reported ACT score, b=-2.512, se=0.622, t(151)=-4.039, p<.001, β=-0.275; participant who self-reported cheating on the CAT during session one scored, on average, 2.512 points lower on self-reported ACT than participants who did not self-report cheating on the CAT during session one.

Next, an interaction term was created by multiplying CAT score during session one by the dummy coded variable representing self-reported cheating behavior, and this term was added to the model. This interaction term represented the moderating effect of self-reported cheating behavior on the relationship between CAT score during session one and self-reported ACT score. This model was significant, R=.566, R^2=.309, F(3, 150)=22.361, p<.001, however the additional variance in self-reported ACT score explained by the model was not, ΔR^2=.004, ΔF(1, 150)=0.797, p=.374 (see Table 66). The term representing the interaction between CAT score during session one and self-reported cheating behavior did not significantly moderate the relationship between CAT score during session one and self-reported ACT score, b=-0.083, se=0.093, t(150)=-0.893, p=.374, β=-0.219.

The results of the above analyses provide no evidence in support of hypothesis 7c: self-reported cheating behavior did not significantly moderate the relationship between CAT score during session one and self-reported ACT score in the student sample.

mTurk Sample

Hypotheses 7a, 7b, and 7c: CAT score will be significantly, positively related to self-reported ACT; this relationship will be moderated by experimental condition and by self-reported cheating behavior. In the mTurk sample, I used linear regression to test hypotheses 7a, (i.e., that participants' CAT score would be significantly related to self-reported ACT) and 7b (i.e., that experimental condition would moderate the relationship between participants' CAT scores and self-reported ACT).

To test hypotheses 7a, I used simple linear regression (SLR). A scatter-plot confirmed that a linear relationship appeared to be an appropriate fit for the data. After the model was run, the residuals were saved and plotted on a histogram; the histogram approximated a normal distribution. Residuals were plotted against the predictor variable (i.e., CAT score), and the plot revealed no detectable pattern. Because all of the assumptions were met, I proceeded with the SLR analysis.

Studentized deleted residuals were calculated for all cases and used to detect outliers. Cases with a studentized deleted residual with an absolute value greater than two were removed from the analyses. This led to the removal of 7 cases.

CAT score explained a significant amount of variance in self-reported ACT, R=.669, R^2=.447, $F(1,142)$=114.918, p<.001. Participants' CAT score significantly predicted self-reported ACT (b=0.595, se=0.056, p<.001, β=0.669). Thus, I was able reject the null hypothesis, providing support for hypothesis 7a in the mTurk sample (see Table 67).

To test hypothesis 7b, I used multiple linear regression (MLR). I requested a histogram and p-plot of the residuals, both of which provided evidence in support of the assumption of normality. Predicted values were plotted against the unstandardized residuals, and the scatter plot

revealed no pattern, supporting the assumption of homogeneity of variance. None of the part and partial plots revealed a detectable pattern. Because all of the assumptions were met, I proceeded with the MLR analysis.

To test for main effects of experimental condition on self-reported ACT, dummy coded variables representing each of the three experimental factors (i.e., the subjective evaluation of the probability of passing the test with cheating, the subjective evaluation of the probability of being caught cheating, the subjective value of being caught cheating) were added to the model along with participants' CAT score. A "1" on the dummy coded variables represents the higher version of the variable (e.g., high subjective evaluation of the probability of passing the test with cheating) whereas a "0" on the dummy coded variables represents the lower version of the variable (e.g., a low subjective evaluation of the probability of passing the test with cheating). This model was also significant, R=.464, R^2=.449, F(4, 139)=30.114, p<.001, however the additional amount of variance explained by the experimental conditions was not, ΔR^2=.017, ΔF(3, 139)=1.467, p=.226 (see Table 68).

In this model, participants' CAT score again significantly predicted self-reported ACT, b=0.636, se=0.059, t(139)=10.840, p<.001, β=0.714. The dummy coded variable representing the subjective evaluation of the probability of passing the test with cheating was also significant, b=-1.208, se=0.583, t(139)=-2.074, p=.040, β=-0.136; participants in the condition in which the subjective likelihood of passing the test with cheating was high self-reported significantly lower ACT scores. The dummy coded variable representing the subjective evaluation of the probability of being caught cheating was non-significant, b=-0.153, se=0.552, t(139)=-0.278, p=.782, β=-0.017; self-reported ACT did not significantly differ as a function of the subjective evaluation of the probability of being caught cheating. The dummy coded variable representing the subjective

value of being caught cheating was non-significant, b=0.094, se=0.554, t(139)=0.170, p=.866, β=0.011; self-reported ACT did not significantly differ as a function of the subjective value of being caught cheating.

Next, interaction terms were created by multiplying each of the three experimental factors by participants' CAT score, and then these terms were added to the model. These interaction terms represent the moderating effect of each experimental factor on the relationship between participants' CAT score and self-reported ACT. This model was significant, R=.684, R^2=.468, F(7, 136)=17.104, p<.001, however the additional amount of variance explained by the interaction terms was not, ΔR^2=.004, ΔF(3, 136)=0.335, p=.800 (see Table 69). The interaction term representing the moderating effect of subjective evaluation of the probability of passing the test with cheating on the relationship between CAT score and self-reported ACT was non-significant, b=-0.105, se=0.124, t(136)=-0.853, p=.395, β=-0.324; the relationship between participants' CAT score and self-reported ACT was not significantly moderated by the subjective evaluation of the probability of passing the test with cheating.

The interaction term representing the moderating effect of the subjective evaluation of the probability of being caught cheating on the relationship between CAT score and self-reported ACT was non-significant, b=0.002, se=0.117, t(136)=0.019, p=.985, β=0.006; the relationship between participants' CAT scores and self-reported ACT was not significantly moderated by the subjective evaluation of the probability of being caught cheating.

The interaction term representing the moderating effect of the subjective value of being caught cheating on the relationship between CAT score and self-reported ACT was non-significant, b=0.042, se=0.122, t(136)=0.341, p=.734, β=0.120; the relationship between

participants' CAT scores and self-reported ACT was not significantly moderated by the subjective value of being caught cheating.

The results suggest that experimental condition did not moderate the relationship between CAT score and self-reported ACT, providing no evidence in support of hypothesis 7b in the mTurk sample.

To test hypothesis 7c (i.e., that the relationship between CAT score and self-reported ACT will be moderated by self-reported cheating behavior), I added a dummy coded variable representing self-reported cheating (0=participant did not self-report cheating; 1=participant did self-report cheating) to the model regressing self-reported ACT on CAT score (i.e., hypothesis 7a). Studentized deleted residuals were calculated for all cases and used to detect outliers. Cases with a studentized deleted residual with an absolute value greater than two were removed from the analyses. This led to the removal of six cases.

The model including the dummy coded cheating term was significant, R=.662, R^2=.438, $F(2, 142)$=55.243, p<.001, explaining no additional variance in self-reported ACT, ΔR^2=.000, $\Delta F(1, 142)$=0.020 p=.888 (see Table 70). Self-reported cheating behavior was not a significant predictor of self-reported ACT, b=0.117, se=0.829, $t(142)$=0.141, p=.888, β=0.009; there was no significant difference in self-reported ACT scores between participants who self-reported and did not self-report cheating on the CAT.

Next, an interaction term was created by multiplying CAT score by the dummy coded variable representing self-reported cheating behavior, and this term was added to the model. This interaction term represented the moderating effect of self-reported cheating behavior on the relationship between CAT score and self-reported ACT. This model was significant, R=.663, R^2=.440, $F(3, 141)$=36.872, p<.001, but the additional amount of variance explained by the

interaction terms was not, ΔR^2=.002, ΔF(1, 141)=0.511, p=.476 (see Table 71). The term representing the interaction between CAT score and self-reported cheating behavior did not significantly predict self-reported ACT, b=-0.102, se=0.142, t(141)=-0.715, p=.476, β=-0.192.

The results of the above analyses provide no evidence in support of hypothesis 7c, that self-reported cheating behavior significantly moderated the relationship between CAT score and self-reported ACT in the mTurk sample.

DISCUSSION

Overview

The current study employed an experimental protocol to test the effectiveness of utility theory as a framework for understanding the decision making process in regard to cheating on online tests. Online testing is an increasingly popular selection tool, but human resource managers are rightfully worried about the negative impact that cheating has on the validity and utility of their selection systems (Tippins, 2006). The purpose of the current study was to apply a theoretical framework to understand cheating behavior, to estimate the prevalence of cheating in two different samples, and to estimate the impact of cheating on the validity of an online test.

This study relied heavily on deception. If participants knew the true purpose of the study was to research cheating behavior, that knowledge might have influenced cheating behavior. For instance, if participants knew the only reason they were required to take a long, boring vigilance task was to encourage them to cheat, and that there was no punishment for cheating, they may have been more likely to cheat. Conversely, if participants knew that cheating behavior was being investigated, they may have been less likely to cheat because of their own self-image as a person that does not cheat, even if there were no direct consequences of being caught. Thus, effective deception was vital to the outcome of this study.

The experimental protocol seemed successful in achieving two goals integral to the study. First, only a small percentage of participants were able to guess the true purpose of the study prior to debriefing: five of 518 (1.0%) participants in the mTurk sample, and 10 of 162 (6.2%) participants in the student sample reported that they believed the study was actually about cheating. One possible reason that a significantly greater portion of participants from the student sample suspected deception in the study is that these participants were involved in several

psychological studies in addition to the current study, and many of them were students in *Introduction to Psychology*, so it is likely that they were more familiar with the methods used in psychological research, and the possibility of deception, than participants in the mTurk sample.

The experimental protocol was also successful in providing an opportunity to cheat for those participants who so desired. In total, 86 of 680 (12.7%) participants self-reported cheating behavior. Not everyone who cheated necessarily admitted to it, meaning that 12.7% represents the minimum number of individuals who cheated in this sample.

The evidence discussed above suggests that the experimental protocol was effective at obscuring the true purpose of the study and providing participants with opportunities to cheat. Below, we discuss the theoretical and practical contributions of the study, as well as the strengths and limitations, and directions for future research.

Theoretical Contributions

One of the main goals of this study was to provide a theoretical framework to help understand why some individuals decide to cheat on online tests. The theoretical framework chosen for this purpose was utility theory, which, in brief, states that individuals consider the probability and value of possible outcomes when making decisions (Rettinger, 2007). In accordance with this theory, three experimental factors were manipulated in attempt to influence the decision making process: the probability of passing the test with cheating, the probability of being caught cheating, and the value of being caught cheating. The results of these manipulations are discussed below.

Although the experimental protocol was effective at creating a scenario in which participants were unaware that the true purpose of the study was to investigate cheating and in which opportunities for cheating existed, the experimental manipulations were less successful.

The only manipulation that was significant across any of the analyses was the probability of passing the test with cheating (i.e., providing access to an answer key). In the mTurk sample, participants in conditions in which they had access to the answer key scored higher on the CAT, on average, than participants in conditions without access to the answer key. The eta-squared for the factor representing the probability of passing the test with cheating was .073, considered a medium effect size (Cohen, 1988). In the student sample, probability of passing the test with cheating moderated the relationship between CAT scores during session one and CAT scores during session two, such that the relationship was weaker for participants who had access to the answer key. This can be interpreted as evidence that a greater number of participants in those conditions cheated, moderating the relationship between the two CAT scores. The standardized slope for the interaction term was .579, generally considered to be a large effect size (Cohen, 1988). These results provided partial support for the appropriateness of the application of utility theory to understanding cheating behavior in online testing scenarios.

