TEST COACHING ON ASSESSMENTS OF COGNITIVE CONSTRUCTS

BY

BEN-ROY DO

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2010

Urbana, Illinois

Doctoral Committee:

Professor Fritz Drasgow, Chair
Professor Hua-Hua Chang
Professor James Rounds
Associate Professor Deborah Rupp
Assistant Professor Alan Mead, Illinois Institute of Technology

ABSTRACT

In continuously administered employment tests, test security may be compromised by examinees revealing test items to future test candidates. Unlike academic testing, employee selection tests generally have longer test windows and are less likely to implement new test items or have multiple parallel forms for the sake of test security. Therefore, it is important to study whether test takers can benefit from item sharing, and whether integrity and personality inventories can explain such behaviors. Results showed that casual item sharing increased test scores relative to the control group, with small to moderate effect sizes. Conscientiousness and Emotional Stability were the best predictors, with conscientiousness more influential on numerical reasoning and Emotional Stability on verbal reasoning.

To my family and friends

TABLE OF CONTENTS

INTRODUCTION

Item sharing refers to the sharing of previously administered test items among past and future test takers, such that future test takers may have the opportunity to benefit from such information, which is called item preknowledge. While item sharing may occur in academic and licensing exams and pose a threat to test security, the issue is particularly serious in employment testing, as continuous testing and unproctored Internet testing are frequently used in such situations.

As many organizations hire job applicants on a rolling basis throughout the year, employment and selection tests are often administered on a continuing basis. Consequently, test items may be shared among past and future test takers. On the one hand, item preknowledge may help future job applicants to prepare themselves better for the employment test. They may become more familiar with the knowledge, skill, and abilities needed on the job, and therefore facilitate the process of matching applicants and jobs. On the other hand, such activities may compromise the integrity of the test and influence hiring decisions. As it takes considerable resources to develop alternate forms and equate them, employment batteries are often limited to a few forms or even just a single form. As a result, there can be considerable overlap in test items administered to past and future applicants, which increases the vulnerability of the testing program to compromise.

In addition to giving selection tests in a continuous testing environment, more and more organizations utilize unproctored Internet testing as part of the selection process. Job applicants can complete the selection tests at any time and in any location, as long as there is an available Internet enabled computer (Tippins et al., 2006). If the stakes for the

employment or promotion are high and there are no serious consequences (e.g., being expelled from the organization), job applicants may be motivated to cheat, and unproctored Internet testing provides an opportunity to do so. Therefore, in a continuous unproctored Internet testing environment, items may be shared with little effort, and it is possible that test items can be compromised easily and quickly.

Terminologies

Prior empirical research on item sharing is almost non-exist and thus it may be useful to explore, define, and distinguish different terms related to compromised test items. I will first try to clarify and differentiate the concepts of test preparation, test score pollution, test coaching, and test compromise.

Test preparation usually refers to preparation that increases a person's knowledge in a defined domain that will later be tested (Clause, Delbridge, Schmitt, Chan, & Jennings, 2001). According to Haladyna, Nolen, and Haas (1991), there is a continuum of test preparation activities from ethical to highly unethical. Ethical activities include training in testwiseness skills, checking answer sheets, and increasing student motivation. In addition, developing a curriculum to match the test, and using commercial materials specifically designed to improve test performance could and should increase test scores and would be considered ethical test preparation. Other forms of test preparation are unethical. For example, it is improper to present items verbatim from the test to future test takers. This poses a huge threat to the validity and usefulness of test scores, because scores do not reflect a test taker's knowledge of a domain, but rather reflect item

preknowledge. This later type of test preparation activity is considered a form of test score pollution.

Test score pollution refers to factors affecting the truthfulness of a test score's interpretation (Messick, 1984). As pollution influences test performance in a way that is unrelated to the construct measured by the test, the process can unfold before or during the actual test. One example of test score pollution during the actual test is answer copying, which is generally regarded as a form of cheating. Test score pollution can occur before the actual test, and such activities usually fall within the domain of test coaching.

Test coaching is "intensive, concentrated drill or 'cramming' on sample test questions" (Anastasi, 1981). In order to perform well on the test, the purpose of test coaching is not to teach or train test takers on the content of the broad construct domains that the test attempts to measure (i.e., ethical test preparation). Quite the opposite, the focus is to allow future test takers to understand a very limited set of information that is covered by the test questions. While test coaching can refer to general test preparation that is related to testwiseness and "teaching to the test", at the extreme, test coaching refers to simply getting answers to test questions.

Item sharing is a specific, extreme form of test coaching. Here, test takers may obtain "item preknowledge" from previous test takers who disclosed previously administered items (McLeod, Lewis, & Thissen, 2003). Test takers may benefit from item sharing through information about questions and answers on a test, and may obtain a better test score. It is a form of unethical test preparation that may compromise the validity and integrity of the test quickly.

Test compromise is an important issue for test developers. To increase test security, test developers can increase the size of the item pool and replace or update the pool frequently such that it is less likely that the test will be compromised. However, such approaches may not be practical or cost-effective (Segall, 2004). Test compromise also affects the fairness among test takers. It is unfair for some examiners to have item preknowledge when others do not.

To clarify, item sharing is an unethical activity of test preparation and one instance of test score pollution. Item sharing is also a specific, extreme form of test coaching, and is an example of test compromise. In a continuous testing environment, test coaching through the use of item preknowledge is one type of test compromise that is particularly likely to occur. For example, test compromise can arise from test takers sharing item content with a future test taker, or from an organized effort to reconstruct a bank of operational test questions, (Stocking, Ward, & Potenza, 1998). While test preparation activities are usually centered on learning the material to be tested, the extreme form of test coaching (e.g., item sharing) focuses on item preknowledge.

This paper will first discuss the issue of test security in academic, licensing, and employee selection settings, followed by why it is important to consider test score pollution in terms of process. One approach to controlling test compromise, item exposure control, will be discussed. Then, prior research on test coaching will be reviewed in the context of employee selection. Those studies may help us better understand the influence of item sharing.

As cognitive ability and personality inventories are often used for employment selection decisions (Schmidt & Hunter, 1998), this study examined whether unorganized

item sharing may have an effect on test results. To understand the process of item sharing, personality inventories and covert integrity tests were used to see how each measure is related to benefiting from item sharing. It is hoped that in this casual, unorganized item sharing situation, we are able to see who is more likely to benefit from item sharing.

Test security in academic setting

Many tests used for academic admissions, such as the Graduate Record Examination (GRE) and Graduate Management Admission Test (GMAT), are administered on a continuing basis. Examinees may take a test at varying times, and they may receive coaching about items that have been previously administered in the same testing window. Answers to such overlapping items are more likely to be memorized, which compromises test security.

In academic testing, there have been several known breaches of test security. For instance, the Kaplan Education Centers were able to use 20 test takers to reconstruct a significant portion of the computer-based GRE (Honan, 1995). In 2002, the Educational Testing Service (ETS) found some Chinese and Korean websites offering questions from live exams of the computerized GRE. Those websites included both questions and answers that were obtained by previous test takers who memorized and reconstructed questions to share with other test takers. As a consequence, ETS made plans to introduce the Revised GRE in late 2007, which was designed to not only change the item types such that it would be more difficult to recall items, but also switch from an adaptive test to a linear test format. The Revised GRE was planned to only be administered on fixed

administration dates, limiting the possibility of item theft. Even though ETS eventually decided to scrap the Revised GRE, test security is clearly an important problem in continuous testing.

There can be serious consequences if a test taker is found to be using one of the web sites that provide item preknowledge. In 2008, the Graduate Management Admission Council (GMAC) sued Scoretop.com for distributing copyrighted GMAT-related materials without GMAC's permission (Lavelle, 2008). GMAC was granted the site's domain name and said "any students found to have used the Scoretop site will have their test scores canceled, the schools that received them will be notified, and the student will not be permitted to take the test again". While it is not certain what the MBA programs will do with applicants, current students, or even graduates who have used the site and information, it is clear that test compromise can have a great impact on test developers and test takers.

Test security in licensing and certification

Licensing and certification exams are given for many occupations and professions to ensure that candidates have mastered the necessary knowledge, skill, and ability to perform the job. While some licensing exams such as the Uniform Certified Public Accountancy Examination are administered on a continuing basis, many are given only a few times each year. For example, the National Conference of Bar Examiners (NCBE) administers the Multistate Bar Examination (MBE) twice every year in several jurisdictions. The MBE is a 200-item multiple-choice exam, with only 50 new questions added each year (*Law Student Journal*, 2006).

Even though a test is only administered on fixed administration dates, test security can still be breached if items are reused or even if items are pretested by being embedded in operational forms: Test takers can engage in an organized item memorization scheme to compromise the item bank. For instance, in 2005, NCBE filed a federal law suit against Multistate Legal Studies Inc. (MLSI) for illegally copying questions from the Multistate Bar Exam for its test preparation course. The judge found that "evidence of copying practically leaps from the page", ruled in favor of the plaintiff (NCBE), and ordered MLSI to pay $11.9 million to NCBE (Duffy, 2006).

In another case, Microsoft filed a law suit against Test King (Johnston, 2006). An unusual twist in the case was that Microsoft asked the Court to identify the people behind the site (the website is registered in the United Kingdom). Test King is one of many "braindump" web sites that reveal actual test items from certification exams, such as Microsoft Certified Systems Engineer (MCSE), Cisco Certified Network Associate (CCNA), Certified Information Systems Security Professionals (CISSP), etc. These web sites provide an interface that not only allows a test taker to submit items (a.k.a. brain dumps), but also vote for the best answer and provide a rationale for each option. In an empirical study, it was found that items posted on those web sites were "surprisingly accurate", and most of the item bank was available after only in about 8 months (Smith, 2004).

Notice that the MBE and MCSE exam incidents are based on large scale item theft. That is, a company (e.g., MLSI, Test King) was engaged in organized activities to profit from test preparation materials, and the goal was to benefit many future test takers. These large scale operations to compromise a test may be more likely to be detected by

test developers, and copyright infringement suits can be filed against test preparation companies or web sites. However, if such conspiracy is small scale and the aim is to benefit a small group of people, it is difficult to identify test takers who posted memorized items or benefited from the posted items.

A recent incident at University of Medicine and Dentistry of New Jersey illustrates how small scale test compromise conspiracies work. A number of students at the dental school organized a "clandestine operation to cull questions from the exams" (Sherman & Margolin, 2007). They instructed other members of the class to memorize specified questions and to contribute questions that might be included on future exams. The directions included setting up e-mail accounts under false names, posting the items on a web site, and giving the composite exams as a gift to the next year's class. If participating members can keep such operations secret, it is difficult to detect this type of small scale test compromise.

Test security in employee selection

Many companies utilize standardized test batteries, and test scores often play an important part in hiring and promotion decisions. For instance, predictors of job performance, such as general cognitive ability and personality inventories, are often used for employment selection decisions (Schmidt & Hunter, 1998; Murphy, 2002). While human resources departments utilize other assessment tools such as job interviews and background information to make hiring decisions, if a job applicant did not perform well enough on a selection test to be considered among the top applicants, the person may not even have a chance to be invited to a job interview. Thus, job applicants may be

motivated to engage in activities that can cause test score pollution. Prior research has found cheating in many fields, including pilots, marines, cosmetologists, teachers, police officers, and even clergy (Cizek, 1999).

In academic testing, a particular form of a test or an item pool may be in use for as little as one month or less. Here, the chance of items being memorized and exposed is lower. Organizations such as the College Board and ETS collect test registration fees to not only perform test equating and generate score reports, but also recover the costs of developing new forms of the tests. However, in employee selection, job applicants do not pay a fee for testing. As the costs to develop and administer tests are not covered by job applicants, companies are less likely to create new selection test items. Therefore, the test windows are usually longer, item pool sizes are smaller, and item rotation is less frequent. All of these factors may affect test security.

In addition, many companies are moving towards unproctored Internet testing to reduce costs and better accommodate their hiring needs. As the top three reasons for companies to use technologies in recruiting, screening, and selection processes are increased efficiency, availability of new assessment tools, and reduced cost (Chapman &Webster, 2003), unproctored Internet testing has several advantages over traditional proctored test administrations for both test administrators and job applicants. For instance, the company no longer needs to schedule job applicants on-site for in-person test sessions (Wheeler, Foss, & Handler, 2001), and the test content can be updated more frequently. Job applicants' responses are centralized and collected faster with fewer human mistakes (Beaty, Fallon, Shepherd, & Barrett, 2002). While some previous research has shown measurement equivalence across proctored and unproctored Internet testing in field and

lab studies (Do, Shepherd, & Drasgow, 2005), it should be noted that there were test score differences, and test compromise may cause a selection test to become invalid and useless. In such cases, companies may have to use other, more costly tools to select employees. Therefore, it is important to study test security and test score pollution in the context of employee selection.

Process of test score pollution

As mentioned previously, test score pollution can unfold before or during the actual test. Test score pollution during the actual test is generally regarded as cheating, and is often studied in terms of academic dishonesty (Jensen, Arnett, Feldman, & Cauffman, 2002) or academic integrity (Chapman, Davis, Toy, & Wright, 2004). For example, previous research has suggested that cheating during college ranges from 3% of students (Karlins, Michaels, & Podlogar, 1988) to 87% (McCabe, 1992). While this wide range may reflect confusion about what cheating is and may be overestimated due to peer pressure or researcher expectations (Chapman et al., 2004), it is nonetheless clear that cheating does occur in proctored settings. One interesting finding is that 57% of students would look at another student's exam during a test, but only 12% would change their answers (Nonis & Swift, 1998). This means answer copying during a test may reflect different processes from using item preknowledge obtained from other test takers prior to the test.

Some work has already been conducted to understand the characteristics of people who cheat on tests (Whitley, 1998) and why they cheat (Smith, Ryan, & Diggins, 1972; Baird, 1980; Stevens & Stevens, 1987). For instance, it has been found that students who

are high in achievement motivation are more likely to cheat (Johnson, 1981; Eisenberger & Shank, 1985). Prior research also found lenient attitudes toward cheating are positively related to tolerance of deviance, and negatively related to self-restraint (Jensen et al., 2002).

Nonetheless, these studies have not attempted to understand and address how test score pollution can occur before the actual test. Historically, this has not been a key issue because there is a myriad of ways to cheat. Thus, instead of studying test score pollution in terms of people's attitudes and opinions toward cheating, we propose to examine test security and test score pollution issues in terms of process. The process consists of three stage, item memorization, item preknowledge, and item responding during the actual testing.

Item memorization. Test score pollution may start with item memorization by previous test takers, also known as sources. Sources memorize the items they are presented and provide information about items to other test takers, known as beneficiaries.

Item preknowledge. The beneficiaries may then acquire the information from the sources, called item preknowledge. For this information to be most useful, the beneficiaries need to take the test during the same test window as the source. When test items in a test window are selected from a single item pool, the beneficiaries may gain an unfair advantage over other test takers.

Item responding during testing. Finally, during the actual testing, the beneficiaries utilize their item preknowledge to answer some items. While answer copying may be prevented in a proctored setting, a test proctor has no way at knowing if an examiner is

utilizing item preknowledge.

As the process of test score pollution may start before the test is administered to a beneficiary, prior research has taken at least two approaches to understanding test security: item exposure algorithms, and simulations based on item preknowledge states. While each approach has its merits and has important practical implications, each has its shortcomings as well.

Item exposure control algorithms

The most frequently used psychometric method to reduce the impact of item preknowledge on computerized adaptive tests (CATs) is to use item exposure control methods, such that items with the greatest psychometric information will not be selected too frequently and thus quickly compromised. There are several exposure control methods that can be used, including McBride and Martin's (MM) technique (McBride & Martin, 1983), the Sympson-Hetter (SH) procedure (Hetter & Sympson, 1997; Sympson & Hetter, 1985), the Davey and Parshall (DP) methodology (Davey & Parshall, 1995; Parshall, Davey, & Nering, 1998), and the Stocking and Lewis unconditional multinomial (SL) and conditional multinomial (SLC) procedure (Stocking & Lewis, 1998). There are also approaches to increase item pool utilization and therefore reduce item exposure rate, such as the *a*-stratified (Chang & Ying, 1999) and *a*-stratified with *b* blocking (Chang, Qian, & Ying, 2001) methods.

While previous research has compared some of the item exposure control strategies, in practice, if items are exposed, memorized, and provided to future test takers, item exposure control may be inefficient in some circumstances. If organizations do not

have large item banks or utilize multiple forms in unproctored Internet testing, any item exposure control method will be limited in effectiveness.

