Spring 2018

Predicting Score Change: An Empirical Investigation of Cheating on Unproctored Employment Tests

Katelyn J. Cavanaugh
Old Dominion University, kjcavanaugh@mdanderson.org

PREDICTING SCORE CHANGE: AN EMPIRICAL INVESTIGATION OF CHEATING ON UNPROCTORED EMPLOYMENT TESTS

by

Katelyn J. Cavanaugh
B.A. May 2008, The University of Delaware
M.S. August 2013, Old Dominion University

A Dissertation Proposal Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY
PSYCHOLOGY
OLD DOMINION UNIVERSITY
May 2018

Approved by:

Richard N. Landers (Director)

Xiaoxiao Hu (Member)

Ryan Klinger (Member)

ABSTRACT

PREDICTING SCORE CHANGE: AN EMPIRICAL INVESTIGATION OF CHEATING ON UNPROCTORED EMPLOYMENT TESTS

Katelyn J. Cavanaugh
Old Dominion University, 2018
Director: Dr. Richard N. Landers

Unproctored internet testing (UIT) is used widely to administer employment tests (Fallaw, Solomonson, & McClelland, 2009), although cognitively loaded tests delivered by UIT are suspected to offer test takers greater opportunities to cheat and increase the risk of test taker cheating (Chapman & Webster, 2003; Tippins et al., 2006; Tippins, 2009). Despite the wide use and suspected cheating concerns, there is a dearth of research investigating cheating on cognitively loaded UITs (Naglieri et al., 2004; Beaty et al., 2011). Based on the lack of theoretically-grounded empirical studies, the current study had two goals: (1) identify which cheating methods are used by test takers to effectively raise test scores and (2) investigate the roles of general cognitive ability and effective cheating methods in raising test scores. To test the specific hypotheses, 340 adult participants recruited from Amazon MTurk completed a UIT used for employee selection first under honest conditions and then under cheating conditions. Results indicated that not all test takers were able to increase their scores by cheating; cheating effectiveness depended upon the interaction between cognitive ability and the use of effective cheating methods. These results suggest that increased cognitive ability may lead to increased cheating effectiveness on selection tests, but that score change is contingent on applicant awareness of appropriate cheating methods for those tests.

ACKNOWLEDGEMENTS

When looking for research relevant to a project, I often come across a dissertation and spend far too long getting to know the author through their acknowledgement section. It is surreal to be writing my own and feels less like a “mic drop” than I hoped it would. As with most three-and-a-half year endeavors, this one was not taken on alone, and I am glad to start this off by thanking my key players.

First and foremost, thank you to my advisor, Richard Landers, for sharing your unending enthusiasm and talent for I/O Psychology with me and for making the last almost-decade of my life my richest in learning and hopefully my poorest financially. Thank you to my dissertation committee, Ryan and Xiaoxiao, for your thoughtful feedback and insightful suggestions, and for the time you have both taken to improve this dissertation.

I came into graduate school as a cohort of one, and although I suffered more than my fair share of imposter syndrome because of it, I cherish the close relationships with more senior students I developed because of it. Kristina, Rachel, Craig, and many others outside our lab, thank you for giving your time so willingly to guide me, reassure me, and most of all, have fun with me. Thank you especially to Craig for being my research partner in crime on this project and for the many hours you spent on my dissertation instead of your own.

Thank you to my best girlfriends, Emily and Rachel, who have let me complain to them for 7.5 years, and who remind me as often as I need to hear it “who run the world”. Thank you to my parents, Susan and Jim, for everything. I am so grateful for your unending support, your belief in me, the sacrifices you made to ensure my success, and the values with which you raised me. Thank you to my almost husband, Jonas, for being a saint. You have been my cheerleader and my voice of reason when I could not be my own. You have taught me patience and tolerance by example almost every day, and I hope that it one day starts to stick.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

Unproctored internet-based tests (UITs) are widely used to administer selection tests to job applicants despite practical concerns surrounding both faking and cheating (Coyne & Bartram, 2006; Tippins et al., 2006). All Fortune 500 companies use some form of online recruiting or applications (Younger, 2011), and two-thirds of all organizations use UITs for application processes (Fallaw, Solomonson, & McClelland, 2009). Cheating on UITs in selection contexts has been defined 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" (Lievens & Burke, 2011, p. 818) and "assistance from others who have knowledge of the items before the test, assistance from others during the test, [or] substitution of test takers" (Tippins, 2009, p. 5) on a cognitively loaded test (i.e., a test of knowledge, skill, ability, or achievement; Sackett, Schmitt, Ellingson, & Kabin, 2001). In contrast, faking in selection contexts has been defined as "intentional distortion" on a personality measure (Hough, Eaton, Dunnette, Kamp, & McCloy, 1990, p. 581). Faking on personality measures is an issue across all personality measures used for selection, no matter the medium (Richman, Kiesler, Weisband, & Drasgow, 1999), although cognitively loaded tests delivered by UIT offer test takers greater opportunities to cheat and increase the risk of test taker cheating (Chapman & Webster, 2003; Tippins et al., 2006; Tippins, 2009).

The use of UITs for employee selection brings both advantages and disadvantages in comparison to traditional proctored in-person testing (Chapman & Webster, 2003; Lievens & Harris, 2003; Naglieri, et al., 2004; Tippins et al., 2006; Tippins, 2009). The potential advantages of internet-based tests in general over traditional paper-and-pencil selection tests include increased consistency, efficiency, ease of delivery and administration, increased security of test taker data, and reduced missing data. The unproctored nature of UITs additionally allows a) expansion of the applicant pool in size and geographic disbursement, leading to improved selection ratios and utility if highly qualified candidates are identified by the test; b) cost reduction from the lack of candidate travel and designated test-taking equipment, space, and administration; c) increased accessibility for candidates currently employed and special populations for whom travelling is burdensome; and d) reduced bias from characteristics unrelated to job performance such as age, race, and weight. Potential challenges introduced by the use of UITs include difficulty identifying test takers, risk of test item security, and a lack of control over the testing environment. Finally, increased opportunities for cheating is an often-cited reason to limit the use of UITs for employee selection (e.g., Tippins et al., 2006), although this phenomenon has not yet been thoroughly researched.

Researchers have not yet proposed theoretical models of applicant cheating during UIT nor conducted empirical directed-cheating studies. Instead, there are four recent studies estimating the prevalence and magnitude of cheating on UITs in naturalistic employee selection contexts (Arthur, Glaze, Villado, & Taylor, 2010; Do, Shepherd, & Drasgow, 2005; Nye, Do, Drasgow, & Fine, 2008; and Wright, Meade & Gutierrez, 2015), but each lack the rigorous methodological control needed to investigate open questions surrounding cheating on UITs; namely, how does applicant cheating affect test scores, and how can cheating be predicted? Further, researchers have put forth informed opinions of how people may cheat on UITs, e.g., assistance from others (Tippins, 2009) and using prohibited materials (Lievens & Burke, 2011). However, neither exploratory nor controlled studies have been conducted to determine whether these specific methods are used by test takers, the effectiveness of these methods to increase test scores, or whether there are other potential cheating methods.

To address these gaps, the purpose of the present dissertation is to propose and experimentally test a model explaining UIT score increases due to cheating. General cognitive ability is proposed as both a predictor of the use of effective cheating methods and as a moderator of the use of effective cheating methods to predict cheating effectiveness, which is measured as the latent test score change due to cheating. Using a controlled experimental research design, adult research participants will complete a cognitively loaded UIT under both honest and cheating instruction conditions in order to isolate the effect of cheating on score change. The outcomes of this study will be used to inform theory regarding the potential for test score change due to cheating and to inform applied decisions regarding the use of UITs for employee selection, such as the implementation of cheating prevention and deterrence methods.

Cheating on Unproctored Internet Tests

A problem commonly considered by practitioners when administering cognitively loaded UITs is increased cheating potential, the prevalence and impact of which has not yet been thoroughly explored theoretically. Among researchers and practitioners, such use for employee selection is considered controversial (Sackett & Lievens, 2008, p. 437). Although there is consensus among experts that administering cognitively loaded UITs is risky, there are a range of opinions regarding which alternative is most appropriate, including only using UITs to screen out candidates unlikely to be effective, always following up a UIT with a proctored assessment, or never using UITs for high stakes tests of ability (Tippins et al., 2006). This is in part because there are so many unknowns related to cheating, including the potential impact of cheating on test validity (Naglieri et al., 2004; Tippins et al., 2006), the types of cheating possible, valid evidence of cheating, and the rates of cheating in unproctored and proctored testing environments (Tippins, 2009).

What is known in this domain comes from three distinct research literatures: cheating in UIT selection contexts, cheating in academic contexts, and faking in selection contexts. First, researchers that have studied cheating in selection have generally focused upon the measurement and detection of faking and cheating rather than its effects**,** using descriptive research designs in applicant samples to compare test scores between-subjects or within-subjects in high stakes and low stakes contexts, and in both proctored and unproctored settings (Arthur et al., 2010; Do, et al., 2005; Nye et al., 2008; Wright et al., 2014). Second, researchers in academic contexts also tend to focus on measurement but not detection, typically employing descriptive self-report research designs to estimate the frequency of cheating on components of courses or the frequency of using various cheating methods (Franklyn-Stokes & Newstead, 1995; Gallant & Drinan, 2006; Teixeira & Rocha, 2010). There are no published directed cheating studies in either of these research literatures, **likely in part due to the practical limitations of asking applicants to cheat on a test used for employment selection or students to cheat on a test that impacts course grades.** Third, researchers studying faking in selection, which makes up the largest of these literatures by a substantial margin, have generally employed either descriptive designs to compare test scores between applicant and non-applicant test takers (Birkeland, Manson, Kisamore, Brannick, & Smith, 2006) or controlled experimental designs to compare test scores between-subjects or within-subjects under “honest” and “fake good” instruction conditions or “fake good” and “fake bad” instruction conditions in student or research volunteer samples (Viswesvaran & Ones, 1999).

In all three literatures, researchers have also investigated the prevalence of cheating and faking and the magnitude of score change achievable from them, although this evidence is sparser and varies widely in both approach and quality. First, in cheating in UIT selection contexts, two studies of applicant samples reported fairly low prevalence of cheating as implied by differences between proctored and unproctored settings: Nye and colleagues (2008) found only 1% of applicants (four out of 856) decreased scores on a proctored follow-up test compared to their unproctored scores as part of a selection process (Nye et al., 2008), and Arthur and colleagues (2010) found 7.7% of applicants (eighteen out of 239) reached their threshold of scoring lower on a low stakes follow-up test (Arthur et al., 2010). Two other studies of applicant samples reported a range of magnitudes; Do and colleagues (2005) reported slightly higher scores on unproctored tests compared to proctored scores (d = 0.09) whereas Wright and colleagues (2014) found higher unproctored scores compared to proctored scores in one sample (d = 0.51) but higher proctored scores compared to unproctored scores in a second sample (d = -0.11). Second, in the academic cheating literature, prevalence estimates are entirely based upon self-report; studies published between 1963 and 1996 provide estimates ranging from 44% to 82% of undergraduate students cheating (McCabe, Treviño, & Butterfield, 2001). No research has been published in the academic cheating literature that investigates the magnitude of test score impact due to cheating by measuring it directly (Teixeira & Rocha, 2010). Third, in faking in selection contexts, prevalence of faking in high stakes situations has been most directly measured by tracking the endorsement of bogus items, such as reporting experience using equipment that does not exist. Pannone (1984) reported 35% of applicants endorse these items whereas Anderson and colleagues (1984) reported 45%. Meta-analytic estimates of score change magnitude due to faking reveal moderate to very large effect sizes in score change due to instructions (sample size weighted mean *d* ranging from 0.47 to -3.66; Viswesvaran and Ones, 1999) and in comparisons of applicant and non-applicant scores (δ ranging from 0.11 to 0.45; Birkeland et al., 2006).

Several prevention and detection methods have been proposed, which rely on assumptions about methods people use to cheat or fake. First, single-use URLs, passwords, warnings of identity checks and consequences of cheating, speeded tests, follow-up tests, Computer Adaptive Testing, and remote proctoring (Tippins et al., 2006 Guo & Drasgow, 2010; Fetzer & Grelle, 2010; Reynolds & Dickter, 2010) are used to discourage and detect cheating in online selection contexts. Second, academic cheating scholars advocate a holistic approach to academic integrity; for example, Gallant and Drinan (2006) suggest a multi-strategy approach involving both school-wide policies and classroom norms. Third, in faking research, forced choice items, subtle items, and warnings are commonly used to discourage faking (Hough et al., 1999), whereas detection scales, eye-tracking, and response latencies have been use to detect it, although in lab contexts (Holden & Hibbs, 1995; van Hooft & Born, 2012). Little data is currently available on how often these methods are actually used in organizations. Interestingly, there is little overlap between these methods; warnings are used for both faking and cheating in selection contexts but all other methods used are unique to either faking or cheating. Faking discouragement and detection methods involve changes directly to some or all test items, whereas methods related to cheating typically involve limiting cheating opportunities for the entire test.

Existing research on cheating methods thus provides some insight into cheating on UITs which might be applied to the selection context, but there is no empirical evidence exploring this or the efficacy of these methods to increase scores. First, Tippins and colleagues (2006) propose that the methods used to cheat on selection tests involve receiving information from others and using “forbidden” resources (p. 210). Second, academic researchers have attempted to directly survey students on which methods of cheating they have used by asking them to endorse various cheating methods on behavioral lists created from discussions with students and alumni (e.g., Franklyn-Stokes & Newstead, 1995). Third, it is suggested that faking involves comprehending test items, understanding situational requirements, and then selecting the best answer (Pauls & Crost, 2005). Although these potential methods put forth are informative, they cannot serve as definitive taxonomies of all potential methods used to cheat, nor can they be used to determine which of those methods can be used to effectively cheat.

Research Question 1. What cheating methods can be used to effectively increase scores on a UIT, and how do these methods compare in effectiveness?

Need for Empirically Supported Theories to Explain Cheating on Cognitively Loaded UITs

In general, research focusing upon cheating on UITs for selection has been limited. Empirical studies feature two untested assumptions: first, that cheating does not occur in proctored and low stakes comparison groups, and second, that cheating is detectable by score change or score differences alone and not in combination with the measurement or self-report of cheating (e.g., Arthur et al., 2010, Nye et al., 2008). However, it is unlikely that all opportunities to cheat and all instances of cheating are eliminated by proctoring or low stakes testing, so the control groups in these studies (i.e., Do et al., 2005; Wright et al., 2014) could be contaminated by cheating. Because cheating is operationalized solely by score differences and not actual test taker behavior, this indicates that all test takers who cheat must be effective in raising their test scores, though it is unlikely that all applicants using prohibited methods raise their test scores. Finally, research samples in this area may suffer from self-selection bias, limiting the studies' ability to detect score change due to cheating. A mere 7% of applicants were invited (due to their high unproctored test scores) to take the proctored follow up test in Nye and colleague's (2008) investigation, and only 3% of the unproctored high stakes test takers voluntarily completed the low stakes follow up test in Arthur and colleague's (2010) investigation. If applicants who cheated were less likely to complete a follow-up test, these studies underestimate the prevalence and magnitude of score change due to cheating.

These empirical limitations highlight a more significant problem; these previous investigations of UIT cheating are generally atheoretical, relying on descriptive research designs in applicant samples in addition to unconfirmed assumptions about cheating detection. Research studies that are not grounded in theory but guided solely by accumulated empirical data tend to lead to inconsistent and confusing conclusions, instead of making sense of phenomena (Landy, 1993). In descriptive research designs, variables are recorded as they naturally occur and no variables are controlled or manipulated (Sackett & Larson, 1990). Effectively testing potential causal explanations is difficult in this type of design; only when well-defined theories guide the equations of causal patterns and all relevant variables are measured with little measurement error can conclusions of causation be interpreted in confidence (Sackett & Larson, 1990). Given that these conditions have not yet been met in investigation of cheating, the causal relationships underlying cheating have not yet been discerned, given the lack of experimental research. One expert went so far as to say, "the most pressing need [for UIT research] 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," (Tippins et al., 2006, p. 218). No such model currently exists in the literature.

Overview of the Theoretical Model to be Tested

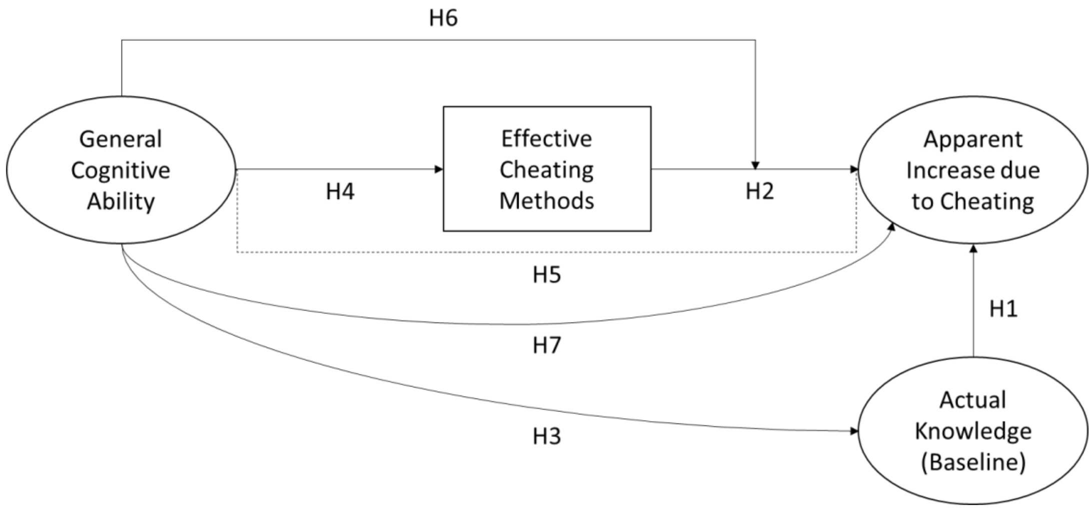


Figure 1. Proposed Theoretical Model of Cheating on UITs.