The other experimental manipulations (i.e., the probability of being caught cheating and the value of being caught cheating) did not significantly influence CAT score, self-reported cheating behavior, or the relationship between CAT score and other measures of cognitive ability in either of the samples. There are two potential explanations for this: the application of utility theory was correct, but the manipulations were unsuccessful, or the manipulations were successful, but utility theory does not provide a useful framework for understanding judgment and decision-making in regard to cheating on online tests. Both explanations are discussed below.

At the end of the second session, once the true purpose of the study was revealed to participants, participants who reported not having cheated on the exam were asked why they

decided not to cheat. Several of the participants in the mTurk sample responded that they were worried about not getting paid, or about having their submissions rejected (which negatively affects their worker rating on mTurk, and thus their ability to complete subsequent assignments and earn money). Although these qualitative data were not formally analyzed, and thus should be interpreted with caution, this does imply that at least some of the workers were considering the consequences of the decision to cheat, which is in agreement with utility theory. This suggests that utility theory may provide a useful framework for understanding judgment and decision-making in regard to cheating on online tests, but that the experimental manipulations may not have been successful at affecting participants perceptions of the probability of passing the test with cheating, the probability of being caught cheating, and/ or the value of being caught cheating. Perhaps the "low value of being caught cheating" condition (in which participants were told that if they were caught cheating they would have to retake the cognitive ability test) actually represented a high value to mTurk workers. These workers are paid by the number of assignments they complete, and having to invest time in retaking the cognitive ability test would have equated to losing the opportunity to complete another assignment and earn pay, which may have actually represented a large value to these workers.

Across both samples, however, several participants cited personal characteristics in their explanations for not cheating, including references to integrity and personal beliefs against cheating. Utility theory does not take individual beliefs such as these into account when describing the judgment and decision making process. If decisions regarding cheating behavior are driven mainly by personal beliefs, then perhaps utility theory is not the most appropriate framework to understand judgment and decision making in regard to cheating on online tests.

I tend to agree with both interpretations. On the one hand, situational factors, specifically access to the answer key, did affect CAT scores in the expected direction in the mTurk sample. This suggests that participants were weighing some aspects of the situation when deciding whether or not to cheat. It is possible then that the other manipulations were simply not effective at affecting participants' perceptions of the likelihood and value of the outcomes of their decisions. I think the manipulations may have been based too strongly in the mathematical aspect of utility theory. For example, in the "high probability of being caught cheating" condition, participants were told there was a 90% chance of being caught cheating, as opposed to 10% in the "low probability of being caught cheating" condition. Discussing the possibility of cheating at all, however, may have made participants aware that there was a possibility they would be caught cheating. A more effective manipulation might have been to emphasize the probability of being caught cheating in the "high probability of being caught cheating" condition vs. not mention that possibility at all in the "low probability of being caught cheating" condition. The percentage chance of being caught cheating may have had little effect compared to merely bringing up the chance that participants might be caught cheating.

On the other hand, even when presented with the answers and the opportunity to use them to improve their scores, the majority of participants did not self-report cheating. This implies that there is a limit to the percent of variance in cheating behavior that utility theory can explain. Cheating decisions might be based much more in individual differences, for example. However, from a practical perspective, organizations probably have a limited ability to influence individual differences such as moral beliefs before or during online testing. Even if utility theory is capable of explaining only a small percentage of the variance in cheating behavior, it might be the most

practical framework for organizations to conceive of and design anti-cheating interventions during online testing.

As Tippins (2006; 2009b) noted, within the organizational literature there is little research into why people cheat and which conditions facilitate or prevent cheating. Within the educational literature, Rettinger (2007) argued that cheating can best be understood as a decision, and that a judgment and decision making theory, specifically utility theory, could be helpful in understanding cheating behavior. One of the main goals of the current study was, following Rettinger, to apply utility theory to understand individuals' decisions regarding whether to cheat in UIT.

The results of this study add to the literature by highlighting the limitations of utility theory, and arguably decision-making theories in general, for understanding cheating behavior. The mixed, and largely non-significant findings, of the current study call into question the appropriateness of utility theory for understanding cheating decisions in UIT. As discussed earlier, one of the reasons may be that individual characteristics, such as personality, have a greater influence on cheating decisions than aspects of the environment. Another limitation of utility theory often discussed by researchers is that utility theory presents decision-making as a non-emotional, mathematical process, although research has shown that emotions have a relatively large impact on decision-making processes (e.g., Weber & Johnson, 2009). The results of this study can be interpreted as a call to other researchers to understand cheating behavior from a more holistic perspective, including not only decision-making processes, but personality traits and emotional states.

As Rettinger (2007) argued, utility theory seems to provide a useful framework for understanding cheating behavior. It would also seem to offer a useful framework for

understanding differences between cheating behaviors in proctored and unproctored settings. For example, utility theory posits that, all else being equal, when the probability of being caught cheating is high, or when the probability of cheating successfully is low, individuals will be less likely to cheat. Compared to traditional, proctored tests, UIT implicitly creates a situation where participants are less likely to be caught cheating, and more likely to cheat successfully, theoretically increasing the likelihood that test-takers will cheat. Although I was not very effective at using UIT to manipulate cheating behavior in this study, UIT might still be a useful tool for understanding differences in cheating behavior in proctored and unproctored settings. Thus, the results of this study can be used to provide direction to practitioners comparing the benefits and drawbacks of traditional, proctored testing versus UIT.

Another limitation regularly discussed in both the educational and organizational literatures is detecting cheating (Guo & Drasgow, 2010; Haney & Clark, 2007). The current study added to the literature by testing several methods of detecting cheating, though largely at the group level. In line with previous research in both the educational and organizational literature (e.g., Arthur, Glaze, Villado, & Taylor, 2009; Beaty, Fallon, Shepherd, & Barrett, 2002; Haney & Clark, 2007) mean differences in test scores were found between groups in which cheating was easier versus groups in which cheating was harder. Differences were also found in validity coefficients for the CAT depending on condition, as well as differences in the proportion of self-reported cheaters versus non-self-reported cheaters who "passed" the cognitive ability test. The use of mean differences as an indication of cheating is questionable in previous research, namely because differences in scores might be due not just to differences in the prevalence of cheating, but due to differences in testing conditions, motivation, or practice effects (when studies utilize a within-person design; e.g. Do, 2005). This study uniquely

contributed to the literature by including a condition in which participants were proctored at both times one and two.

The current research also contributed to the literature by utilizing a novel method for detecting cheaters, that is, the use of impossible items that could only be answered correctly through guessing or the use of the answer key. Although the analyses which used the fake items as an outcome variable were all non-significant, many of them were marginally significant, implying that this method might be useful for detecting cheating at the group or individual level.

Finally, the current study contributed to the literature by providing evidence that self-reported cheating can provide at least a rough estimate of cheating behavior. Although it is impossible to know what percentage of actual cheaters self-reported cheating behavior, requesting self-reported cheating behavior was shown to be a useful tool at least for estimating minimum rates of cheating within the samples.

Practical Contributions

Although two of the theoretical factors that were manipulated did not significantly affect cheating behavior in the current study, several practical findings were gleaned from the results.

Manipulating the probability of passing the test with cheating significantly affected CAT scores in the mTurk sample. Though it may seem obvious that providing participants with access to an answer key will increase test performance, this situation is analogous to the situation in which job candidates complete selection tests online. When taking a test in an unproctored situation, job candidates potentially have access to nearly unlimited resources that could provide answers to selection tests, such as the Internet, reference books, and knowledgeable acquaintances (Tippins, 2006). The results of this study suggest that access to these types of resources inflates test scores and undermines test validity, at least with cognitive ability tests

(some of the most popular selection instruments; Lievens & Burke, 2011). Other types of tests, such as projective tests or personality tests, might be more difficult to cheat on because the "correct" answer is less objective, and may be less obvious to test takers.

Another practical contribution of this study is providing an estimate of the prevalence of cheating on online tests in two different samples. In the mTurk sample, 10.8% of participants self-reported cheating, and in the student sample, 15.6% of participants self-reported cheating. As noted earlier, this is not the percentage of participants that actually cheated, but the percentage of participants that admitted cheating, representing a lower end of the prevalence of cheating in these samples. Past research has shown that individuals are often unwilling to admit to deviant behavior, even when guaranteed anonymity, especially in regard to cheating. For example, a study by Erickson and Smith (1974) compared actual cheating behavior (as measured through direct observation) to self-reported cheating behavior in a sample of 118 undergraduate students. In this study, 10.2% of participants self-reported cheating behavior, whereas 43.2% were directly observed cheating; only 23.5% of cheaters admitted to cheating. These results imply that cheating may have been substantially underreported in the current study. This is supported by the finding that, in the mTurk sample, access to the answer key was significantly related to higher CAT scores, whereas access to the answer key was not significantly related to self-reported cheating, implying that some participants may have cheated (and inflated their CAT scores with the answer key) without reporting it.

The prevalence of self-reported cheating in this study, especially when interpreted as a minimum estimate of the true prevalence of cheating, has important practical implications. Many organizations use online selection tests as a first hurdle, and participants who perform well enough on these initial tests are then invited to take a shorter form of the test in a proctored

situation in order to validate their original scores (Tippins, 2006). The results of this study imply that, at least in some samples, the prevalence of cheating is high enough to support the use of these validation tests for differentiating true high performers from those job candidates who may have cheated.

Cheating only poses a problem for identifying qualified candidates if cheating is effective at falsely inflating test scores. For many of the participants, cheating was effective; self-reported cheaters in both samples were significantly more likely to "pass" the cognitive ability test (i.e., perform well enough to be excused from the vigilance task, saving 35 minutes of additional work). In the student sample, self-reported cheating moderated the relationship between CAT performance during session one and CAT performance during session two, analogous to the relationship between selection test scores and job performance in selection scenarios. This provides evidence that cheating undermines the validity of selection tests, a finding that organizations should consider before deciding to use online tests as part of a selection system. These results suggest that cheating does occur in online testing situations, and that it poses a potential problem for employers searching for the most qualified candidates.

Another interesting finding is that many of the self-reported cheaters were unsuccessful. In both samples, only a small minority of self-reported cheaters actually performed well enough on the CAT to be excused from the vigilance task (7.1% of self-reported cheaters in the mTurk sample; 14.9% in the student sample). Furthermore, in neither condition did self-reported cheaters perform significantly better on the CAT, on average, than participants who did not self-report cheating behavior. In light of the evidence that only a minority of cheaters are actually admitting cheating behavior (Ericksom & Smith, 1974), it might be that only ineffective cheaters self-reported their cheating behavior. An alternative explination has to do with participants'

motivation to cheat; several participants reported looking up answers to the CAT items because they were curious or frustrated, not necessarily to "pass" the CAT. Another explanation is that cheating effectively on cognitively ability tests is itself a task that requires high cognitive ability. Participants can search the Internet for the definition of words used in an anology, for example, but they cannot search the Internet for the relationship between the words. Results of the chi-square test of independence revealed no significant difference in the number of self-reported cheaters across the experimental conditions in either sample. Cheaters in conditions in which the answer key was provided may have had an easy time of cheating, but cheaters in conditions without access to the answer key may have had to rely on other, less effective cheating strategies (e.g., searching the Internet). Perhaps only individuals with higher cognitive ability were able to cheat effectively in these conditions.