Simulation studies based on item preknowledge states

Simulation studies have been used to examine test security by simulating test compromise on a certain percentage of items. For example, Stocking, Ward, and Potenza (1998) used the weighted deviations model (WDM) adaptive testing paradigm to simulate a worst case scenario: with 10% or 20% of the item pool compromised, they assumed that all test takers would have perfectly correct performance on all disclosed items they encounter. They found that with 10% of the items compromised, the increases in test scores may be rather small, but if 20% of the item pool is disclosed, the effects may be large enough to have practical consequences.

Another approach assumes the item preknowledge state is dichotomous, where an item is either completely compromised or not. McLeod, Lewis, and Thissen (2003) proposed a modified 3PL model, where the probability of a correct response to an item is the combination of the test taker's underlying proficiency (without) preknowledge of the specific item, and the probability of answering the item correctly based on the test taker's preknowledge of the item. They proposed a Bayesian posterior log odds ratio index to detect the use of item preknowledge.

While both of these approaches are promising, assuming that 10% of items are compromised or modeling item preknowledge states dichotomously may be an oversimplification. On the one hand, with organized item theft, actual item compromise rates may be much higher than 10%. For instance, each 200-item Multistate Bar Exam

(MBE) administration includes about 60 questions from earlier tests (Nathan, 2006), which means about 30% of the items were compromised before the test was given. On the other hand, with unorganized test coaching, previous test takers may imperfectly describe items and consequently fail to help cheaters. Here, a dichotomous item preknowledge state may not be the best way to describe how test coaching effects test performance.

Item sharing obviously may improve a test taker's performance, but it may also impair performance. This distinction turns on whether the exposed item was compromised with the correct answer or an erroneous response. Unless each and every source has extremely high ability and very good memorization skills, the information provided by them will include both correct and erroneous responses. As the information shared may contain only partial information about the test items and options, it may be difficult for the beneficiaries to determine whether the information is accurate or incorrect. While correct answers will help beneficiaries, if the beneficiaries receive erroneous answers, their score may actually be lowered.

Test coaching in employee selection

No empirical study has been conducted on item sharing in the employee selection setting. One major distinction between academic testing and employee selection test preparation is that job applicants may consider whether spending many hours learning about broad domains of job knowledge is worth it or not. While job applicants may spend a great deal of time studying a general content domain, if specific item preknowledge was available and there is no serious consequences preventing them to benefit from item

sharing, they may be willing to spend a small amount of time reviewing items they obtained from previous applicants. Therefore, the characteristics of a smaller conspiracy focusing on the benefiting from item sharing should be examined.

One approach to examining test coaching in employee selection settings is to investigate whether retaking identical selection tests results in significant test score increases between the first and second or further test administrations. In this scenario, candidates obtain preknowledge through the first test administration. Hausknecht, Trevor, and Farr (2002) studied 4726 candidates who took the same annual verbal cognitive ability test and oral communication ability test two or more times. They found statistically significant score increases, regardless of the number of tests taken. They also found the number of tests taken was positively related to training performance, and negatively related to turnover. While they did find practice effects, it is difficult to tell whether the effect was due to test familiarity, decreased test anxiety and stress, or actual ability increases. As the validity coefficient was significantly smaller for repeat test takers, it is possible that applicants learned specific item content that produced higher test scores, yet did not actually improve on the underlying construct presumed to predict performance.

The above example illustrates why it is important to study item preknowledge. When the identical selection test is administered to the same person two or more times, those test takers have item preknowledge and score significantly higher upon retesting. Similarly, if item preknowledge was provided to prospective job applicants, it is possible that their test scores would increase significantly, and reduce the validity of the test. However, as item sharing may improve or impair a test taker's performance, the amount

of benefiting from item sharing may depend on various factors such as the test taker's personal characteristics.

Unfortunately, there is no known study examining item sharing in terms of employee selection settings. A recent review by Sackett, Borneman, and Connelly (2008) showed that test coaching was studied primary in the educational admissions domain, particularly the SAT. While test coaching on the SAT may produce 0.1 standard deviation increases in scores, Kulik, Bangert-Drowns, and Kulik (1984) have found greater score increases for other intelligence and aptitude tests. In addition, while job applicants who retake identical selection tests showed significant test score increases, the specific mechanisms underlying score improvement are unknown. Therefore, it is important to examine the process underlying score gain from item sharing.

The next section describes mechanisms that may underlie benefiting from item sharing.

Theory of planned behavior (TPB)

Why would test takers benefit from item sharing? The theory of planned behavior (Ajzen, 1991) argued that individuals make decisions to engage in specific behavior as a function of three components: behavior beliefs and their valences, normative beliefs and subjective norms, and control beliefs and perceived behavior control. Based on their own beliefs about the specific behavior and the expectation of a positive outcome after having engaged in the behavior, they are more likely to perform the specific behavior. Beck and Ajzen (1991) have found the model was able to predict most of the systematic variance in student decisions to cheat.

As item sharing may improve or impair a test taker's performance, the amount of benefiting from item sharing may depend on various factors such as the test taker's personal characteristics. In the next section, I will use TPB to explain how each personality factor may be used to explain why the test taker may benefit from item sharing. Here, the three major components of TPB will be described first and examples will be provided to show how each component may be related to item sharing.

Behavior beliefs and attitude toward the behavior.

First, the test taker may utilize behavior beliefs, which is the subjective belief that the particular behavior will produce a given outcome. In this study, a test taker may first believe that item sharing will produce a given outcome, which is increased test performance. Once such a behavior belief is established, the test taker will need to evaluate whether the behavior belief is positive or negative. If the valences of benefiting from item preknowledge is positive, the test taker is more likely to engage in item sharing. If the valence is negative, then the test taker is less likely to engage in item sharing.

An example of evaluating behavior beliefs as positive or negative may be illustrated to explain why some students would look at another student's exam during a test (Nonis & Swift, 1998). Here, looking at another student's exam might be expected to produce a given outcome, which is an increased test score. If their valence toward such an outcome is positive, they are more likely to look at another student's answer.

Normative beliefs and subjective norm.

The second component of TPB is normative belief, the individual's perception about the particular behavior judged by the person's referent others such as classmates, friends, and family members. These normative beliefs, with the motivation to comply

with others, determine the prevailing subjective norm, which is the perceived social pressure to engage or not to engage in a behavior. A test taker may perceive item sharing as socially acceptable. For example, whether item sharing is considered ethical or unethical by significant others, and the motivation to comply with these others will influence the focal individual's behavior.

When the test taker evaluates the suggested behavior (i.e., item sharing) as positive and the subjective norms encourage the test taker to perform such behaviors, s/he is more likely to do so. Indeed, many users of Scoretop.com argued that they thought the site's questions were legitimate (Levy, 2008) and that "Nowhere in the Web site does it say you would be violating the rules".

Control beliefs and perceived behavior control.

Finally, control beliefs assess factors that may facilitate or impede performance of the behavior. With several control beliefs, the person may develop perceived behavioral control, which is the test taker's evaluation of how easy or difficult it is to perform the particular behavior. With the help of Internet and search engines, access to test information is much easier than word-of-mouth. Some users on web sites such as PostYourTest.com believe that "making them available to everyone may level out the playing field," and doing so facilitates perceived behavioral control (Veroff, 2008).

To sum up, behavioral beliefs link the behavior of interest (item sharing) to expected outcomes (increased test performance). The test taker may use normative beliefs to evaluate whether item sharing is positive or negative. When factors present (e.g., item preknowledge) to facilitate the performance of the behavior, test takers may develop perceptions of behavioral control between item sharing and test performance. A test

taker's intention to benefit from item sharing is based on the positive attitude toward the behavior, favorable subjective norm, and the perceived behavioral control of factors that facilitate the behavior of interest.

The theory of planned behavior (Ajzen, 1991) is one of the most influential theories to link intention and behavior. Meta-analytic research has found the theory can account for about 50% of the variance in intention in self-report surveys, and 20% of the variance of observed behavior (Armitage & Conner, 2001). Prior research on unethical behavior of engineering and humanities students has showed that the TPB model was able to predict cheating regardless of context such as gender, discipline, and high school (Harding, Mayhew, Finelli, & Carpenter, 2007). As many studies have been conducted and provide strong support for the theory, this study will not test the theory of planned behavior itself. We are not interested in measuring the test taker's subjective evaluation of item sharing, subjective norms about item sharing, or the perceived behavioral control of factors that facilitate the behavior. Rather, the theory is used to provide a rationale to explain why several personal characteristic may influence test takers to item sharing behavior.

While it may be possible to ask test takers about their behavior beliefs and attitudes, their normative beliefs and subjective norms, as well as their control beliefs and perceived behavioral control toward benefiting from item sharing, it is unlikely that such questions will be present in an employee selection situation. These direct questions may not be job related and improper to use as part of the selection hurdle. Instead, organizations may utilize valid predictors of job performance such as personality and integrity tests as a proxy to such sensitive questions. Given that personality and integrity

items are frequently used during the hiring process, it may be possible to use these variables to explain item sharing.

Using personality to explain benefiting from item sharing

The Big Five model provides a widely accepted taxonomy for personality trait variables. It includes Extraversion, Conscientiousness, Agreeableness, Neuroticism (or Emotional Stability), and Openness to Experience. While some have argued that the Big Five model is not comprehensive enough, and its constructs are confounded, Barrick and Mount's (1991) meta-analysis showed that the Big Five personality dimensions are valid predictors of job performance in some circumstances. Poropat (2009) argued that personality is a component of an individual's willingness to perform and may be used to predict socially valued behaviors. As such, we are not concerned about prior research (Schmit & Ryan, 1993; Viswesvaran & Ones, 1999; Alliger & Dwight, 2000) on the susceptibility of personality measures to faking. In the present study, personality dimensions are used as explanatory variables, not dependent variables. Specifically, we are not interested in the attitude or opinions of test takers toward item sharing. We are interested in what personality dimensions are related to the degree a person may engage in item sharing and hence benefit. In other words, the study examines whether benefiting from item sharing can be explained though personality dimensions.

To begin with, it is possible that different personality dimensions will have different magnitudes of relationship with the degree a person may be benefit from item sharing. The remainder of this section will provide a rationale on how personality may be related to benefiting from item sharing.

Extraversion.

People who are extraverted are more likely to be assertive and enthusiastic, and they enjoy interpersonal interaction and engage in community activities. They are often motivated by reward or success, compared to introverts who are primarily motivated to avoid punishment and failure (Dodson, 2000). Instead of being reserved and depressed, they express their energy and tend to be sociable, dominant, and positive (Watson & Clark, 1997). As such, one might expect that highly extraverted people will perform relatively well in learning contexts. However, reviews by De Raad and Schouwenburg (1996) showed that extraverted students were only superior to introverted students at school before the age of 11-12, and successful university students usually score low on extraversion.

Nonetheless, according to Barrick and Mount (1991), extraversion is a valid predictor for managers and sales representatives (ρ = .18 and .15, respectively). Extraversion was also a significant predictor of training proficiency (ρ = .26). With regards to the relationship between extraversion and motivation, Judge and Ilies (2002) found a moderate relationship with self-efficacy motivation (ρ = .33) and low correlations with goal-setting motivation and expectancy motivation (ρ = .15 and .10, respectively).

Based on the above correlations, it is possible to infer the relationship between extraversion and benefiting from item sharing through motivation and managerial performance in a team environment where training is used. Prior research showed that team members who are extraverted are more likely to seek help from other team members (Porter, Hollenback, Ilgen, Ellis, West, & Moon, 2003). As test coaching can be considered one type of training, the effectiveness of using item preknowledge may

depends on one's motivation. As such, people who are extraverted may be motivated to seek help from others.

However, while extraverted people are motivated to seek help from others, it does not mean they are necessary open to receiving help from others. Prior research showed that the correlation between extraversion and motivation was strongest for self-efficacy motivation (Judge & Ilies, 2002). As people with strong perceived self-efficacy are more likely to believe in their own capabilities to reach designated levels of performance (Bandura, 1994), it is possible that they may consider any benefits from item sharing to be inferior to what they might do based on their own ability, and render the relationships between extraversion and benefiting from item sharing to be weak.

Using the theory of planned behavior, it is possible that people who are extraverted are enthusiastic and believe that item sharing will produce a given outcome, which is increased test performance. As extraverts are more assertive than introverts, they may perceive benefiting from item sharing as positive. Given that their attitude toward benefiting from item preknowledge is positive, they may be more likely to engage in item sharing.

With regards to the second component of TPB, normative belief, it is possible that extraverts evaluate the judgment of significant others about benefiting from item sharing the same way as themselves. In a study on extraversion and psychological closeness, Beck and Cartwright (1998) found that more extraverted people were more likely to assume that other people were similar to themselves. As such, people who are extravert may perceive item sharing as socially acceptable, and believe it is a legitimate way to improve test performance.

Given that extraversion has a low yet significant correlation with training proficiency, and a moderate relationship with self-efficacy motivation yet low correlation with goal-setting motivation and expectancy motivation, it seems necessary to examine the relationship between extraversion and benefiting from item sharing in an empirical study.

Conscientiousness.

Although conscientiousness may be conceptualized as a single construct, it includes various subcomponents such as competence, order, dutifulness, achievement striving, self-discipline, deliberation (Robertson, Baron, Gibbons, MacIver, & Nyfield, 2000). Conscientiousness people are usually hardworking and goal oriented. They are usually motivated and achievement oriented (Barrick & Mount, 1991). Using the theory of planned behavior, as people who are conscientiousness are goal oriented and motivated to perform well, they may believe that item sharing can produce a given outcome, which is increased test performance. Among the Big Five personality dimensions, prior research has showed that only the conscientiousness dimension was found to be consistently related to intentions and behavior (Conner & Abraham, 2001). As such, it is possible that people who are conscientiousness could engage in item sharing and obtain higher test scores.

How will conscientiousness people evaluate item sharing as positive or negative? Stober (2001) developed a social desirability scale and correlate the inventory with the NEO Five Factor Inventory (FFI) (Costa & McCrae, 1992). Among the five factors, conscientiousness has the highest correlation with social desirability. In other words, conscientiousness people who view item sharing as positive may also believe that such

behavior is socially acceptable by others. With normative beliefs to comply with others, their subjective norms and perceived social support help the conscientious test taker to engage in the behavior, which is engaging in item sharing. When the test taker evaluates the suggested behavior (i.e., item sharing) as positive and socially desirable, the test taker is encouraged and more likely to to perform such behavior.

The last component of TPB involves control beliefs, which are factors that may facilitate or impede the performance of the behavior. If item preknowledge is not difficult to obtain and the evaluation of item sharing is easy, the test taker will have perceived behavioral control. Given the consistency between behavior beliefs, normative beliefs, and control beliefs, conscientiousness could be the best predictor of benefiting from item sharing.

Agreeableness.

People who are agreeable tend to be considerate, trusting, friendly, and better at interpersonal facilitation (Hurtz & Donovan, 2000). They are more likely to cooperate with people and be compassionate to help out others. Meta-analysis (Barrick & Mount, 1991) results show that the correlation between agreeableness and job performance is small ($\rho < .10$). In another meta-analysis (Judge & Ilies, 2002) on agreeableness and three types of motivation, except for a moderate correlation with goal-setting motivation ($\rho = -.29$), the correlations with expectancy motivation and self-efficacy motivation were small ($\rho < .13$). As such, it is questionable whether people who are agreeable will be highly motivated to perform well.

Can we predict benefiting from item sharing from agreeableness? It is hard to say. On the one hand, prior research has shown that agreeableness may have some positive

impact on academic performance by facilitating cooperation with learning processes (De Raad & Schouwenburg, 1996). As people who are agreeable are more trusting and cooperative, they could perceive item preknowledge as truth and accept it to help their test performance.

On the other hand, accepting help in the case of agreeableness is not the same as actively seeking help that might be expected from people who are extraverts. Prior research has shown that the relationship between agreeableness and organizational citizenship behaviors is weak (Organ & McFall, 2004), meaning that agreeable people may not actively engage in discretionary behavior such as item sharing. In addition, while agreeableness is correlated with goal-setting motivation ($\rho = -.29$), it was a negative correlation (Judge & Ilies, 2002), meaning agreeable people may not be motivated to engage in item sharing.

We may use the framework of the theory of planned behavior (Ajzen, 1991) to explain how agreeable people will react when they are exposed to item preknowledge. As agreeable people are more trusting with others, their behavioral belief may expect item sharing to produce a desirable outcome, namely improved test performance. When they evaluate whether such behavior as positive or negative, their cooperative tendencies may result in evaluations of such behavior as positive.

When normative beliefs are considered by agreeable people, given their tendency to agree and cooperate may lead to perception that item sharing is socially acceptable. In other words, item sharing is cooperating with others. As such, they may be more likely to accept the item preknowledge provided and engage in item sharing.