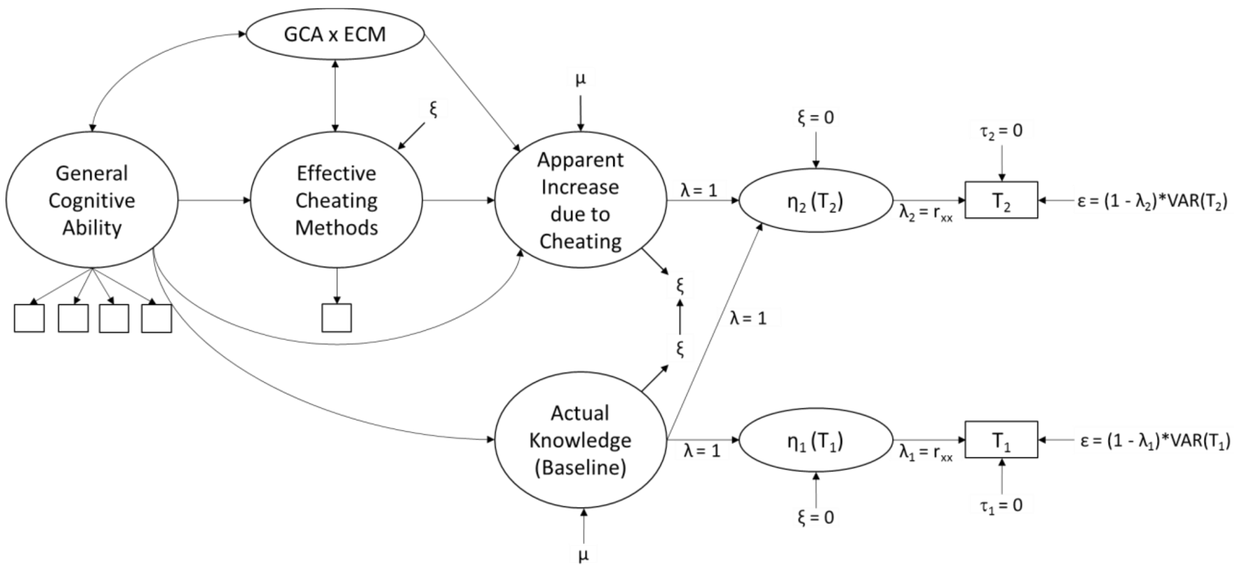


Figure 2. Proposed Empirical Model of Cheating on UITs.

To begin to remedy this gap in theory, and to improve the foundation of future empirical work on cheating, I have developed a theoretical model of score change due to UIT cheating, providing specific paths by which individual differences in cognitive ability and the use of effective cheating methods are expressed in scores where cheating has occurred, which is depicted in Figure 1. This model incorporates apparent increase as a latent change from actual knowledge. Figure 2 incorporates the empirical relationships to be tested with Latent Change Analysis (LCA), which involves the statistical modeling of change between two or more observations (McArdle & Nesselroade, 2014). The latent change construct (apparent increase due to cheating, in this case) is not a latent Time 2 score; instead, it represents the latent rate of change between Time 1 and Time 2, interpreted much like similarly defined latent variables in SEM-based latent growth modeling. Scores are observed twice (T_1 and T_2) and their latent variances are estimated (η_1 and η_2). One new latent construct is then defined as loading on both Time 1 and Time 2 latent variances (both constrained to 1) which can be interpreted as a latent measure of scores at baseline. A second latent construct is defined as loading on only the Time 2 latent variance (also constrained to 1), which can be interpreted as a latent measure of change. These baseline and change constructs can then be modeled at will, as in any other SEM. Model fit also can be interpreted by comparing the hypothesized theoretical model with the observed data, as with other SEM models (McArdle & Nesselroade, 2014). Unlike traditional longitudinal and repeated measures analyses, LCA does not assume Time 1 and Time 2 score are equivalent or that the change between Time 1 and Time 2 is linear. Latent change will be modelled because LCA incorporates the many benefits of Structural Equation Modeling (SEM), allows investigations of true score change and predictors affecting individual rates of change, and addresses limitations of traditional repeated measures analyses used in previous UIT research. Unlike the difference scores used in previous studies (e.g., Arthur et al., 2010), LCA does not control for Time 1 score; when difference scores are used, this eliminates the effects of predictors except for those predicators that predict changes in rank order. Instead, LCA assumes that Time 1 mean scores contain useful statistical information needed to estimate inter- and intraindividual differences (McArdle & Nesselroade, 2014). This approach has been used recently to study change in a variety of Psychology research areas; changes in the social desirability of job seekers with unhealthy alcohol use (Haberecht, Schnuerer, Gaertner, Johns, & Freyer-Adam, 2015), changes in proactive personality and work attributes (Li, Fay, Frese, Harms, & Gao, 2014), and changes in intentions, planning, and self-efficacy to predict latent change in health behaviors (Reuter et al., 2010). Finally, the within-person research design used here, will increase the information used for each participant and reduce error arising from variability in individual differences between subjects by using each participant as their own control (Maxwell & Delaney, 2004).

Using the framework of latent change modeling, I will furthermore propose that individual differences in general cognitive ability are a critical predictor of the use of effective cheating methods as well as a moderator of the relationship between effective cheating methods and the apparent latent change attributable to cheating. Much of the empirical work in UIT cheating has already examined the impact of motivation on scores (e.g., high stakes versus low stakes testing; Arthur et al., 2010), so ability represents a more sizable gap in the literature. Further, the influence of salient high stakes situational prompts (e.g., a job application process) constrains the impact of motivational individual differences on score change due to cheating in real-world scenarios, whereas the effect of situational influences on ability's impact on score change is likely to be smaller. The importance of abilities for performance of cognitively loaded tasks has been widely researched in Psychology; the capacity to perform is necessary for effective performance (Hirschfeld, Lawson, & Mossholder, 2004; Perry, Hunter, Witt, & Harris, 2010) and objective ability is critical to enable motivated individuals to perform (Lawler & Suttle, 1973; Locke, 1978). Results of numerous studies of faking provide evidence that it is possible to fake personality tests when directed to do so or in the context of high stakes situations, but that people do not homogenously alter their scores (Hough et al., 1990; McFarland & Ryan, 2000; Ryan & Sackett, 1987; Ones, Viswesvaran, & Reiss, 1996). Neither experimental manipulations nor comparisons of high and low stakes groups induce perfectly consistent patterns of faking as expected by manipulation or group (Zickar, Gibby, & Robie, 2004), leading to much discussion and investigation of the causes of this variability. Given this literature, I contend a similar pattern is likely for cheating; there is likely to be variability in score change magnitude as a result of individuals varying in their cheating ability.

Although previous research provides compelling evidence that people can increase their scores on non-cognitive UIT measures when directed to appear more desirable for hiring, a lack of similarly designed studies on cognitively loaded measures has created a research gap. As the literature currently stands, no researchers have hypothesized or investigated potential predictors of cheating on UITs (Cavanagh, 2014) despite numerous calls for research in this area citing a dearth of research on the Psychological causes of cheating (Pulfrey & Butera, 2013). It is currently unknown who cheats, why they cheat, and how cheating is facilitated or prevented (Tippins, et al., 2006; Tippins, 2009). The current study will address this gap by testing the effects of cheating instructions on within-person test score change in order to determine the extent to which people can increase their scores when cheating, compared to their scores on that same test when they are not cheating. Assessing such differences in the framework of latent change is ideal, because LCA can isolate the effects of change while incorporating the effects of measurement error, enabled by change being predicted with other latent variables. Given how poorly prior literature has isolated the effects of cheating, this makes LCA the most appropriate approach to test a causal model of score differences and their determinants experimentally.

Prior Knowledge Predicts Apparent Knowledge Change due to Cheating

Previous research indicates those with greater existing knowledge within a subject area should be better able to cheat on a test in that area; specifically, actual knowledge should lead to greater increases when cheating. Support for this hypothesis comes from two research literatures. First, several studies on trainee learning have shown that trainees who begin a training program with more knowledge or experience as measured by a pre-training knowledge test tend to score higher on post-tests than those with less pre-training knowledge (Brown, 2001; Calisir, Eryazici, & Lehto, 2008; Pieschl, Stahl, & Bromme, 2008; Hannafin & Sullivan, 1996; Orvis, Fisher, & Wasserman, 2009). This effect may exist because those with more experience and knowledge have different and superior mental models, information processing, and information storage capacity for a particular subject area compared to those with less experience and knowledge (Salas & Rosen, 2010). It is likely that this extends to cheating on a UIT; greater existing knowledge enables increased scores due to cheating because cheating likely involves the use of mental models, information processing, and storage. Second, recent research investigating practice effects due to re-testing has shifted focus to the learning benefits of taking tests, suggesting that a test is not solely a measure of learning but rather is itself a learning experience (Karpicke & Blunt, 2011). The practice of knowledge retrieval and reconstruction that occurs during test taking has been shown to produce increased gains in learning over traditional studying methods focused on encoding knowledge (Karpicke & Blunt, 2011). In the current study design, participants will complete a knowledge-based UIT at Time 1 with no incentive to cheat. Given this prior research, those with higher Time 1 scores should be able to increase their scores more when instructed to cheat than those with lower Time 1 scores.

Hypothesis 1. Baseline knowledge positively predicts apparent increase in knowledge due to cheating.

Predictors of Latent Change in Apparent Knowledge due to Cheating

Individual differences in ability have been theorized to influence how cheating leads to test score increases. The influence of ability has not generally been investigated in the academic cheating literature in relation to cheating effectiveness, which may be a reflection of the research methods typically used (i.e., descriptive, post-test self-reports) and the research questions typical of the domain; generally, education scholars are more interested in motivation to cheat than the ultimate effect on test scores. Whitly (1998) described this as a limitation when speculating that some students reporting cheating behaviors are not effective in increasing their grades, as students that are motivated to cheat may not be able to do so effectively. In the faking literature, where test validity is a greater concern, this has been theorized and tested directly. Numerous faking researchers have suggested that to distort responses, beyond motivation or intentions, cognitive abilities are necessary to comprehend test items, recognize specific situational requirements and opportunities to distort, then respond accordingly (i.e., aligned with situational expectations; Austin, Hofer, Deary, & Eber, 2000; Ellingson & McFarland, 2011; McFarland & Ryan, 2000; Pauls & Crost, 2005). When empirically tested, results support the idea that ability predicts response bias (Bing, Whanger, Davison, & VanHook, 2004; Furnham, 1986; Grubb & McDaniel, 2007; Mersman & Shultz, 1998; Pauls & Crost, 2005), although researchers have not yet explored the process by which ability allows test takers to distort their responses and effectively fake.

Although job applicants are highly motivated to perform well on selection tests, few applicants reach score change thresholds to be identified as obvious cheating. Identifying cheaters on tests using methods such as outlier analysis and searches for answer patterns signifying cheating (e.g., Landers, Sackett, & Tuzinski, 2011) have demonstrated some success in identifying which test takers may be cheating based only on their responses to personality test items. However, because people likely vary in their ability to cheat, these methods are inconclusive to identify who is cheating and how effective they are at cheating (i.e., was an individual able to score higher because of cheating or because they have an exceptionally high true score on the target construct?). The small research literature investigating cheating in selection has shed some light on the prevalence and magnitude of score change in applicant samples, but no theory or empirical evidence is currently available to link those score changes to actual test taker cheating behavior; the score changes detected may be due entirely to factors other than cheating. It is also currently unknown whether score changes due to cheating are high enough in magnitude to be detected as cheating, as it is possible that many test takers cheat but few are able to substantially raise their score by cheating. Finally, the methods used to cheat are unknown, as are the effectiveness of those methods.

One key individual difference, the application of effective cheating methods, should predict test score increases due to cheating. Expert ratings have long been used to determine effectiveness, e.g., ratings of training methods (Carroll, Paine, & Ivancevich, 1972) and ratings of occupational stress management interventions (Bellarosa & Chen, 1997). There are many different ways a test taker could try to cheat on a test, but each method is not guaranteed effectiveness in selecting the right answer. Prior to testing the theoretical model of test score increase due to cheating, testing experts will be asked to rate potential cheating methods for their efficacy to increase test scores. Test takers who use effective methods, as rated by experts, should be more likely to effectively increase their test scores. Test takers who use methods that experts rate as less effective will be less likely to enable test takers to raise their scores.

Hypothesis 2. Effective cheating methods positively predicts apparent increase in knowledge due to cheating.

General Cognitive Ability Predicts Baseline Knowledge Scores

General cognitive ability (GCA) has a well-known influence on the performance of all cognitive tasks. GCA is a "general mental capability that… involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly, and learn from experience" (Gottfredson, 1997, p.13). The impact of cognitive ability on the performance of cognitive tasks is widely supported by evidence in many domains, including job task performance (Hunter & Hunter, 1984) and learning (Colquitt, LePine, & Noe, 2000). The Cattell-Horn-Theory of Intelligence (CHC) is the current "consensus…for understanding the structure of human intelligence" (McGrew, 2009, p. 1), and serves as the framework for understanding the effects of cognitive ability in the current study. CHC is a hierarchical taxonomy of intelligence in which a general "g" factor consists of 9 broad abilities; fluid reasoning, crystallized intelligence, visual processing, auditory processing, processing speed, short-term memory, long-term retrieval, quantitative knowledge, and correct decision speed, each of which consists of several even narrower facets of ability (McGrew, 2009).

Given the many potential stimuli encountered in selection tests, effectively choosing correct answers on an online knowledge test involves several of these abilities, such as fluid reasoning, visual processing, processing speed, and long-term memory retrieval, to effectively choose correct answers. Numerous empirical studies have shown that GCA is strongly and positively related to performance on a variety of cognitively loaded tests, such as the SAT and

ACT (Condon & Revelle, 2014; Schmidt & Ford, 2003). Thus, higher GCA broadly conceptualized should lead to increased baseline knowledge scores in comparison to lower GCA.

Hypothesis 3. General cognitive ability directly and positively predicts baseline knowledge.

General Cognitive Ability Predicts the Use of Effective Cheating Methods

Because choice of cheating methods is a complex cognitive task, GCA is also likely required to choose effective cheating methods. Higher GCA will lead to increased apparent knowledge when cheating in comparison to lower GCA, given the varied reasoning involved in choosing the best method to accomplish the task of cheating effectively for a given situation and its particular restraints. Various components of GCA, such as sensory discrimination (Acton & Schroeder, 2001) and complex problem solving (Stadler, Becker, Gödker, Leutner, & Greiff, 2015), should be involved in decision-making processes to determine how to cheat, as these decisions are made based on a number of situational factors while experiencing novel stimuli. Thus, GCA broadly conceptualized should predict the use of more effective cheating methods, which in turn should predict apparent increase in knowledge due to cheating.

Hypothesis 4. General cognitive ability positively predicts the use of effective cheating methods.

Hypothesis 5. The relationship between general cognitive ability and test score increase is mediated by the use of effective cheating methods.

General Cognitive Ability Moderates the Relationship between Effective Cheating Methods and Apparent Knowledge Change due to Cheating

In addition to the influence of cognitive ability on the selection of effective cheating methods, general cognitive ability should influence the effective use of these behaviors. Those higher in GCA will be able to more effectively carry out these methods in order to raise their test scores, so their use of those methods will be more strongly related to score change. Individuals lower in general cognitive ability will be less effective in their use of these methods and will not be able to raise their scores as easily, so their use of those methods will be less strongly related to score change. The methods people use to cheat likely involve understanding novel test items, interpreting the item accurately to search or contact one or more outside sources, interpreting and filtering new information encountered from those sources, and then using that to correctly answer a test item in which the individual did not previously know the correct answer. This moderation relationship should exist because using effective cheating methods to effectively cheat on a test is a novel and complex task in which creative problem solving must be used, and there is ample empirical evidence that individuals higher in general cognitive ability are more successful in these types of tasks (Hunter & Hunter, 1984; Ones et al., 2005). Thus, the effective use of effective cheating behaviors should depend on cognitive ability.

Hypothesis 6. General cognitive ability moderates the relationship between the use of effective methods and apparent increase in knowledge due to cheating.

Additionally, GCA should predict leftover variance in test score increases due to cheating not explained by the use of effective cheating methods or the interaction between the use of those methods and GCA. The effects of GCA on retesting effects (i.e., score changes after previous exposure to a test; Lievens, Buyse, & Sackett, 2005) have been studied by Lievens and colleagues (2007) in an investigation of the effects of both memory and GCA on a retest of medical and dental school entry exams. The researchers found that memory was a stronger predictor of retest scores, though the testing sessions were held approximately two months apart (Lievens, Reeve, Heggestad., 2007). For a test retaken in a single session, individual differences in memory should have a weaker effect on Time 2 scores because of the comparatively short time between test-taking sessions and cognitive ability should have a greater effect on retest scores. Evidence from the GCA literature has identified relationships to constructs directly involved in retesting, such as information processing (Sheppard & Vernon, 2008), memory, and reading comprehension (James & Carretta, 2002; Ree & Earles, 1994).

Hypothesis 7. General cognitive ability directly predicts apparent increase in knowledge due to cheating.