Results of an informal analysis support this explanation. The SAT and ACT scores of successful mTurk cheaters (i.e., self-reported cheaters who performed well enough on the CAT to be excused from the vigilance task) were compared to those of unsuccessful cheaters (i.e., self-reported cheaters who failed to perform well enough on the CAT to be excused from the vigilance task). If cheating successfully requires higher cognitive ability, we would expect successful cheaters to have, on average, higher cognitive ability as compared to unsuccessful cheaters. Although the sample size was much too small to yield significant results (e.g., only four participants self-reported cheating and performed well enough to be excused from the vigilance task in the mTurk sample), the absolute differences on self-reported SAT and ACT between the groups were quite large and in the expected direction. In the mTurk sample, the average self-reported SAT score (expressed as a percentage) was 80.0% for successful cheaters (n=4); for unsuccessful cheaters (n=52) it was nearly 10% lower: 71.5%. A similar pattern was found for

ACT scores: for successful cheaters, the mean ACT score was 31, compared to 26.6 for unsuccessful cheaters. In the student sample, the average self-reported SAT score was slightly higher for successful cheaters (m=63.0%, n=4) compared to unsuccessful cheaters (m=62.1%; n=3), as was the average self-reported ACT score, m=24.1 for successful cheaters (n=7) and m=22.5 for unsuccessful cheaters (n=18). Although underpowered, these results do provide evidence of a trend that successful cheaters have higher cognitive ability than unsuccessful cheaters, supporting the contention that a certain level of cognitive ability is necessary for effective cheating.

Websites and forums do exist where job candidates share questions and answers from selection tests, so there are situations in selection scenarios in which job candidates essentially have access to "answer keys," (Tippins, 2009). For many other selection tests, however, job candidates must rely on their own cognitive ability and ingenuity to cheat. Because many of the self-reported cheaters failed to perform well enough on the cognitive ability test to be excused from the vigilance study, the results of the current study imply that many of these cheating strategies are ineffective.

Strengths

There were several strengths to this study that represented an improvement over earlier research on cheating on online tests.

One of the major strengths of this study, as compared to earlier studies on cheating on online tests, was the amount of control provided by a lab study. Most of the previous research on cheating on online tests has been field research (e.g., Arthur, 2009; Do, 2005; Nye, 2008). Although these studies are useful for understanding cheating in applied contexts, they lack the control available in lab studies. For example, in field research it is usually unethical or

impractical to manipulate aspects of the testing situation to make cheating more or less attractive. In this study, I manipulated the probability of passing the test with cheating, the likelihood of being caught cheating, and the value of being caught cheating in an attempt to change participants' cheating behavior. I was also able to gather self-reports of cheating behavior in a situation in which participants were relatively free of fear of reprisal (at least compared to employment scenarios). The estimates of cheating in this study were substantially higher than in the field studies mentioned above. For example, Arthur (2009) estimated that only 7.77% of participants cheated, and Nye (2008) identified only four candidates out of 856 as likely cheaters; because thist study provided a relatively safe and anonymous way to self-report cheating behavior, the higher percentage of self-reported cheaters in this study may be a more accurate estimate of the true prevalence of cheating, though likely still represents an under-estimation.

Another benefit of doing this type of research within the context of a lab study is that we were able to collect data on self-reported cheating. Most job candidates would probably avoid admitting to a potential employer that they cheated on a selection test. Within the context of a research study, however, we were able to assure participants of their anonymity and protection. Although it is impossible to know what percentage of cheaters actually admitted to cheating in this study, the willingness of approximately 13% of the sample to admit to cheating provides evidence that at least some participants had faith in the researchers' assurances of anonymity and protection. These data on self-reported cheating allowed us to test theoretically and practically significant hypotheses, such as the prevalence of cheating in the samples, which conditions affected cheating behavior, how cheating affected performance, and how cheating affected the relationship between the unproctored, online CAT and other measures of cognitive ability.

Collecting data on several measures of cognitive ability (i.e., self-reported SAT and ACT scores, CAT scores in a proctored setting), as well as the presence of a true control condition, represent other strengths of this study. One of the main concerns that organizations have about cheating is not whether cheating effects mean scores (although this is important), but whether, and by how much, cheating affects the validity of tests (Lievens & Burke, 2011).

Because I collected data on other measures of cognitive ability, we were able to estimate the impact of cheating on the relationship between the online cognitive ability test and these measures. Furthermore, the presence of a true control condition (i.e., the condition which took the test twice under proctored circumstances) provided a benchmark estimate of the relationship between the CAT scores during sessions one and two. Test performance can change over time for a number of reasons (e.g., differences in versions of the test, changes in participant motivation, and practice effects; Nye, 2008). Earlier studies (e.g., Nye, 2008) have relied on statistical estimates of regression to the mean and practice effects to correct for these changes, which, although useful, may not accurately reflect the influence of these factors on a particular sample's changes in performance.

Another strength of this study, compared to earlier field studies, is that everyone in the student sample who took the initial CAT was retested. Often in field studies, only those individuals who perform well enough on the initial selection test to be considered for employment are retested, restricting the range of scores available for analysis (e.g., Do, 2005). In our study, all participants in the student sample retook the CAT, regardless of their initial performance, maintaining a range of scores across both testing sessions.

Of course, one of the major criticisms of lab studies is that they are unrealistic, especially when they rely on student samples. A situation in which candidates are completing a selection

test in order to be considered for employment is fundamentally different from a situation in which student participants are completing a test for research credits. To overcome that criticism, we tested our study with two samples: a student sample and an mTurk sample. Although an mTurk sample is not a perfect analog to a sample of candidates completing selection tests, there are similarities. mTurk provides a sample of workers who know they are being evaluated on their performance, much as potential job candidates know they are being evaluated by their potential employers. Previous research has shown mTurk samples to be as reliable as student samples (Buhrmester, Kwang, & Gosling, 2011; Sprouse, 2010), as well as more ethnically diverse (Buhrmester et al., 2011). mTurk is particularly appropriate for a study investigating the behavior of employees (or potential employees), as the majority tend to be employed (Ipeirotis, 2010).

Limitations

Despite its strengths, this study did have several limitations. A control condition and a proctored CAT score were only available for the student sample, because it was impossible to test mTurk workers in a proctored setting. As discussed above, there are limitations to using a student sample, which calls into question the generalizability of the findings for the control condition.

Another limitation is that the experiment differed from selection scenarios in several meaningful ways. In this study, participants were motivated to perform well in order to avoid punishment (i.e., having to complete a long, boring vigilance task). In selection scenarios, job candidates are motivated to achieve a reward (i.e., getting a job). Within the framework of utility theory, this difference should not be meaningful; people make decisions in order to maximize the likelihood of valued outcomes, whatever those outcomes might be. However, this difference could meaningfully impact decision-making processes.

Another meaningful difference between this study and selection scenarios is that, for both samples, this study represented a close-ended interaction: students completed the study and received credit; mTurk workers completed the study and were paid. In a selection scenario, the application procedure is often the beginning of a long and potentially complex relationship between employers and employees. Cheating someone in a close-ended interaction might represent a different situation to people than cheating someone with whom they are beginning a potentially long-term relationship. The outcome for this study was payment received at the end of the task; the outcome for job candidates is potentially a psychological contract between employer and employee, in which mutual beliefs, obligations, and expectations are established (Rousseau, 1989).

Future research

The effectiveness of the probability of passing the test with cheating manipulation provides partial evidence that utility theory is a useful framework for understanding the decision-making process in regard to cheating on online tests. The lack of significant findings in regard to the other manipulations, however, calls into question the applicability of this theory for this topic. Future research should attempt to test utility theory using different, and potentially more effective, manipulations of both the value and probability of different outcomes in regard to cheating on online tests. For example, a stern warning that cheaters will most likely be caught might be more effective than mentioning what percentage of participant responses will be analyzed for cheating. When choosing a "high value of being caught cheating" manipulation, it is important to make sure the punishment actually does represent the loss of something valuable to participants.

It is likely that individual differences have an impact on decision-making processes in regard to cheating. Some individuals are probably likely to cheat regardless of how difficult it is, how severe the consequences are, or how likely it is that they will get caught; other individuals would probably refuse to cheat regardless of how easy it is, how superficial the consequences are, or how unlikely it is that they will get caught. The results of this study, showing that the majority of people who had access to the answer key did not use it (or at least, did not admit to using it), support the proposition that individual differences do play a role in cheating behavior. Future research should investigate what, and how malleable, these individual differences are. This research should be driven by a theoretical orientation, but has very practical applications. If it were possible to give a short intervention prior to testing that increased or decreased the level of these individual differences (e.g., ethical orientation), this could be a very useful tool for preventing cheating behavior on online tests. McTernan, Love, and Rettinger (2014), for example, found evidence that sensation seeking and impulsivity were positively related to cheating behavior, whereas empathetic perspective taking was negatively related to cheating. Another study by Ejei, Shahabi, and Alibazi (2012) found a negative relationship between both agreeableness and conscientiousness and cheating behavior.

Another area for future research would be to investigate which behaviors people consider "cheating." In this study, cheating was explicitly explained to participants as the use of outside materials such as the Internet and reference books, having another person take the test in lieu of the legitimate participant, or going back and changing answers once participants were shown the correct answer. It is likely, however, that in a selection context individuals probably have very different ideas of what constitutes acceptable and unacceptable cheating behavior. For example, most people would probably agree that searching the Internet for answers is cheating, but what

about using cognitive enhancing drugs such as amphetamines? At least initially, these studies should be exploratory in nature, asking participants to describe strategies they use while test taking, and then asking them to rate them on whether or not they consider these strategies to be cheating. Studies could also be designed to present participants with several different options for cheating, and observe how the use of these strategies influences self-reported cheating behavior. It is possible that some participants do not self-report cheating because they do not believe their behavior constitutes cheating, even if potential employers do consider these strategies dishonest.

Finally, in my opinion, one of the most interesting findings of this study was that the vast majority of cheaters in both samples did not perform well enough on the CAT to be excused from the vigilance task. Utility theory assumes that the decision making process is a rational calculation of the value and probability of different outcomes. Why would a person risk the potential consequences of cheating without obtaining the positive outcome? Was it that they were unable to cheat effectively enough to "pass" the CAT and be excused from the vigilance task, or did they consider looking up a few answers as a lesser form of cheating? Did they think it was acceptable to look up answers to items they felt were "unfair"? This finding represents a paradox in cheating behavior, and should be further investigated.