The third component of TPB is about beliefs concerning factors that may facilitate or impede performance of the behavior. When item preknowledge is presented to people who are agreeable, they may develop perceptions of higher self-efficacy. Given the consistency between behavior beliefs, normative beliefs, and control beliefs, agreeableness could also predict item sharing and consequent score increases.

However, in a meta-analysis examining the relationships of personality dimensions to performance motivation (Judge & Ilies, 2002), unlike conscientiousness that is positively correlated with all three types of motivation, agreeableness was only weakly or even negatively correlated with performance motivation. This could indicate that the degree of item sharing explained by agreeableness will be much smaller than what can be predicted through conscientiousness. As such, it is uncertain whether agreeableness will predict item sharing and score increases.

Neuroticism.

Individuals who are neurotic are more likely to experience negative emotions such as anxiety, anger, and depression. They may respond poorly to stress and have difficulty controling emotions. As people who are neurotic are more anxious and tend to focus on their emotional state, their academic performance is usually reduced due to interference (De Raad & Schouwenburg, 1996). Neuroticism is also a dispositional driver of employee attitudes, affect, and behavior (Judge, Van Vianen, & De Pater, 2004). Prior research using a numerical reasoning test showed that in a stressful situation, neurotic individuals performed significantly less well than stable individuals, while in a less stressful, relaxed condition, there was no significant score difference between neurotic and stable participants (Dobson, 2000).

On the opposite end of the construct, people who are emotionally stable are usually calm and secure. Emotional Stability is positively associated with self-efficacy (Judge & Bono, 2002), and self-efficacy is considered to be a strong predictor of grade point average (Robbins, Lauver, Le, Davis, Langley, & Carlstrom, 2004). Emotional Stability generally contributes positively to performance (Mount, Barrick, & Stewart, 1998).

By applying the theory of planned behavior (Ajzen, 1991), it can be argued that test takers who are emotionally stable may first believe that item sharing will produce a given outcome, which is increased test performance. When they evaluate whether such behavior is positive or negative, they may perceive benefiting from item sharing as positive because doing so is consistent with their tendency to perceive the world in a way evolves little stress and anxiety.

With regard to the second component of TPB, normative beliefs, emotionally stable individuals may consider benefiting from item sharing as acceptable from referent others, as usually people do not consider item preknowledge as a way of cheating. When the test taker evaluates benefiting from item sharing as positive and the perceived subjective norms seems to indicate the behavior is socially acceptable, s/he is encouraged to perform such behavior and more likely to benefit from item sharing.

Finally, emotionally stable people are less likely to experience negative feelings and anxiety, and more able to maintain stable emotional adjustment (McCrae & Costa, 1996). As such, when they assess the factors that may facilitate or impede performance of the behavior, they may perceive greater control over their test performance. As a result,

their control beliefs may regard item preknowledge as a way to improve their test performance, and thereby benefit from item sharing.

Openness to experience.

According to Costa and McCrae (1992), elements of openness to experience include fantasy, aesthetics, feelings, actions, ideas, and values. People who are open to experience are usually more creative and unconventional (McCrae, 1996). They are more likely to be attentive, imaginative, and possess intellectual curiosity (Costa & McCrae, 1992).

According to Barrick and Mount (1991), individuals high in openness to experience have more positive attitudes towards learning new things and are more likely to engage in learning experiences. De Raad and Schouwenburg (1996) also argued that openness to experience seems to reflect "the idea student" because of the relationship with foresighted, intelligent, and resourceful.

Given the above mentioned findings, it is possible that people who are open to experience may consider item preknowledge to be novel and interesting. However, whether openness to experience can translate to improved test performance is questionable. While prior research showed openness to experience is a valid predictor of training proficiency ($\rho = .25$), the predictive relationship with job performance is rather weak (Barrick & Mount, 1991). Thus, the question whether they will benefit from item sharing should be addressed empirically.

In addition, openness to experience is probably the least studied construct in the Big Five Model of personality. When the correlation between openness to experience and motivation is considered, the relationship is small and inconsistent: $\rho = .18$ for goal-

setting motivation, ρ = -.08 for expectancy motivation, and ρ = .20 for self-efficacy motivation (Judge & Ilies, 2002). As such, it is uncertain whether people who are open to experience may be motivated to benefit from item sharing.

How will people who are open to experience perceive the normative beliefs from significant others with regards to item preknowledge and item sharing? It's hard to say. On the one hand, those relevant others may be accustomed to the innovative and imaginative thinking about the individual who is open to experience, and agree with the individual's ideas, values, actions, and feelings. On the other hand, the individual may have already experienced the negative feedback from significant others about his/her novel behaviors. Thus, it is difficult to evaluate the perceived social pressure to engage or not to engage in item sharing. Therefore, it is uncertain whether openness to experience will be a good predictor of item sharing.

Using integrity scales to explain benefiting from item sharing

Personality based integrity tests are usually correlated with conscientiousness, agreeableness, and emotional stability (Berry, Sackett, & Wiemann, 2007). They are considered to be "compound traits (Hough & Schneider, 1996)" such that the combination of basic traits is able to predict a specific criterion in specific contexts and achieve higher criterion-related validity. As a result, they are considered to be "criterion-focused" such that items are designed to predict a criterion, such as counterproductive behavior (Berry, Sackett, & Wiemann, 2007). Here, again, we examine integrity as an explanatory variable for predicting benefiting from item sharing.

Will covert integrity tests be more likely to explain benefiting from item sharing than personality scales? It is hard to say, because people may consider the use of item preknowledge differently. Some may consider it is unethical to use item preknowledge and wrong to be benefiting from item sharing to help themselves, but others may think item preknowledge helps to prepare themselves to perform well and present themselves as ideal employees and that has nothing to do with one's integrity. As such, it may be necessary to examine whether personality based integrity measures may be able to explain a specific behavior, benefiting from item sharing.

Test taker motivation

In Campbell's basic model of job performance (Campbell, 1990), job performance is determined by declarative knowledge, procedural knowledge and skill, and motivation. As such, it is possible that the test taker's motivation may influence the effect of item sharing. According to Salgado, Remeseiro, and Iglesias (1996), test-taking motivation can be defined as the positive or negative attitudes for answering the tests. Prior research has found that with test taker ability controlled, job applicants showed higher test-taking motivation than actual employees (Arvey, Strickland, Drauden, & Martin, 1990). Test takers may also consider themselves differently when responding to the same assessment labeled for personnel assessment versus anonymous assessment for research (Schmit & Ryan, 1993). However, it is uncertain whether test taking motivation may be different between test takers with or without item preknowledge.

In high stakes testing, test takers may be motivated to engage in test score pollution activities such as copying answers from other test takers or even asking more

knowledgeable colleagues or friends to take the test. If the stakes of the testing are not considered high, they are less likely to engage in those unethical activities. However, if item preknowedge is presented to test takers, test takers with different degrees of motivation may have different responses. Therefore, motivational factors concerning benefiting from item sharing could be studied.

It should be noted that there are correlations among the Big Five personality dimensions and test-taking motivation. In an empirical study, Salgado, Remeseiro, and Iglesias (1996) found Neuroticism and Agreeableness were significantly correlated with test anxiety and poor control (r=.37 and .17, respectively), and Extraversion and Conscientiousness were significantly correlated with motivation and confidence in tests (r =.16 and .17, respectively). When performance motivation in general is concerned, Judge and Ilies (2002) found Extraversion and Openness to Experience were significant predictors of goal-setting and self-efficacy motivation, Conscientiousness was a significant predictor of performance motivation across the three motivational perspectives, Agreeableness was a significant negative predictor of goal-setting motivation, and Neuroticism was moderately negatively correlated with performance motivation. They found the average multiple correlations between the five-factor model and performance motivation was .49. This indicates that a substantial percentage of test taker motivation can be explained by personality dimensions.

However, it should be noted that in personnel assessment situations, test takers may be motivated to present themselves as ideal candidates and bias their answers to self-report measures of motivation (Salgado, Remeseiro, & Iglesias, 1996). Such distortion may be related to positive test taker motivation, but it is questionable how test taker

motivation can be translated into performance motivation such as goal-setting motivation, expectancy motivation, and self-efficacy motivation. As such, test taker motivation will not be used as part of the independent variables, but as a manipulation check to see whether test takers may have different levels of test-taking motivation given the presence or absence of item preknowledge.

Effects of item sharing: Cognitive constructs

It is not clear that item sharing will increase score equally across different cognitive abilities. While there are several types of cognitive ability measures, reasoning was selected in this study for its frequent usage in academic and occupational testing. As reasoning can be expressed in various contexts such as numerical, verbal, pictorial, or spatial, the study will use two facets of reasoning, numerical and verbal. Thus, measures of these two facets will be used to examine the effects of item sharing.

Numerical reasoning.

The Numerical Reasoning test (CareerHarmony, 2003) asks test takers to examine a series of numbers arranged according to a certain rule. A sample item looks like:

5, 7, 14, 16, 32, 34, 68

Then test takers are asked to choose among a set of 5 possible options to determine which option whose corresponding number best continues the series.

Numerical reasoning relies on strategies for determining computational rules, and there is no need to memorize numbers in order to solve the questions. As far as test coaching is concerned, if the strategy for finding the rule has been disclosed through information exchange and item sharing, test performance will likely be improved from

the baseline condition. Therefore, the influence of general test coaching (e.g., strategy) could be greater than obtaining specific item preknowledge in benefiting from item sharing.

Verbal reasoning.

The Verbal Reasoning test (CareerHarmony, 2003) asks test takers to first observe a pair of words that are related to one another. A sample item looks like this:

hospital : nurse

Test takers will first need to determine the relationship of the pair (e.g., a nurse works in the hospital). Then they are presented with five additional pairs of words, and they must select the pair that best expresses the relationship conveyed by the original pair. Here, the order of the words in the pairs is important. For example, the option "waiter : restaurant" is not as good as the option "bank : teller", since the former pair has an incorrect ordering of words, even though the relationship of the pair makes sense.

Verbal reasoning focuses on prior knowledge of words, as well as the relationship between pairs of words. A test taker will need to put the words into working memory, and could potentially be more likely to disclose such information to future test takers. As far as test coaching is concerned, if the strategy to find the relationship such as "pay attention to the order of words" has been disclosed through information exchange and item sharing, test performance will likely be improved from baseline condition. In this case, both general test coaching and specific item information can contribute to benefiting from item sharing.

Effects on Numerical and Verbal Reasoning

Will the scores of numerical and verbal reasoning items increase the same way? Probably not. Previous research on test coaching, particularly SAT coaching studies (see Appendix A), suggests that SAT-M items are more likely to be influenced by test coaching than SAT-V items. One possible explanation is that coaching familiarizes test takers with rules for mathematical algorithms, while the meaning of words can only be learned through prior knowledge. If a test taker knows a reasonable number of rules for numerical reasoning items, it may be easier to determine the next number in a series. For instance, participants may learn that the relationship in the series depends on the pairs in the series, and hence improve their test scores.

Another possible explanation for a larger coaching effect on numerical items is that each number represents only a single computational expression. For verbal items, a word can have several meanings, and picking the wrong interpretation may result in establishing an invalid word relation. Therefore, information about verbal items obtained through test coaching may be less helpful, if the test taker has a limited vocabulary, or lacks prior knowledge about the words. While it is possible to guess the meaning of a word through the primary lexical unit of a word such as its root, the lexical rules are more difficult to interpret than numerical computational rules. Therefore, it is hypothesized that the magnitude of numerical reasoning test score increases will be larger than verbal reasoning test score increases.

Study variables

The dependent variable will provide an index of the degree to which a research participant learned from item sharing. It will be inferred from baseline and coached test scores for the Numerical Reasoning and Verbal Reasoning tests. As such, two parallel forms will be assembled and given before and after information exchange. Both forms will have items that are comparable in difficulty. The difference score between the two parallel forms will be the dependent variable of interest.

As the study is interested in casual benefiting from item sharing, using a comprehensive "cheat sheet" to provide item preknowledge would not be feasible. As it is unlikely that previous test takers will have perfect memory, a casual benefiting from item sharing scenario is probably closer to real-life situations.

As the study utilizes a within-subjects design, it could be argued that any score increases might be due to practice, as opposed to benefiting from item sharing. However, prior research (Powers, 1986) using meta-analysis found within-test practice effect sizes were generally smaller than test preparation studies. As participants will respond to two parallel forms within an hour, any score increase is more likely to be due to benefiting from item sharing, as opposed to simply practice in a short period of time. Importantly, the difference scores obtained from control group will be used to examine whether the experimental manipulation is effective.

The independent variables include the Big Five personality variables, namely Extraversion, Conscientiousness, Agreeableness, Emotional Stability, and Openness to Experience, an integrity scale measuring Honesty, Integrity, and Authenticity, and the experimental manipulation. Personality and Integrity tests will be administered in a way

that they do not look like they are related to the target behavior of interest (i.e., item sharing).

Personality and integrity tests are often used in employee selection, and they are good predictors of job performance (Schmidt & Hunter, 1998; Barrick & Mount, 1991). It is very likely that research participants may have responded to such questionnaires in the past and are familiar with the types of questions asked. As the study is designed to simulate a real job application situation, measures that are not valid predictors of job performance are excluded.

.

Research questions

The first research question is a manipulation check on benefiting from item sharing, to see whether there is a difference between the experiment group and the control group.

Research Question 1: Does item sharing influence test performance?

The second research question involves asking which predictor variable best explains benefiting from item sharing. Items measuring the Five-Factor Model of personality are broad and intended to measure general personality constructs. Personality based integrity tests are designed to measure the tendency to behave in unethical ways (Berry, Sackett, & Wiemann, 2007). As such, it is important to see whether the broad, general measures of personality or the specific, personality based integrity test is more closely related to benefiting from item sharing. It is hypothesized that the integrity test will have greater predictive power than personality measures.

Research Question 2: Are integrity test items more likely to explain benefiting from item sharing than personality measures?

As the constructs of the Big Five model of personality may have different degrees of correlations with the willingness to engage in item sharing, it is important to see how they explain benefiting from item sharing. Based on the literature review, it is hypothesized that Conscientiousness and Emotional Stability will be better predictors than Extraversion, Agreeableness, and Openness to Experience.

Research Question 3: Which personality construct(s) best explain benefiting from test coaching?

In addition, it would also be of particular interest to examine whether those independent variables can identify participants in the experimental group from the control group.

Research Question 4: Is benefiting from item sharing related to Integrity and personality measures?

METHODS

Sample

A total of 400 University of Illinois undergraduates from the Psychology 100 subject pool were recruited for the study. Among them, 247 were in the experimental condition, with 45.2% male, 64% white, 11.7% Asian, 5.7% Hispanic, and 5.3% black. Their average undergraduate GPA was 3.24. The remaining 153 participants were in the control group, with 37.3% male, 47.7% white, 30.7% Asian, 10.5% Hispanic, and 2% black. Their average undergraduate GPA was 3.19. They received course credit in return for their participation.

Measures

Questionnaire 1 consisted of a 9-item self-report measure of integrity. The scale was obtained from the Honesty/Integrity/Authenticity scale of the International Personality Item Pool (IPIP; Goldberg et al, 2006). It takes less than 3 minutes to complete Questionnaire 1.

Two parallel forms of a cognitive ability test, Form A and Form B, were administered. Each form consists of two subtests, Numerical Reasoning and Verbal Reasoning. Each subtest has 15 items, and about 12 minutes in total were needed to complete the two subtests. The cognitive ability test has been administered to applicants around the world and has been found to have good criterion-related validity for a variety of jobs (CareerHarmony, 2003).

One form of the personality inventory, obtained from the IPIP (Goldberg et al., 2006), was also administered before the information exchange. The scale is composed of

50 items measuring the Big Five personality factors, and takes about 5 minutes to complete.

The Psychology undergraduate subject pool is ideal for the purpose of the study because items on the cognitive ability test are similar to ACT/SAT questions and students should be familiar with these types of test materials and have prior experience with such items.

Procedure

For the experimental group, participants were randomly assigned to groups of 2 to 6 after signing the informed consent. Within each group, about half of the participants were randomly assigned to receive Form A in Exercise 1, and the remaining participants received Form B in Exercise 1.

All participants first responded to Questionnaire 1, which includes the integrity measure. They continued to Exercise 1, which included one form of the 15-item numerical reasoning test, one form of the 15-item verbal reasoning test, and a 50-item personality inventory.

The order of presenting cognitive and personality items was counterbalanced, such that approximate 1/3 of participants took the numerical reasoning test first, another 1/3 took the verbal reasoning test first, and the remaining 1/3 of participants took the personality inventory first. The order of these 3 measures was counterbalanced to reduce the order effect on memory recall.