CHAPTER 2

CHEATING METHOD SCALE DEVELOPMENT STUDY

This study was conducted to answer Research Question 1: What cheating methods can be used to effectively increase scores on a UIT? A sample of Subject Matter Experts (SMEs) in employee selection tests and/or cheating methods will revise an existing measure of cheating methods for use in a UIT context and rate each method for effectiveness of raising test scores. These ratings will be used to compile a list of UIT cheating methods, weighted by effectiveness, which will be used to develop a scale assessing effective cheating methods.

Method

Participants. A group of SMEs with expertise in employee selection tests was identified to participate in a scale revision and rating task of cheating methods. Twelve SMEs from the large business consulting firms ICF and CEB, as well as selection experts from IBM, were invited via email to volunteer 30 minutes for scale revision and rating tasks regarding UIT cheating methods. All SMEs had experience in employee selection test construction, scoring, and/or validation, or experience in identifying and/or reducing cheating on employee selection tests.

A power analysis was conducted to determine how many SMEs would be needed for a target inter-rater reliability of .95. Meta-analytic research exploring the reliability of supervisor ratings of performance quality (defined as "quality of tasks completed, lack of errors, accuracy to specifications, thoroughness, and amount of wastage"), a single supervisor rater's reliability for rating employee performance quality was .63 (Viswesvaran, Ones, & Schmidt, 1996). This value was used as the baseline for the power analysis to determine inter-rater reliability across raters. I adjusted this value using the Spearman-Brown prophecy formula to predict the number of SME raters required to reach a .95 threshold of reliability. The results of these calculations indicated that 11 SME raters were needed to obtain an inter-rater reliability of .95.

Materials. A list of potential cheating methods was distributed to SMEs, which can be found in Appendix A, along with instructions for the two tasks. This list of cheating methods was derived from Yardley et al.'s (2009) list of cheating behaviors compiled from discussions with undergraduate students and recent graduates, then used as a measure of undergraduate student self-reported cheating frequencies (Gaskill, 2014). Based on a review of all available student cheating measures, Gaskill's (2014) list is the most applicable to the types of behaviors test takers may use to cheat on UITs. Several other measures were identified (e.g., Newstead et al., 1996; Franklyn-Stokes & Newstead, 1995; Graham et al., 1994; McCabe & Trevino, 1993) but most of those methods were specific to in-person testing (e.g., "copying from another student during a test without his/her knowledge"; McCabe & Trevino, 1993) or open-ended tests ("Paraphrasing material from another source without acknowledging the original author"; Newstead et al., 1996).

Procedure. SMEs were first asked to revise the list of methods for relevance to an online unproctored multiple choice test by removing irrelevant or impossible items, revising existing items, and brainstorming additional potential methods test takers might use to cheat that were not already contained in the list, returning an updated list of methods. All updated lists were compiled into a final list of cheating methods by conceptually pooling the methods and removing overlap, and this final list was then sent back to each SME, who rated each method's *potential effectiveness* for raising test scores. SMEs rated effectiveness for each method on a Likert scale (1 = *potentially not at all effective*, 5 = *potentially very effective*).

Analyses. Mean effectiveness ratings for each cheating method were calculated. These means were used as weights indicating the potential effectiveness for each method to increase test scores. These methods and associated weights were then used to create a scale to capture self-reported cheating method effectiveness, which was used in the main study. Specifically, the SME-derived effectiveness weights were used to create a weighted mean score that assesses the mean effectiveness of each participant's cheating methods (e.g., if a participant reported engaging in two methods with weights of 3.5 and 4.0, their effectiveness score would be considered 3.75).

Results

Seventeen methods for cheating were identified by the development process. These methods and their mean effectiveness weights can be found in Table 1. The final list included solicitation of help from other people and other sources in a variety of ways, including asking people physically present, asking others using electronic messaging, and searching the internet. Four methods (18-21) were specific to the test used in the main study, and involved using the program itself to cheat. The two methods rated the most effective by SMEs (at 4.08 out of 5.00) were "Purchase test content online (i.e., test cheating websites)" and "Solicit help (e.g., ask for help, advice, or answers) from someone physically present while you take the test". The method rated the least effective by SMEs (at 2.50 out of 5.00) was "Inspect source code of test for indications of correct answers". Reliability of the scale mean was investigated by calculating an intraclass correlation coefficient (ICC(2,12) = 0.669) on rater means, which was near the conventional standard of 0.70 (Nunnally, 1978). Multivariate rater outliers were inspected, and none were found.

Table 1

Cheating Methods and Effectiveness Weights

Method	*Min*	*Max*	*M*	*SD*
1. Use electronic notes stored on devices during test pertaining to test content (e.g., computer, tablet, cellphone)	2.00	5.00	3.50	1.00
2. Purchase test content online (i.e., test cheating websites)	2.00	5.00	4.08	1.08
3. Search the Internet for test content	2.00	5.00	3.33	1.30
4. Pause test to search for answers online	2.00	5.00	3.58	1.00
5. Pause test to consult with another person	2.00	4.00	3.67	0.65
6. Inspect source code of test for indications of correct answers	1.00	5.00	2.33	1.37
7. Take the test once using a login created with false information, then re-take the test with your own login	1.00	5.00	3.58	1.24
8. Ask a contact within the organization or test company to request new assessment sessions to allow you to re-take the test	1.00	5.00	2.50	1.38
9. Ask a contact within the organization or test company to gain insight into the test content	1.00	5.00	2.42	1.51

(Table 1 continued)

Method	*Min*	*Max*	*M*	*SD*
10. Solicit help (e.g., ask for help, advice, or answers) from someone physically present while you take the test	3.00	5.00	4.08	0.67
11. Solicit help (e.g., ask for help, advice, or answers) from someone over the phone while you take the test	2.00	5.00	3.58	0.90
12. Solicit help from someone via electronic messaging (e.g., email, text message, Google chat, Facebook messenger) while you take the test				
12a. by typing text about the test content	2.00	4.00	2.92	0.90
12b. by screenshotting test content	3.00	5.00	3.67	0.65
12c. by sending a picture of test content	3.00	5.00	3.67	0.65
12d. by copying and pasting test content	2.00	4.00	3.25	0.75
13. Ask someone to take the test for you	2.00	5.00	3.42	1.00
14. Hire/pay someone to take the test for you	2.00	5.00	3.58	1.08
15. Receive answers from someone else completing the test at the same time	1.00	5.00	3.33	1.23

(Table 1 continued)

Method	*Min*	*Max*	*M*	*SD*
16. Post test content on social media (e.g., Facebook, Twitter) to solicit help or answers from others	1.00	5.00	2.58	1.38
17. Post test content on an online discussion board (e.g., Yahoo Answers, Microsoft Answers, Turk Opticon) to solicit help or answers from others	1.00	5.00	2.83	1.34
18. Use the program being tested (Excel) to determine the correct answer on the same computer you completed the test on	1.00	4.00	2.67	1.07
19. Use the program being tested (Excel) to determine the correct answer on a different computer	2.00	5.00	3.67	0.89
20. Search for test content in the Help section of the program being tested (Excel) on the same computer you completed the test on	1.00	4.00	2.58	0.79
21. Search for test content in the Help section of the program being tested (Excel) on a different computer	2.00	5.00	3.42	1.00

*n = 12 ratings for each method's effectiveness rating.

CHAPTER 3

METHOD

Pilot Study

A pilot study was conducted to select the most appropriate UIT for the main study. Four potential knowledge and skill UITs, currently used for employee selection by I/O psychologists in a large US consulting firm, were administered to Amazon Mechanical Turk (MTurk) Workers to ensure they were appropriate to test the hypothesized model in the main study. Specifically, this pilot study was necessary, because these tests were constructed for and used in employee selection and had not been validated with an MTurk sample. A ceiling effect when completing the test honestly would limit MTurk Workers' ability to increase their scores when cheating during the main study, which would limit the investigation of the causes of those score increases. There also needed to be additional room at the high end of the scale such that it would be unlikely to exhibit a ceiling effect on scores when cheating. The test chosen therefore needed to have a difficulty level among MTurk Workers likely to produce a normal distribution of scores both before and after cheating.

Participants. Participants were recruited from the crowdsourcing website, Amazon MTurk. Within MTurk, "Requestors" outsource job task requests or "Human Intelligence Tasks" (HITs) to "Workers" to complete at their convenience, in exchange for monetary compensation (Kleeman, Vob, & Reider, 2008). MTurk is increasingly used as a viable approach for research participant recruitment and data collection in a variety of research topics (e.g., Cole et al., 2009; Strickland, Reynolds, & Stoops, 2016). Investigations of data quality reveal that MTurk Workers and undergraduate participants do not differ significantly in terms of completeness, quality, completion time, or word count on open-ended questions (Behrend, Sharek, Meade, & Weibe, 2011).

Considering the relative novelty of online samples, the specific merits and drawbacks of this sample were considered, per Landers and Behrend's (2015) recommendations. No prior theory or research suggests MTurk Workers possess unique capabilities to cheat on a test differently from a typical job applicant, nor should they score higher or lower than another adult sample on a GCA test. The majority of MTurk Workers are Caucasian, indicating potential for range restriction in GCA test scores, given the wide support for the relationship between GCA and ethnicity (e.g., Roth, Bevier, Bobko, Switzer, & Tyler, 2001). Compared to undergraduate students, a potential alternative sample, MTurk Workers are more likely to possess similar demographic characteristics to job seekers and applicants, including age, education attainment, and employment experience, according to Behrend and colleagues (2011).

Because the goal of this pilot was only to estimate the mean, standard deviation, and reliability of each test, participants included a sample of only 30 MTurk Workers; 15 (50%) of which were female, 12 (40%) male, and 3 (10%) preferred not to answer. Twenty-one (71%) reported their ethnicity as Caucasian, 1 (3%) as Black, 5 (17%) as Asian, 1 (3%) as Hispanic, 1 (3%) as "Two or more races," and 1 (3%) preferred not to answer. Participants selected their age range; 18 (60%) reported they were younger than 40 years old, 9 (30%) reported they were 40 years old or older, and 3 (10%) preferred not to answer. Participants were compensated with $3 for completion of the study. This rate was determined based upon suggested rates for MTurk study compensation (75 cents for a 30 minute task; Barger, Behrend, Sharek, & Sinar, 2011), as well as recent research noting diminishing returns in work quality and pay satisfaction for overpaying beyond those accepted rates (Behrend, 2016).

Measures. A research agreement was created between CEB Inc and me for the use of four UITs used for employee selection, in a manner consistent with the procedure detailed below. These measures are proprietary and are currently in use as selection tests for clients of CEB. Thus, descriptions and example items are given for each test, but full items could not be included in the manuscript of this dissertation. Each test has a generous time limit and was not designed as a speeded test.

Basic Computer Literacy. This assessment evaluated knowledge of general computer terms, ability to manage files and accomplish tasks in a Windows operating system and application software, and access the internet. There was a 35 minute time limit and a total of 30 items; 16 simulation and 14 multiple choice items of various skill levels (15 basic, 11 intermediate, 14 advanced). See Appendix B for an example item.

General Clerical Grammar. This assessment evaluated skill using various parts of speech in written communication, including subject-verb agreement, sentence structure, and punctuation. There was a 20-minute time limit and a total of 30 multiple choice items of various skill levels (10 basic, 10 intermediate, 10 advanced). See Appendix B for an example item.

Microsoft Excel 2010. This assessment evaluated skill using Microsoft Excel, including sorting and filtering data, applying functions and formulas, modifying cell formatting and content, creating and labeling charts and pivot tables, and using conditional formatting and statements. There was a 35-minute time limit and a total of 30 simulation items of various skill levels (10 basic, 10 intermediate, 10 advanced). See Appendix B for an example item.

Microsoft PowerPoint 2010. This assessment evaluated skill using Microsoft PowerPoint, including creating and saving presentations, adding and arranging multimedia elements, formatting slides and content, modifying the layout, and reviewing and delivering presentations. There was a 25-minute time limit and a total of 20 simulation items of various skill levels (10 basic, 10 intermediate). See Appendix B for an example item.

Procedure. After signing up for the study and completing an online informed consent document, participants followed a link to CEB's online testing platform containing the four tests. Participants were instructed that it is important for the conclusions of the study that they take each test to the best of their ability but do not use any outside resources to do so. Participants did not see their scores or receive any compensation based upon their scores to minimize the likelihood of attempting to cheat on the tests. Participants completed each test in a random order to decrease the potential for error due to order and fatigue.

Results. Descriptive statistics for each test were examined, including means, standard deviations, and reliability estimates (KR-20), which can be found in Table 2. The purpose of these analyses was to examine mean scores and distribution of scores in an MTurk sample for each test. I began with a strategy of finding the test with a normal distribution of scores among MTurk Workers and a mean score approaching or exceeding a difference of 4 standard deviations from the total possible score. If one test had been found to have an appropriate mean score and variability, that test would be used for the main study. If none of the tests met those requirements, one test would be tailored (i.e., items dropped) until more desirable psychometric characteristics were achieved.

All four test score distributions were normally distributed for the 30 pilot participants. The Microsoft Excel test was the only test to meet the mean score requirements; 4 standard deviations (5.72) above the mean score (10.62) was 33.48 (3.48 greater than the total possible score). Therefore, this test was chosen as the test for the main study. Item statistics for the

Microsoft Excel test in the pilot study can be found in Table 3, sorted from most difficult to least difficult. The easiest item was dropped from the test, creating a new 29 item test with a mean score of 9.69 (33% correct; SD = 5.63) in order to increase the difficulty for the main study.

Table 2

Descriptive Statistics of Four Pilot Tests

Variable	*Min*	*Max*	*M*	*%*	*SD*	*Skew*	*Kurtosis*
1. Basic Computing (30 items)	17	30	26.48	88.28%	3.27	-1.22	1.09
2. Clerical Grammar (30 items)	7	24	16.37	54.56%	4.60	-0.29	-0.89
3. Microsoft Excel (30 items)	1	21	10.62	35.40%	5.72	0.07	-1.06
4. Microsoft PowerPoint (20 items)	2	18	11.66	58.28%	4.56	-0.473	-0.90

n = 30

Table 3

Item Difficulty for the Microsoft Excel Pilot Test

Item	*M*	*SD*	Item	*M*	*SD*
1.	0.03	0.19	16.	0.34	0.48
2.	0.03	0.19	17.	0.34	0.48
3.	0.07	0.26	18.	0.34	0.48
4.	0.07	0.26	19.	0.38	0.49
5.	0.07	0.26	20.	0.38	0.49
6.	0.10	0.31	21.	0.38	0.49
7.	0.14	0.35	22.	0.45	0.51
8.	0.21	0.41	23.	0.52	0.51
9.	0.24	0.44	24.	0.52	0.51
10.	0.24	0.44	25.	0.59	0.50
11.	0.24	0.44	26.	0.66	0.48
12.	0.28	0.46	27.	0.69	0.47
13.	0.28	0.46	28.	0.76	0.44
14.	0.28	0.46	29.	0.79	0.41
15.	0.34	0.48	30.	0.93	0.26

n = 30

Participants

Mplus 7 was used to conduct a power analysis for the main study using a Monte Carlo simulation for stability of path coefficients for the hypothesized model relationships (Muthén & Muthén, 2002). The full Mplus code created for this power analysis can be found in Appendix C. In a meta-analysis examining effect size magnitudes published across I/O Psychology, Bosco and colleagues (2015) reported a 50th percentile effect size of 0.21 for relationships between performance and knowledge, skills, and abilities (Bosco et al., 2015). Given the widely reported strong and positive effects of GCA on performance-related processes (James & Carretta, 2002) and outcomes (e.g., Ree, Earles, and Teachout, 1994; Schmidt, 2002), a median effect was chosen as a conservative estimate of hypothesized direct effects from GCA and the use of effective methods. Results for this power analysis with a significance criterion of 0.05 and 80% power to detect all effects at anticipated effect sizes indicated that 340 participants were needed to detect the hypothesized effects. This was primarily driven by the hypothesized mediation effect, which was the most demanding in terms of sample size; it required 340 participants to reach 80% power, whereas all direct relationships exhibited greater power (between 91% - 97%). Thus, a sample of 400 participants was recruited for the study to account for listwise deletion of participant data where careless responding or inattention to instructions was observed, to be described later.

Participants were recruited from MTurk. They were given $3 for completing the study and were informed during the study that the top 25% of scorers when cheating would receive an additional $3 bonus to motivate Workers to try to achieve higher Time 2 UIT retest scores. The bonus amount was determined based on a recent study investigating the effects of MTurk compensation on indicators of data quality; effort, persistence, and satisfaction (Behrend, 2016).

In that study, a sample of 360 MTurk Workers received a payment of $.50, $1, or $2 for completing a 30-minute Time 1 survey and were invited to return for a follow-up Time 2 survey to receive either the same (100%) payment rate at Time 2 or increased Time 2 payment (by 200% or 400%). Completion time was not affected by Time 1 pay rates or the Time 2 increase, although Workers receiving Time 1 pay rates meeting accepted payment standards passed more attention checks, were more likely to return for the follow-up Time 2 survey, and reported higher pay satisfaction. No effects were found for Time 2 increase rates on data quality indicators; raising base pay by 200% or 400% did not affect Worker behavior. These results indicate that base pay is more salient to MTurk Workers and there are diminishing returns on increasing follow-up pay (Behrend, 2016). Thus, a 100% pay rate was chosen for the high score bonus, as increasing this further would likely not lead to increased effortful responding by Workers.