Summary and Conclusions

The purpose of this study was to investigate the usefuleness of utility theory as a framework for understanding why some individuals decide to cheat. The manipulations were laregely ineffective, although some evidence was found that providing answers to participants does increase cheating behavior, and that cheating behavior can undermine the validity of cognitive ability tests. Despite the ineffectiveness of the manipulations, the experimental protocol appeared to be sound for investigating cheating behavior. Only a small percentage of

participants were able to identify the true purpose of the study, and a relatively large percentage of participants reported cheating. Future research should investigate whether utility theory is actually ineffective at explaining cheating decisions, or whether the manipulations used in this study were ineffective. Future research should also investigate other possible frameworks for understanding cheating behavior, including an emphasis on individual differences such as personality.

Table 1

Student: Descriptive statistics for variables of interest.

	N	*M*	*SD*	*Min*	*Max*
ACT	318	25.1	3.5	14.0	34.0
SATper	91	0.66	0.14	0.201	0.97
CAT	384	21.0	6.5	4.0	35.0
Fake sum	384	2.1	1.1	0.0	5.0
CAT2	244	20.4	5.8	5.0	32.0

Note: SATper= SAT score expressed as a percentage; *CAT*= cognitive ability test score during session 1; *Fake sum*= number of fake items answered "correctly"; *CAT2*= cognitive ability test score during session 2.

Table 2

Student: Frequencies for dichotomous variables

	0		*1*		
	Raw	*%*	*Raw*	*%*	*Total*
Gender	50	28.7	124	71.3	174
Suspect	228	95.8	10	4.2	238
Cheat	200	84.4	37	15.6	237
CATdummy	322	83.9	62	16.1	384

Note: Gender=a dummy coded variable reflecting whether participants were male (0) or female (1); *Suspect*= a dummy coded variable reflecting whether participants suspected the study was about cheating (1) or not (0)*; CHEAT*= a dummy coded variable reflecting whether participants self-reported cheating (1) or not (0); *CATdumm*y=a binary variable reflecting whether participants scored well enough to be excused from the vigilance task (1) or not (0).

Table 3

Student: Correlation matrix for variables of interest.

	ACT	SATper	CAT	Fake sum	CAT2	Suspect	Gender	CATdummy	
ACT	-								
SATper	.56**	-							
CAT	.46**	.19	-						
Fake Sum	.07	.15	.11*	-					
CAT2	.56**	.35*	.72**	.12	-				
Suspect	.07	-.16	.01	.03	.00	-			
Gender	-.06	.24	.00	.14	-.01	-.14	-		
CATdummy	.28**	.15	.' 64**	.11*	.46**	.03	-.10	-	
Cheat	-.23**	-.13	.14	-.17*	.03	.08	.05	.18*	

Note: SATper= SAT score expressed as a percentage; *Fake sum*= number of fake answers participants answered correctly; *CAT2*=CAT score during session two; *Suspect*= a dummy coded variable reflecting whether participants suspected the study was about cheating (1) or not (0); Gender=a dummy coded variable reflecting whether participants were male (0) or female (1); *CATdummy*=a binary variable reflecting whether participants scored well enough to be excused from the vigilance task (1) or not (0); *CHEAT*= a dummy coded variable reflecting whether participants self-reported cheating (1) or not (0).

Note 2: *=p<.05, **=p<.01.

Table 4

mTurk: Descriptive statistics for variables of interest.

	N	*M*	*SD*	*Min*	*Max*
CAT	518	23.9	5.6	7	32
Fake sum	518	0.8	0.9	0	5
SATper	197	0.76	0.18	0.27	1.34
ACT	151	27.3	4.7	11	35
Age	515	37.2	12.6	18	76

Note: Fake sum= number of fake items answered "correctly"; *SATper*= SAT score expressed as a percentage.

Table 5

mTurk: Frequencies for dichotomous variables

	0		*1*		
	Raw	*%*	*Raw*	*%*	*Total*
Gender	181	34.9	337	65.1	518
Suspect	513	99.0	5	1.0	518
Cheat	462	89.2	56	10.8	518
CATdummy	511	98.6	7	1.4	518

Note: Gender=a dummy coded variable reflecting whether participants were male (0) or female (1)*; Suspect*= a dummy coded variable reflecting whether participants suspected the study was about cheating (1) or not (0)*; CHEAT*= a dummy coded variable reflecting whether participants self-reported cheating (1) or not (0); *CATdummy*=a binary variable reflecting whether participants scored well enough to be excused from the vigilance task (1) or not (0).

Table 6

mTurk: Correlation matrix for variables of interest.

	CAT	CATdummy	SATper	ACT	GEN	Age	Cheat	Suspect
CAT	-							
CATdummy	.16**	-						
SATper	.14	.02	-					
ACT	.61**	.03	.27	-				
Gender	-.12**	-.02	-.02	-.12	-			
Age	.05	-.03	.00	-.02	.29	-		
Cheat	.04	.18**	-.07	-.03	-.02	-.13**	-	
Suspect	.02	-.01	.03	-.02	.03	.09*	-.03	-
Fake sum	.25**	.54**	.13	.12	-.08	.04	.09*	.07

Note: CATdummy=a binary variable reflecting whether participants scored well enough to be excused from the vigilance task (1) or not (0); *SATper*= SAT score expressed as a percentage; *ACT*=ACT score, with outliers removed; *Gender*=a dummy coded variable reflecting whether participants were male (0) or female (1); *CHEAT*= a dummy coded variable reflecting whether participants self-reported cheating (1) or not (0); *Suspect*= a dummy coded variable reflecting whether participants suspected the study was about cheating (1) or not (0); *Fake sum*= number of fake items answered correctly.

Note 2: *=$p<.05$, **=$p<.01$.

Table 7

Mean ACT, SAT, and CAT scores for mTurk and student samples.

	Sample									
	0			1						
	n	*m*	*sd*	*n*	*m*	*sd*	*t*	*df*	*p*	*d*
ACT	151	27.3	4.7	318	25.1	3.5	5.59	233.8	<.001	0.531
SATper	197	0.76	0.18	91	0.66	0.14	4.32	286	<.001	0.577
CAT	518	23.9	5.6	384.0	21.0	6.5	6.770	749.2	<.001	0.478

Note. For the variable *sample,* 0=mTurk sample, 1=student sample; *SATper*= self-reported SAT expressed as a percentage.

Table 8

Mann-Whitney U test of differences in number of fake items answered correctly between participants in the mTurk and student samples.

Sample	*N*	*Mean Rank*	*Sum of Ranks*	*Mann-Whitney U*	*p*	*r*
mTurk	518	326.625	169192	34771.000	<.001	0.650
Student	384	619.951	238061			
Total	902					

Table 9

Proportion of participants who suspected the study was actually about cheating prior to debriefing in the mTurk and student samples.

	Suspect						
Sample	No	Yes	Total	χ^2	*df*	*p*	ϕ
mTurk	513	5	518	15.514	1.000	<.001	0.151
Student	152	10	162				
Total	665	15	680				

Table 10

Proportion of participants who performed well enough on CAT to be excused from the vigilance task in the mTurk and student samples.

	Excused						
Sample	No	Yes	Total	χ^2	*df*	*p*	ϕ
mTurk	511	7	518	68.324	1.000	<.001	0.275
Student	322	62	384				
Total	833	69	902				

Table 11

Proportion of self-reported cheaters in the mTurk and student samples.

	Self-reported cheating						
Sample	No	Yes	Total	χ^2	*df*	*p*	ϕ
mTurk	462	56	518	6.636	1.000	.010	0.099
Student	132	30	162				
Total	594	86	680				

Table 12

Student: Effect of experimental factors on CAT score.

	df	*F*	*p*	η^2
Model	8	11.344	<.001	
Intercept	1	0.075	.785	
ACT	1	82.706	<.001	.218
A	1	11.303	<.001	.030
P	1	0.519	.472	.001
V	1	0.006	.937	.000
A*P	1	0.336	.563	.001
A*V	1	0.318	.574	.001
P*V	1	0.188	.665	.000
A*P*V	1	0.647	.422	.002
Error	284			
Total	293			

Note. *A*=probability of passing the test with cheating; *P*=probability of being caught cheating; *V*=value of being caught cheating.

Table 13

Student: Estimated marginal means of CAT score during session one by experimental condition.

	A		P		V	
	0	1	0	1	0	1
m	22.8	20.5	21.9	21.4	21.7	21.6
se	0.5	0.5	0.5	0.5	0.5	0.5

Note. A=probability of passing the test with cheating; *P*=probability of being caught cheating; *V*=value of being caught cheating.

Table 14

mTurk: Effect of experimental factors on CAT score.

	df	*F*	*p*	η^2
Model	8	19.841	<.001	
Intercept	1	19.175	<.001	
ACT	1	106.217	<.001	.395
A	1	19.639	<.001	.073
P	1	0.652	.421	.002
V	1	1.679	.197	.006
A*P	1	0.001	.982	.000
A*V	1	0.723	.397	.003
P*V	1	4.157	.043	.015
A*P*V	1	0.164	.686	.001
Error	136			
Total	145			

Note. *ACT*= ACT score with 6 outliers removed; *A*=probability of passing the test with cheating; *P*=probability of being caught cheating; *V*=value of being caught cheating.

Table 15

mTurk: Estimated marginal means of CAT score during session one by experimental condition.

	A		P		V	
	0	1	0	1	0	1
m	23.5	26.2	24.6	25.1	25.2	24.5
se	0.4	0.5	0.4	0.4	0.4	0.4

Note. A=probability of passing the test with cheating; *P*=probability of being caught cheating; *V*=value of being caught cheating.

Table 16

mTurk: Pairwise comparisons for each combination of high/ low probability of being caught and high/ low value of being caught. A Bonferroni correction was used to compensate for experimentwise error rate.

	Mean difference	*SE*	*p*	*d*	*CI low*	*CI high*
pv vs. pV	3.0	0.9	0.004	3.452961672	0.7	5.3
pv vs. Pv	2.0	0.9	0.198	2.151679307	-0.5	4.5
pv vs. PV	1.0	0.9	1	1.109965636	-1.4	3.3
pV vs. Pv	-1.0	0.9	1	-1.057877814	-3.5	1.5
pV vs. PV	-2.0	0.9	0.15	-2.26440678	-4.4	0.4
Pv vs. PV	1.0	0.9	1.000	1.078389831	-1.5	3.5

Note. *p*=low probability of being caught cheating; *P*=high probability of being caught cheating*; v*=low probability of being caught cheating; *V*=high probability of being caught cheating.

Table 17

mTurk: Means and standard errors of combined experimental conditions, controlling for self-reported ACT.

	M	*SE*	*CI low*	*CI high*
pv	26.1	0.6	25.0	27.3
pV	23.2	0.6	22.0	24.4
Pv	24.2	0.7	22.8	25.5
PV	25.2	0.6	23.9	26.4

Note. p=low probability of being caught cheating*; P*=high probability of being caught cheating; *v*=low probability of being caught cheating*; V*=high probability of being caught cheating.

Table 18

Student: Chi-square test of independence for CHEAT by experimental condition.

	CHEAT						
Group	0	1	Total	χ^2	*df*	*p*	ϕ_c
Pro	8	1	9				
apv	18	2	20				
apV	18	1	19				
aPv	17	1	18				
aPV	19	6	25				
Apv	12	2	14				
ApV	19	3	22				
APv	21	10	31				
APV	22	5	27				
Total	154	31	185	10.834	8	.211	0.185

Note: CHEAT= self-reported cheating behavior, 0= did not report, 1=did report; *Pro*= proctored conditions; *a*= conditions in which the probability of passing the test with cheating was low; *A*=conditions in which the probability of passing the test with cheating was high; *p*=conditions in which the probability of being caught cheating was low; *P*=conditions in which the probability of being caught cheating was high; *v*=conditions in which the value of being caught cheating was low; *V*=conditions in which the value of being caught cheating was high.