After 20 minutes, participants were asked to stop Exercise 1 and to start exchanging information about the test they just completed. They were asked to recall any

information that may help other students do well on the test. They were told that they could talk about anything they wanted, for instance, they could talk about questions and answers that appeared on Exercise 1, or strategies they used. They had about 5 minutes to exchange information.

Then they spent approximately 20 minutes working on Exercise 2. Participants who took Form A in Exercise 1 worked on parallel Form B in Exercise 2, and vice versa. When they finished Exercise 2, they continued to Questionnaire 2. It included questions about their confidence in answering items correctly, the quality of the information exchange, their motivation for passing out information, demographic variables, and why or why not test information was exchanged. It took less than 3 minutes to complete the last part of the experiment. The total time commitment for the participants was less than 50 minutes.

The control group participants responded to the same materials as in the experimental group. As the control group was designed to measure the effectiveness of the experimental manipulation, they did not engage in information exchange and simply took the parallel forms.

RESULTS

Descriptive statistics for cognitive ability measures

Before examining the effects of item sharing, means and standard deviations of the baseline test scores were first compared within each study condition to examine the equivalence of parallel forms. Table 1 presents the reliabilities, means, standard deviations, and *t*-test comparisons of means before item sharing (i.e., Exercise 1). Both forms had comparable reliability, ranging between .62 and .87 for the experimental condition, and between .75 and .95 for the control condition.

Before item sharing.

For the experimental condition, both forms of the Numerical Reasoning scale had comparable means and standard deviations (M = 13.72, SD = 2.36; and M = 13.65, SD = 2.20). The score difference between the two parallel forms was not significant (t = -.24, p = .81), and the effect size was small (d = -.03). Similarly, both forms of the Verbal Reasoning scale had comparable means and standard deviations (M = 12.16, SD = 2.37; and M = 12.15, SD = 2.32). The score difference between the two parallel forms was not significant (t = -.03, p = .97), with a very small effect size (d = -.01). The composite General Cognitive Ability measure (= Verbal Reasoning + Numerical Reasoning) also had comparable means and standard deviations (M = 25.88, SD = 4.02; and M = 25.80, SD = 3.52). Again, the score difference between the two parallel forms was not significant (t = -.17, p = .87), with a very small effect size (d = -.02). Therefore, in the experimental condition, both forms had similar means and standard deviations.

Next, the control condition test scores were compared between the two parallel forms. Both forms of the Numerical Reasoning scale had comparable means and standard

deviations (M = 13.04, SD = 3.05; and M = 13.57, SD = 2.99). The score difference between the two parallel forms was not significant (t = 1.09, p = .28), and the effect size was small (d = .18). Similarly, both forms of the Verbal Reasoning scale had comparable means and standard deviations (M = 11.49, SD = 2.87; and M = 11.82, SD = 3.00). The score difference between the two parallel forms was also not significant (t = .70, p = .49), with a small effect size (d = .11). The composite General Cognitive Ability measure also had comparable means and standard deviations (M = 24.53, SD = 4.91; and M = 25.38, SD = 4.75), with non-significant score differences (t = 1.09, p = .28) and a small effect size (d = .18). Therefore, in the control condition, both forms had similar means and standard deviations. Insert Table 1 around here.

After item sharing.

Table 2 presents the reliabilities, means, standard deviations, and *t*-test comparisons of means after benefiting from item sharing (i.e., Exercise 2). After benefiting from item sharing, both forms had comparable reliability, ranging between .66 and .85 for the experimental condition, and between .84 and .91 for the control condition.

For the Numerical Reasoning test administered to the experimental group, both forms still had comparable means and standard deviations (M = 13.99, SD = 1.61; and M = 13.87, SD = 2.31). The score difference between the two parallel forms was not significant (t = -.47, p = .64), and the effect size was small (d = -.06). Similarly, after item sharing, both forms of the Verbal Reasoning scale had slightly increased means and slightly decreased standard deviations (M = 12.43, SD = 2.07; and M = 12.49, SD = 2.11). The score difference between the two parallel forms was not significant (t = .23, p = .82), and the effect size was also small (d = .03). The composite General Cognitive Ability

measure also had comparable means and standard deviations (M = 26.42, SD = 3.16; and M = 26.36, SD = 3.65). Again, the score difference between the two parallel forms was not significant (t = -.14, p = .89), with a very small effect size (d = -.02). Therefore, in the experimental condition, even with the manipulation of test coaching, both forms had similar means and standard deviations.

Similar to the experimental condition, parallel forms of the Numerical Reasoning test in the control condition had comparable means and standard deviations (M = 12.79, SD = 3.98; and M = 13.54, SD = 3.05). The score difference between the two parallel Numerical Reasoning forms was not significant (t = 1.31, p = .19), and the effect size was small (d = .21). For the Verbal Reasoning scale, the means and standard deviations were comparable (M = 11.05, SD = 3.46; and M = 10.88, SD = 3.67), with a non-significant score difference (t = -.29, p = .77) and a small effect size (d = -.05). Insert Table 2 around here.

Test score differences.

As participants in both conditions had comparable means and standard deviations with small effect sizes, test score differences between the two parallel forms were computed. Table 3 presents the means, standard deviations, and t-test comparisons of means of difference scores.

For the experimental condition, the mean score differences for the Numerical Reasoning scales were not significantly larger than 0 (M = .27, SD = 2.07, t = 1.48, p = .07; and M = .22, SD = 2.07, t = 1.17, p = .12), and the effect sizes were small (d = .19 and .15). Similarly, for the Verbal Reasoning scales, the score differences were not significant larger than 0 for both conditions of the experimental group (M = .27, SD =

2.30, t = 1.29, p = .10; and M = .34, SD = 2.37, t = 1.60, p = .06), with small effect sizes (d = .16 and .20). However, the composite General Cognitive Ability measure was significantly greater than 0 (M = .54, SD = 3.50, t = 1.72, p = .05; and M = .56, SD = 3.11, t = 2.00, p = .03), with small effect sizes (d = .22 and .26) for one-tailed tests of the score increase.

For the control group, there was no experimental manipulation. Interestingly, scores declined in Exercise 2. For the Numerical Reasoning scales, the score difference was not significantly different from 0 for both groups (M = -.25, SD = 4.46, t = -.49, p = .63; and M = -.03, SD = 2.91, t = -.08, p = .94), and the effect sizes were small (d = -.08 and -.01). However, for the Verbal Reasoning scales, results were inconsistent: One condition saw no difference from zero and the other was significantly different from zero (M = -.44, SD = 3.56, t = -1.09, p = .28; and M = -.93, SD = 3.60, t = 2.26, p = .03), and the effect sizes ranged from small to moderate (d = -.18 and -.37). Results from the composite General Cognitive Ability measure were similar to the Verbal Reasoning scales (M = .69, SD = 5.56, t = -1.09, p = .28; and M = -.96, SD = 3.49, t = -2.40, p = .02), with small to moderate effect sizes (d = -.18 and -.39). These results suggest a possible fatigue effect. Insert Table 3 around Here.

Test taker motivation.

In this study, test taker motivation was evaluated using the common items of the post-test questionnaire given to the experimental and control groups. For the experimental condition, the mean and standard deviations were slightly smaller than the control group (M = 12.1, SD = 1.73, and M = 12.48, SD = 2.24, respectively), and the test taker motivation score difference was not significantly different from 0 (t = 1.83, p = .07),

with a small effect size (d = .19). This indicates that the test taker motivation was similar between the experimental and control groups.

Pooled test scores.

Given the extent of parallelism from the two forms, scores were pooled. Table 4 presents the reliabilities, means, and standard deviations of the dependent and independent variables used in this study for the experimental and control conditions. The first three rows presents the pooled results after benefiting from item sharing (i.e., Exercise 2). Participants in the experimental condition had significantly higher pooled test scores than the control condition in the composite General Cognitive Ability (t = 4.99, p = .00), Numerical Reasoning (t = 2.78, p = .01), and Verbal Reasoning scales (t = 5.26, p = .00), with small to moderate effect sizes (d = .27 to .51).

The next three rows present the pooled results of comparing the difference scores between the two exercises across the experimental and control conditions. Here, only the pooled composite General Cognitive Ability (t = 3.43, p = .00), and Verbal Reasoning scales (t = 3.31, p = .00) had significantly difference scores, with small to moderate effect sizes (d = .34 and .32). The score difference in Numerical Reasoning (t = 1.33, p = .18) was not significant, with a small effect size (d = .13). These results suggest that test coaching effected scales differently, with the effect on Verbal Reasoning items more than twice as large as the effect on Numerical Reasoning items. When the composite General Cognitive Ability scale is considered, the effect of benefiting from item sharing is similar to the effect on the Verbal Reasoning scale. Insert Table 4 around here.

Research Question 1: Does item sharing influence test performance?

To answer the first research question, test scores from the experimental condition and control condition were compared using pooled test scores. For the Numerical Reasoning scale, the effect was significant but small ($t = 2.78$, $p = .01$, $d = .27$). For the Verbal Reasoning scale, the effect was larger ($t = 5.26$, $p = .00$, $d = .51$). A similar result was found on the composite General Cognitive Ability scale ($t = 4.99$, $p = .00$, $d = .48$). As such, research question 1 was answered affirmatively.

Descriptive statistics for integrity and personality measures

The 9-item Integrity scale had a mean of 37.46 (about 7.20 positively endorsed items) and standard deviation of 4.37 for the experimental group, and a mean of 35.99 (about 6.76 positively endorsed items) and standard deviation of 5.13 for the control group. The score difference was significant ($t = 3.06$, $p < .01$), with a moderate effect size ($d = .31$). Among the Big Five constructs, Agreeableness was most frequently endorsed ($M = 39.72$, $SD = 5.44$ for the experimental group and $M = 38.52$, $SD = 5.79$ for the control group), followed by Openness to Experience ($M = 36.92$, $SD = 4.98$ for the experimental group and $M = 36.39$, $SD = 5.88$ for the control group). All five personality scales had small effect sizes for the difference between the control and experimental groups, ranging from -.04 to .23.

Scale correlations

Table 5 shows the correlations for dependent and independent variables for the experimental and control groups. Undergraduate GPA was not correlated with the

General Cognitive ability score difference (GD). Integrity was significantly correlated with General Cognitive ability score difference (r = .13, p = .04). This means the effect of benefiting from item sharing may be partially explained by Integrity.

For the control group, among the Big Five personality scales, only Conscientiousness and Agreeableness were significantly correlated with the General Cognitive ability difference score (r = .16, p = .01; and r = .16, p = .01). For the control group, none of the independent variables were significantly correlated with the dependent variable. Insert Table 5 around here.

Multiple regression with the experimental group

The first set of regression analyses was performed using the experimental sample only. As participants in the experimental condition were subject to various degrees of test coaching, results presented here could be similar to those in settings with low to moderate stakes, e.g., unproctored Internet testing, where some levels of cheating could occur easily yet be difficult to detect. Table 6 displays the unstandardized regression coefficients (B) and intercept, the standardized regression coefficients (β), R^2, and adjusted R^2. Insert Table 6 around here.

First regression model: General Cognitive score difference.

The first model used the General Cognitive score difference (GD) as the dependent variable. As specific personality constructs have different degrees of correlations with the dependent variable, it is of particular interest to examine all 5 personality subscales, namely Extraversion (EXT), Conscientiousness (CON),

Agreeableness (AGR), Emotional Stability (EMO), and Openness to Experience (OPE). Thus, the first regression equation included Undergraduate GPA, Integrity, EXT, CON, AGR, EMO, and OPE as independent variables.

The first regression model showed significant predictability, with the *R* significantly different greater than zero ($R = .31$, $F = 3.36$, $p = .01$), with $R^2 = .10$. The adjusted R^2 value of .07 indicates that approximate 7% of the variability was predicted by Undergraduate GPA, Integrity, Extraversion, Conscientiousness, Agreeableness, Emotional Stability, and Openness to Experience combined.

Among the independent variables, Emotional Stability showed the largest and most significant contribution to the model, with $t = -2.55$ and $p = .01$. Conscientiousness was also significant, with $t = 2.36$ and $p = .02$. Both had large standardized coefficients. The remaining independent variables were not significant.

As General Cognitive score difference (GD) consists of the Numerical Reasoning score difference (ND) and the Verbal Reasoning score difference (VD), two additional regression equations were formulated to use ND and VD as the dependent variable, respectively, to explore which independent variable best explains benefiting from item sharing.

Second regression model: Numerical Reasoning score difference.

Model 2 used the Numerical Reasoning score difference (ND) as the dependent variable, and Undergraduate GPA, Integrity, EXT, CON, AGR, EMO, and OPE as independent variables. This regression model showed no significant predictability and

had a low multiple correlation (R = .22, F = 1.67, p = .12), although Emotional Stability again had the largest standardized coefficient that was significant (t = -2.52, p = .01).

Third regression model: Verbal Reasoning score difference.

Model 3 used the Verbal Reasoning score difference (VD) as the dependent variable and the same set of independent variables in Model 1 and 2. Here, the third regression model had an overall R = .32, which was significant (F = 3.46, p < .01). Conscientiousness had the largest standardized coefficient (t = 3.42, p < .01).

Research Question 2: Are integrity test items more likely to explain benefiting from item sharing than personality measures?

Across all three regression models in the experimental group, the Integrity scale was not a significant predictor of benefiting from item sharing.

Research Question 3: Which personality construct best explains benefiting from item sharing?

Using the General Cognitive score difference (GD) as the dependent variable, both Emotional Stability and Conscientiousness contributed significantly to the model. As such, Emotional Stability and Conscientiousness better explain benefiting from item sharing than other personality constructs that were not significant. Further analysis showed that Conscientiousness was the best predictor of the Verbal Reasoning score difference (VD), while Emotional Stability was the best for explaining the Numerical Reasoning score difference (ND).

Multiple regression with the control group

Another set of regression analyses was performed using the control group only, using Undergraduate GPA, Integrity, EXT, CON, AGR, EMO, and OPE as independent variables. While both the experimental group and control group had similar descriptive statistics, the independent variables were unable to explain the score differences observed in the control group.

Regardless of which dependent variable was used, none of the independent variables was able to contribute to the regression model significantly. The unstandardized regression coefficients (*B*) and intercept, the standardized regression coefficients (β), R^2, and adjusted R^2 for the control group are shown in Table 7. Insert Table 7 around here.

Multiple regression with the experimental and control groups combined

The third set of regression analyses was performed using both the experimental and control group samples. Results presented here could be used to see whether the independent variables are able to differentiate participants with or without item sharing. That is, these analyses consider whether the effects of item sharing can be predicted using Integrity and personality measures.

Table 8 shows the correlations for dependent and independent variables for the experimental and control groups combined. The General Cognitive ability score difference (GD) and the Verbal Reasoning score difference (VD) were significantly correlated with Extraversion, Agreeableness, Conscientiousness, and Openness to

Experience. The Numerical Reasoning score difference (ND) was not significantly correlated with any of the independent variables. Insert Table 8 around here.

First regression model: General Cognitive score difference.

The first regression model utilized General Cognitive score difference (GD) as the dependent variable and Undergraduate GPA, Integrity, Extraversion, Conscientiousness, Agreeableness, Emotional Stability, Openness to Experience, and Group Membership as independent variables. The regression was significant (R = .23, F = 2.77, p = .01), with R^2 = .05 and an adjusted R^2 value of .03. The only significant regression coefficient was Group Membership (t = - 3.14 and p < .01), which also had the largest standardized coefficient β, followed by Emotional Stability, and Conscientiousness. Table 9 displays the unstandardized regression coefficients (B) and intercept, the standardized regression coefficients (β), R^2, and adjusted R^2. Insert Table 9 around here.

Second regression model: Numerical Reasoning score difference.

The second regression model utilized the Numerical Reasoning score difference (ND) as the dependent variable. Results showed that the independent variables had a low multiple correlation (R = .16) with the dependent variable, and the regression was not significant (F = 1.29, p = .25). Here, Emotional Stability had the largest standardized coefficient β, followed by Openness to Experience, Undergraduate GPA, and Group Membership. However, none of the regression coefficients was significant.

Third regression model: Verbal Reasoning score difference.

Finally, the third regression model utilized the Verbal Reasoning score difference (VD) as the dependent variable. The regression was significant ($R = .26$, $F = 3.53$, $p < .01$), with $R^2 = .07$ with an adjusted R^2 value of .05. The regression coefficient for Group Membership was significant ($t = -3.06$ and $p < .01$) with the largest standardized coefficient β, followed by Conscientiousness ($t = 2.43$ and $p = .02$).