Measures

General Cognitive Ability (GCA). General cognitive ability was measured using Condon and Revelle's (2014) International Cognitive Ability Resource (ICAR) sample test of GCA. This test was developed in response to the need for a valid, reliable, secure, yet freely available measure of general cognitive ability for research. The ICAR sample test contains a 16-item subset of the full 60 item ICAR test. The shorter sample measure was chosen instead of the full measure to reduce participant time and cognitive resources spent on completing the GCA measure, while maintaining an effective representation of GCA. There are four subscales; verbal reasoning, letter and number series, matrix reasoning, and three-dimensional rotation. The authors report KR-20 reliability of 0.81 for the sample test in a sample of 96,958 individuals from 199 countries. IRT analyses for sample test items were similar to the full test in respect to the relationships between subscales and the spread of item difficulty across latent trait levels (Condon & Revelle, 2014). ICAR sample test scores correlate with self-reported achievement test scores at 0.59 for the SAT and 0.52 for the ACT (correlations corrected for reliability; Condon & Revelle, 2014). In a separate sample, ICAR sample test scores correlated with two Shipley-2 subscale scores, a commercial measure of cognitive functioning and impairment (Shipley, Gruber, Martin, & Klein, 2009, 2010), at 0.81 and 0.82 when corrected for range restriction and reliability (Condon & Revelle, 2014). The full ICAR sample test can be found in Appendix D. In the current study, the KR-20 estimate of reliability was strong (α = 0.80), and participants scored between 1 and 16 items correct (*m* = 8.72; 55% correct, *SD* = 3.60). Because the fit of individual ICAR items was not relevant to study hypotheses, and because the modeling of dichotomous indicators would introduce estimation challenges related to mediation testing, continuous subscale means were created and used as indicators of a latent general cognitive ability factor during modeling. This also enabled the use of more common fit statistics, such as the Standardized Root Mean Residual (SRMR) instead of Weighted Root Mean Square Residual (WRMR), which would have been necessitated by the inclusion of dichotomous items.

UIT (both baseline and after cheating): Microsoft Excel. The Microsoft Excel measure developed in the pilot study was completed by participants, first under honest instructions and retaken under cheating instructions. The KR-20 estimate of reliability was strong with honest instructions (α = 0.89) and with cheating instructions (α = 0.94). Participants scored between 0 and 27 items correct with honest instructions (*m* = 10.16 out of 29; 35% correct, *SD* = 6,29) and between 0 and 29 items correct with cheating instructions (*m* = 16.02; 55% correct, *SD* = 8.07).

Effective cheating methods. The scale developed in the cheating methods scale development study was used as a self-report measure of cheating methods. After the completion of the UIT after cheating instructions, participants were asked to report whether they used each cheating method ("yes" or "no"). Additionally, participants were given an option to describe a method they used outside the measure of cheating methods. They were required to select a minimum of one method (which could include the open-ended response option). They were also asked to estimate the percentage of time spent on each method they indicated using out of their total time spent attempting to raise their test score. Number and percent of participant responses for endorsing each method can be found in Table 4, along with the mean percentage of time used for that method, for those participants using that method. The effectiveness weight derived in the cheating methods scale development study and time percentage weights were used to create a measure of overall effectiveness of cheating methods. This mean score was finally multiplied by 100 to represent the score as a percentage, which was used in all modeling.

Table 4

Endorsement of use and percentage of time spent for each cheating method used

Method	*n endorsed*	*% endorsed*	*m % time**
1. Use electronic notes stored on devices during test pertaining to test content (e.g., computer, tablet, cellphone)	12	3.5%	60.17%
2. Purchase test content online (i.e., test cheating websites)	2	0.6%	18.50%
3. Search the Internet for test content	247	72.4%	85.70%
4. Pause test to search for answers online	88	25.8%	66.19%
5. Pause test to consult with another person	18	2.3%	32.06%
6. Inspect source code of test for indications of correct answers	1	0.3%	17.00%
7. Take the test once using a login created with false information, then re-take the test with your own login	2	0.6%	25.00%
8. Ask a contact within the organization or test company to request new assessment sessions to allow you to re-take the test	2	0.6%	100.00%
9. Ask a contact within the organization or test company to gain insight into the test content	1	0.3%	12.00%

(Table 4 continued)

Method	*n endorsed*	*% endorsed*	*m % time*
10. Solicit help (e.g., ask for help, advice, or answers) from someone physically present while you take the test	20	5.9%	49.42%
11. Solicit help (e.g., ask for help, advice, or answers) from someone over the phone while you take the test	5	1.5%	15.00%
12. Solicit help from someone via electronic messaging (e.g., email, text message, Google chat, Facebook messenger) while you take the test			
12a. by typing text about the test content	6	1.8%	34.00%
12b. by screenshotting test content	4	1.2%	40.75%
12c. by sending a picture of test content	1	0.3%	100.00%
12d. by copying and pasting test content	0	0.0%	--
13. Ask someone to take the test for you	6	1.8%	47.33%
14. Hire/pay someone to take the test for you	1	0.3%	100.00%
15. Receive answers from someone else completing the test at the same time	2	0.6%	64.00%

(Table 4 continued)

Method	*n endorsed*	*% endorsed*	*m % time*
16. Post test content on social media (e.g., Facebook, Twitter) to solicit help or answers from others	3	0.9%	11.00%
17. Post test content on an online discussion board (e.g., Yahoo Answers, Microsoft Answers, Turk Opticon) to solicit help or answers from others	0	0.0%	--
18. Use the program being tested (Excel) to determine the correct answer on the same computer you completed the test on	31	9.1%	33.55%
19. Use the program being tested (Excel) to determine the correct answer on a different computer	8	2.3%	33.00%
20. Search for test content in the Help section of the program being tested (Excel) on the same computer you completed the test on	28	8.2%	37.79%
21. Search for test content in the Help section of the program being tested (Excel) on a different computer	7	2.1%	57.71%

*Of the participants who endorsed using each method

Careless responding. Several methods of detecting careless responding were used, per the recommendations of Meade and Craig (2012). First, bogus items were included in the GCA measure**.** These were created by adding four exceptionally easy items, one for each section of the ICAR sample test, that appeared similar in style to items already on the ICAR. These items appear in Appendix E. Second, the total time spent completing each assessment was automatically recorded by the survey software to identify participants who spend very little time answering study items. Finally, each variable was regressed on each case number to check for the existence of outliers on any variable.

Demographics. Age, gender, ethnicity, job status, and experience taking UITs for employee selection were assessed.

Procedure

Participants who signed up for the main study first viewed an online notification statement, followed by the demographic items. Second, they completed both the GCA measure and the Microsoft Excel UIT honestly. Third, instructions were given to participants explaining they would retake the same knowledge test but would be encouraged to use any resources of their choosing to raise their scores. They were also informed that the top 25% of scorers on this test would receive a $3 bonus payment. These instructions can be found in Appendix F. During administration of the Microsoft Excel test (both honest and cheating), participants were limited to the time allotted for the original UIT to more closely reflect a typical employee selection testing situation. Finally, participants completed the effective cheating methods scale to report which cheating methods they used and the proportion of their time spent on each method.

CHAPTER 4

RESULTS

To identify careless responding, histograms of the number of bogus items endorsed, Mahalanobis distances, and time taken by each participant on study measures were examined, per the recommendations of Meade and Craig (2012). Based upon these analyses, individual cases were removed from the dataset for the following reasons: five participants failed all three bogus items, six respondents exhibited consistent responding on six or more consecutive items in a row, and three participants spent less than two minutes on the 29-item Excel test. Additionally, 37 individuals reported that they chose not to cheat on the second administration of the Excel test and were removed from the dataset. This left a total of 341 individuals completing the study.

Prior to hypothesis testing, the dataset was screened for missing data, univariate and multivariate outliers, and non-normality. Missing data was expected to be minimal due to the nature of online data collection, and no variables in the path model were missing. Outliers were investigated using scatterplots and boxplots of the data and measures of leverage, discrepancy, and influence. No extreme univariate or multivariate outliers were detected for any study variables. Non-normality was assessed by inspecting scatterplots of the data and examining skewness and kurtosis estimates. The cognitive ability test and both Excel test scores appeared normally distributed. The effective cheating methods variable did exhibit non-normality (skewness = -4.03, kurtosis = 17.477; see histogram in Appendix G); however, the variable was not transformed or centered because the maximum likelihood estimator used by Mplus is robust to significant departures from normality for endogenous variables, and the leptokurtic distribution was due to a large proportion of participants reporting the use of a particular method (i.e., "Search the Internet"); thus, this distribution appeared to be a natural reflection of the construct.

Investigation of study variables

Descriptive statistics and bivariate correlations were calculated for each study variable and can be found in Table 5. A CFA was conducted for the general cognitive ability measurement model and showed excellent fit according to the cutoff values recommended by Hu and Bentler (1999; $\chi^2(2) = 2.558$, $p = 0.278$, RMSEA = 0.029, CFI = 0.998, SRMR = 0.014). Standardized item loadings for each subscale can be found in Table 6. The other three variables, effective cheating methods and both Excel administrations, were modeled as single-item measures.

Table 5

Descriptive Statistics and Correlations for Study Variables

Variable	*m*	*SD*	1	2	3	4
1. General Cognitive Ability (16 items)	8.72	3.60	--			
2. Effective Cheating Methods	13.52	2.60	0.04	--		
3. Excel (honest instructions; 29 items)	10.16	6.29	0.46**	0.12*	--	
4. Excel (cheating instructions; 29 items)	16.02	8.07	0.50**	0.12*	0.82**	--

$n = 341$ *$p < 0.05$, **$p < 0.01$

Table 6

Item Loadings for General Cognitive Ability Measurement Model

Subscale	β	S.E.	*t*	*p*
Letter and Number Series	0.697	0.045	15.586	<0.000
Matrix Reasoning	0.645	0.046	13.892	<0.000
Three-Dimensional Rotation	0.522	0.051	10.275	<0.000
Verbal Reasoning	0.644	0.046	14.106	<0.000

n = 341

Hypothesis Testing

The hypothesized latent change model was tested using Mplus 7 (Muthén & Muthén, 2015), using bias corrected bootstrapping with 1,000 replications, as recommended by Preacher and Hayes (2008) for tests of direct and indirect mediation effects. Overall model fit is reported in order to examine how well the variances and covariances of the model are predicted by the theoretical relationships. Multiple global fit indices were employed; the chi-square fit index, Standardized Root Mean Residual (SRMR), Root Mean Square Error of Approximation (RMSEA), and Comparative Fit Index (CFI), using cutoff values recommended by Hu and Bentler (1999). These fit indices were chosen because of their common use in published empirical articles employing SEM, as well as the results of Monte Carlo simulations, such as Fan, Thompson, and Wang's (1999) study, indicating minimal influence from sample size and random variation.

First, a model was fit without the interaction between general cognitive ability and effective cheating methods in order to assess model fit using the global fit indices as well as interpret hypotheses tests for Hypotheses 1-5 and Hypothesis 7. Fit for this model was excellent ($\chi^2(12) = 18.943$, $p = 0.090$, RMSEA = 0.041, CFI = 0.991, SRMR = 0.029). The interaction was added to the model in order to test Hypothesis 6. Although the models with and without the interaction can be compared directly because they are nested, traditional fit statistics cannot be calculated for models using maximum likelihood estimation with robust standard errors (i.e., MLR), which is necessitated by the inclusion of a latent interaction term. The interaction was statistically significant, and improved model fit according to a chi square difference test of log likelihood values, using Satorra and Bentler's (2010) equation for difference testing using loglikelihood values (model with interaction loglikelihood H0 value = -4945.34, H0 scaling correction factor = 1.26, *df* = 24 versus model without interaction loglikelihood H0 value = -4949.22, H0 scaling correction factor = 1.36, *df* = 23 yielded a chi square value of 6.017, significant at $p = 0.02$). Thus, all hypotheses were tested by examining standardized path coefficients first for statistical significance and then for magnitude of effects within the full hypothesized latent change model, which can be found in Table 7 and Figure 3. Because of the novelty of the effects being tested, general guidelines for interpreting relationship strength are used to interpret effect sizes. Bosco and colleagues (2015) reported the following ranges of effect size benchmarks in their meta-analysis for relationships between performance and knowledge, skills, and abilities: small effects between .08 and .12, moderate effects between .13 and .30, and large effects .31 and greater (Bosco et al., 2015).

Table 7

Standardized Path Coefficients for Hypothesized Latent Change Model

	β	S.E.	*t*	*p*
Without Interaction				
Change on				
General Cognitive Ability (H7)	0.357	0.088	4.041	< 0.001
Baseline Knowledge (H1)	-0.073	0.081	-0.903	0.366
Effective Cheating Methods	0.030	0.074	0.397	0.691
Methods on				
General Cognitive Ability (H4)	0.069	0.060	1.161	0.246
Baseline on				
General Cognitive Ability (H3)	0.552	0.054	10.281	< 0.001
Mediation (H5)	0.009	0.027	0.318	0.751

(Table 7 continued)

	β	S.E.	*t*	*p*
With Interaction				
Change on				
General Cognitive Ability	-0.753	0.507	-1.485	0.138
Baseline Knowledge	-0.091	0.079	-1.153	0.249
Effective Cheating Methods (H2)	0.123	0.068	1.816	0.069
General Cognitive Ability x Effective Cheating Methods (H6)	0.214	0.095	2.245	0.025
Methods on				
General Cognitive Ability	0.071	0.058	1.217	0.224
Baseline on				
General Cognitive Ability	0.549	0.052	10.477	< 0.001
Mediation	0.009	0.018	0.480	0.632

n = 341

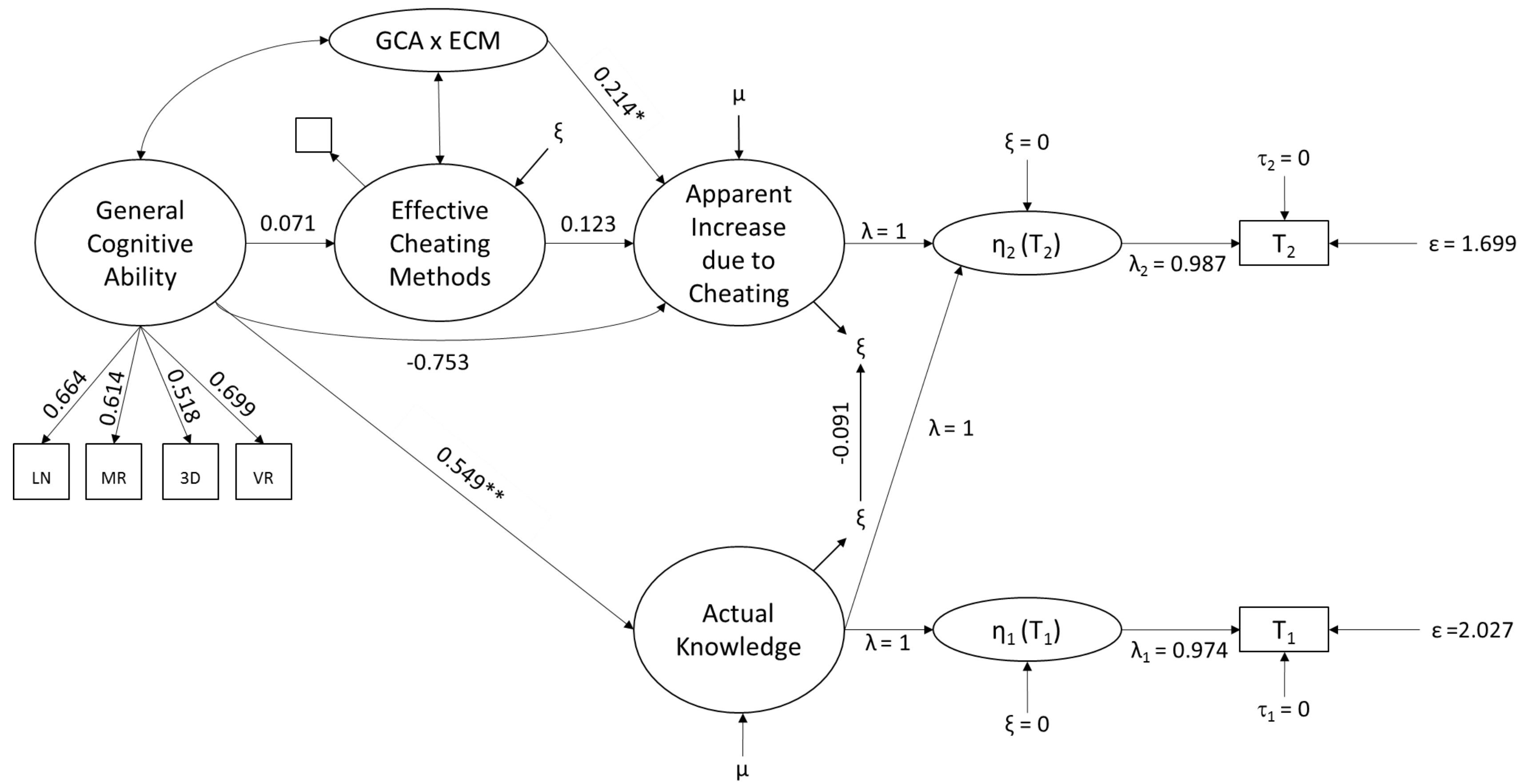


Figure 3. Standardized Path Coefficients for Full Hypothesized Latent Change Model with Interaction.

Hypothesis 1. Hypothesis 1 stated that baseline knowledge would predict apparent increase in knowledge due to cheating. In the interactive model, baseline knowledge did not statistically significantly predict apparent increase in knowledge due to cheating ($\beta = -0.091$, $p = 0.249$). Post hoc, another model was run containing only the latent change model in order to examine the relationship between baseline knowledge and increase without the controlling effects of general cognitive ability. When the influence of general cognitive ability was removed, the relationship did emerge; baseline knowledge statistically significantly predicted apparent increase in knowledge due to cheating ($\beta = 0.128$, $p = 0.037$). Thus, there was mixed support for this hypothesis; baseline knowledge does predict apparent increase, but this effect may be better explained by cognitive ability.