Table 19

Student: Proportion of self-reported cheaters in conditions in which the probability of passing the test with cheating was low vs. conditions in which it was high.

	Self-reported cheating						
Probability of passing the test with cheating	No	Yes	Total	χ^2	*df*	*p*	ϕ
Low	80	11	91	2.799	1.000	.094	0.123
High	74	20	94				
Total	154	31	185				

Table 20

Student: Proportion of self-reported cheaters in conditions in which the probability of being caught cheating was low vs. conditions in which it was high.

	Self-reported cheating						
Probability of being caught cheating	No	Yes	Total	χ^2	*df*	*p*	ϕ
Low	75	9	84	4.027	1.000	.045	0.148
High	79	22	101				
Total	154	31	185				

Table 21

Student: Proportion of self-reported cheaters in conditions in which the value of being caught cheating was low vs. conditions in which it was high.

Value of being caught cheating	Self-reported cheating						
	No	Yes	Total	χ^2	*df*	*p*	ϕ
Low	76	16	92	0.053	1.000	.818	-0.017
High	78	15	93				
Total	154	31	185				

Table 22

mTurk: Chi-square test of independence for CHEAT by experimental condition.

	CHEAT						
Group	0	1	Total	χ^2	*df*	*p*	ϕ_c
apv	55	7	62				
apV	67	8	75				
aPv	68	5	73				
aPV	67	9	76				
Apv	57	8	65				
ApV	50	4	54				
APv	48	7	55				
APV	50	8	58				
Total	462	56	518	2.833	7	.900	0.074

Note: CHEAT= self-reported cheating behavior, 0= did not report, 1=did report; *a*= conditions in which the probability of passing the test with cheating was low; *A*=conditions in which the probability of passing the test with cheating was high; *p*=conditions in which the probability of being caught cheating was low; *P*=conditions in which the probability of being caught cheating was high; *v*=conditions in which the value of being caught cheating was low; *V*=conditions in which the value of being caught cheating was high.

Table 23

mTurk: Proportion of self-reported cheaters in conditions in which the probability of passing the test with cheating was low vs. conditions in which it was high.

	Self-reported cheating						
Probability of passing the test with cheating	No	Yes	Total	χ^2	*df*	*p*	ϕ
Low	257	29	286	0.298	1.000	.585	0.024
High	205	27	232				
Total	462	56	518				

Table 24

mTurk: Proportion of self-reported cheaters in conditions in which the probability of being caught cheating was low vs. conditions in which it was high.

	Self-reported cheating						
Probability of being caught cheating	No	Yes	Total	χ^2	*df*	*p*	ϕ
Low	229	27	256	0.037	1.000	.848	0.008
High	233	29	262				
Total	462	56	518				

Table 25

mTurk: Proportion of self-reported cheaters in conditions in which the value of being caught cheating was low vs. conditions in which it was high.

Value of being caught cheating	Self-reported cheating						
	No	Yes	Total	χ^2	*df*	*p*	ϕ
Low	228	27	255	0.026	1.000	.872	0.007
High	234	29	263				
Total	462	56	518				

Table 26

Student: Kruskal Wallis test for Fake Sum by experimental condition.

Group	*N*	*Mean Rank*	*H*	*df*	*p*
apv	37	179.040			
apV	44	177.940			
aPv	37	210.810			
aPV	53	185.420			
Apv	40	173.790			
ApV	46	180.620			
APv	58	159.540			
APV	45	186.330			
Total	360		6.424	7	.491

Note: a= conditions in which the probability of passing the test with cheating was low; *A*=conditions in which the probability of passing the test with cheating was high; *p*=conditions in which the probability of being caught cheating was low; *P*=conditions in which the probability of being caught cheating was high; *v*=conditions in which the value of being caught cheating was low; *V*=conditions in which the value of being caught cheating was high.

Table 27

Student: Mann-Whitney U test of differences in number of fake items answered correctly between participants in conditions in which the probability of passing the test with cheating was high vs. low.

Probability of passing the test with cheating	*N*	*Mean Rank*	*Sum of Ranks*	*Mann-Whitney U*	*p*	*r*
Low	195	201.344	39262	16703	0.100	0.094
High	189	183.376	34658			
Total	384					

Table 28

Student: Mann-Whitney U test of differences in number of fake items answered correctly between participants in conditions in which the probability of being caught cheating was high vs. conditions in which it was low.

Probability of being caught cheating	*N*	*Mean Rank*	*Sum of Ranks*	*Mann-Whitney U*	*p*	*r*
Low	191	192.463	36760.5	18425	0.995	0.000
High	193	192.536	37159.5			
Total	384					

Table 29

Student: Mann-Whitney U test of differences in number of fake items answered correctly between participants in conditions in which the value of being caught cheating is high vs. conditions in which it was low.

Value of being caught cheating	*N*	*Mean Rank*	*Sum of Ranks*	*Mann-Whitney U*	*p*	*r*
Low	196	192.543	37738.5	18416	0.994	0.000
High	188	192.455	36181.5			
Total	384					

Table 30

mTurk: Kruskal-Wallis test for Fake Sum by experimental condition.

Group	*N*	*Mean Rank*	*H*	*df*	*p*
apv	62	250.660			
apV	75	239.310			
aPv	73	271.120			
aPV	76	240.650			
Apv	65	265.160			
ApV	54	290.220			
APv	55	260.140			
APV	58	269.570			
Total	518		6.955	7	.434

Note: a= conditions in which the probability of passing the test with cheating was low; *A*=conditions in which the probability of passing the test with cheating was high; *p*=conditions in which the probability of being caught cheating was low; *P*=conditions in which the probability of being caught cheating was high; *v*=conditions in which the value of being caught cheating was low; *V*=conditions in which the value of being caught cheating was high.

Table 31

mTurk: Mann-Whitney U test of differences in number of fake items answered correctly between participants in conditions in which the probability of passing the test with cheating was high vs. conditions in which it was low.

Probability of passing the test with cheating	*N*	*Mean Rank*	*Sum of Ranks*	*Mann-Whitney U*	*p*	*r*
Low	286	250.25	71571	30530	0.089	0.080
High	232	270.91	62850			
Total	518					

Table 32

mTurk: Mann-Whitney U test of differences in number of fake items answered correctly between participants in conditions in which the probability of being caught cheating is high vs. conditions in which it is low.

Probability of being caught cheating	*N*	*Mean Rank*	*Sum of Ranks*	*Mann-Whitney U*	*p*	*r*
Low	256	259.36	66397	33501	0.982	0.001
High	262	259.63	68024			
Total	518					

Table 33

mTurk: Mann-Whitney U test of differences in number of fake items answered correctly between participants in conditions in which the value of being caught cheating is high vs. conditions in which it is low.

Value of being caught cheating	N	Mean Rank	Sum of Ranks	*Mann-Whitney U*	*p*	*r*
Low	255	262.26	66876	32829	0.653	0.021
High	263	256.83	67545			
Total	518					

Table 34

Student: Chi-square test of independence for number of participants who performed well enough on the CAT to be excused from the vigilance task by experimental condition.

	CHEAT						
Group	0	1	Total	χ^2	*df*	*p*	ϕ_c
Pro	21	2	23				
apv	33	5	38				
apV	37	7	44				
aPv	33	4	37				
aPV	46	7	53				
Apv	31	9	40				
ApV	37	9	46				
APv	44	14	58				
APV	40	5	45				
Total	322	62	384	7.480	8	.486	0.074

Note: Pro= proctored conditions; *a*= conditions in which the probability of passing the test with cheating was low; *A*=conditions in which the probability of passing the test with cheating was high; *p*=conditions in which the probability of being caught cheating was low; *P*=conditions in which the probability of being caught cheating was high; *v*=conditions in which the value of being caught cheating was low; *V*=conditions in which the value of being caught cheating was high.

Table 35

Student: Proportion of participants who performed well enough on the CAT to be excused from the vigilance task in conditions in which the probability of passing the test with cheating was low vs. conditions in which it was high.

Probability of passing the test with cheating	Excused from vigilance task						
	No	Yes	Total	χ^2	*df*	*p*	ϕ
Low	170	25	195	3.236	1.000	.072	.092
High	152	37	189				
Total	322	62	384				

Table 36

Student: Proportion of participants who performed well enough on the CAT to be excused from the vigilance task in conditions in which the probability of being caught cheating was low vs. conditions in which it was high.

	Excused from vigilance task						
Probability of being caught cheating	No	Yes	Total	χ^2	*df*	*p*	ϕ
Low	159	32	191	0.104	1.000	.747	-.016
High	163	30	193				
Total	322	62	384				

Table 37

Student: Proportion of participants who performed well enough on the CAT to be excused from the vigilance task in conditions in which the value of being caught cheating was low vs. conditions in which it was high.

	Excused from vigilance task						
Value of being caught cheating	No	Yes	Total	χ^2	*df*	*p*	ϕ
Low	162	34	196	0.427	1.000	.514	-.033
High	160	28	188				
Total	322	62	384				

Table 38

mTurk: Proportion of participants who performed well enough on the CAT to be excused from the vigilance task in conditions in which the probability of passing the test with cheating was low vs. conditions in which it was high.

Probability of passing the test with cheating	Excused from vigilance task				
	No	Yes	Total	*Fisher's exact sig.*	*OR*
Low	285	1	286	0.049	0.210
High	226	6	232		
Total	511	7	518		

Table 39

mTurk: Proportion of participants who performed well enough on the CAT to be excused from the vigilance task in conditions in which the probability of being caught cheating was low vs. conditions in which it was high.

Probability of being caught cheating	Excused from vigilance task				
	No	Yes	Total	*Fisher's exact sig.*	*OR*
Low	252	4	256	0.722	1.297
High	259	3	262		
Total	511	7	518		

Table 40

mTurk: Proportion of participants who performed well enough on the CAT to be excused from the vigilance task in conditions in which the value of being caught cheating was low vs. conditions in which it was high.

	Excused from vigilance task				
Value of being caught cheating	No	Yes	Total	*Fisher's exact sig.*	*OR*
Low	252	3	255	1.000	0.730
High	259	4	263		
Total	511	7	518		

Table 41

Student: Mean CAT scores of self-reported cheaters compared to participants who did not self-report cheating behavior.

	Self-reported cheating									
	0			1						
	n	*M*	*SD*	*n*	*M*	*SD*	*t*	*df*	*p*	*d*
CAT	154	20.9	6.1	31	23.1	6.6	-1.808	183	0.072	0.346

Table 42

mTurk: Mean CAT scores of self-reported cheaters compared to participants who did not self-report cheating behavior.

	Self-reported cheating									
	0			1						
	n	*M*	*SD*	*n*	*M*	*SD*	*t*	*df*	*p*	*d*
CAT	462	23.8	5.5	56	24.5	6.2	-0.876	516	0.382	0.119

Table 42

Student: Proportion of self-reported cheaters who performed well enough on the CAT to be excused from the vigilance task.