Research Question 4: Can benefiting from item sharing be identified from Integrity and personality measures?

When the experimental and control group data were combined, regression models using the General Cognitive score difference (GD) or the Verbal Reasoning score difference (VD) were significant, with Group Membership having the largest standardized regression coefficient in both cases. Conscientiousness was significant in the regression equation for the Verbal Reasoning score difference (VD). No Integrity or personality measure was related to the General Cognitive score difference (GD) or the Numerical Reasoning score difference (ND). Research question 4 was partially answered affirmatively. In other words, only the personality construct of Conscientiousness was able to identify test takers who benefited from item sharing, and only on the Verbal Reasoning score difference (VD) when the regression model included a dummy variable for the experimental manipulation.

DISCUSSION

Manipulation of item sharing

In this study, item sharing was operationalized as a casual information exchange about questions and answers on a test in the hope of getting a better test score. As previous studies on item preknowledge utilized either organized item theft or simulations based on various assumptions, it is of particular interest to see if this casual item sharing scenario has an effect on test performance. Results showed that there were significant score differences between the experimental and the control conditions, with moderate effect sizes around .42 (averaged across parallel forms). Such between group score differences showed that casual item sharing through the use of item preknowledge can have an effect on test performance, even though participants were not asked specifically to memorize and share the test information.

In a repeated measure design, there is always a concern about within-subjects score differences. For the control group participants, test scores decreased slightly (on average, -.55). This could due to external factors such as fatigue and decreased motivation. As such a score decrease was not significant and the effect sizes were small (on average, d = -.20), there is support for the notion that within-test practice effects played a very minor role in score change (Powers, 1986) in this study. For the experimental group participants, while their test scores increased slightly (on average, .37), the score change was not significant, and the effect size was small as well (on average, d = .20). While it can be argued that such score fluctuation within subjects may represent practice effects, it should be noted that item sharing contributed to the score difference. Indeed, when test scores of the experimental group and the control group are

compared, a significant score difference and small to moderate effect sizes showed that item sharing did indeed play a role.

Casual versus organized item sharing

Organized item sharing refers to systematic item theft in which a large portion of the item pool is exposed such that future test takers can benefit from the item preknowledge disclosed by previous test takers. In a previous study on an organized item theft (Do, Brummel, Chuah, & Drasgow, 2006), when participants were given a cheat sheet with either correct or incorrect responses, their scores were significantly different from the control group, with moderate to large effect sizes. It was also found that regardless of item difficulty, participants relied more on the item preknowledge provided, as opposed to utilizing their own ability to answer questions. However, in occupational settings, unless the stakes for the selection are high, it appears unlikely that test takers will engage in much effort toward organized item theft. Therefore, it seems more likely that casual item sharing may occur in employee selection.

The current study utilized a casual item sharing scenario and found small to moderate effect sizes on score differences. Participants were not given a cheat sheet, and the amount of information shared was totally up to the participants. During the study, participants in the experimental condition were asked what items were disclosed to them during the information exchange. Results showed that on average, 2.57 numerical reasoning items and 2 verbal reasoning items were shared. This translates to approximately 15% of the items being compromised. What is the practical implication of this item exposure rate? Simulations performed by Stocking, Ward, and Potenza (1998)

found that with 10% of the items compromised, the increases in test scores were rather small, but when 20% of the item pool was disclosed, the effects were large enough to have practical consequences. In academic and licensing exams, as the stakes are much higher than employment tests, it is more likely to observe large scale, organized item theft. In employee selection, unless applicants are motivated to perform well, organized item theft seems unlikely. However, even in this casual item sharing scenario, test scores were increased significantly and approximate 15% of the items were disclosed. Thus, benchmarking against Stocking et al.'s results, a casual conspiracy of the sort studied here may lead to serious consequences. With the growing popularity of unproctored Internet testing, it is even more important to pay attention to item sharing activities. Methods to reduce item exposure to less than 10%, include increasing the size of the item pool, and shortening the length of test windows.

Numerical versus verbal reasoning

While the cognitive processes underlying numerical reasoning and verbal reasoning are beyond the scope of this paper, it should be noted that item sharing had somewhat different effects on the two constructs. With regards to the number of items exposed, there were slightly more disclosed items for the Numerical Reasoning scale than the Verbal Reasoning scale (2.57 versus 2). Upon first look, it appears that the score difference would be greater for Numerical reasoning scale, as about 29% more items were disclosed. However, that was not the case.

When comparing experimental and control groups, on average, the score difference was .77 for the Numerical Reasoning scale and 1.50 for the Verbal Reasoning

scale. In other words, the Verbal Reasoning scale score increase (d = .51) was about twice as large as the Numerical Reasoning scale (d = .26). As mentioned previously, there is no need to memorize numbers in order to solve numerical reasoning questions; moreover it may be more difficult to recall an exact number series. Because numerical reasoning questions rely on strategies for determining computational rules, item preknowledge such as "think of the number series in pairs as opposed to triplets or quads" may benefit future test takers more than the simple help given to answer verbal items. Providing exact answers for numerical reasoning questions could be of little help to future test takers, as it would be difficult for them to match the answer such as 24 to the question (unless there is only one option of 24 in the whole test). Strategies provided by previous test takers might be more beneficial to future test takers, and this is probably why there were more Numerical Reasoning items disclosed during information exchange than Verbal Reasoning items.

Verbal Reasoning, on the other hand, focuses on prior knowledge of words, as well as the relationship between pairs of words. Because a test taker needs to put the words into working memory, it could be easier for them to disclose accurate item information to future test takers. For example, information such as "the right answer about an airplane is cruise ship" provides specific information that could be more beneficial than strategies such as "think about not only the relationship between the pairs of words, but also the order of the word pairs". Therefore, both general test coaching as well as specific item preknowledge can lead to benefiting from item sharing.

However, when within-subject score differences are considered, the uneven score increase between the Numerical Reasoning and Verbal Reasoning disappeared. On

average, the Verbal Reasoning scale had a slightly larger score increase (d = .18) than the Numerical Reasoning scale (d = .17). While this appears to be opposite to what was hypothesized earlier in the paper, the mean differences between the Numerical Reasoning and Verbal Reasoning score differences were quite small. It appeared that if one is good at the test before item sharing, s/he will continue to be good at the test after item sharing. Item sharing helps performance a little bit, but not enough to create a huge score jump. This notion is consistent with the findings for the SAT that coaching yields about 15 to 25 point each on the Verbal and the Mathematical portions of the SAT. Such score increases add little improvement to a "typical" test taker's standing (Powers, 1993).

Benefiting from item sharing

Compared to academic and licensing exams, employment testing can utilize various selection measures to make a hiring decision (Anastasi & Urbina, 1997). This strength allow us to use other valid predictors of job performance to explain benefiting from item sharing behavior without asking overt questions about ethics or motivation that are less likely to be answered honestly in an occupational testing setting. In addition to undergraduate GPA, this study utilized integrity and five personality factors to explain benefiting from item sharing. While prior research has shown that personality based integrity tests can measure the tendency to behave in unethical ways (Berry, Sackett, & Wiemann, 2007), integrity scale scores did not correlate highly with the general cognitive score difference. As such, integrity items were able to explain very little about the dependent variable. It seems that benefiting from item sharing is not an effect that can be

explained through integrity tests, at least in the context of experiments along the lines at the one described here.

The reason why personality can explain benefiting from item sharing is probably due to the broad range of behaviors that can be predicted from the general constructs of Conscientiousness and Emotional Stability. Conscientiousness is usually considered to be the best predictor of contextual performance (Borman, Klimoski, & Ilgen, 2001), and it is correlated with goal-setting, expectancy, and self-efficacy motivation (Judge & Ilies, 2002). Neuroticism, however, was not highly related to job performance (Salgado, 1997; Judge & Bono, 2001). Nonetheless, prior research has found Neuroticism to be moderately correlated with three central aspects of motivation (Judge & Ilies, 2002). It seems that the results found in this study resemble what Judge and Ilies (2002) found for personality and performance motivation. In their meta-analysis, they found that Neuroticism and Conscientiousness were the strongest correlates of performance motivation. In employee selection, overt questions about motivation can be easily faked because there are obvious right or wrong answers to motivational questions. However, personality inventories maybe more covert and the general public usually believes that there are no right or wrong answers to personality items. Thus, it makes sense to utilize personality to predict the effect of benefiting from item sharing.

As benefiting from item sharing is best explained by Conscientiousness and Neuroticism, it is important to consider practical implications of these two constructs. Why would people engage in item sharing to improve their test performance? It could be argued that they are trying to present themselves as "ideal employees" such that their chance of being selected would be better (Schmit & Ryan, 1993). Because

conscientiousness people are usually hard working and goal oriented, they may be motivated to utilize item preknowledge to their advantage. Among the Big Five, Conscientiousness is also the most highly related to intentions and behavior (Conner & Abraham, 2001). While this study only explores casual item sharing scenarios, when several test takers realize the "true benefit" of item sharing in an employee selection context, they may become more organized and engage in item theft. In other words, "too conscientiousness" may result in individuals actively seeking test information as well as using it to present themselves as ideal employees who know all the ropes of the job even before being hired. While conscientiousness people are usually very motivated and perform well on the job, if they do not consider item sharing to be unethical and wrong, they may engage in item sharing to their advantage.

It is also interesting to see that Conscientiousness was able to explain benefiting from item sharing better for the Verbal Reasoning than the Numerical Reasoning scale. Conscientiousness people may be very motivated to perform well on the test to get the job and thus the specific item preknowledge of the Verbal Reasoning yielded a greater benefit than knowledge of the Numerical Reasoning items. As mentioned previously, in Verbal Reasoning, prior knowledge of the words could be much more useful than preknowledge of strategies to solve Numerical Reasoning items. Therefore, if Verbal Reasoning items are necessary parts of a screening tool, it might be necessary to generate word pairs in a way that more than two questions have shared words as part of the options, and thus reduce the utility of specific item information.

In general, neuroticism is negatively correlated with job performance and performance motivation. As explained by the theory of planned behavior (Ajzen, 1991),

test takers who are emotionally stable may regard using item preknowledge as a way to facilitate their performance and perceive benefiting from item sharing as a means for greater control over their test performance. They are interested in performing well to reduce negative attitudes, affect, and behavior (Judge, Van Vianen, & De Pater, 2004). In other words, the reason they benefit from item sharing is to help them gain control in the selection process.

In the experimental group, as Emotional Stability was able to explain the General Cognitive ability score difference as well as Numerical Reasoning, it seems people who are emotionally stable were less able to benefit more from item information that is related to strategies to solve questions, as opposed to specific knowledge of words as in Verbal Reasoning. From an organizational perspective, emotionally stable people may be more likely to seek to understand rules and strategies to help them perform well, as opposed to relying on a quick fix and short term solutions. As they are usually motivated, their emotional stability in general is beneficial to performance.

Finally, the current study found that benefiting from item sharing can be detected from Integrity and personality measures only when General Cognitive score difference (GD) or Verbal Reasoning score difference (VD) was used as the dependent variable, but not on Numerical Reasoning score difference (ND). Among the independent variables, Conscientiousness was the only one that contributed significantly to VD. As such, using Conscientiousness as a predictor of benefiting from item sharing in Verbal Reasoning tests might be warranted.

Practical implications of Conscientiousness

As benefiting from item sharing is best explained by Conscientiousness, yet Conscientiousness is also a good predictor of job performance, the practical implications of using Conscientiousness as part of a selection hurdle seem to send out mixed signals. On the one hand, employers like to hire or promote employees who are conscientious because they are more proficient on the job (Barrick & Mount, 1991), more likely to have career success (Judge, Higgins, Thoresen, & Barrick, 1999), more likely to performance to contribute to the social and psychological core of the organization (Borman, Klimoski, & Ilgen, 2001), and more likely to possess performance motivation such as goal-setting, expectancy, and self-efficacy motivation (Judge & Ilies, 2002). On the other hand, if the test taker is highly motivated and engages in response distortion (Salgado, Remeseiro, & Iglesias, 1996) or item sharing, it renders the validity of the test questionable.

It should be noted that prior research has showed response distortion has little effect on the validity of the test. If the instrument is construct oriented, intentional distortion of self-descriptions in a desirable way does not affect criterion-related validities (Hough, Eaton, Dunnette, Kamp, & McCloy, 1990) or the construct validity of personality measures used in real-world selection contexts (Smith & Ellingson, 2002), at least in some studies. However, when faking was high, social desirability would interact with the personality predictor and could reduce validity close to zero (White, Young, & Rumsey, 2001). Nevertheless, a meta-analysis showed impression management has been found to have little effect on the prediction of job performance for managerial jobs where interpersonal interactions are important (Viswesvaran, Ones, & Hough, 2001).

However, while the validity of the test may be unaffected by response distortion, strategies used by applicants can affect the rank ordering of job applicants and selection decisions (Levin & Zickar, 2002). For instance, overt tests have been found to be more susceptible to fake good coaching conditions (Alliger, Lilienfeld, & Mitchell, 1996). Alliger and Dwight (2000) argued that in overt tests, faking can inflate test scores by approximately one standard deviation above honest instruction (d =1.02), and coaching can inflate overt scores by a half standard deviation above fake good conditions (d =1.54). Nevertheless, on personality based covert integrity tests, faking has been found to inflate the scores by about one half standard deviation (d =.59). Comparing coaching on covert personality based integrity tests versus honest conditions, the effect size was smaller (d =.36) than fake good versus honest conditions. This indicates that with covert tests, response distortion due to faking or coaching is much less.

What are the practical implications if conscientiousness people are very motivated and benefit much from item sharing? First, as conscientiousness contribute to job performance and yet the effect of response distortion from faking or coaching conditions are moderate to small, it can be argued that the benefits of "too conscientiousness" outweighs the risks of response distortion. Second, even with the concern of response distortion, in general conscientiousness is not the single most important hurdle in personnel selection. As such, scores from other tests can be considered at the same time, and it is possible to utilize select-out rather than select-in strategies to remove obviously unqualified test takers (Mueller-Hanson, Heggestad, & Thornton, 2003). Finally, it is also possible to ask test takers to sign a non-disclosure agreement as well as endorsing an honor code to reduce the likelihood of item sharing. While the usefulness of warning and

the threat of retesting in a proctored environment are questionable, if a test taker agrees with the pre-testing consent and later violates the agreement, the company could pursuit legal action if desired.

Is this study actually about how well participants follow instructions?

While it can be argued that the score change observed on the dependent variables may simply due to how well participants followed the experimental instructions, I can propose at least three ways to argue this was not the case. First, it is possible to examine this factor by looking at test taker motivation. It should be noted that there was no significant score difference on test taker motivation between the experimental and control groups, and the effect size was small. Thus, participants in the experimental condition were not particularly motivated to engage in item sharing.

Another possible way to study how well participants follow study instruction is to give out a questionnaire at the end of the study to ask them to self-report the quality of participation. However, whether test takers will respond to such questions honestly is questionable. It is possible that some participants may respond favorably for the fear that the experimenter might punish them if they responded to the question unfavorably, because a research study only ends when the debriefing is given, and it is not uncommon for psychological studies to utilize deception. Thus, asking participants on how well they follow research instructions is not an ideal solution.

Furthermore, while it can be argued that participants in the experimental condition followed the research instruction well, it can also be argued that they did not follow the instructions well. In the study, participants were given about five minutes to exchange

information. If they follow the instructions well, they would utilize the full five minutes to share item information. However, none of the participants fully utilized the five minutes. In other words, unless participants are very motivated and followed the instructions well, it can be argued that how well they followed instructions is questionable. It should be noted that the study was designed to examine casual item sharing. If participants were very motivated and recalled several items, the present study would be measuring organized or semi-organized item sharing, not casual item sharing.

Can other variables explain benefiting from item sharing?

Even though this study found Conscientiousness to be the best predictor of benefiting from item sharing in Verbal Reasoning, it should be noted that specific facets of conscientiousness could explain more. The facets of conscientiousness include orderliness, self-control, industriousness, responsibility, traditionality, decisiveness, punctuality, formality, and virtue (Roberts, Chernyshenko, Stark, & Goldberg, 2005). However, using multiple items to measure all facets of conscientiousness may be too time consuming in a multiple hurdle selection approach, and doing so may not necessarily detect test takers who benefit from item sharing. Further research could be performed to identify key facets of conscientiousness related to item sharing

Similarly, while it is possible to utilize other variables such as justice and ethics as part of the selection package, additional study needs to be done to see how other variables may be utilized to explain item sharing. For instance, prior research has found academic dishonesty is negatively related to self-concept, cognitive development, and attitude toward cheating (Arvidson, 2004). However, whether it is appropriate to use those

instruments measuring the above mentioned constructs in a selection context is also questionable and requires further study.