Hypothesis 2. Hypothesis 2 stated that effective cheating methods would predict apparent increase in knowledge due to cheating. This hypothesis was not supported; cheating methods did not directly predict apparent increase ($\beta = -0.123$, $p = 0.069$).

Hypothesis 3. Hypothesis 3 stated that general cognitive ability would directly predict baseline knowledge. This hypothesis was supported; general cognitive ability predicted baseline knowledge in the non-interactive model ($\beta = 0.552$, $p < 0.001$).

Hypothesis 4. Hypothesis 4 stated that general cognitive ability would predict the use of effective cheating methods. This hypothesis was not supported; general cognitive ability did not directly predict the use of effective cheating methods ($\beta = 0.069$, $p = 0.246$).

Hypothesis 5. Hypothesis 5 stated that the relationship between general cognitive ability and test score increase would be mediated by use of effective cheating methods. This hypothesis was not supported; the bias-corrected bootstrapped confidence interval around the indirect effect estimate contained zero, indicating the indirect effect was not statistically significant ($\beta = 0.038$, $p = 0.363$, 95% BS CI [-0.014, - 0.086]).

Hypothesis 6. Hypothesis 6 stated that general cognitive ability would moderate the relationship between the use of effective methods and apparent increase in knowledge due to cheating. This hypothesis was supported; the inclusion of the interaction term led to a statistically significant change in model fit, and the statistically significant interaction term exhibited a moderate effect ($\beta = 0.214$, $p = 0.025$). To understand the nature of this interaction, simple slopes were calculated to test whether each slope differs from zero (Cohen, Cohen, West, & Aiken, 2003). The simple slope for individuals 1 SD below the mean of GCA was not statistically significant ($\beta = -0.151$, $p = 0.349$), but the standardized simple slope for individuals 1 SD above the mean of GCA was positive and statistically significant ($\beta = 0.563$, $p = 0.018$). A graph of this interaction can be found in Figure 4, and depicts change in test scores, when controlling for baseline scores. Individuals lower in general cognitive ability are better able to raise their test scores because their baseline test scores were lower to begin with, but using more effective cheating methods does little to increase their scores. Conversely, individuals higher in general cognitive ability are less able to raise their test scores because of their high baseline scores, but the more effective cheating methods used, the higher they raise their test scores when they are cheating. The presence of this interaction also makes the coefficients tested for Hypotheses 2 and 7 more difficult to interpret.

Hypothesis 7. Finally, Hypothesis 7 stated that general cognitive ability would directly predict apparent increase in knowledge due to cheating. This hypothesis supported; general cognitive ability exhibited a strong positive relationship with apparent increase in knowledge due to cheating in the model without the interaction ($\beta = 0.357$, $p < 0.001$).

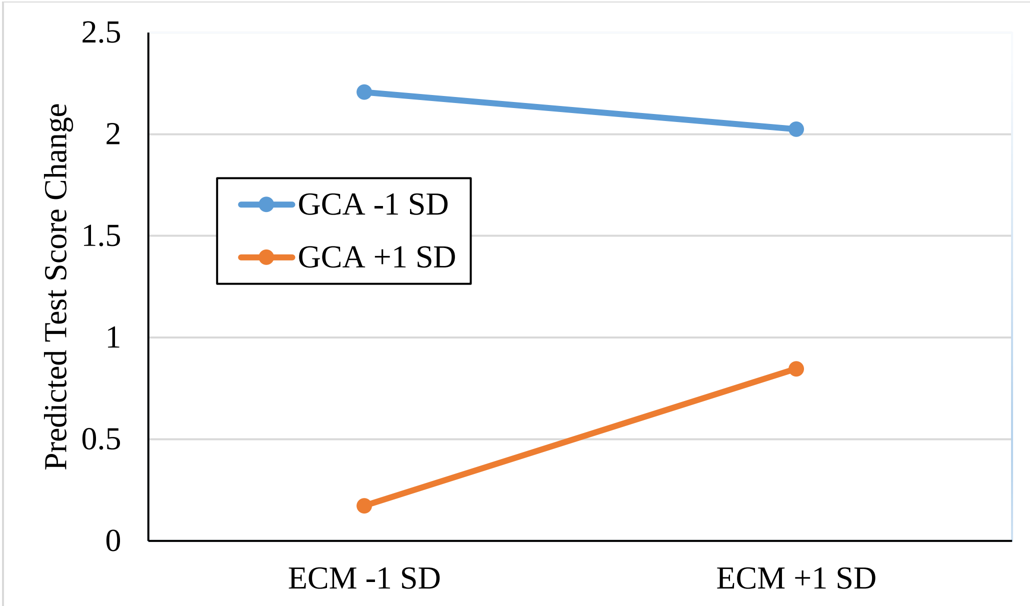


Figure 4. Interaction between General Cognitive Ability and Effective Cheating Methods to Predict Test Score Change due to Cheating.

CHAPTER 5

DISCUSSION

The present study investigated a model explaining latent test score change due to cheating and makes important contributions to our theoretical understanding of the methods used to cheat on UITs and the role of general cognitive ability in effective cheating. This study also addressed several open questions relevant to practitioners using or considering the use of UITs for hiring. Its three most important contributions are described below.

First, a wide variety of methods can be used by individuals attempting to cheat, which involves the identification and use of online informational resources, online or collocated person resources, or physical resources. Inspecting source code and soliciting help from employees in the assessor organization were viewed as least effective by SMEs; safeguards placed on technological and personnel resources were believed to be effective in reducing the potential for score change using these methods. The most effective methods for effective cheating according to SMEs were purchasing test content online and soliciting help from another person physically present while taking a test.

Second, general cognitive ability plays a key role in cheating effectiveness such that among people who cheat, those with higher cognitive ability will be better able to cheat if they identify effective methods for doing so. Although general cognitive ability neither predicted the use of effective cheating methods nor suggested effective cheating methods as a mediator of its effect, it did predict baseline knowledge and apparent increase in knowledge. Therefore, previous research on the importance of cognitive ability as a predictor of skill test scores was replicated, but furthermore, cognitive ability was shown as a likely cause of greater increases in skill test scores by cheating. This supports previous research showing that general cognitive ability is necessary for performance on cognitive tests and complex problem solving (e.g., Condon & Revelle, 2014; Stadler, et al., 2015), and the cognitive process involved in cheating could be reasonably considered a type of problem solving. Since individuals higher in general cognitive ability were not more likely to choose more effective cheating methods than individuals lower in cognitive ability, unmeasured knowledge factors, such as previous experience cheating on tests, might better explain the choice of cheating methods used. Those higher in cognitive ability were still better able to use effective cheating methods to increase their test scores (i.e., the effective use of those methods depends upon cognitive ability). Those lower in cognitive ability were unable to take advantage of cheating resources as well as those higher in cognitive ability, even though those resources were generally believed to be effective by experts. Instructing participants to cheat was not sufficient to raise test scores; while all participants attempted to cheat, only some participants had the cognitive resources to effectively cheat using the methods they decided to adopt.

Third, neither pre-test knowledge of test material nor the use of effective cheating methods directly predicted apparent test scores within the context of the full hypothesized model. Individuals using more effective cheating methods were not all able to increase their scores more greatly by cheating than those using ineffective cheating methods. From these results, it appears that among people who are already cheating, learning and employing more generally effective cheating methods will not necessarily be effective; other explanatory antecedents, including but likely not limited to general cognitive ability, must be needed to effectively raise test scores. Similarly, when controlling for general cognitive ability, existing knowledge of the test subject did not affect how well a person could cheat. One potential explanation for this is that there may be a test or item difficulty threshold across which cheating is differentially effective; for example, easier items may be easier to cheat than difficult items. In the present study, internet searchers were a common cheating behavior, and individuals with lower baseline knowledge or skill who missed both easy and difficult items initially might have been able to more easily find the answers to those easier items. An individual with greater baseline knowledge might have only missed difficult items, which would be more difficult to cheat on. Thus, the relationship between item difficulty and cheating effectiveness emerged as a priority for future research.

Additionally, due to the theorized effect of general cognitive ability on both baseline knowledge and cheating effectiveness, the relationship between baseline knowledge and apparent score increase was tested without the controlling effects of general cognitive ability post hoc. Because control variables are an aggressive approach to removing construct-irrelevant covariance in a model, effectively assuming that all variance explained by other variables in the model is irrelevant, Spector and Brannick (2007) advised testing models both with and without those variables. Thus, this analysis served as an empirical probe into alternative explanations for the lack of the hypothesized effect. Without controlling for general cognitive ability, baseline knowledge did predict latent test score increase among people cheating, so greater pre-existing skills or knowledge did predict cheating effectiveness. This finding is supported by training literature theory that individuals with more experience and knowledge have superior information processing in that subject area and are able to learn more from training (Salas & Rosen, 2010); however, it is unclear if the bias introduced by omitting general cognitive ability or the bias introduced by overcontrolling for general cognitive ability is a better reflection of the true relationship between these three constructs. Thus, this also emerged as a key area for future research.

Contributions to UIT Theory

Existing cheating research provided some insight into what methods might be used in a UIT selection context, but there was no empirical evidence exploring this, nor the efficacy of those methods to increase test scores (Tippins et al., 2006). This gap was filled by revising an existing list of cheating methods for use in a UIT context and rating those methods for effectiveness, completed by a sample of SMEs in employee selection tests and/or cheating methods. Results of this effort produced a comprehensive taxonomy of the methods used by applicants to cheat on UITs for selection, and the perceived effectiveness of those methods for raising test scores. The methods identified in the present study aligned with methods cited as potential cheating behaviors in the literature (i.e., receiving information from others and using "forbidden" resources (p. 210; Tippins et al., 2006)), and extended them further with additional specific details about how individuals cheating elicit or gain access to these resources.

Previous studies estimated the prevalence of cheating on UITs in selection contexts (Arthur et al., 2010; Do et al., 2005; Nye et al., 2008; and Wright et al., 2015) by comparing test scores from situations in which the presence of cheating is unlikely with situations in which the presence of cheating is likely (e.g., non-applicant versus applicant samples). The results of the present study challenge that measurement of cheating; people differ in their ability to cheat when they are trying to cheat and only effective cheating would be detected by this method, as supported by both an overall moderation effect and analysis of simple slopes. Score change should not be used alone to detect cheating if the purpose of the investigation is to detect the presence of ***all*** cheating, not only effective cheating.

Because only effective cheaters were detected in these studies, and given the range of ability and skill levels in selection pools, individuals who cheat on selection tests could score in any range of test scores. These studies may have accurately detected effective cheating, yet concluded that the presence of all cheating is quite low. This distinction between cheating behaviors (i.e., performance) and cheating effectiveness is similar to the distinction made by Campbell and colleagues (1993) in defining job performance. Employees may exhibit the right behaviors, methods, or actions but they may or may not translate into effective results. Similarly, cheating may be thought of as a task in which performance and effectiveness differ in theoretically meaningful ways. Participants may have used effective methods, but the use of those methods only translated into the desired outcome for some of them.

Although the existence of effective and willful cheating is low in the selection contexts in which it has been studied; ineffective and willful cheating likely occurs at a much more frequent rate but cannot be detected by examining differences between unproctored and proctored test scores alone. Practitioners might only be concerned with detecting effective cheaters if a UIT is used as one hurdle a selection process, assuming that ineffective cheaters will be dropped due to lower test scores in later proctored hurdles. However, the presence of cheating in any test score, regardless of where that score falls in the sample distribution would likely impact both construct and predictive validity of those test scores, a critical concern for any test used for selection. This is explored further in the following section.

Practical Implications

The results of this experiment show that some individuals instructed to cheat, depending upon their ability and the methods they use, can impact their test scores. This study empirically demonstrated that willful cheating can increase test scores, so cheating and potential types of cheating should be considered when designing and evaluating tests within a selection system. However, because of the moderating effect of cognitive ability, applicants do not increase their test scores consistently. Controlling for baseline scores, those test takers cheating more effectively (as measured by test score change) are likely to be of lower cognitive ability. In a hiring context, volition must also be considered. Intuitively, individuals lower in cognitive ability are probably more likely to choose to cheat. Because using more effective methods benefits individuals lower in cognitive ability less so than individuals higher in cognitive ability, whose baseline test scores are already high, it is possible that the real-world impact of cheating on test scores is smaller than was found in the context of the current study. Thus, the costs and utility of cheating mitigation should be considered carefully. Although there is significant research on the discouragement of faking on non-cognitive measures, the extant literature on cheating detection and prevention is sparse. Timed tests are frequently suggested and used to decrease cheating opportunities within a test (Arthur et al., 2010; Lievens & Burke, 2011; Nye et al., 2008), though participants in the current study were able to increase their scores regardless of the same time restrictions being used as when this test is used in authentic selection contexts. Further, participants did so overwhelmingly by searching the internet because of the widely and freely available online information about the test content. Identifying which cheating methods could be used and which would be effective for a given test are the first considerations in ascertaining how to prevent or minimize the impacts of cheating. If organizational leadership is worried that hiring applicants with the highest scores on selection tests may lead to hiring more dishonest individuals, covert or personality based integrity measures could be used in tandem with such tests to attempt to identify these individuals.

The present study also demonstrates that cheating can impact the construct validity of a test. If applicants cheat on a UIT, test scores contain construct information related to the methods they adopted. For example, test scores likely contain variance related to information retrieval skills among those using a search-the-Internet strategy to cheat. The exact combination of each test score in terms of information retrieval skills versus pre-existing domain knowledge or skill would be unique to each test taker, depending on how much they cheated versus used their own knowledge. If information retrieval skills are an important aspect of the job, they should be measured directly, not by encouraging applicants to cheat but by designing a test specifically for measuring information retrieval skills. If an organization hires based on unproctored test scores alone without considering or addressing cheating, the test might reward undesirable and non-job related individual differences. For example, if individuals lower in integrity are more likely to cheat and were able to increase their scores, the organization could hire more individuals with low integrity, and over time the organization could become more populated with lower integrity employees. This would be undesirable as integrity is associated with outcomes such as job performance and counterproductive work behaviors (Ones, Viswesvaran, & Schmidt, 1993). Further, the organization could reject honest individuals with similar levels of skills or knowledge but were unable to score as high as applicants who cheated.

Criterion-related validity of a test may be altered by the presence of cheating; no research that I could identify examines this in the cheating literature, but research in the faking literature provides insight into two possibilities. Faking on personality tests in a selection context may introduce more error into those tests, reducing predictive validity as it disrupts the rank ordering of true scores (Rosse et al., 1998; Smith & Ellingson, 2002). It is possible that the same process exists with cheating; true scores are altered and predictive validity is reduced. Alternatively, other research shows that the ability to fake on personality items in a selection context may reflect similar self-presentation abilities in a work environment. Individuals who are better able to fake on pre-hire personality tests can also exhibit similar socially desirable behavior later on the job (Ones et al., 1996), and that faking can actually augment the criterion-related validity of conscientiousness (Hough et al., 1990; Komar, Brown, Komar, & Robie, 2008). It is possible that this same process exists with cheating; individuals better able to cheat on tests will be better able to locate relevant outside resources and use them to solve new problems on the job.

Limitations and Future Directions

The present study examined several open questions related to the antecedents of test score increase due to cheating, including methods used to cheat. The experimental methodology employed, in which all participants were asked to cheat and attempted to raise their test score by cheating, was critical for maximizing the chances that change scores could be directly attributable to cheating. However, this methodology precluded the investigation of some other questions relevant to cheating on UITs in selection contexts, outside the scope of the current study.

There are three specific limitations attributable to choice of methodology. First, the methodology employed could not perfectly represent a selection context, as it is unlikely an organization would allow their applicants to be specifically be instructed to cheat, and then those test scores used to select employees. Deviations from an actual selection context were made in the testing process, the sample used, and the outcomes of testing. Test-takers were asked to re-take the same test a second time and cheat on it to increase their score, which is also a deviation from typical selection procedures – normally candidates cannot re-take a selection test to increase their score, although there are notable exceptions (e.g., Landers, Sackett & Tuzinksi, 2011). By asking participants to re-take the test, the additional administration may have increased test fatigue which may have reduced the effects of cheating. Second, the data collected was collected from mTurk Workers, not a sample of job applicants. Previous research shows that mTurk Workers are likely more similar to a sample of job applicants compared to most undergraduate samples in terms of age and job experience (e.g., Behrend et al., 2011), but an mTurk Worker sample has not yet been directly compared to a sample of job applicants using these measures and in this particular study context. Thus, although this maximized internal validity, it is unknown whether or to what extent a job applicant sample differs from an mTurk sample in terms of their motivation or ability to cheat. To combat this, the monetary reward was used to motivate participants as similarly to a job application context as possible given limited resources. Third, there was no investigation of outcomes past the immediate impact on test scores, which I/O practitioners must consider carefully in real-world selection contexts.

Beyond methodological limitations, it is important to note that this study examined the capacity of individuals to cheat, not their volition. This leads to three specific cautions regarding the generalizability of these results. First, all participants were instructed to cheat and actively attempted to cheat, thus, this study did not investigate which individuals would be more likely to choose to cheat on a UIT in a selection context, nor the impact of previous cheating on present cheating effectiveness. Second, this study did not examine the impact of test taker motivation, which is likely a strong influence on cheating behavior decisions in a selection context (e.g., Hough et al., 1990), because it was experimentally controlled. Efforts were taken to motivate participants to cheat to the best of their ability using a monetary reward and clear instructions during the cheating condition, and there was little motivation for participants to cheat during the control condition. Yet, because there was no job at stake for participants, lower levels of motivation in this context may have reduced the effects of cheating on test scores, attenuating observed effects in relation to those among real world cheaters. Third, this study did not investigate systematic differences among individuals experienced in cheating and not experienced in cheating regarding the choice of cheating method used or the effectiveness in using those methods. Given that previous experience can lead to better test score outcomes (e.g., Brown, 2001), cheating experience and effectiveness may be related, so experience may impact UIT scores more directly in selection contexts than general cognitive ability. Experienced and effective cheaters may be a greater organizational threat than other types of cheaters.