Self-reported cheating	Excused from vigilance task						
	No	Yes	Total	χ^2	*df*	*p*	ϕ
0	134	20	154	5.026	1.000	.025	0.165
1	22	9	31				
Total	156	29	185				

Table 44

mTurk: Proportion of self-reported cheaters who performed well enough on the CAT to be excused from the vigilance task, compared to participants who did not self-report cheating.

	Excused from vigilance task				
Self-reported cheating	No	Yes	Total	*Fisher's exact sig.*	*OR*
No	459	3	462	0.003	6.620
Yes	52	4	56		
Total	511	7	518		

Table 45

Student: Mann-Whitney U test of differences in number of fake items answered correctly between participants who self-reported cheating and those who did not.

Self-reported cheating	*n*	*Mean Rank*	*Sum of Ranks*	*Mann-Whitney U*	*p*	*r*
No	154	96.292	14829	1880.0	0.053	0.212
Yes	31	76.645	2376			
Total	185					

Table 46

mTurk: Mann-Whitney U test of differences in number of fake items answered correctly between participants who self-reported cheating and those who did not.

Self-reported cheating	*n*	*Mean Rank*	*Sum of Ranks*	*Mann-Whitney U*	*p*	*r*
No	462	257.97	119181.5	12228.5	0.466	0.055
Yes	56	272.13	15239.5			
Total	518					

Table 47

Student: CAT score during session two regressed on CAT score during session one.

	b	*SE*	*β*	*t*	*p*	*CI (low)*	*CI (high)*
Intercept	6.740	1.011		6.670	p<.001	4.747	8.734
CAT1	0.655	0.046	0.719	14.335	p<.001	0.565	0.746
R^2	.517						
F	205.480				p<.001		
degrees of freedom regression	1						
degrees of freedom residual	192						
ΔR^2	205.48						
ΔF	1				p<.001		
degrees of freedom regression	192						
degrees of freedom residual	0						

Note. CAT1= CAT scores during session 1.

Table 48

Student: CAT score during session two regressed on CAT score during session one, as well as each of the experimental conditions.

	b	*SE*	*β*	*t*	*p*	*CI (low)*	*CI (high)*
Intercept	6.960	1.221		5.700	p<.001	4.551	9.368
CAT1	0.655	0.046	0.719	14.124	p<.001	0.564	0.747
A	0.863	0.594	0.076	1.452	.148	-0.309	2.035
P	-0.677	0.599	-0.060	-1.130	.260	-1.860	0.505
V	-0.493	0.588	-0.044	-0.839	.403	-1.652	0.666
Pro	-0.531	1.297	-0.023	-0.410	.683	-3.090	2.027
R^2	.526						
F	41.802				<.001		
degrees of freedom regression	5						
degrees of freedom residual	188						
ΔR^2	0.01						
ΔF	0.943				.440		
degrees of freedom regression	4						
degrees of freedom residual	188						

Note. *CAT1*= CAT scores during session 1; *A*=probability of passing the test with cheating; *P*=probability of being caught cheating; *V*=value of being caught cheating; *Pro*=proctored session 1.

Table 49

Student: CAT score during session two regressed on CAT score during session one, each of the experimental conditions, as well as the interactions between CAT score during session one and each of the experimental conditions.

	b	*SE*	*β*	*t*	*p*	*CI (low)*	*CI (high)*
Intercept	2.735	2.284		1.197	.233	-1.772	7.241
CAT1	0.851	0.101	0.934	8.466	p<.001	0.653	1.049
A	3.635	2.112	0.321	1.721	.087	-0.532	7.802
P	2.068	2.140	0.183	0.967	.335	-2.153	6.290
V	2.226	2.081	0.197	1.069	.286	-1.881	6.332
Pro	-0.870	4.870	-0.037	-0.179	.858	-10.479	8.739
CAT1*A	-0.132	0.095	-0.262	-1.393	.165	-0.319	0.055
CAT1*P	-0.130	0.096	-0.258	-1.356	.177	-0.319	0.059
CAT1*V	-0.130	0.094	-0.260	-1.378	.170	-0.315	0.056
CAT1*Pro	0.031	0.230	0.028	0.137	.892	-0.422	0.485
R^2	.543						
F	24.316				p<.001		
degrees of freedom regression	9						
degrees of freedom residual	184						
ΔR^2	0.017						
ΔF	1.691				.154		
degrees of freedom regression	4						
degrees of freedom residual	184						

Note. CAT1= CAT scores during session 1; *A*=probability of passing the test with cheating; *P*=probability of being caught cheating; *V*=value of being caught cheating; *Pro*=proctored session 1; *CAT1*A*=interaction between *CAT1* and *A*; *CAT1*P*=interaction between *CAT1* and *P*; *CAT1*V*=interaction between *CAT1* and *V*; *CAT1*Pro*=interaction between *CAT1* and *Pro.*

Table 50

Student: CAT score during session two regressed on CAT score during session one and self-reported cheating.

	b	*SE*	*β*	*t*	*p*	*CI (low)*	*CI (high)*
Intercept	7.297	1.023		7.133	p<.001	5.278	9.316
CAT1	0.651	0.047	0.731	13.891	p<.001	0.558	0.743
CHEAT	-1.734	0.783	-0.117	-2.215	.028	-3.279	-0.189
R^2	.524						
F	96.512				p<.001		
degrees of freedom regression	2						
degrees of freedom residual	175						
ΔR^2	0.013						
ΔF	4.906				.028		
degrees of freedom regression	1						
degrees of freedom residual	175						

Note. *CAT1*= CAT scores during session 1; *CHEAT*=dummy coded variable representing self-reported cheating, 0=did not report cheating, 1=reported cheating.

Table 51

Student: CAT score during session two regressed on CAT score during session one, self-reported cheating, and the interaction between CAT score during session one and self-reported cheating.

	b	*SE*	*β*	*t*	*p*	*CI (low)*	*CI (high)*
Intercept	6.262	1.114		5.620	p<.001	4.063	8.461
CAT1	0.701	0.052	0.787	13.604	p<.001	0.599	0.802
CHEAT	4.194	2.782	0.282	1.507	.134	-1.298	9.685
CAT1*CHEAT	-0.262	0.118	-0.426	-2.218	.028	-0.495	-0.029
R^2	.538						
F	67.422				p<.001		
degrees of freedom regression	3						
degrees of freedom residual	174						
ΔR^2	0.013						
ΔF	4.919				.028		
degrees of freedom regression	1						
degrees of freedom residual	174						

Note. *CAT1*= CAT scores during session 1; *CHEAT*=dummy coded variable representing self-reported cheating, 0=did not report cheating, 1=reported cheating; *CAT1*CHEAT*=interaction between *CAT1* and *CHEAT*.

Table 52

Student: Self-reported SAT regressed on CAT score.

	b	*SE*	*β*	*t*	*p*	*CI (low)*	*CI (high)*
Intercept	0.580	0.048		11.991	p<.001	0.484	0.677
CAT1	0.004	0.002	0.189	1.815	.073	0.000	0.008
R^2	.036						
F	3.293				.073		
degrees of freedom regression	1						
degrees of freedom residual	89						
ΔR^2	3.293						
ΔF	1				.073		
degrees of freedom regression	89						
degrees of freedom residual	0.073						

Note. CAT1= CAT scores during session 1.

Table 53

Student: Self-reported SAT regressed on CAT score and each experimental condition.

	b	*SE*	*β*	*t*	*p*	*CI (low)*	*CI (high)*
Intercept	0.570	0.062		9.119	p<.001	0.446	0.694
CAT1	0.004	0.002	0.194	1.777	.079	0.000	0.008
A	-0.003	0.030	-0.011	-0.097	.923	-0.063	0.057
P	0.010	0.030	0.036	0.331	.742	-0.049	0.069
V	0.002	0.030	0.009	0.083	.934	-0.057	0.062
Pro	0.072	0.076	0.109	0.946	.347	-0.079	0.223
R^2	.047						
F	0.842				.524		
degrees of freedom regression	5						
degrees of freedom residual	85						
ΔR^2	0.012						
ΔF	0.257				.905		
degrees of freedom regression	4						
degrees of freedom residual	85						

Note. *CAT1*= CAT scores during session 1; *A*=probability of passing the test with cheating; *P*=probability of being caught cheating; *V*=value of being caught cheating; *Pro*=proctored session 1.

Table 54

Student: Self-reported SAT regressed on CAT score, each experimental condition, and the interaction between CAT score and experimental condition.

	b	*SE*	*β*	*t*	*p*	*CI (low)*	*CI (high)*
Intercept	0.552	0.128		4.328	p<.001	0.298	0.806
CAT1	0.005	0.006	0.235	0.862	.391	-0.006	0.016
A	0.009	0.111	0.033	0.080	.937	-0.213	0.231
P	0.098	0.104	0.363	0.945	.348	-0.109	0.306
V	-0.056	0.109	-0.205	-0.513	.609	-0.272	0.161
Pro	-0.199	0.332	-0.301	-0.599	.551	-0.860	0.462
CAT1*A	-0.001	0.005	-0.044	-0.110	.913	-0.010	0.009
CAT1*P	-0.004	0.005	-0.348	-0.870	.387	-0.013	0.005
CAT1*V	0.002	0.005	0.218	0.519	.605	-0.007	0.012
CAT1*Pro	0.014	0.016	0.422	0.853	.396	-0.018	0.045
R^2	.072						
F	0.694				.713		
degrees of freedom regression	9						
degrees of freedom residual	81						
ΔR^2	0.024						
ΔF	0.532				.713		
degrees of freedom regression	4						
degrees of freedom residual	81						

Note. CAT1= CAT scores during session 1; *A*=probability of passing the test with cheating; *P*=probability of being caught cheating; *V*=value of being caught cheating; *Pro*=proctored session 1; *CAT1*A*=interaction between *CAT1* and *A*; *CAT1*P*=interaction between *CAT1* and *P*; *CAT1*V*=interaction between *CAT1* and *V*; *CAT1*Pro*=interaction between *CAT1* and *Pro.*

Table 55

Student: Self-reported SAT regressed on CAT and self-reported cheating.

	b	*SE*	*β*	*t*	*p*	*CI (low)*	*CI (high)*
Intercept	0.554	0.065		8.504	p<.001	0.423	0.684
CAT	0.006	0.003	0.273	1.985	.053	0.000	0.012
CHEAT	-0.072	0.054	-0.185	-1.346	.184	-0.180	0.036
R^2	.089						
F	2.442				p<.001		
degrees of freedom regression	2						
degrees of freedom residual	50						
ΔR^2	0.033						
ΔF	1.812				.184		
degrees of freedom regression	1						
degrees of freedom residual	50						

Note. *CHEAT*=dummy coded variable representing self-reported cheating, 0=did not report cheating, 1=reported cheating.