Limitations and directions for future research

While this study examined two cognitive constructs, certainly there are other domains that may be subject to item sharing. For example, in the Do et al. (2006) study, a critical reasoning test with long question stems and response options was used. It would be interesting to see whether item sharing would have different effects on different types of cognitive constructs and assessment formats.

Furthermore, there can be multiple screening and selection tools utilized before a hiring decision is made. As such, different assessment tools could be used to predict benefiting from item sharing on different types of cognitive constructs. As the current independent variables explained small to moderate amounts of variance of benefiting from item sharing, further research could examine other predictors.

Items used in the current study had a short item length and could be answered quickly. However, it is possible that a longer item length may result in better item recall due to much more information about the context of the question being available to provide additional help for item memorization and recall. Future research could explore the relationships of item length and item recall.

It is possible that there are alternative ways to share information. For instance, by using a confederate in the group to induce item sharing. However, doing so may also introduce experimenter bias by focusing exclusively on the item sharing. In this study, participants in the experimental condition were asked to exchange information about the

test they just completed. They were also told that they could talk about anything they wanted. Further research could be conducted to explore alternative ways of verbal item sharing.

The study manipulated item sharing by matching future test takers with past test takers. However, it is possible that there could be other forms of information exchange. For example, it is not necessary for information exchange to be conducted verbally. As many test booklets are reused, it is possible that previous test takers may intentionally or unintentionally leave evidence of their responses. For instance, some test takers may prefer to mark their responses on the test booklet and transcribe them onto a bubble sheet later, or write the steps involved in a calculation. Such test coaching may be worth studying if test booklets are reused.

In addition, it is possible that the effect of benefiting from item sharing may be influenced by how information is exchanged. Future test takers may consider verbal information exchange as a way to "trick me" and be afraid to "get caught" and thus rely less on item preknowledge. Information obtained through erased responses on test booklets or postings on the Internet may provide a sense of anonymity and thus be considered a "safer" way to cheat than talking to previous test takers. It is unknown whether there will be differences with regards to different information exchange media.

In a computerized testing scenario, it becomes much easier for test developers to change the order of response options, as well as replacing words with synonyms or related words. Such watermark-like changes could be useful in tracking item exposure as well as detecting aberrant responses. Further research on this topic may be helpful to understand how items are memorized and recalled.

Conclusion

In continuously administered employment tests, test security may be compromised by examinees revealing test items to future test candidates. A specific type of test coaching involving disclosing items, called item sharing, can result in significant score increases, with small to moderate effect sizes. Conscientiousness and Emotional Stability were found to be the best predictors of benefiting from test coaching. Conscientiousness was able to predict the Verbal Reasoning score difference well, while Emotional Stability was better in explaining the General Cognitive ability score difference, as well as the Numerical Reasoning score difference. When test takers with or without item preknowledge were combined, the only variable other than group membership that can identify benefiting from item sharing was Conscientiousness. Limitations of the current study and directions for further research such as the facets of Conscientiousness were discussed and suggested in the hope of helping us better understand the effects of item sharing.

TABLES

Table 1. Cognitive scale descriptive statistics, before item sharing (Exercise 1)

Before item sharing			Form A				Form B					
Experimental Condition	*N*	*α*	*M*	*SD*	*N*	*α*	*M*	*SD*	*M dif*	*t*	*p*	*d*
General Cognitive	124	.76	25.88	4.02	123	.81	25.80	3.52	-.08	-.17	.87	-.02
Numerical reasoning	124	.71	13.72	2.36	123	.87	13.65	2.20	-.07	-.24	.81	-.03
Verbal reasoning	124	.62	12.16	2.37	123	.63	12.15	2.32	-.02	-.03	.97	-.01
Control Condition												
General Cognitive	77	.90	24.53	4.91	76	.91	25.38	4.75	.85	1.09	.28	.18
Numerical reasoning	77	.95	13.04	3.05	76	.92	13.57	2.99	.53	1.09	.28	.18
Verbal reasoning	77	.75	11.49	2.87	76	.80	11.82	3.00	.33	.70	.49	.11

Table 2. Cognitive scale descriptive statistics, after item sharing (Exercise 2)

After item sharing			Form B				Form A					
Experimental Condition	*N*	*α*	*M*	*SD*	*N*	*α*	*M*	*SD*	*M dif*	*t*	*p*	*d*
General Cognitive	124	.76	26.42	3.16	123	.83	26.36	3.65	-.06	-.14	.89	-.02
Numerical reasoning	124	.81	13.99	1.61	123	.85	13.87	2.31	-.12	-.47	.64	-.06
Verbal reasoning	124	.66	12.43	2.07	123	.68	12.49	2.11	.06	.23	.82	.03
Control Condition												
General Cognitive	77	.89	23.84	6.15	76	.89	24.42	5.10	.58	.64	.53	.10
Numerical reasoning	77	.91	12.79	3.98	76	.88	13.54	3.05	.75	1.31	.19	.21
Verbal reasoning	77	.85	11.05	3.46	76	.84	10.88	3.67	-.17	-.29	.77	-.05

Table 3. Cognitive scale descriptive statistics of the difference scores

Experimental Condition		*N*	*M*	*SD*	*t*	*p*	*d*
General Cognitive	Form B-A	124	.54	3.50	1.72	.09	.22
	Form A-B	123	.56	3.11	2.00	.05*	.26
Numerical reasoning	Form B-A	124	.27	2.07	1.48	.14	.19
	Form A-B	123	.22	2.07	1.17	.24	.15
Verbal reasoning	Form B-A	124	.27	2.30	1.29	.20	.16
	Form A-B	123	.34	2.37	1.60	.11	.20
Control Condition							
General Cognitive	Form B-A	77	-.69	5.56	-1.09	.28	-.18
	Form A-B	76	-.96	3.49	-2.40	.02*	-.39
Numerical reasoning	Form B-A	77	-.25	4.46	-.49	.63	-.08
	Form A-B	76	-.03	2.91	-.08	.94	-.01
Verbal reasoning	Form B-A	77	-.44	3.56	-1.09	.28	-.18
	Form A-B	76	-.93	3.60	-2.26	.03*	-.37

Table 4. Descriptive statistics comparing the experimental and control conditions.

	Control				Experimental							
Variables	*N*	*α*	*M*	*SD*	*N*	*α*	*M*	*SD*	*M dif*	*t*	*p*	*d*
General Cognitive pooled	153		24.13	5.65	247		26.39	3.41	2.26	4.99	.00	.48
Numerical Reasoning pooled	153		13.16	3.55	247		13.93	1.99	.77	2.78	.01	.27
Verbal Reasoning pooled	153		10.97	3.57	247		12.46	2.09	1.49	5.26	.00	.51
General Cognitive difference	153		-.82	4.65	247		.55	3.31	1.37	3.43	.00	.34
Numerical Reasoning difference	153		-.14	3.77	247		.25	2.07	.39	1.33	.18	.13
Verbal Reasoning difference	153		-.68	3.58	247		.30	2.34	.99	3.31	.00	.32
Undergraduate GPA	153		3.19	.48	247		3.24	.51	.05	.97	.33	.10
Integrity	153	.74	35.99	5.13	247	.73	37.46	4.37	1.47	3.06	.00	.31
Composite Personality	153	.83	175.04	19.99	247	.83	178.25	19.18	3.21	1.60	.11	.16
Extraversion	153	.75	33.63	7.36	247	.75	34.27	4.37	.64	1.09	.28	.11
Agreeableness	153	.67	38.52	5.79	247	.69	39.72	5.44	1.20	2.09	.04	.21
Conscientiousness	153	.66	32.87	6.28	247	.75	34.33	6.46	1.46	2.22	.03	.23
Emotional Stability	153	.76	33.32	7.52	247	.73	33.02	6.61	-.30	-.42	.68	-.04
Openness to Experience	153	.66	36.39	5.88	247	.68	36.92	4.98	.53	.96	.34	.10

Table 5. Correlations.

	1. GD	2. ND	3. VD	4. GPA	5. INT	6. EXT	7. AGR	8. CON	9. EMO	10. OPE
1. General Cognitive difference (GD)	-	.66**	.61**	-.04	-.04	.06	.04	.02	.04	.14
2. Numerical Reasoning difference (ND)	.71**	-	-.20*	-.11	-.03	-.07	-.02	-.06	-.05	.04
3. Verbal Reasoning difference (VD)	.78**	.12	-	-.06	-.02	.15	.07	.09	.10	.13
4. Undergraduate GPA (GPA)	.01	-.06	.07	-	.11	.07	-.07	.11	.06	.05
5. Integrity (INT)	.15*	.12	.11	.07	-	.21*	.26**	.27**	.18*	.25*
6. Extraversion (EXT)	.14*	.06	.14*	.04	.35**	-	.54**	.41**	.31**	.60**
7. Agreeableness (AGR)	.18**	.08	.19**	.09	.35**	.47**	-	.51**	.32**	.48**
8. Conscientiousness (CON)	.19**	.01	.26**	.20**	.40**	.25**	.47**	-	.20*	.40**
9. Emotional Stability (EMO)	-.04	-.10	.03	.07	.26**	.33**	.38**	.31**	-	.28**
10. Openness to Experience (OPE)	.09	.06	.08	.09	.25**	.46**	.48**	.33**	.30**	-

Note. Sample correlations for the experimental condition ($N = 247$) are below the diagonal and correlations for the control condition ($N = 153$) are above the diagonal. * $p < .05$, ** $p < .01$

Table 6. Regression models for the experimental condition

Model 1: DV: General Cognitive difference (GD)

Variables	*B*	*SE B*	*β*	*t*	*p*
Constant	-4.69	2.45		-1.91	.06
Undergraduate GPA	-.24	.45	-.04	-.53	.60
Integrity	.05	.05	.07	.95	.34
Extraversion	.04	.04	.08	1.02	.31
Agreeableness	.08	.05	.15	1.77	.08
Conscientiousness	.10	.04	.18	2.36	.02*
Emotional Stability	-.10	.04	-.19	-2.55	.01*
Openness to Experience	-.02	.05	-.03	-.35	.73

$R = .31$, $R^2 = .10$, *Adjusted* $R^2 = .07$, $F\ (7, 222) = 3.36$, $p = .01$**

Model 2: DV: Numerical Reasoning difference (ND)

Variables	*B*	*SE B*	*β*	*t*	*p*
Constant	-.71	1.56		-.45	.65
Undergraduate GPA	-.30	.29	-.07	-1.03	.31
Integrity	.06	.03	.13	1.71	.09
Extraversion	.01	.02	.03	.32	.75
Agreeableness	.04	.03	.10	1.18	.24
Conscientiousness	-.00	.03	-.01	-.09	.93
Emotional Stability	-.07	.03	-.19	-2.52	.01*
Openness to Experience	.01	.03	.03	.36	.72

$R = .22$, $R^2 = .05$, *Adjusted* $R^2 = .02$, $F\ (7, 222) = 1.67$, $p = .12$

Model 3: DV: Verbal Reasoning difference (VD)

Variables	*B*	*SE B*	*β*	*t*	*p*
Constant	-3.98	1.73		-2.31	.02*
Undergraduate GPA	.06	.32	.01	.18	.86
Integrity	-.01	.04	-.01	-.19	.85
Extraversion	.03	.03	.09	1.16	.25
Agreeableness	.05	.03	.12	1.44	.15
Conscientiousness	.11	.03	.26	3.42	.00*
Emotional Stability	-.04	.03	-.10	-1.36	.18
Openness to Experience	-.03	.03	-.06	-.82	.41

$R = .32$, $R^2 = .10$, *Adjusted* $R^2 = .08$, $F\ (7, 222) = 3.46$, $p = .00$**

Table 7. Regression models for the control condition

Model 1: DV: General Cognitive difference (GD)

Variables	*B*	*SE B*	*β*	*t*	*p*
Constant	-.67	4.42		-.15	.88
Undergraduate GPA	-.39	.82	-.04	-.48	.63
Integrity	-.08	.09	-.08	-.86	.39
Extraversion	-.02	.08	-.03	-.23	.82
Agreeableness	-.01	.08	-.01	-.10	.92
Conscientiousness	.01	.08	-.01	-.14	.89
Emotional Stability	.01	.06	.02	.21	.84
Openness to Experience	.13	.08	.18	1.66	.10

$R = .17$, $R^2 = .03$, *Adjusted* $R^2 = -.02$, $F\ (7, 145) = .59$, $p = .77$

Model 2: DV: Numerical Reasoning difference (ND)

Variables	*B*	*SE B*	*β*	*t*	*p*
Constant	3.34	3.58		.93	.35
Undergraduate GPA	-.74	.66	-.09	-1.11	.27
Integrity	-.01	.08	-.02	-.17	.86
Extraversion	-.07	.06	-.12	-1.01	.28
Agreeableness	.01	.06	.01	.12	.90
Conscientiousness	-.04	.07	-.05	-.55	.59
Emotional Stability	-.02	.05	-.04	-.46	.65
Openness to Experience	.09	.07	.15	1.41	.16

$R = .17$, $R^2 = .03$, *Adjusted* $R^2 = -.02$, $F\ (7, 145) = .64$, $p = .72$

Model 3: DV: Verbal Reasoning difference (VD)

Variables	*B*	*SE B*	*β*	*t*	*p*
Constant	-4.01	3.40		-1.18	.24
Undergraduate GPA	.35	.63	.05	.55	.69
Integrity	-.07	.07	-.08	-.93	.35
Extraversion	.05	.06	.09	.85	.40
Agreeableness	-.02	.06	-.03	-.26	.80
Conscientiousness	.02	.06	.04	.39	.69
Emotional Stability	.03	.05	.07	.75	.46
Openness to Experience	.04	.06	.07	.68	.50

$R = .19$, $R^2 = .04$, *Adjusted* $R^2 = -.01$, $F\ (7, 145) = .77$, $p = .62$

Table 8. Correlations for the experimental and control groups combined.

	1. GD	2. ND	3. VD	4. GPA	5. INT	6. EXT	7. AGR	8. CON	9. EMO	10. OPE
1. General Cognitive difference (GD)	-									
2. Numerical Reasoning difference (ND)	.67**	-								
3. Verbal Reasoning difference (VD)	.69**	-.07	-							
4. Undergraduate GPA (GPA)	-.00	-.08	.07	-						
5. Integrity (INT)	.08	.05	.06	.09	-					
6. Extraversion (EXT)	.10*	-.00	.14**	.05	.30**	-				
7. Agreeableness (AGR)	.12*	.03	.14**	.03	.32**	.50**	-			
8. Conscientiousness (CON)	.12*	-.02	.19**	.17**	.36**	.31**	.49**	-		
9. Emotional Stability (EMO)	-.00	-.07	.06	.06	.23**	.32**	.35**	.26**	-	
10. Openness to Experience (OPE)	.12*	.05	.11*	.08	.26**	.52**	.48**	.36**	.29**	-

Note. Sample correlations ($N = 400$). * $p < .05$, ** $p < .01$

Table 9. Regression models for the experimental and control groups combined

Model 1: DV: General Cognitive difference (GD)

Variables	*B*	*SE B*	*β*	*t*	*p*
Constant	-2.85	2.25		-1.27	.21
Undergraduate GPA	-.25	.42	-.03	-.60	.55
Integrity	.01	.05	.01	.25	.81
Extraversion	.02	.04	.04	.62	.53
Agreeableness	.03	.04	.05	.69	.49
Conscientiousness	.05	.04	.07	1.20	.23
Emotional Stability	-.04	.03	-.07	-1.21	.12
Openness to Experience	.04	.04	.06	1.03	.30
Group Membership	-1.26	.40	-.16	-3.14	.00*

$R = .23$, $R^2 = .05$, *Adjusted* $R^2 = .03$, $F\ (8, 391) = 2.77$, $p = .01$

Model 2: DV: Numerical Reasoning difference (ND)

Variables	*B*	*SE B*	*β*	*t*	*p*
Constant	.79	1.65		.48	.63
Undergraduate GPA	-.46	.31	-.08	-1.51	.13
Integrity	.04	.03	.06	1.17	.25
Extraversion	.02	.03	-.04	-.61	.54
Agreeableness	.02	.03	.04	.63	.53
Conscientiousness	-.02	.03	-.05	-.80	.42
Emotional Stability	-.04	.03	-.09	-1.66	.10
Openness to Experience	.04	.03	.08	1.31	.19
Group Membership	.35	.29	-.06	-1.21	.23

$R = .16$, $R^2 = .03$, *Adjusted* $R^2 = .01$, $F\ (8, 391) = 1.29$, $p = .25$

Model 3: DV: Verbal Reasoning difference (VD)

Variables	*B*	*SE B*	*β*	*t*	*p*
Constant	-3.65	1.67		-2.20	.03*
Undergraduate GPA	.21	.31	.04	.69	.49
Integrity	-.03	.03	-.05	-.83	.41
Extraversion	.04	.03	.09	1.46	.15
Agreeableness	.01	.03	.02	.31	.76
Conscientiousness	.07	.03	.14	2.43	.02*
Emotional Stability	.00	.03	.00	.02	.99
Openness to Experience	.00	.03	.01	.10	.92
Group Membership	-.90	.30	-.15	-3.06	.00*

$R = .26$, $R^2 = .07$, *Adjusted* $R^2 = .05$, $F\ (8, 391) = 3.53$, $p = .00$

REFERENCES

Abraham, C. J., & Conner, M. (2001). Efficacy of the Theory of Planned Behaviour: A meta-analytic review. *British Journal of Social Psychology, 40*, 471-499.