In addition to motivation and general cognitive ability, other individual differences may impact which applicants cheat and which are successful at doing so. Individuals higher in integrity may be less likely to cheat when given the opportunity to do so, as they tend to be generally more honest and trustworthy (Ones, Viswesvaran, & Schmidt, 1993). Individual differences in information retrieval skills (i.e., recognizing information needs and effectively identifying and evaluating new information in order to answer a question or solve a problem; Bruce, 1999) may be an additional ability-related moderator of cheating success. Future investigation into the impact of these individual difference antecedents on cheating is warranted, given their potential impact on cheating and cheating success and that they can be measured during a selection process.

Because the present study was the first investigation of this type into cheating methods on UITs, its design was limited to the most fundamental questions surrounding them in a design carefully constructed to maximize internal validity. First, the SME effectiveness ratings used to develop the cheating methods scale were slightly below conventional reliability standards. It appeared SMEs were consistent for some methods but ranged widely for others, and the reasons for this pattern are unclear. Effectiveness rating variation may have stemmed from limited directions guiding ratings or differing mental models of unproctored tests. Future work should focus on investigating cheating on specific types of tests and include qualitative feedback regarding the situations in which these methods would be effective. The UIT used in this study is currently being used to select job applicants, providing the means in which to realistically investigate test score change due to cheating. However, the Microsoft Excel test used here is just one of many types of test formats and subject areas tested for employee selection. Microsoft Excel is a test subject in which free online resources are plentiful, and most participants did search the internet for test content. Test takers might use different methods to cheat on tests measuring knowledge or skills in other areas; for example, fewer test takers would effectively search the internet when cheating on a test for which there are few or no freely available online resources. The list of selection methods and their corresponding effectiveness weights was created to broadly apply to any UIT, but differences in the rate of use may differ among participants based on characteristics of the test, including format and subject area. Additionally, the participants in this study were not given advanced notice of test content or topic, as is common in applied settings. If participants knew they would be completing this particular test with higher stakes, they might have used cheating methods differently and might even have been able to increase their scores more successfully. Finally, the most widely used method in this study was searching the internet and there are a myriad of ways test-takers could be doing this, which in turn could impact that method's effectiveness for raising test scores. Cheating behaviors have also never been directly measured in a UIT selection context, instead, researchers have relied on proxies for these behaviors, such as test score change (e.g., Arthur et al., 2010) or post-test self-report in the current study. Measuring these behaviors directly (e.g., implementing remote proctoring methods that simply measure but do not diminish the behaviors) could provide useful details in how methods are used and represent an important future direction.

The effectiveness of methods used to prevent or detect cheating was not specifically addressed by the current study to provide all participants with a consistent platform in which to attempt to raise their test scores by cheating. The Microsoft Excel test was administered during the study with the same time limit used for selection to more realistically represent a testing scenario. Time limits are implemented because they are thought to minimize cheating opportunities (e.g., Arthur et al., 2010). Although most participants in this study finished within the time limit, and spent less time on the second test administration while cheating, it is possible that the time limit did reduce some participants' ability to cheat. The efficacy of the time limit to reduce cheating was not directly investigated, nor were other methods of cheating prevention or detection used, nor could a comparison group without any prevention methods be used. Future research is needed to address these open questions which remain largely unanswered.

Conclusion

This study was the first to propose and test a theory of individual differences and behavior to investigate the impact of cheating on test scores. This study extended previous research by investigating the methods people use to cheat, the effectiveness of those methods, and both the antecedents and impact of cheating on test scores. In summary, the presence of cheating does impact test scores, and knowledge of effective cheating methods enables high-ability cheaters to increase their scores even further. Given these findings, selection decisions made based on UIT scores in the presence of cheating will favor who are better at cheating over other test takers. Because cheating will likely never be completely preventable in UIT, organizations must consider both discouragement and detection of cheating, as well as the ultimately validity and ethical implications of hiring decisions made wherever cheating is possible at individual and organizational levels.

REFERENCES

Acton, G. S., & Schroeder, D. H. (2001). Sensory discrimination as related to general intelligence. *Intelligence, 29*(3), 263-271. doi: 10.1016/S0160-2896(01)00066-6

Anderson, C. D., Warner, J. L., & Spencer, C. C. (1984). Inflation bias in self-assessment examinations: Implications for valid employee selection. *Journal of Applied Psychology*, *69*(4), 574–580. doi: 10.1037/0021-9010.69.4.574

Austin, E. J., Hofer, S. M., Deary, I. J., & Eber, H. W. (2000). Interactions between intelligence and personality: Results from two large samples.*Personality and Individual Differences*, *29*(3), 405-427. doi: 10.1016/S0191-8869(99)00202-0

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. doi: 10.1111/j.1468-2389.2010.00476.x

Barger, T., Behrend, T.S., Sharek, D. J., & Sinar, E. F. (2011). I-O and the crowd: Frequently asked questions about using Mechanical Turk for research. *The Industrial-Organizational Psychologist, 49*(2), 11-18.

Behrend, T.S., (2016, May). The responsible use of crowdsourcing in behavioral research. Paper presented at the 27th Annual Convention for the Association for Psychological Science, New York, NY.

Behrend, T. S., Sharek, D. S., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavior Research Methods, 43,* 800-813. doi: 10.3758/s13428-011-0081-0

Bellarosa, C., & Chen, P. Y. (1997). The effectiveness and practicality of occupational stress management interventions: A survey of subject matter expert opinions. *Journal of Occupational Health Psychology, 2*(3), 247-262. doi: 10.1037/1076-8998.2.3.247

Bing, M. N., Whanger, J. C., Davison, H. K., & VanHook, J. B. (2004). Incremental validity of the frame-of-reference effect in personality scale scores: a replication and extension. *Journal of Applied Psychology, 89*(1), 150-157. doi: 10.1037/0021-9010.89.1.150

Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T., & Smith, M. A. (2006). A Meta-Analytic Investigation of Job Applicant Faking on Personality Measures. *International Journal of Selection and Assessment, 14*(4), 317-335. doi: 10.1111/j.1468-2389.2006.00354.x

Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. *Journal of Applied Psychology*, *100*(2), 431-449. doi: 10.1037/a0038047

Brown, K. G. (2001). Using computers to deliver training: Which employees learn and why? *Personnel Psychology*, *54*, 271-296. doi: 10.1111/j.1744-6570.2001.tb00093.x

Bruce, C. S. (1999). Workplace experiences of information literacy. *International Journal of Information Management*, *19*(1), 33-47. doi: 10.1016/S0268-4012(98)00045-0

Calisir, F., Eryazici, M., & Lehto, M. R. (2008). The effects of text structure and prior knowledge of the learner on computer-based learning. *Computers in Human Behavior, 24*(2), 439-450. doi: 10.1016/j.chb.2007.01.032

Campbell, J. P., McCloy, R. A., Oppler, S. H., & Sager, C. E. (1993). A theory of performance. N. Schmitt & W.C. Borman (Eds.), *Personnel Selection in Organizations* (pp. 35-

70).Carroll, S. J., Paine, F. T., & Ivancevich, J. J. (1972). The relative effectiveness of training methods—expert opinion and research. *Personnel Psychology*, *25*(3), 495-509. doi: 10.1111/j.1744-6570.1972.tb00833.x

Cavanagh, T. M. (2014). *Cheating on online assessment tests: Prevelance and impact on validity* (Doctoral Dissertation). Retrieved from ProQuest Dissertations & Theses Full Text. (1615129366)

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(2-3*), 113-120. doi: 10.1111/1468-2389.00234

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). Applied multiple regression/correlation analysis for the behavioral sciences, 3rd ed. Hillsdale: Erlbaum.

Cole, F., Sanik, K., DeCarlo, D., Finkelstein, A., Funkhouser, T., Rusinkiewicz, S., Singh, M. (2009). How well do line drawings depict shape? *ACM Transactions on Graph*, *28,* 1-9, doi: 10.1145/1531326.1531334

Colquitt, J. A., LePine, J. A., & Noe, R. A. (2000). Toward an integrative theory of training motivation: A meta-analytic path analysis of 20 years of research. *Journal of Applied Psychology*, *85*(5), 678-707. doi: 10.1037/0021-9010.85.5.678

Condon, D. M., & Revelle, W. (2014). The International Cognitive Ability Resource: Development and initial validation of a public-domain measure. *Intelligence, 43*, 52-64. doi: 10.1016/j.intell.2014.01.004

Cook, T. D., Campbell, D. T., & Peracchio L. (1990). Quasi experimentation. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of Industrial and Organizational psychology, Vol. 1 (2nd ed.)* (pp. 491-576). Palo Alto, CA, US: Consulting Psychologists Press.

Coyne, I., & Bartram, D. (2006). Design and development of the ITC guidelines on computer-based and internet-delivered testing. *International Journal of Testing, 6*(2), 133-142. doi: 10.1207/s15327574ijt0602_3

Do, B. R., Shepherd, W. J., & Drasgow, F. (2005, April). *Measurement equivalence across proctored and unproctored administration modes of web-based measures*. Paper presented at the 20th Annual Conference of the Society for Industrial and Organizational Psychologists, Los Angeles, CA.

Ellingson, J. E., & McFarland, L. A. (2011). Understanding faking behavior through the lens of motivation: An application of VIE theory. *Human Performance*, *24*(4), 322-337. doi: 10.1080/08959285.2011.597477

Fallaw, S. S., Solomonson, A. L., & McClelland, L. (2009, April). Current trends in assessment use: A multi-organizational survey. *Poster presented at the 24th annual conference of the Society for Industrial and Organizational Psychology, New Orleans, LA*.

Fan, X., Thompson, B., & Wang, L. (1999). Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 56-83. doi: 10.1080/10705519909540119

Fetzer, M., & Grelle, D. (2010). PreVisor ConVerge: The best practice for unproctored/unsupervised internet testing. *Alpharetta, GA: White paper, SHL*.

Franklyn-Stokes, A., & Newstead, S. E. (1995). Undergraduate cheating: Who does what and why? *Studies in Higher Education*, *20*, 159-172. doi: 10.1080/03075079512331381673

Furnham, A. (1986). Response bias, social desirability and dissimulation. *Personality and Individual Differences*, *7*(3), 385-400. doi: 10.1016/0191-8869(86)90014-0

Gallant, T. B., & Drinan, P. (2006). Organizational theory and student cheating: Explanation, responses, and strategies. *The Journal of Higher Education*, *77*(5), 839-860. doi: 10.1353/jhe.2006.0041

Gaskill, M. C. (2014). *Cheating in business online learning: Exploring students' motivation, current practices and possible solutions.* (Doctoral Dissertation). Retrieved from ProQuest Dissertations & Theses Full Text.

Gottfredson, L. S. (1997). Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. *Intelligence, 24,* 13-23. doi: 10.1016/S0160-2896(97)90011-8

Graham, M. A., Monday, J., O'Brien, K., & Steffen, S. (1994). Cheating at small colleges: An examination of student and faculty attitudes and behaviors. *Journal of College Student Development, 35*, 255-260.

Grubb III, W. L., & McDaniel, M. A. (2007). The fakability of Bar-On's emotional quotient inventory short form: Catch me if you can. *Human Performance*, *20*(1), 43-59. doi: 10.1080/08959280709336928

Guo, J., & Drasgow, F. (2010). Identifying cheating on unproctored internet tests: The Z-test and the likelihood ratio test. *International Journal of Selection and Assessment, 18(4),* 351-364. doi: 10.1111/j.1468-2389.2010.00518.x

Haberecht, K., Schnuerer, I., Gaertner, B., John, U., & Freyer-Adam, J. (2015). The stability of social desirability: A latent change analysis. *Journal of Personality, 83*(4), 404-412. doi: 10.1111/jopy.12112

Hannafin, R. D., & Sullivan, H. J. (1996). Preferences and learner control over amount of instruction. *Journal of Educational Psychology, 88,* 162-173. doi: 10.1037/0022-0663.88.1.162

Hirschfeld, R. R., Lawson, L., & Mossholder, K. W. (2004). Moderators of the relationship between cognitive ability and performance: General versus context–specific achievement motivation. *Journal of Applied Social Psychology*, *34*(11), 2389-2409. doi: 10.1111/j.1559-1816.2004.tb01983.x

Holden, R. R., & Hibbs, N. (1995). Incremental validity of response latencies for detecting fakers on a personality test. *Journal of Research in Personality*, *29*(3), 362-372. doi: 10.1006/jrpe.1995.1021

Hough, L. M., Eaton, N. K., Dunnette, M. D., Kamp, J. D., & McCloy, R. A. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities. *Journal of Applied Psychology, 75*(5), 581-595. doi: 10.1037/0021-9010.75.5.581

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*, 1-55. doi: 10.1080/10705519909540118

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological bulletin*, *96*(1), 72. doi: 10.1037/0021-9010.79.2.184

James, M., & Carretta, T. R. (2002). g2k. *Human Performance, 15*(1-2), 3-23. 10.1080/08959285.2002.9668081

Karim, M., Kaminsky, S., & Behrend, T. (2014). Cheating, reactions, and performance in remotely proctored testing: An exploratory experimental study. *Journal of Business & Psychology, 29*(4), 555-572. doi: 10.1007/s10869-014-9343-z

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval Practice Produces More Learning than Elaborative Studying with Concept Mapping. *Science, 331*(6018). doi: 10.1126/science.1199327

Kleemann, F., Voß, G. G., & Rieder, K. (2008). Un(der)paid innovators: The commercial utilization of consumer work through crowdsourcing. *Science, Technology & Innovation Studies*, *4*, 5-26.

Komar, S., Brown, D. J., Komar, J. A., & Robie, C. (2008). Faking and the validity of conscientiousness: A Monte Carlo investigation. *Journal of Applied Psychology*, *93*(1), 140-154. doi: 10.1037/0021-9010.93.1.140

Landers, R. N., Sackett, P. R., & Tuzinski, K. A. (2011). Retesting after initial failure, coaching rumors, and warnings against faking in online personality measures for selection. *Journal of Applied Psychology, 96*(1), 202-210. doi: 10.1037/a0020375

Landy, F. J. (1993). Early influences on the development of industrial/organizational psychology. In T. K. Fagan & G. R. VandenBos (Eds.), *Exploring applied psychology: Origins and critical analyses* (pp. 79-118). Washington, DC, US: American Psychological Association.

Lawler, E. E., & Suttle, J. L. (1973). Expectancy theory and job behavior. *Organizational Behavior and Human Performance, 9*(3), 482-503. doi: 10.1016/0030-5073(73)90066-4

Locke, E. A. (1978). The ubiquity of the technique of goal setting in theories of and approaches to employee motivation. *Academy of Management Review*, *3*(3), 594-601. doi: 10.5465/AMR.1978.4305786

Li, W. D., Fay, D., Frese, M., Harms, P. D., & Gao, X. Y. (2014). Reciprocal relationship between proactive personality and work characteristics: A latent change score approach. *Journal of Applied Psychology*, *99*(5), 948. doi: 10.1037/a0036169

Lievens, F., & Burke, E. (2011). Dealing with the threats inherent in unproctored Internet testing of cognitive ability: Results from a large-scale operational test program. *Journal of Occupational and Organizational Psychology*, *84(4*), 817-824. doi: 10.1348/096317910x522672

Lievens, F., Buyse, T., & Sackett, P. R. (2005). Retest effects in operational selection settings: Development and test of a framework. *Personnel Psychology, 58*, 981–1007. doi: 10.1111/j.1744-6570.2005.00713.x

Lievens, F., & Harris, M.M. (2003). Research on internet recruiting and testing: Current status and future directions. In C.L. Cooper & I.T. Robertson (Eds.), *International Review of Industrial and Organizational Psychology*, vol. 16. (pp. 131-165). Chicester: John Wiley & Sons, Ltd.

Lievens, F., Reeve, C. L., & Heggestad, E. D. (2007). An examination of psychometric bias due to retesting on cognitive ability tests in selection settings. *Journal of Applied Psychology*, *92*(6), 1672. doi: 10.1037/0021-9010.92.6.1672

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective*. New York, N.Y.: Psychology Press.

McArdle, J. J., & Nesselroade, J. R. (2014). *Longitudinal data analysis using structural equation models*. Washington, DC: American Psychological Association. doi: 10.1037/14440-000

McCabe, D. L., & Trevino, L. K. (1993). Academic dishonesty: Honor codes and other contextual influences. *Journal of Higher Education*, 522-538. doi: 10.2307/2959991

McCabe, D. L., Treviño, L. K., & Butterfield, K. D. (2001). Cheating in academic institutions: A decade of research. *Ethics &Behavior*, *11*(3), 219-232. doi: 0.1207/S15327019EB1103_2

McFarland, L. A., & Ryan, A. M. (2000). Variance in faking across noncognitive measures. *Journal of Applied Psychology, 85*(5), 812-821. doi: 10.1037/0021-9010.85.5.812

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence*, *37*(1), 1-10. doi: 10.1016/j.intell.2008.08.004

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437. doi: 10.1037/a0028085

Mersman, J. L., & Shultz, K. S. (1998). Individual differences in the ability to fake on personality measures. *Personality and Individual Differences*, *24*(2), 217-227. doi: 10.1016/S0191-8869(97)00160-8

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, *9*(4), 599-620. doi: 10.1207/S15328007SEM0904_8

Muthén, L. K., & Muthén, B. O. (2015). *Mplus: Statistical analysis with latent variables: User's guide*. Los Angeles: Muthén & Muthén.