Table 56

Student: Self-reported SAT regressed on CAT, self-reported cheating, and the interaction between CAT and self-reported cheating.

	b	*SE*	*β*	*t*	*p*	*CI (low)*	*CI (high)*
Intercept	0.536	0.070		7.634	p<.001	0.395	0.677
CAT	0.007	0.003	0.003	2.086	.042	0.000	0.013
CHEAT	0.068	0.212	0.212	0.319	.751	-0.359	0.494
CAT*CHEAT	-0.006	0.008	0.008	-0.681	.499	-0.022	0.011
R^2	.098						
F	1.765				.166		
degrees of freedom regression	3						
degrees of freedom residual	49						
ΔR^2	0.009						
ΔF	0.464				.499		
degrees of freedom regression	1						
degrees of freedom residual	49						

Note. *CHEAT*=dummy coded variable representing self-reported cheating, 0=did not report cheating, 1=reported cheating; *CAT***CHEAT*=interaction between *CAT* and *CHEAT*.

Table 57

mTurk: Self-reported SAT regressed on CAT score.

	b	*SE*	*β*	*t*	*p*	*CI (low)*	*CI (high)*
Intercept	0.586	0.049		12.006	p<.001	0.489	0.682
CAT	0.007	0.002	0.269	3.732	p<.001	0.003	0.011
R^2	.073						
F	13.925				p<.001		
degrees of freedom regression	1						
degrees of freedom residual	178						
ΔR^2	0.073						
ΔF	13.925				p<.001		
degrees of freedom regression	1						
degrees of freedom residual	178						

Table 58

mTurk: Self-reported SAT regressed on CAT score.

	b	*SE*	*β*	*t*	*p*	*CI (low)*	*CI (high)*
Intercept	0.588	0.051		11.467	p<.001	0.487	0.689
CAT	0.008	0.002	0.298	4.115	p<.001	0.004	0.011
A	-0.040	0.018	-0.162	-2.234	.027	-0.076	-0.005
P	-0.020	0.018	-0.082	-1.133	.259	-0.056	0.015
V	0.019	0.018	0.075	1.041	.299	-0.017	0.054
R^2	.110						
F	5.393				p<.001		
degrees of freedom regression	4						
degrees of freedom residual	175						
ΔR^2	0.037						
ΔF	2.436				.066		
degrees of freedom regression	3						
degrees of freedom residual	175						

Note. *A*=probability of passing the test with cheating; *P*=probability of being caught cheating; *V*=value of being caught cheating.

Table 59

mTurk: Self-reported SAT regressed on CAT score.

	b	*SE*	*β*	*t*	*p*	*CI (low)*	*CI (high)*
Intercept	0.575	0.123		4.684	p<.001	0.333	0.818
CAT	0.008	0.005	0.315	1.703	.090	-0.001	0.018
A	-0.115	0.102	-0.462	-1.121	.264	-0.316	0.087
P	0.008	0.103	0.033	0.081	.936	-0.195	0.211
V	0.101	0.099	0.406	1.018	.310	-0.095	0.297
CAT*A	0.003	0.004	0.319	0.729	.467	-0.005	0.011
CAT*P	-0.001	0.004	-0.121	-0.286	.775	-0.009	0.007
CAT*V	-0.003	0.004	-0.337	-0.844	.400	-0.011	0.004
R^2	.117						
F	3.244				.003		
degrees of freedom regression	7						
degrees of freedom residual	172						
ΔR^2	0.007						
ΔF	0.447				.719		
degrees of freedom regression	3						
degrees of freedom residual	172						

Note. A=probability of passing the test with cheating; *P*=probability of being caught cheating; *V*=value of being caught cheating; *CAT***A*=interaction between *CAT* and *A*; *CAT***P*=interaction between *CAT* and *P*; *CAT***V*=interaction between *CAT* and *V*.

Table 60

mTurk: Self-reported SAT regressed on CAT and self-reported cheating.

	b	*SE*	*β*	*t*	*p*	*CI (low)*	*CI (high)*
Intercept	0.603	0.050		12.104	p<.001	0.504	0.701
CAT	0.007	0.002	0.246	3.422	p<.001	0.003	0.010
CHEAT	-0.064	0.031	-0.146	-2.038	.043	-0.126	-0.002
R^2	.079						
F	7.651				p<.001		
degrees of freedom regression	2						
degrees of freedom residual	179						
ΔR^2	0.021						
ΔF	4.151				.043		
degrees of freedom regression	1						
degrees of freedom residual	179						

Note. *CHEAT*=dummy coded variable representing self-reported cheating, 0=did not report cheating, 1=reported cheating.

Table 61

mTurk: Self-reported SAT regressed on CAT, self-reported cheating, and the interaction between CAT and self-reported cheating.

	b	*SE*	*β*	*t*	*p*	*CI (low)*	*CI (high)*
Intercept	0.578	0.053		10.996	p<.001	0.474	0.681
CAT	0.008	0.002	0.282	3.721	p<.001	0.004	0.012
CHEAT	0.167	0.162	0.382	1.032	.303	-0.152	0.487
CAT*CHEAT	-0.009	0.006	-0.541	-1.455	.147	-0.021	0.003
R^2	.090						
F	5.838				p<.001		
degrees of freedom regression	3						
degrees of freedom residual	178						
ΔR^2	0.011						
ΔF	2.117				.147		
degrees of freedom regression	1						
degrees of freedom residual	178						

Note. *CHEAT*=dummy coded variable representing self-reported cheating, 0=did not report cheating, 1=reported cheating; *CAT***CHEAT*=interaction between *CAT* and *CHEAT*.

Table 62

Student: Self-reported ACT regressed on CAT score.

	b	*SE*	*β*	*t*	*p*	*CI (low)*	*CI (high)*
Intercept	19.881	0.598		33.246	p<.001	18.704	21.057
CAT1	0.244	0.027	0.456	9.115	p<.001	0.191	0.297
R^2	.208						
F	83.082				p<.001		
degrees of freedom regression	1						
degrees of freedom residual	316						
ΔR^2	83.082						
ΔF	1				p<.001		
degrees of freedom regression	316						
degrees of freedom residual	0						

Note. CAT1= CAT scores during session 1.

Table 63

Student: Self-reported ACT regressed on CAT score, as well as each of the experimental conditions.

	b	*SE*	*β*	*t*	*p*	*CI (low)*	*CI (high)*
Intercept	19.466	0.710		27.417	p<.001	18.069	20.863
CAT1	0.248	0.026	0.464	9.370	p<.001	0.196	0.300
A	1.236	0.360	0.176	3.439	p<.001	0.529	1.944
P	-0.323	0.358	-0.046	-0.900	.369	-1.028	0.383
V	-0.355	0.359	-0.051	-0.990	.323	-1.061	0.351
Pro	0.931	0.806	0.063	1.156	.249	-0.654	2.516
R^2	.245						
F	20.293				p<.001		
degrees of freedom regression	5						
degrees of freedom residual	312						
ΔR^2	0.037						
ΔF	3.847				.005		
degrees of freedom regression	4						
degrees of freedom residual	312						

Note. *CAT1*= CAT scores during session 1; *A*=probability of passing the test with cheating; *P*=probability of being caught cheating; *V*=value of being caught cheating; *Pro*=proctored session 1.

Table 64

Student: Self-reported ACT on CAT score, each of the experimental conditions, as well as the interactions between CAT score and each of the experimental conditions.

	b	*SE*	*β*	*t*	*p*	*CI (low)*	*CI (high)*
Intercept	17.570	1.450		12.120	p<.001	14.718	20.422
CAT1	0.337	0.064	0.629	5.226	p<.001	0.210	0.464
A	4.065	1.248	0.579	3.258	p<.001	1.610	6.521
P	-1.048	1.216	-0.149	-0.863	.389	-3.440	1.343
V	0.702	1.218	0.100	0.577	.565	-1.695	3.099
Pro	2.599	3.184	0.176	0.816	.415	-3.666	8.865
CAT1*A	-0.133	0.056	-0.439	-2.381	.018	-0.243	-0.023
CAT1*P	0.035	0.054	0.114	0.650	.516	-0.072	0.142
CAT1*V	-0.052	0.055	-0.171	-0.939	.348	-0.160	0.056
CAT1*Pro	-0.078	0.142	-0.119	-0.551	.582	-0.357	0.201
R^2	.262						
F	12.126				p<.001		
degrees of freedom regression	9						
degrees of freedom residual	308						
ΔR^2	0.016						
ΔF	1.692				.152		
degrees of freedom regression	4						
degrees of freedom residual	308						

Note. CAT1= CAT scores during session 1; *A*=probability of passing the test with cheating; *P*=probability of being caught cheating; *V*=value of being caught cheating; *Pro*=proctored session 1; *CAT1*A*=interaction between *CAT1* and *A*; *CAT1*P*=interaction between *CAT1* and *P*; *CAT1*V*=interaction between *CAT1* and *V*; *CAT1*Pro*=interaction between *CAT1* and *Pro.*

Table 65

Student: Self-reported ACT regressed on CAT and self-reported cheating.

	b	*SE*	*β*	*t*	*p*	*CI (low)*	*CI (high)*
Intercept	19.429	0.791		24.552	p<.001	17.865	20.992
CAT	0.265	0.036	0.507	7.439	p<.001	0.194	0.335
CHEAT	-2.512	0.622	-0.275	-4.039	p<.001	-3.740	-1.283
R^2	.305						
F	33.187				p<.001		
degrees of freedom regression	2						
degrees of freedom residual	151						
ΔR^2	0.075						
ΔF	16.310				p<.001		
degrees of freedom regression	1						
degrees of freedom residual	151						

Note. CHEAT=dummy coded variable representing self-reported cheating, 0=did not report cheating, 1=reported cheating.

Table 66

Student: Self-reported ACT regressed on CAT, self-reported cheating behavior, and the interaction between CAT and self-reported cheating behavior.

	b	*SE*	*β*	*t*	*p*	*CI (low)*	*CI (high)*
Intercept	19.118	0.865		22.099	p<.001	17.408	20.827
CAT	0.279	0.039	0.535	7.120	p<.001	0.202	0.357
CHEAT	-0.635	2.192	-0.070	-0.290	0.772	-4.967	3.696
CAT*CHEAT	-0.083	0.093	-0.219	-0.893	0.374	-0.268	0.101
R^2	.309						
F	22.361				p<.001		
degrees of freedom regression	3						
degrees of freedom residual	150						
ΔR^2	0.004						
ΔF	0.797				0.374		
degrees of freedom regression	1						
degrees of freedom residual	150						

Note. CHEAT=dummy coded variable representing self-reported cheating, 0=did not report cheating, 1=reported cheating; *CAT*CHEAT*=interaction between *CAT* and *CHEAT*.

Table 67

mTurk: Self-reported ACT regressed on CAT score.

	b	*SE*	*β*	*t*	*p*	*CI (low)*	*CI (high)*
Intercept	12.397	1.410		8.792	p<.001	9.609	15.184
CAT	0.595	0.056	0.669	10.720	p<.001	0.486	0.705
R^2	.447						
F	114.918				p<.001		
degrees of freedom regression	1						
degrees of freedom residual	142						
ΔR^2	0.447						
ΔF	114.918				p<.001		
degrees of freedom regression	1						
degrees of freedom residual	142						

Table 68

mTurk: Self-reported ACT regressed on CAT score, as well as each of the experimental conditions.

	b	*SE*	*β*	*t*	*p*	*CI (low)*	*CI (high)*
Intercept	11.962	1.489		8.035	p<.001	9.019	14.906
CAT	0.636	0.059	0.714	10.840	p<.001	0.520	0.752
A	-1.208	0.583	-0.136	-2.074	.040	-2.361	-0.056
P	-0.153	0.552	-0.017	-0.278	.782	-1.246	0.939
V	0.094	0.554	0.011	0.170	.866	-1.001	1.189
R^2	.464						
F	30.114				p<.001		
degrees of freedom regression	4						
degrees of freedom residual	139						
ΔR^2	0.017						
ΔF	1.467				.226		
degrees of freedom regression	3						
degrees of freedom residual	139						

Note. *A*=probability of passing the test with cheating; *P*=probability of being caught cheating; *V*=value of being caught cheating.