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes, 50*, 179-211.

Allalouf, A., & Ben-Shakhar, G. (1998). The effect of coaching on the predictive validity of scholastic aptitude tests. *Journal of Educational Measurement, 35*, 31-47.

Alliger, G. M., & Dwight, S. A. (2000). A meta-analytic investigation of the susceptibility of integrity tests to faking and coaching. *Educational and Psychological Measurement, 60*, 59-72.

Alliger, G. M., Lilienfeld, S. O., & Mitchell, K. E. (1996). The susceptibility of overt and covert integrity tests to coaching and faking. *Psychological Science, 7*, 32-39.

Anastasi, A. (1981). Coaching, test sophistication, and developed abilities. *American Psychologist, 36*, 1086-1093.

Anastasi, A., & Urbina, S. (1997). *Psychological Testing*, 7th ed. Upper Saddle River, NJ: Prentice-Hall.

Arvey, R. D., Strickland, W., Dauden, G., & Martin, C. (1990). Motivational components of tests taking. *Personnel Psychology, 43*, 695-716.

Arvidson, C. J. (2004). *The anatomy of academic dishonesty: Cognitive development, self-concept, neutralization techniques, and attitudes toward cheating*. Available from ProQuest Dissertations and Theses database. (UMI No. 3144972).

Baird, J. S., Jr. (1980). Current trends in college cheating. *Psychology in the Schools, 17,* 515-522.

Bandura, A. (1994). Self-efficacy. In V. S. Ramachaudran (Ed.), *Encyclopedia of human behavior* (Vol. 4, pp. 71-81). New York: Academic Press.

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*, 1-26.

Barrick, M. R., Mount, M. K., & Strauss, J. P. (1993). Conscientiousness and performance of sales representatives: Test of the mediating effects of goal setting. *Journal of applied Psychology, 78*, 715-722.

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

Beck, L., & Ajzen, I. (1991). Predicting dishonest actions using the Theory of Planned Behavior. *Journal of Research in Personality*, 25, 285-301.

Becker, B. J. (1990). Coaching for the Scholastic Aptitude Test: Further synthesis and appraisal. *Review of Educational Research, 60*, 373-417.

Berry, C. M., Ones, D. S., & Sackett, P. R. (2007). Interpersonal deviance, organizational deviance, and their correlates: A review and meta-analysis. *Journal of Applied Psychology, 92,* 410-424.

Berry, C. M., Sackett, P. R., & Wiemann, S. (2007). A review of recent developments in integrity test research. *Personnel Psychology, 60,* 271-301.

Briggs, D. C. (2005). Meta-analysis: A case study. *Evaluation Review, 29*, 87-127.

Borman, W. C., Klimoski, R. J., & Ilgen, D. R. (2003). Stability and change in Industrial-Organizational Psychology. In W. C. Borman, D. R. Ilgen, and R. J. Klimoski (Eds.), *Comprehensive handbook of psychology, Volume 12: Industrial and organizational psychology*. New York: Wiley.

Borman, W. C., & Motowidlo, S. J. (1993). Expanding the criterion domain to include elements of contextual performance. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection on organizations* (pp.71-98). San Francisco: Jossey-Bass.

Borman, W. C., Penner, L. A., Allen, T. D., & Motowidlo, S. J. (2001). Personality predictors of citizenship performance. *International Journal of Selection and Assessment, 9*, 52-69.

Brief, A. P., & Motowidlo, S. J. (1986). Prosocial organizational behaviors. *Academy of Management Review, 11,* 710-725.

Campbell, J. P. (1990). Modeling the performance prediction problem in industrial and organizational psychology. In M. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. 1, pp.687-732). Palo Alto, CA: Consulting Psychologists Press.

CareerHarmony (2003). *Test Information Catalog.* CareerHarmony, Ltd.

Chang, H.-H., & Ying, Z. (1999). *a*-stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 23*, 211-222.

Chang, H. H., Qian, J., & Ying, Z. (2001). *a*-stratified multistage computerized adaptive testing with *b* blocking. *Applied Psychological Measurement, 25*, 333-341.

Chang, S.-W., & Ansley, T. N. (2003). A comparative study of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 40*, 71-103.

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

Chapman, K. J., Davis, R., Toy, D., & Wright, L. (2004). Academic integrity in the business school environment: I'll get by with a little help from my friends. *Journal of Marketing Education, 26*, 236-249.

Cizek, G. J. (1999). *Cheating on tests: How to do it, detect it, and prevent it*. New Jersey, Lawrence Erlbaum Associates.

Clause, C. S., Delbridge, K., Schmitt, N., Chan, D., & Jennings, D. (2001). Test preparation activities and employment test performance. *Human Performance, 14*, 149-167.

Colquitt, A. C., & Shaw, J. C. (2005). How should organizational justice be measured? In J. Greenberg & J. A. Colquitt (Eds.), Handbook of Organizational Justice (pp. 113-152). Mahwah, NJ: Lawrance Erlbaum Associates.

Costa, P. T. Jr., & McCrae, R. R. (1992). Four ways five factors are basic. *Personality and Individual Difference, 13*, 653-665.

Dalal, R. S. (2005). A meta-analysis of the relationship between organizational citizenship behavior and counterproductive work behavior. *Journal of Applied Psychology, 90,* 1241-1255.

Davey, T., & Parshall, C. G. (1995, April). *New algorithms for item selection and exposure control with computerized adaptive testing*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

De Raad, B. & Schouwenburg, H. C. (1996). Personality in learning and education: a review. *European Journal of Personality, 10,* 303-336.

DerSimonian, R., & Laird, N. M. (1983). Evaluating the effect of coaching on SAT scores: A meta-analysis. *Harvard Educational Review, 53*, 1-15.

Do, B.-R., Brummel, B. J., Chuah, S. C., & Drasgow, F. (2006). Item preknowledge on test performance and item confidence. Paper presented at the annual conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

Do, B.-R., Shepherd, W., & Drasgow, F. (2005). *Measurement equivalence across proctored versus unproctored testing with job incumbents*. Paper presented at the annual conference of the Society for Industrial and Organizational Psychology, Los Angeles, CA.

Dobson, P. (2000). An investigation into the relationship between neuroticism, extraversion and cognitive test performance in selection. *International Journal of Selection and Assessment, 8*, 99-109.

Duffy, S. P. (2006, August 28). *Bar prep co. ordered to pay $11.9 million for copying multistate exam questions*. The Legal Intelligencer. Retrieved from http://www.law.com/jsp/law/LawArticleFriendly.jsp?id=1156511800192.

Ellis, A. P. J., & Ryan, A. M. (2003). Race and cognitive-ability test performance: The mediating effects of test preparation, test-taking strategy use and self-efficacy. *Journal of Applied Social Psychology, 33*, 2607-2629.

Eisenberger, R., & Shank, D. M. (1985). Personal work ethic and effort training affect cheating. *Journal of Personality and Social Psychology, 49*, 520-528.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. C. (2006). The International Personality Item Pool and the future of public-domain personality measures. *Journal of Research in Personality, 40*, 84-96.

Guion, R. M. (1998). *Assessment, Measurement, and Prediction for Personnel Decisions*. Mahwah, NJ: Lawrence Erlbaum.

Haladyna, T. M., Nolen, S. B., & Haas N. S. (1991). Raising standardized achievement test scores and the origins of test score pollution. *Educational Researcher, 20 (5)*, 1-7.

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

Hausknecht, J. P., Trevor, C. O., & Farr, J. L. (2002). Retaking ability tests in a selection setting: Implications for practice effects, training performance, and turnover. *Journal of Applied Psychology, 87*, 243-254.

Hetter, R. D., & Sympson, J. B. (1997). Item exposure control in CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBribe (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 141-144). Washington, DC: American Psychological Association.

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

Holden, R. R., Wood, L. L., & Tomashewski, L. (2001). Do response time limitations counteract the effect of faking on personality inventory validity? *Journal of Personality and Social Psychology, 81,* 160-169.

Holden, R. R., Kroner, D. G., Fekken, G. C., & Popham, S. M. (1992). A model of personality test item response dissimulation. *Journal of Personality and Social Psychology, 63,* 272-279.

Honan, W. H. (1995, January 4). *Computer admissions test to be given less often*. The New York Times, p. A16.

Hough, L. M. (1998). Effects of intentional distortion in personality measurement and evaluation of suggested palliatives. *Human Performance, 11,* 209-244.

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,* 581-595.

Hough, L. M., & Schneider, R. J. (1996). Personality traits, taxonomies, and applications in organizations. In K. R. Murphy (Ed.), *Individual differences and behavior in organizations* (pp. 31-38). San Francisco: Jossey-Bass.

Humphreys, L. G. (1979). The construct of general intelligence. *Intelligence, 3*, 105-120.

Hunt, S. T. (1996). Generic work behavior: An investigation into the dimensions of entry-level, hourly job performance. *Personnel Psychology, 49*, 51-83.

Hurtz, G. M., & Alliger, G. M. (2002). Influence of coaching on integrity test performance and unlikely virtues scale scores. *Human Performance, 12*, 255-273.

Hurtz, G. M., & Donovan, J. J. (2000). Personality and job performance: The Big Five revisited. *Journal of Applied Psychology, 85*, 869-879.

Jensen, L. A., Arnett, J. J., Feldman, S. S., & Cauffman, E. (2002). It's wrong, but everybody does it: Academic dishonesty among high school and college students. *Contemporary Educational Psychology, 27*, 209-228.

Johnson, P. B. (1981). Achievement motivation and success: Does the end justify the means? *Journal of Personality and Social Psychology, 40*, 374-375.

Johnston, M. J. (2006, August 17). Microsoft sues testing materials vendor. Microsoft Certified Professional Magazine Online. Retrieved from http://www.mcpmag.com/news/article.asp?EditorialsID=1015.

Judge, T. A, & Bono, J. E. (2001). Relationship of core self-evaluations traits—self-esteem, generalized self-efficacy, locus of control, and emotional stability—with job satisfaction and job performance: a meta-analysis. *Journal of Applied Psychology, 86*, 80-92.

Judge, T. A., & Bono, J. E. (2002). A rose by any other name: Are self-esteem, generalized self-efficacy, neuroticism, and locus of control indicators of a common construct? In B. W. Roberts & R. T. Hogan (Eds.), *Personality psychology in the workplace* (pp. 93–118). Washington, DC: American Psychological Association.

Judge, T. A., Higgins, C. A., Thoresen, C. J., & Barrick, M. R. (1999). The big five personality traits, general mental ability, and career success across the life span. *Personnel Psychology, 52*, 621-652.

Judge, T. A., & Ilies, R. (2002). Relationship of personality to performance motivation: A meta-analytic review. *Journal of Applied Psychology, 87*, 797-807.

Judge, T. A., Van Vianen, A. E. M, & De Pater, I. E. (2004). Emotional stability, core self-evaluations, and job outcomes: A review of the evidence and an agenda for future research. *Human Performance, 17*, 325-346.

Karlins, M., Michaels, C., & Podlogar, S. (1988). An empirical investigation of actual cheating in a large sample of undergraduates. *Research in Higher Education, 29*, 359-364.

Kingsbury, G. G., & Zara, A. R. (1984). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, 2, 359-375.

Kulik, J. A., Bangert-Drowns, R. L., & Kulik, C.-L. C. (1984). Effectiveness of coaching for aptitude tests. *Psychological Bulletin, 95*, 179-188.

Lavelle, L. (2008, June 27). GMAT cheating controversy grows. *Businessweek*. Retrieved from http://www.businessweek.com/bschools/content/jun2008/bs20080627_391632.htm.

Law Student Journal (2006). PMBR Ordered to Pay $11.9M for Copying Multistate Exam Questions. *Law Student Journal, 12 (1),* 1-9.

Levin, R. A., & Zickar, M. J. (2002). Investigating self-presentation, lies, and bullshit: Understanding faking and its effects on selection decisions using theory, field research, and simulation. In J. M. Brett & F. Drasgow (Eds), *The psychology of work: Theoretically based empirical research* (pp. 253-276). Mahwah, NJ: Lawrence Erlbaum Associates.

Levy, F. (2008, July 1). GMAT scandal has MBA students sweating. *Businessweek*. Retrieved from http://www.businessweek.com/bschools/content/jul2008/bs2008071_278439.htm.

Lounsbury, J. W., Steel, R. P., Loveland, J. M., & Gibson, L. W. (2004). An investigation of personality traits in relation to adolescent school absenteeism. *Journal of Youth and Adolescence, 33*, 457–466.

Maurer, T., Solamon, J., & Troxtel, D. (1998). Relationship of coaching with performance in situational employment interviews. *Journal of Applied Psychology, 83*, 128-136.

McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 223-236). New York: Academic Press.

McCabe, D. L. (1992). The influence of situational ethics on cheating among college students. *Sociological Inquiry, 62*, 365-374.

McCrae, R. R. (1996). Social consequences of experiential openness. *Psychological Bulletin, 120*, 323-337.

McLeod, L., Lewis, C., & Thissen, D. (2003). A Bayesian method for the detection of item preknowledge in computerized adaptive testing. *Applied Psychological Measurement, 27*, 121-137.

Mead, A. D., & Drasgow, F. (1993). *Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis*. Psychological Bulletin, 114, 449-458.

Messick, S., & Jungblut, A. (1981). Time and method in coaching for the SAT. *Psychological Bulletin, 89*, 191-216.

Motowidlo, S. J. (2003). Job performance. In W. C. Borman, D. R. Ilgen, and R. J. Klimoski (Eds.), *Comprehensive handbook of psychology, Volume 12: Industrial and organizational psychology*. New York: Wiley.

Mount, M. K., Barrick, M. R., & Stewart, G. L. (1994). Validity of observer rating of the Big five personality factors. *Human Performance, 11*, 145-165.

Mount, M. K., Witt, L. A., & Barrick, M. R. (2000). Incremental validity of empirically keyed biodata scales over GMA and the five factor personality constructs. *Personnel Psychology*, 53, 299-323.

Mueller-Hanson, R., Heggestad, E. D., & Thornton, G. C. (2003). Faking and selection: Considering the use of personality from select-in and select-out perspectives. *Journal of Applied Psychology, 88,* 348-355.

Murphy, K. R. (1989). Dimensions of job performance. In R. Dillon and J. Pelligrino (Eds.), *Testing: Applied and theoretical perspectives*. New York: Praeger.

Murphy, K. R. (2002). Can conflicting perspectives on the role of g in personnel selection be resolved? *Human Performance*, 15, 173-186.

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,* 150-162.

Nathan, D. (2006, Sepetember 7). Bar review prep company must pay $12 million for infringement. *Intellectual Property Litigation Reporter*, *13* (11). Andrews Publications. Retrieved from http://news.findlaw.com/andrews/bt/int/20060907/20060907_barreview.html.

Nonis, S. A., & Swift, C. O. (1998). Deterring Cheating Behavior in the Marketing Classroom: An Analysis of the Effects of Demographic, Attitudes, and In-Class Deterrent Strategies. *Journal of Marketing Education, 20,* 188-199.

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*, 656-664.

Organ, D. W. (1988). *Organizational citizenship behavior: The good soldier syndrome*. Lexington, MA: Lexington Books.

Organ, D. W. (1997). Organizational citizenship behavior: It's construct clean-up time. *Human Performance, 10*, 85-97.

Organ, D. W, McFall, J. B. (2004). Personality and Citizenship Behavior in Organizations. In B. Schneider, D. B. Smith, and D. Brent (Eds.), *Personality and organizations* (pp. 291-314). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Organ, D. W., & Ryan, K. (1995). A meta-analytic review of attitudinal and dispositional predictors of organizational citizenship behavior. *Personnel Psychology, 48*, 775-802.