Naglieri, J. A., Drasgow, F., Schmit, M., Handler, L., Prifitera, A., Margolis, A., & Velasquez, R. (2004). Psychological Testing on the Internet: New Problems, Old Issues. *American Psychologist*, *59(3*), 150-162. doi: 10.1037/0003-066x.59.3.150

Newstead, S. E., Franklyn-Stokes, A., & Armstead, P. (1996). Individual differences in student cheating. *Journal of Educational Psychology*, *88*(2), 229-241.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

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*(2), 112-120. doi: 10.1111/j.1468-2389.2008.00416.x

Ones, D. S., Viswesvaran, C., & Dilchert, S. (2005). Cognitive Ability in Selection Decisions. In O. Wilhelm & Engle, R. W. (Eds.), *Handbook of Understanding and Measuring Intelligence* (p. 431-468). Thousand Oaks, CA, US: Sage Publications, Inc.

Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology*, *81*(6), 660-679. doi: 10.1037/0021-9010.81.6.660

Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. *Journal of Applied Psychology*, *78*(4), 679-703. doi: 10.1037/0021-9010.78.4.679

Orvis, K. A., Fisher, S. L., & Wasserman, M. E. (2009). Power to the people: Using learner control to improve trainee reactions and learning in web-based instructional environments. *Journal of Applied Psychology, 94*, 960-971. doi: 10.1037/a0014977

Pannone, R. D. (1984). Predicting test performance: A content valid approach to screening applicants. *Personnel Psychology*, *37*(3), 507-514. doi: 10.1111/j.1744-6570.1984.tb00526.x

Pauls, C. A., & Crost, N. W. (2005). Cognitive ability and self-reported efficacy of self-presentation predict faking on personality measures. *Journal of Individual Differences, 26*(4), 194-206. doi: 10.1027/1614-0001.26.4.194

Perry, S. J., Hunter, E. M., Witt, L. A., & Harris, K. J. (2010). P = f (conscientiousness × ability): Examining the facets of conscientiousness. *Human Performance, 23*(4), 343-360. doi: 10.1080/08959285.2010.501045

Pieschl, S., Stahl, E., & Bromme, R. (2008). Epistemological beliefs and self-regulated learning with hypertext. *Metacognition and Learning*, *3*(1), 17-37. doi: 10.1007/s11409-007-9008-7

Pulfrey, C., & Butera, F. (2013). Why neoliberal values of self-enhancement lead to cheating in higher education: A motivational account. *Psychological Science*, *24*(11), 2153-2162. doi: 10.1177/0956797613487221

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods, 40*(3), 879-891. doi: 10.3758/brm.40.3.879

Ree, M. J., & Earles, J. A. (1994). The ubiquitous predictiveness of g. In M. G. Rumsey, C. B. Walker, & J. B. Harris (Eds.), *Personnel selection and classification* (pp. 127–135). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Ree, M. J., Earles, J. A., & Teachout, M. S. (1994). Predicting job performance: Not much more than g.. *Journal of Applied Psychology*, *79*(4), 518. doi: 10.1037/0021-9010.79.4.518

Reuter, T., Ziegelmann, J. P., Wiedemann, A. U., Geiser, C., Lippke, S., Schüz, B., & Schwarzer, R. (2010). Changes in intentions, planning, and self-efficacy predict changes in behaviors an application of latent true change modeling. *Journal of Health Psychology*, *15*(6), 935-947. doi: 10.1177/1359105309360071

Richman, W. L., Kiesler, S., Weisband, S., & Drasgow, F. (1999). A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires, and interviews. *Journal of Applied Psychology*, *84(5*), 754-775. doi: 10.1037/0021-9010.84.5.754

Reynolds, D. H., & Dickter, D. N. (2010). Technology and employee selection. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of Employee Selection*. Clifton, NJ: Psychological Press.

Rosse, J. G., Stecher, M. D., Miller, J. L., & Levin, R. A. (1998). The impact of response distortion on preemployment personality testing and hiring decisions. *Journal of Applied Psychology, 83*(4), 634–644. doi: 10.1037/0021-9010.83.4.634

Roth, P. L., Bevier, C. A., Bobko, P., Switzer, F. S., & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology*, *54*(2), 297-330. doi: 10.1111/j.1744-6570.2001.tb00094.x

Ryan, A. M., & Sackett, P. R. (1987). Pre-employment honesty testing: Fakability, reactions of test takers, and company image. *Journal of Business and Psychology, 1*(3), 248-256. doi: 10.1007/BF01020813

Sackett, P. R., & Larson Jr, J. R. (1990). Research strategies and tactics in industrial and organizational psychology. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of Industrial and Organizational psychology, Vol. 1 (2nd ed.)* (pp. 419-489). Palo Alto, CA, US: Consulting Psychologists Press.

Sackett, P. R., & Lievens, F. (2008). Personnel selection. *Annual Review of Psychology*, *59*, 419-450. doi: 10.1146/annurev.psych.59.103006.093716

Sackett, P. R., Schmitt, N., Ellingson, J. E., & Kabin, M. B. (2001). High-stakes testing in employment, credentialing, and higher education: Prospects in a post-affirmative-action world. *American Psychologist*, *56*(4), 302-318. doi: 10.1037/0003-066X.56.4.302

Salas, E., & Rosen, M. A. (2010). Experts at work: Principles for developing expertise in organizations. In S. W. J. Kozlowski & E. Salas (Eds.), *Learning, training, and development in organizations* (pp. 99-134). New York: Taylor & Francis Group.

Satorra, A., & Bentler, P. M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika*, *75*(2), 243-248. doi: 10.1007/s11336-009-9135-y

Schmidt, F. L. (2002). The role of general cognitive ability and job performance: Why there cannot be a debate. *Human performance*, *15*(1-2), 187-210. doi: 10.1080/08959285.2002.9668091

Schmidt, A. M., & Ford, J. K. (2003). Learning within a learner control training environment: The interactive effects of goal orientation and metacognitive instruction on learning outcomes. *Personnel Psychology, 56*, 405-429. doi: 10.1111/j.1744-6570.2003.tb00156.x

Sheppard, L. D., & Vernon, P. A. (2008). Intelligence and speed of information-processing: A review of 50 years of research. *Personality and Individual Differences*, *44*(3), 535-551. doi:10.1016/j.paid.2007.09.015

Shipley, W., Gruber, C., Martin, T., & Klein, M. (2009). Shipley-2 professional manual. *Torrance, CA: Western Psychological Services*.

Smith, D. B., & Ellingson, J. E. (2002). Substance versus style: A new look at social desirability in motivating contexts. *Journal of Applied Psychology*, *87*(2), 211-219. doi: 10.1037/0021-9010.87.2.211

Spector, P. E., & Brannick, M. T. (2011). Methodological urban legends: The misuse of statistical control variables. *Organizational Research Methods*, *14*(2), 287-305. doi: 10.1177/1094428110369842

Stadler, M., Becker, N., Gödker, M., Leutner, D., & Greiff, S. (2015). Complex problem solving and intelligence: A meta-analysis. *Intelligence*, *53*, 92-101. doi: 10.1016/j.intell.2015.09.005

Strickland, J. C., Reynolds, A. R., & Stoops, W. W. (2016). Regulation of cocaine craving by cognitive strategies in an online sample of cocaine users. *Psychology of Addictive Behaviors*, *30*(5), 607-612. doi: 10.1037/adb0000180

Teixeira, A. C., & Rocha, M. F. (2010). Cheating by economics and business undergraduate students: An exploratory international assessment. *Higher Education*, *59*, 663–701. doi: 10.1007/s10734-009-9274-1

Tippins, N. T. (2009). Internet alternatives to traditional proctored testing: Where are we now? *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *2*(1), 2-10. doi: 10.1111/j.1754-9434.2008.01097.x

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(1)*, 189-225. doi: 10.1111/j.1744-6570.2006.00909.x

van Hooft, E. A., & Born, M. P. (2012). Intentional response distortion on personality tests: using eye-tracking to understand response processes when faking. *Journal of Applied Psychology*, *97*(2), 301-316. doi: 10.1037/a0025711

Viswesvaran, C., & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement*, *59*(2), 197-210. doi: 10.1177/00131649921969802

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology*,*81*(5), 557. doi: 10.1037/0021-9010.81.5.557

Whitley, B. E. (1998). Factors associated with cheating among college students: A review. *Research in Higher Education*, *39*(3), 235-274. doi: 1023/A:1018724900565

Wright, N. A., Meade, A. W., & Gutierrez, S. L. (2014). Using invariance to examine cheating in unproctored ability tests. *International Journal of Selection and Assessment*, *22*(1), 12-22. doi: 10.1111/ijsa.12053

Yardley, J., Rodriguez, M. D., Bates, S. C., & Nelson, J. (2009). True Confessions?: Alumni's Retrospective Reports on Undergraduate Cheating Behaviors. *Ethics & Behavior*, *19*(1), 1-14. doi: 10.1080/10508420802487096

Younger, J. (2008). Online job recruitment: Trends, benefits, outcomes, and implications. *International Review of Business Research Papers*, *4*(1), 374-383.

Zickar, M. J., Gibby, R. E., & Robie, C. (2004). Uncovering faking samples in applicant, incumbent, and experimental data sets: An application of mixed-model item response theory. *Organizational Research Methods*, *7*(2), 168-190. doi: 10.1177/1094428104263674

APPENDIX A

CHEATING METHODS SCALE DEVELOPMENT STUDY MATERIALS

SME Task 1: Potential Cheating Methods – Scale Revision Task

The purpose of this task is to modify an existing list of student cheating methods to create a comprehensive list of the methods that job applicants might use to cheat on an online unproctored multiple-choice selection test (e.g., cognitive ability, SJT, knowledge of a specific topic); a traditional text-based multiple-choice test or a multiple-choice test containing illustrations (see example items on page 2 of this document). Although there are many types of tests, the purpose is to create a general set of methods to apply broadly. Please do not take into account the effectiveness of these methods; we are interested in capturing both effective and ineffective methods. Part 2 of this task will ask you to rate the full set of SME-created methods for effectiveness. This first task should take about 15 minutes to complete; thank you in advance for sharing your time and your expertise.

Instructions:

1. Read through this list of 18 cheating methods below, originally created for academic (undergraduate) cheating.

2. Keeping Track Changes on,

a) Delete any methods that are irrelevant or impossible to cheating on unproctored multiple-choice employee selection tests.

b) Revise all remaining methods for relevance to cheating on unproctored employee selection tests, as needed.

c) Brainstorm other methods not already included and add them to the list.

List of cheating behaviors (Gaskill, 2014):

1. Any cheating
2. Send pictures of exam questions to others
3. Use electronic notes stored on devices during exam (e.g., cellphone)
4. Taking pictures of exam questions
5. Use notes stored on laptop while taking exam
6. Buy written papers from Websites
7. Copying from Internet without citing sources
8. Receiving e-mail with answers to quizzes
9. Search Internet for answers to exam questions
10. Send pictures of answers to homework questions to friends
11. Sending e-mail with answers to friends
12. Use copy and paste function to copy materials from friends
13. Receiving e-mail with answers to homework
14. Search Internet for answers to quiz question
15. Receive electronic notes on graded assignments or projects
16. Search Internet for answers to homework questions
17. Share personal notes via e-mail to help a friend with homework
18. Copying from Internet citing sources

SME Task 2: Cheating Methods – Rating Effectiveness Task

Purpose (Methods 1-17): Determine the potential effectiveness of a general set of cheating methods (listed below). Although there are many types of tests and testing platforms, this set of ratings will be used to apply broadly to multiple choice-type unproctored online tests used for employee selection.

Instructions (Methods 1-17): Rate each method for potential effectiveness at increasing a test taker's own score on any unproctored online test by typing a whole number between 1 and 5 (1 = potentially not at all effective and 5 = potentially very effective), next to each method.

Purpose (Methods 18-21): Determine the potential effectiveness of a specific set of cheating methods to apply to a particular test you'll read about below.

Instructions (Methods 18-21): Rate each method for potential effectiveness at increasing a test taker's own score on the specific unproctored online test described below by typing a whole number between 1 and 5 (1 = potentially not at all effective and 5 = potentially very effective) next to each method.

Test Information (Methods 18-21):

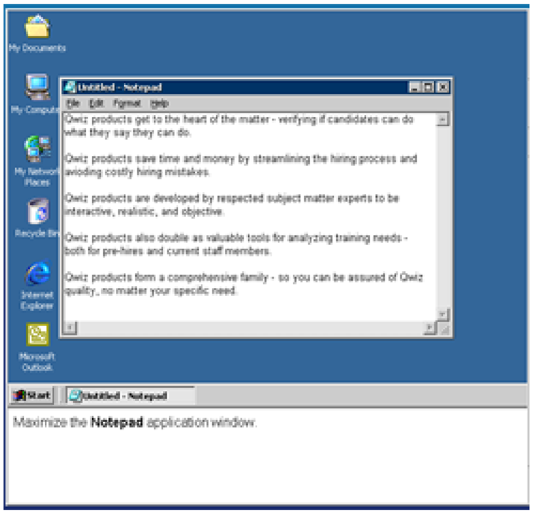
Test Setup	• Participants click on an open link to access a Microsoft Excel 2010 assessment delivered on a commercial testing website.
	• Participants are competing against each other: top scorers receive a monetary incentive, and it is unlikely participants will know other test takers.
Assessment	• Evaluates ability to use Excel: sorting/filtering data, functions and formulas, charts and pivot tables, and conditional formatting/statements.
	• Simulated Excel program presented on participants' browsers with a 35 minute time limit to complete 30 items.
	• Participants interact with the program (i.e., type/click within cells and menus) to complete each item. Items are presented one at a time and each item must be answered to move on to the next question; moving backward to previous items or forward without answering the current item is prohibited.
Constraints	• Pausing and copy/paste functions and are disabled, and colluding with an employee of the company who created/delivers the test is a firable offense, and has never happened. All other methods listed below are possible, but may differ in potential effectiveness.

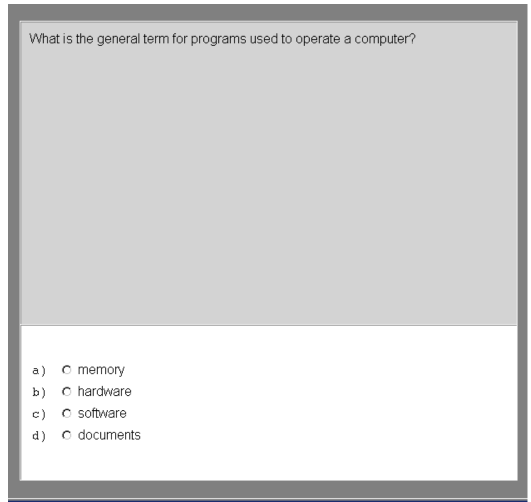
Note (Methods 1-21): The word “content” in the following items refers to the exact test question and/or answer choices, or the general topic of the test question and/or answer.