Table 69

mTurk: Self-reported ACT on CAT score, each of the experimental conditions, as well as the interactions between CAT score and each of the experimental conditions.

	b	*SE*	*β*	*t*	*p*	*CI (low)*	*CI (high)*
Intercept	11.643	3.301		3.527	p<.001	5.115	18.171
CAT	0.652	0.129	0.732	5.073	p<.001	0.398	0.906
A	1.462	3.204	0.165	0.456	.649	-4.875	7.799
P	-0.271	2.995	-0.031	-0.090	.928	-6.194	5.652
V	-0.905	3.132	-0.103	-0.289	.773	-7.100	5.290
CAT*A	-0.105	0.124	-0.324	-0.853	.395	-0.350	0.139
CAT*P	0.002	0.117	0.006	0.019	.985	-0.230	0.234
CAT*V	0.042	0.122	0.120	0.341	.734	-0.200	0.283
R^2	.468						
F	17.104				p<.001		
degrees of freedom regression	3						
degrees of freedom residual	136						
ΔR^2	0.004						
ΔF	0.335				.800		
degrees of freedom regression	3						
degrees of freedom residual	136						

Note. A=probability of passing the test with cheating; *P*=probability of being caught cheating; *V*=value of being caught cheating; *CAT***A*=interaction between *CAT* and *A*; *CAT***P*=interaction between *CAT* and *P*; *CAT***V*=interaction between *CAT* and *V*.

Table 70

mTurk: Self-reported ACT regressed on CAT and self-reported cheating.

	b	*SE*	*β*	*t*	*p*	*CI (low)*	*CI (high)*
Intercept	12.464	1.448		8.607	p<.001	9.601	15.327
CAT	0.594	0.057	0.662	10.494	p<.001	0.482	0.706
CHEAT	0.117	0.829	0.009	0.141	0.888	-1.521	1.755
R^2	.438						
F	55.243				p<.001		
degrees of freedom regression	2						
degrees of freedom residual	142						
ΔR^2	0.000						
ΔF	0.020				0.888		
degrees of freedom regression	1						
degrees of freedom residual	142						

Note. CHEAT=dummy coded variable representing self-reported cheating, 0=did not report cheating, 1=reported cheating.

Table 71

mTurk: Self-reported ACT regressed on CAT, self-reported cheating behavior, and the interaction between CAT and self-reported cheating behavior.

	b	*SE*	*β*	*t*	*p*	*CI (low)*	*CI (high)*
Intercept	11.958	1.614		7.410	p<.001	8.768	15.149
CAT	0.614	0.063	0.685	9.695	p<.001	0.489	0.740
CHEAT	2.575	3.538	0.196	0.728	0.468	-4.419	9.569
CAT*CHEAT	-0.102	0.142	-0.192	-0.715	0.476	-0.382	0.179
R^2	.440						
F	36.872				p<.001		
degrees of freedom regression	3						
degrees of freedom residual	141						
ΔR^2	0.002						
ΔF	0.511				0.476		
degrees of freedom regression	1						
degrees of freedom residual	141						

Note. CHEAT=dummy coded variable representing self-reported cheating, 0=did not report cheating, 1=reported cheating; *CAT*CHEAT*=interaction between *CAT* and *CHEAT*.

REFERENCES

Arthur, W. Jr., Glaze, R. M., Villado, A. J., & Taylor, J. E. (2010). The magnitude and extent of cheating and response distortion effects on unproctored Internet-based tests of cognitive ability and personality. *International Journal of Selection and Assessment, 18,* 1-16.

Arthur, W., Jr., Glaze, R. M., Villado, A. J., & Taylor, J. E. (2009). Unproctored Internet-test based tests of cognitive ability and personality: Magnitude of cheating and response distortion. *Industrial and Organizational Psychology, 2,* 39-45.

Barak, A. (2010). Internet-based psychological testing and assessment. In R. Kraus, G. Stricker & C. Speyer (Eds.), *Online counseling: A handbook for mental health professionals* (2nd ed, pp. 225-256). London, UK: Academic Press.

Beaty, J. C., Fallon, J. D., Shepherd, W. J., & Barrett, C. (2002). *Proctored versus unproctored web-based administration of a cognitive ability test.* Paper presented at the 17th Annual conference of the Society for Industrial and Organizational Psychology, Toronto.

Beaty, J. C., Nye, C. D., Borneman, M. J., Kantrowitz, T. M., Drasgow, F., & Grauer, E. (2011). Proctored versus unproctored Internet tests: Are unproctored noncognitive tests as predictive of job performance? *International Journal of Selection and Assessment, 19*, 1-10.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*, 3-5.

Chapman, D. S., & Webster, J. (2003). The use of technologies in the recruiting, screening, and selection processes for job candidates. *International Journal of Selection and Assessment, 11,* 113–120.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dinges, D. F., & Powell, J. W. (1985). Microcomputer analyses of performance on a portable, simple visual RT task during sustained operations. *Behavior Research Methods, Instruments, & Computers, 17,* 652-655.

Do, Ben-Roy (2009). Research on unproctored Internet testing. *Industrial and Organizational Psychology, 2,* 49-51.

Drasgow, F., Nye, C. D., Guo, J., & Tay, L. (2009). Cheating on proctored tests: The other side of the unproctored debate. *Industrial and Organizational Psychology, 2*, 46-48.

Ejei, J., Shahabi, R., & Alibazi, H. (2012). Relationship between personality traits and self reported academic cheating in high school students. *Journal Of Psychology*, *15*, 412-424.

Erickson, M. L., & Smith W. B. (1974). On the relationship between self-reported and actual deviance: An empirical test. *Humboldt Journal of Social Relations, 1,* 106-113.

Flynn, S., Reichard, M., & Slane, S. (2001). Cheating as a function of task outcome and Machiavellianism. *The Journal of Psychology, 121*, 423-427.

Frey, M. C., & Detterman, D. (2004). Scholastic assessment or g? The relationship between the scholastic assessment test and general cognitive ability. *Psychological Science, 15*, 373-378.

Garavalia, L., Olson, E., Russel, E., & Christensen, L. (2007). How do student cheat? In E. M. Anderman & T. B. Murdock (Eds.) *Psychology of Academic Cheating.* Academic Press: San Diego, 33-55.

Gibby, R. E., Ispas, D., McCloy, R. A., & Biga, A. (2009). Moving beyond challenges to make unproctored Internet testing a reality. *Industrial and Organizational Psychology, 2,* 64-68.

Guo, J., & Drasgow, F. (2010). Identifying cheating on unproctored Internet tests: The Z-tests and the likelihood ratio test. *International Journal of Selection and Assessment, 18*, 351-364.

Haney, W. M., & Clarke, M. J. (2007). Cheating on tests: Prevalence, detection, and implications for online testing. In E. M. Anderman & T. B. Murdock (Eds.) *Psychology of Academic Cheating*. Academic Press: San Diego, 255-288.

Harding, T. S., Mayhew, M. J., Finelli, C. J., & Carpenter, D. D. (2007). The theory of planned behavior as a model of academic dishonesty in engineering and humanities undergraduates. *Ethics and Behavior, 17*, 255-279.

Hense, R., Golden, J. H., & Burnett, J. (2009). Making the case for unproctored Internet testing: Do the rewards outweigh the risks? *Industrial and Organizational Psychology, 2,* 20-23.

Ipeirotis, P. (2010). *Demographics of Mechanical Turk*. CeDER-10–01 working paper, New York University.

Lievens, F., &Burke, E. (2011). Dealing with threats inherent in unproctored Internet testing of cognitive ability: Results from a large-scale operational test program. *Journal of Occupational and Organizational Psychology, 84,* 817-824.

Loh, S., Lamond, N, Dorrian, J., Roach, G., & Dawson, D. (2004). The validity of psychomotor vigilance tasks of less than 10-minute duration. *Behavior Research Methods, Instruments, & Computers, 36*, 339–346.

Mayer, R. E., Stull, A. T., Campbell, J., Almeroth, K., Bimber, B., Chun, D., & Knight, A. (2006). Overestimation bias in self-reported SAT scores. *Educational Psychology Review, 19,* 443-454.

McTernan, M., Love, P., & Rettinger, D. (2014). The influence of personality on the decision to cheat. *Ethics & Behavior*, *24*, 53-72.

Mumford, M. D., Murphy, S. T., Connely, S., Hill, J. H., Antes, A. L., Brown, R. P., Devenport, L. D. (2007). Environmental influences on ethical decision making: Climate and environmental predictors of research integrity. *Ethics and Behavior, 17,* 337-366.

Naglieri, J. A., Drasgow, F., Schmit, M., Handler, L., Prifitera, A., Margolis, A., et al. (2004). *Psychological testing on the Internet: New problems, old issues. American Psychologist, 59,* 150–162.

Nye, C. D., Do, B.-R., Drasgow, F., & Fine, S. (2008). Two-step testing in employee selection: Is score inflation a problem? *International Journal of Selection and Assessment, 16*, 112-120.

Rettinger, D. A. (2007). Applying decision theory to academic integrity decisions. In E. M. Anderman and Tamera B. Murdock (Eds.) *Psychology of Academic Cheating.* Academic Press: San Diego.

Reynolds, D. H., Wasko, L. E., Sinar, E. F., Raymark, P. H., & Jones, J. A. (2009). UIT or not UIT? That is not the only question. *Industrial and Organizational Psychology, 2,* 52-57.

Rousseau, D. M. (1989). Psychological and implied contracts in organizations. *Employee Responsibilities And Rights Journal, 2*, 121-139.

Sackett, P. R., Borneman, J., & Connelly, B. S. (2008). High stakes testing in higher education and employment: Appraising the evidence for validity and fairness. *American Psychologist, 63*, 215-227.

Schmidt, F.L. & Hunter, J. (2004). General mental ability in the world of work: Occupational attainment and job performance. *Journal of Personality and Social Psychology, 86*, 162–173.

Sprouse, J. (2011). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods, 43*, 155-167.

Tippins, N. T. (2009). Internet alternatives to traditional proctored testing: Where are we know? *Industrial and Organizational Psychology, 2,* 2-10.

Tippins, N. T. (2009b). Where is the unproctored Internet testing train headed now? *Industrial and Organizational Psychology, 2,* 69-76.

Tippins, N. T., Beaty, J., Drasgow, F., Gibson, W. M., Pearlman, K., Segall, D. O., & Shepherd, W. (2006). Unproctored Internet testing in employment settings. *Personnel Psychology, 59*, 189-225.

Weber, E. U., & Johnson, E. J. (2009). Mindful judgment and decision making. *Annual Reviews, 60,* 53-85.