APPENDIX B

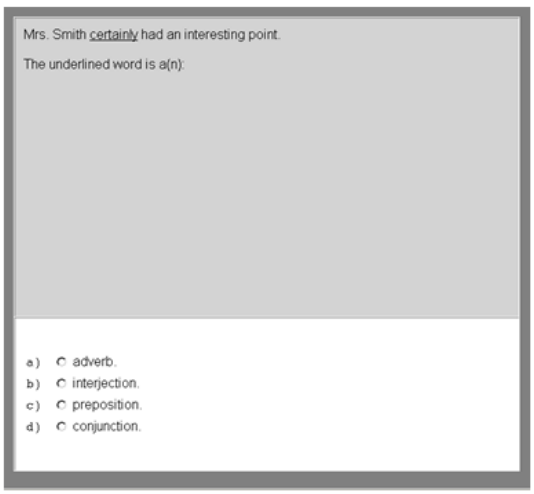
PILOT STUDY MEASURES

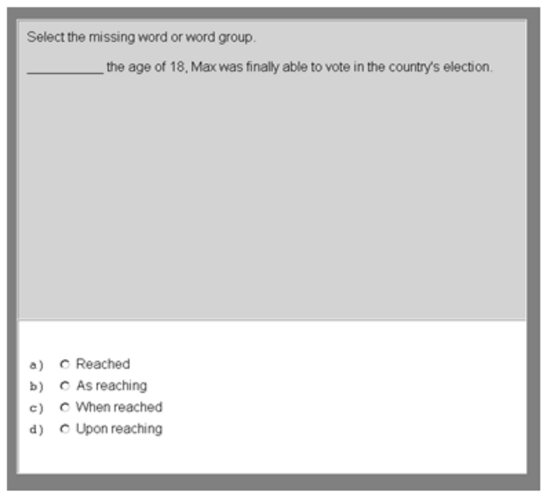
Basic Computer Literacy Example Item



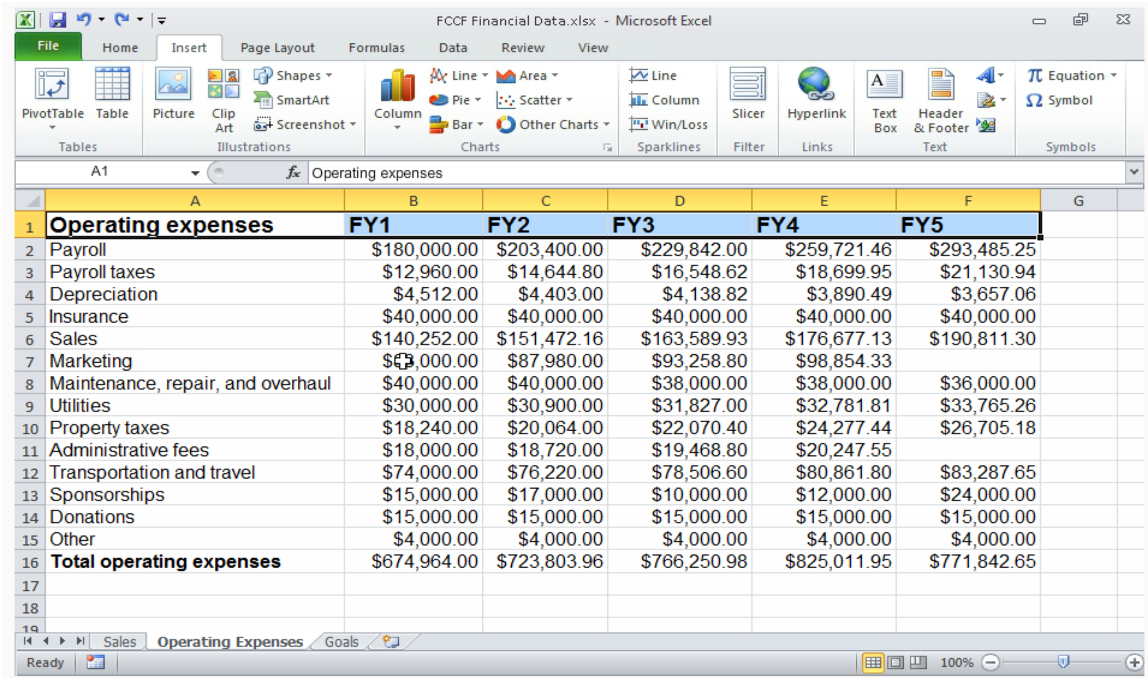


General Clerical Grammar Example Item





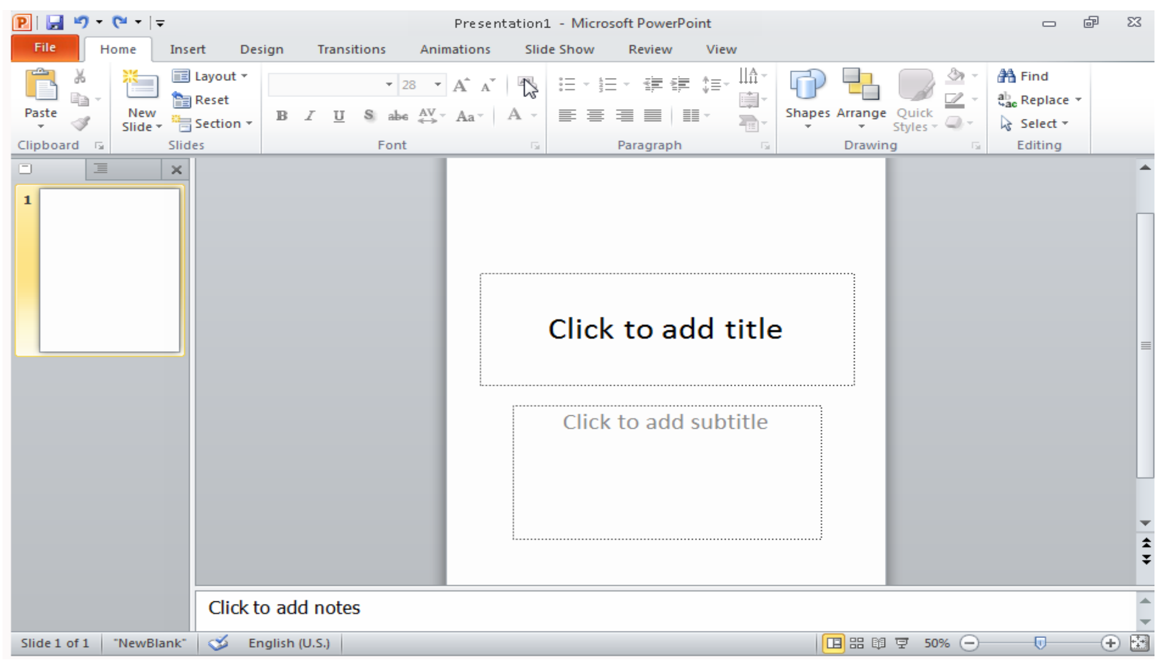
Microsoft Excel Example Item



Operating expenses	FY1	FY2	FY3	FY4	FY5
Payroll	$180,000.00	$203,400.00	$229,842.00	$259,721.46	$293,485.25
Payroll taxes	$12,960.00	$14,644.80	$16,548.62	$18,699.95	$21,130.94
Depreciation	$4,512.00	$4,403.00	$4,138.82	$3,890.49	$3,657.06
Insurance	$40,000.00	$40,000.00	$40,000.00	$40,000.00	$40,000.00
Sales	$140,252.00	$151,472.16	$163,589.93	$176,677.13	$190,811.30
Marketing	$8[illegible],000.00	$87,980.00	$93,258.80	$98,854.33	
Maintenance, repair, and overhaul	$40,000.00	$40,000.00	$38,000.00	$38,000.00	$36,000.00
Utilities	$30,000.00	$30,900.00	$31,827.00	$32,781.81	$33,765.26
Property taxes	$18,240.00	$20,064.00	$22,070.40	$24,277.44	$26,705.18
Administrative fees	$18,000.00	$18,720.00	$19,468.80	$20,247.55	
Transportation and travel	$74,000.00	$76,220.00	$78,506.60	$80,861.80	$83,287.65
Sponsorships	$15,000.00	$17,000.00	$10,000.00	$12,000.00	$24,000.00
Donations	$15,000.00	$15,000.00	$15,000.00	$15,000.00	$15,000.00
Other	$4,000.00	$4,000.00	$4,000.00	$4,000.00	$4,000.00
Total operating expenses	$674,964.00	$723,803.96	$766,250.98	$825,011.95	$771,842.65

Horizontally center the content of the selected cells.

Microsoft PowerPoint Example Item



Set the slide orientation of the selected slide to **Landscape**.

APPENDIX C

MPLUS CODE FOR POWER ANALYSIS

```
TITLE: Full Dissertation Model


MONTECARLO:

        NAMES are gca1-gca4 ecm t1-t2;

        NOBSERVATIONS are 360;

        NREPS are 1000;


MODEL POPULATION

        gca by gca1-gca4*.8;

        gca1-gca4*.05;

        methods by ecm@1;

        ecm@.08;

        eta1 by t1@.9;

        t1@.042;

        eta2 by t2@.9;

        t2@.042;


        gca@1 methods*1;

        [t1@0 t2@0];

        baseline change | eta1@0 eta2@1;

        eta1@0;
```

```
	eta2@0;
	baseline*1 change*1;
	[baseline*0 change*1.5];
	change on baseline*.21;
	change on methods*.21;
	methods on gca*.21;

	gxmethods | gca XWITH methods;

	change on gxmethods*.21;
	change on gca*.21;
	baseline on gca*.21;

MODEL:
	gca by gca1-gca4*.8;
	gca1-gca4*.05;
	methods by ecm@1;
	ecm@.08;
	eta1 by t1@.9;
	t1@.042;
	eta2 by t2@.9;
	t2@.042;
```

```
    gca@1 methods*1;
    [t1@0 t2@0];
    baseline change | eta1@0 eta2@1;
    eta1@0;
    eta2@0;
    baseline*1 change*1;
    [baseline*0 change*1.5];
    change on baseline*.21;
    change on methods*.21 (path1);
    methods on gca*.21 (path2);

    gxmethods | gca XWITH methods;


    change on gxmethods*.21;
    change on gca*.21;
    baseline on gca*.21;
    MODEL CONSTRAINT:
        NEW (mediate*.044);
        mediate = path1*path2;
Analysis:
    TYPE = RANDOM;
    ALGORITHM=INTEGRATION;
OUTPUT: TECH9;
```

APPENDIX D

MAIN STUDY MEASURES

ICAR Sample Test for GCA (random ordering will be used, as recommended by Condon & Revelle, 2014)

Sample test and answer key removed for copyright reasons. Copyright holder is: The International Cognitive Ability Resource Team (2014)

Sample test and answer key removed for copyright reasons. Copyright holder is:

The International Cognitive Ability Resource Team (2014)

Sample test and answer key removed for copyright reasons. Copyright holder is:

The International Cognitive Ability Resource Team (2014)

Sample test and answer key removed for copyright reasons. Copyright holder is:

The International Cognitive Ability Resource Team (2014)

Sample test and answer key removed for copyright reasons. Copyright holder is:

The International Cognitive Ability Resource Team (2014)

Sample test and answer key removed for copyright reasons. Copyright holder is:

The International Cognitive Ability Resource Team (2014)

Sample test and answer key removed for copyright reasons. Copyright holder is:

The International Cognitive Ability Resource Team (2014)

APPENDIX E

MAIN STUDY CARELESS RESPONDING ITEMS

ICAR Sample Test Bogus Items

1. Which of the following numbers are greater than 10? **Correct: (6)**

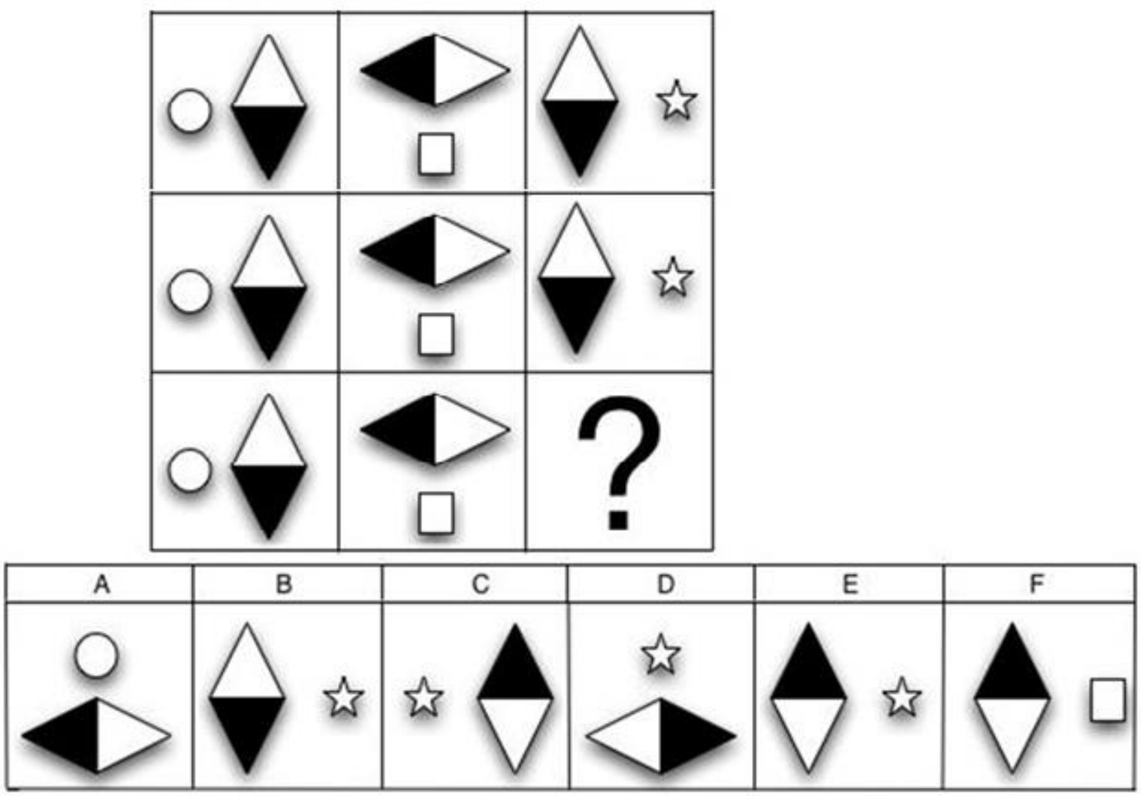
(1) 2 (2) 4 (3) 4 (4) 6 (5) 8 (6) 15

2. In the following alphanumeric series, what letter comes next? A, B, C, D, E, ...

Correct: (1)

(1) F (2) G (3) H (4) I (5) J (6) K

3. **Correct: (2)**



APPENDIX F

STUDY INSTRUCTIONS

Instructions to participants prior to the first administration of the Microsoft Excel test

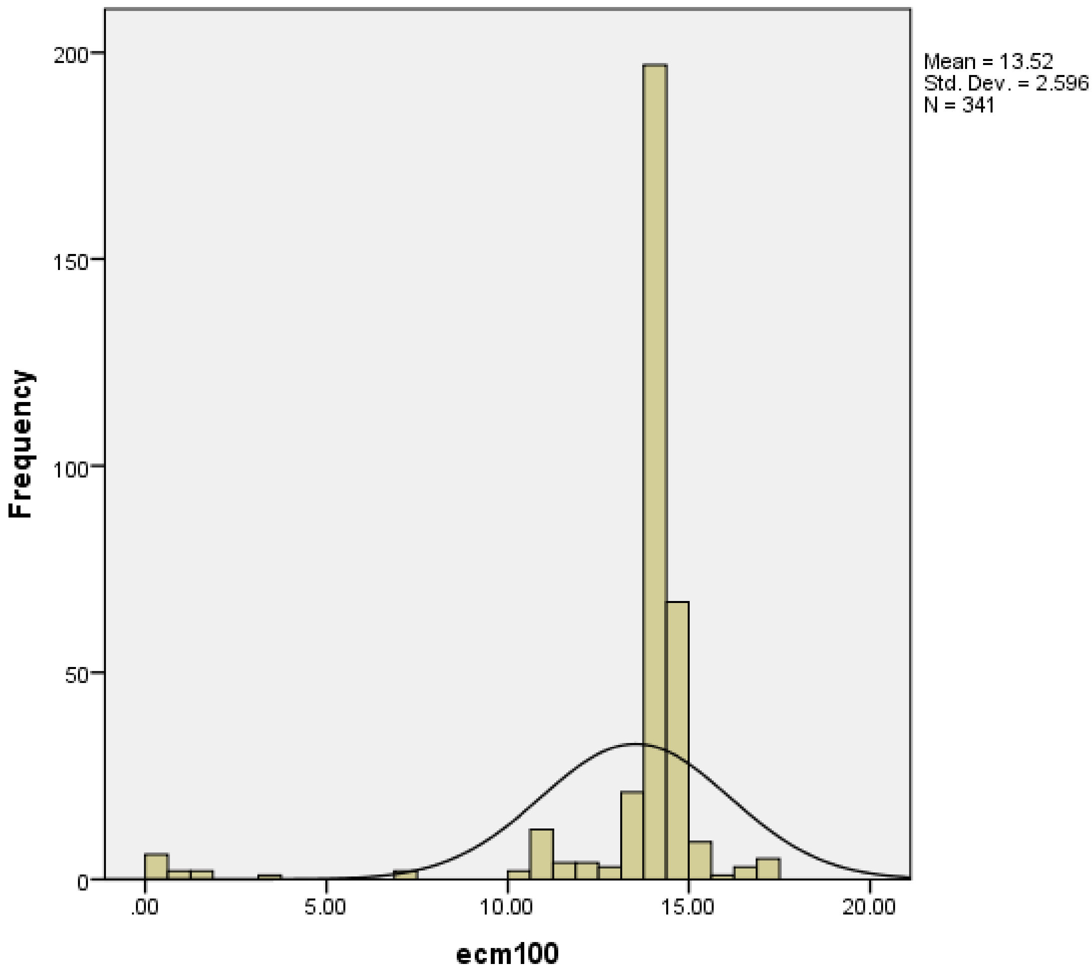
We are interested in your own existing knowledge, so please do not use any outside sources when taking this test (including but not limited to friends or family members, search engines such as Google, or referring to the program itself). You will receive compensation for this HIT based only on your completion of this test, not on your score.

Instructions to participants prior to the second administration of the Microsoft Excel test

You will now be able to re-take the same Microsoft Excel test you just took. This time, we are NOT interested in your own existing knowledge. We are interested in how well you can cheat. YOU ARE BEING INSTRUCTED TO CHEAT ON THIS TEST. Use any outside sources or methods that you think will help you get the highest score possible (including but not limited to friends or family members, search engines such as Google, or refer to the program itself). YOU WILL RECEIVE COMPENSATION FOR CHEATING WELL. The top 25% of scores on this test will receive a $3 bonus.

APPENDIX G

HISTOGRAM OF EFFECTIVE CHEATING METHODS SCORE



VITA

Katelyn J. Cavanaugh
Department of Psychology
Old Dominion University
Norfolk, VA 23529

Education

Old Dominion University, Norfolk, VA; Doctor of Philosophy, Industrial/Organizational Psychology; Expected graduation May 2018

Old Dominion University, Norfolk, VA; Master of Science, Psychology; August 2013

University of Delaware, Newark, DE; Bachelor of Arts, Psychology; May 2008

Professional Experience

Associate, ICF; June 2014 – present

Instructor, Old Dominion University; August 2013 – May 2014

Teaching Assistant, Old Dominion University; August 2011 – May 2015

Consultant, ACCESS AIDS Care; August 2010 – May 2014

Research Assistant, Old Dominion University; August 2010 – May 2018

Research Assistant, Drexel University; August 2008 – July 2010

Publications

Landers, R.N., Brusso, R.C., Cavanaugh, K.J., & Collmus, A.B. (2016). A primer on theory-driven web scraping: Automatic extraction of big data from the internet for use in psychological research. *Psychological Methods, 21*(4), 475-492. doi: 10.1037/met0000081