Parshall, C. G., Davey, T., & Nering, M. L. (1998, April). *Test development exposure control for adaptive testing*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin, 135*, 322-338.

Porter, C. O., Hollenback, J. R., Ilgen, D. R., Ellis, A. P. J., West, B. J., & Moon, H. (2003). Backing up behaviors in teams: The role of personality and the legitimacy of need. *Journal of Applied Psychology, 88*, 391-403.

Powers, D. E. (1986). Relations of test item characteristics to test preparation/test practice effects: A quantitative summary. *Psychological Bulletin, 100*, 67-77.

Powers, D. E. (1993). Coaching for the SAT: A summary of the summaries and an update. *Educational Measurement: Issues and Practice, 12*, 24-30, 39.

Powers, D. E., & Rock, D. A. (1999). Effects of coaching on SAT I: Reasoning test scores. *Journal of Educational Measurement, 36*, 93-118.

Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin, 130*, 271–288.

Roberts, B. W., Chernyshenko, O., Stark, S. & Goldberg, L. (2005). The structure of conscientiousness: An empirical investigation based on seven major personality questionnaires. *Personnel Psychology*, 58, 103-139.

Robertson, I. T., Baron, H., Gibbons, P., MacIver, R., & Nyfield, G. (2000). Conscientiousness and managerial performance. *Journal of Occupational and Organizational Psychology, 73*, 171-180.

Robie, C., Curtin, P. J., Foster, C. T., Phillips, H. L. IV, Zbylut, M., & Ttrick, L. E. (2000). The effect of coaching on the utility of response latencies in detecting fakes on a personality measure. *Canadian Journal of Behavioural Science, 32*, 226-233.

Robinson, S. L., & Bennett, R. J. (1995). A typology of deviant workplace behaviors: A multidimensional scaling study. *Academy of Management Journal, 38*, 555-572.

Roznowski, M., & Bassett, J. (1992). Training test-wiseness and flawed item types. *Applied Measurement in Education, 5*, 35-48.

Sackett, P. R., Schmitt, N., Ellingson, J. E., & Kabin, M. B. (2008). High-stakes testing in employment, credentialing, and higher education. *American Psychologist, 56*, 302-318.

Sackett, P. R. (2002). The structure of counterproductive work behavior: Dimensionality and relationships with facets of job performance. *International Journal of Selection and Assessment, 10*, 5-11.

Salgado, J. F. (1997). The five factor model of personality and job performance in the European community. *Journal of Applied Psychology*, 82, 30-43.

Salgado, J. F., Remeseiro, C., & Iglesias, M. (1996). Personality and test taking motivation. *Psicothema, 8, 553-562*.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262-274.

Schmit, M. J., & Ryan, A. M. (1993). The big five in personnel selection: Factor structure in applicant and nonapplicant populations. *Journal of Applied Psychology*, *78*, 966-974.

Segall, D. O. (2002). An item response model for characterizing test compromise. *Journal of Educational and Behavioral Statistics, 27*, 163-179.

Sherman, T., & Margolin, J. (2007, February 16). *New cheating scheme rocks dental school: UMDNJ plans to discipline eight linked to exam-copying scandal*. The Star-Ledger.

Smith, C. A., Organ, D. W., & Near, J. P. (1983). Organizational citizenship behavior: It's nature and antecedents. *Journal of Applied Psychology, 68*, 653-663.

Smith, C. P., Ryan, E. R., & Diggins, D. R. (1972). Moral decision making: Cheating on examinations. *Journal of Personality, 40,* 640-660.

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

Smith, R. W. (2004). The impact of braindump sites on item exposure and item parameter drift. Paper presented at the annual conference of the American Education Research Association, San Diego, CA.

Sotaridona, L. S., & Meijer, R. R. (2003). Two new statistics to detect answer copying. *Journal of Educational Measurement, 40*, 53-69.

Spector, P. E., Fox, S., Penney, L. M., Bruursema, K., Goh, A., & Kessler, S. (2006). The dimensionality of counterproductivity: Are all counterproductive behaviors created equal? *Journal of Vocational Behavior, 68*, 446-460.

Stevens, G. E. & Stevens, F. W. (1987). Ethical inclinations of tomorrow's managers revisited: How and why students cheat. *Journal of Education for Business, 61,* 24-29.

Stober, J. (2001). The Social Desirability Scale-17 (SDS-17): Convergent validity, discriminant validity, and relationship with age. *European Journal of Psychological Assessment, 17*, 222-232.

Stocking, M. L., & Lewis, C. (1998). Controlling item exposure condition on ability in computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 23*, 57-75.

Stocking, M. L., Ward, W. C. & Potenza, M. T. (1997). Simulating the use of disclosed items in computerized adaptive testing. *Journal of Educational Measurement, 35*, 48-68.

Sympson, J.B., & Hetter, R.D. (1985, October). *Controlling item exposure rates in computerized adaptive testing*. Proceedings of the 27th annual meeting of the Military Testing Association, (pp. 973-977). San Diego CA: Navy Personnel Research and Development Center.

te Nijenhuis, J., Voskuijl, O. F., & Schijve, N. B. (2001). Practice and coaching on IQ tests: Quite a lot of *g*. *International Journal of Selection and Assessment, 9*, 302-308.

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

Veroff, R. (2008, June 12). Web site elicits criticism for allowing old-exam sharing. *The Daily Texan*. Retrieved from http://media.www.dailytexanonline.com/media/storage/paper410/news/2008/06/12/University/Web-Site.Elicits.Criticism.For.Allowing.OldExam.Sharing-3381237.shtml.

Vianello, M., Robusto, E., & Anselmi, P. (2010). Implicit conscientiousness predicts academic performance. *Personality and Individual Differences, 48*, 452-457.

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

Vowell, P. R., & Chen, J. (2004). Predicting academic misconduct: A comparative test of four sociological explanations. *Sociological Inquiry, 74*, 226-249.

Watson, D., & Clark, L. A. (1997). Extraversion and its positive emotional core. In R. Hogan, J. A. Johnson, & S. R. Briggs (Eds.), *Handbook of personality psychology* (pp. 767-793). San Diego, CA: Academic Press.

Wheeler, K. B., Foss, E. E., & Handler, C. A. (2001). *Screening and assessment: Best practices*. Fremont, CA: Global Learning Resources, Inc.

White, L. A., Young, M. C., & Rumsey, M. G. (2001). ABLE implementation issues and related research. In J. P. Campbell & D. J. Knapp (Eds.), *Exploring the limits in personnel selection and classification* (pp. 525–558). Hillsdale, NJ: Erlbaum.

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

APPENDIX

A. Review of test coaching in academic setting

B. Informed Consent

C. Introductory Direction

D. Questionnaire 1

E. Verbal Reasoning

F. Numerical Reasoning

G. Personality Inventory

H. Direction at the End of the Exercise 1

I. Direction for Information Exchange

J. Direction for Exercise 2

K. Questionnaire 2

L. Debriefing

A. Review of test coaching in academic setting[1]

Test coaching has been frequently studied in academic settings, and there have been several studies on the effects of coaching on Scholastic Aptitude Test (SAT) scores. However, many empirical studies on SAT coaching are flawed methodologically (Messick & Jungeblut, 1981). For example, in nonrandomized studies, there is no way to adjust for selection bias, student motivation, and small samples of coached students. But for studies with control groups, after weighting each by experimental sample size, Messick and Jungeblut found an average test coaching effect of 14.3 points for verbal (SAT-V) and 15.1 points for mathematical (SAT-M) portions.

While Messick and Jungeblut used rank order correlation procedures to combine studies with and without control groups, DerSimonian and Laird (1983) used a random effects approach to estimate the overall coaching effect in different situations. In their meta-analysis, they defined "points gained" as the difference in improvement between coached and uncoached groups. In their random effects approach, they parceled out the observed coaching effect for each study into two independent components, the true gain, and the sampling error. They proposed three different models to examine treatment effects. Model 1 assumes the sampling errors are independent and normally distributed, which ignores any dependence or covariance of one coaching program on another. As several coaching studies were actually conducted by the same investigator, Model 2 considers the true gain in terms of program effect and study effect. Model 3 separates uncontrolled, unrandomized evaluations from controlled, and matched or randomized

[1] Note that in this literature, general coaching refers to training on the broad constructs assessed by a test, rather than providing specific information about items on tests. Thus, in this section, the term "test coaching" is used to mean general training on broad constructs, not on specific item preknowledge given in item sharing.

studies. Based on the Model 1, they found the overall coaching effect was about 22 points for both verbal and math scores, with large estimated variances (477 for the verbal and 282 for the mathematical scores). When the correlations between studies were considered, Model 2 showed the average program effect was about 3 to 4 points lower than the results obtained from Model 1. In Model 3 where the effect of the evaluation design was considered, the matched and randomized studies showed mean gains of about 10 points for each of the SAT exams. Controlled studies showed mean gains of 15 points, while uncontrolled studies showed mean gains of 40 and 54 points for verbal and mathematic scores. Considering the three models, they concluded that on average, there is a positive score gain in general test coaching, but the variance across studies is large due to the evaluation design.

Kulik, Bangert-Drowns, and Kulik (1984) examined the effectiveness of coaching on several aptitude tests. With regards to the general coaching effect for the SAT, they found coaching raised scores from pretest to posttest about 0.36 standard deviations for the experimental groups and 0.21 standard deviations for the control groups. They concluded that the effect of coaching alone on the SAT was an average of 0.15 standard deviations.

Another approach to examining general test coaching effect on the SAT is to compare test preparation or test practice effects in terms of item characteristics, such as item format, difficulty level, and complexity of test directions. Powers (1986) used two raters to determine the complexity of directions or task. For instance, antonyms items that ask test takers to choose the option that has the opposite word or phrase as the stem have the lowest rating of complexity in their study. After reviewing five within-test practice

studies (27 effect sizes) and five test preparation studies (15 effect sizes), Powers found the effect sizes for the test preparation studies were generally larger than the within-test practice effect. He found length of directions and complexity of test directions were strongly correlated, where test preparation studies had higher correlations than within-test practice studies. He concluded that complex item formats are more susceptible to coaching, practice, or test preparation than simpler formats.

In another study, Becker (1990) conducted a meta-analysis using a standardized mean-change measure with a generalized least square analysis for published (academic journal articles only) and unpublished studies of coaching for the SAT. Regardless of study design, she found the coaching effects for the SAT-M are consistently larger than those for the SAT-V. When only published studies with control groups were used, she found the overall coaching effect was 0.09 standard deviations on the SAT-V (8 points) and 0.16 on the SAT-M (19 points).

But Becker's (1990) meta-analysis was criticized by Briggs (2005). He argued that using published studies with control groups is problematic when it comes to predicting the results from new studies. Using 16 new coaching reports after Becker's meta-analysis was published, he found the predicted test coaching effect for the SAT-V ranges from 17.4 to 27.7 points, and from 13.6 to 28.6 points for the SAT-M. He suggested that if no treatment and study design characteristics are considered, the expected predictions of the SAT-V coaching effects to be about 17 points, and 29 points for the SAT-M. Therefore, he argued that because Becker's meta-analysis can be criticized for how studies were quantified, a narrative review may be more informative.

Indeed, what Briggs proposed was reflected in Powers's (1993) review. He

reviewed meta-analyses conducted by Messick and Jungeblut (1981), DerSimonian and Laird (1983), Kulik, Bangert-Drowns, and Kulik (1984), and Becker (1990), and summarized the findings from the four meta-analyses by concluding that the effect of general test coaching is about 15 to 25 points each on the verbal and the mathematical portions of the SAT. Using a recent dataset from the College Board, he concluded that in general there is a 10 points increase for the SAT-V and 20 points improvement for the SAT-M, and these score increases add little improvement to a "typical" test taker's standing. That is, for a typical test taker who has SAT-V of 420 and SAT-M of 470, such score increases will only push the test taker's standing from the 48th to the 53rd percentile rank on the SAT-V, and from the 48th to the 54th percentile rank on the SAT-M. At higher and lower score levels, the same score increases change can be only 1% and effect little change to one's standing.

Although the above review shows that general test coaching affects the scores and rankings of test takers, predictive validity seems to be uninfluenced. Allalouf and Ben-Shakhar (1998) randomly assigned participants into coached and uncoached groups to take an Israeli Psychometric Entrance Test, an aptitude test similar to the SAT. They found larger effects for the quantitative test scores (27% of a standard deviation) than with the verbal test scores (18% of a standard deviation), which is consistent with what was found in the above mentioned meta-analysis studies. While they found general test coaching enhanced scores on the SAT-like test, it did not decrease the predictive validity of the test. On the contrary, the predictive validities of the after-coaching scores were slightly higher for both coached and uncoached groups, although not statistically significant. They concluded that test coaching had little effect on predictive validity, and

general test coaching did not create a prediction bias against coached or uncoached students.

To sum up, when general test coaching is considered in the context of the SAT, it seems that there is a 10 to 20 points gain for each of the verbal and mathematical subtests. On average, test coaching had a larger effect for the mathematical section than for the verbal section. However, in a simulated employee selection situation, will we find similar results on specific test coaching, that is, item sharing? We need to first look at how testing is different in educational and occupational settings. More importantly, we need to consider whether general test coaching in educational settings (i.e., training on broad constructs) is different from specific item sharing in employment testing (i.e., benefiting from item preknowledge).

B. Test preparation and test coaching

It should be noted that test preparation is not the same as test coaching. In an article on test preparation activities and employment test performance (Clause et al., 2001), researchers examined the relationship among motivational factors, metacognition and learning strategies, and employment test scores. Participants were given test preparation materials about the nature of the test they would be taking and a procedural manual for the job of state police, and two weeks later their test performance was evaluated using a video-based procedures test. While they found self-efficacy and motivation had effects on test performance through metacognitive processes and learning strategies applicants used to prepare for the test, the study was unable to answer how such test preparation activities effected test scores. As prior knowledge of the applicant about the job was not collected (e.g., they might have previous experience with similar procedure manuals), it would be difficult to link test preparation to score improvement in this setting.

C. Informed consent

Informed Consent

Please read this consent agreement carefully. You must be 18 years old or older to participate.

Purpose of the research:

This research is being conducted by Ben-Roy Do of the Department of Psychology at the University of Illinois under the direction of Dr. Fritz Drasgow. The purpose of this research is to assess how test coaching may effect types of items.

What you will do in this study:

You will be asked to respond to Questionnaire 1 and Exercise 1, exchange information about the exercise, and then take Exercise 2, and respond to a short Questionnaire 2.

Risks:

There are no anticipated risks, beyond those encountered in daily life, associated with participating in this study. .

Compensation:

The study will take under 50 minutes to complete. You will receive 1 Psychology 100 course credit (1 hour credit) for participating in this study. At the end of the study, you will receive an explanation of the study and the hypotheses. We hope that you will learn a little bit about how psychological research is conducted.

Voluntary Withdrawal:

Your participation in this study is completely voluntary, and you may withdraw from the study at any time without penalty (however, you will not receive Psych 100 credit for this study). You may skip over any questions or procedures, or you may withdraw by informing the research associate that you no longer wish to participate (no questions will be asked). Your decision to participate, decline, or withdraw participation will have no effect on your status at or relationship with the University of Illinois.

Confidentiality:

Your participation in this study will remain confidential, and your identity will not be stored with your data. Your responses will be assigned a code number that is not linked to your name or other identifying information. All data and consent forms will be stored in a locked room. Results of this study may be presented at conferences and/or published in books, journals, and/or in the popular media.

Further information:

If you have questions about this study, please contact Fritz Drasgow, Department of Psychology, University of Illinois, Champaign, IL 61820. Email:fdrasgow@s.psych.uiuc.edu; phone: (217)333-2739.

Who to contact about your rights in this study:

If you have any concerns about this study or your experience as a participant, you may contact the Institutional Review Board (IRB) at UIUC at (217) 333-2670 (collect calls will be accepted if you state you are a study participant); email: irb@uiuc.edu

Agreement:

The purpose and nature of this research have been sufficiently explained and I agree to participate in this study. I understand that I am free to withdraw at any time without incurring any penalty. I understand that I will receive a copy of this form to take with me.

Signature: _____ Date: _____

Name (print): _____ IRB-approved template, 12-29-2005

D. Introductory Direction

You will be asked to complete a 3-minute **Questionnaire 1** about **a recent team project you had to work on with your classmates or coworkers**.

Then, you will be asked to spend about 17 minutes on the **Exercise 1**. Please try your best to perform well and answer all of the questions. If you finish the Exercise 1 before the 20 minutes mark, please wait for researcher announcement.

Afterwards, you will spend 5 minutes **information exchange** to talk about the test content you just took in the first-stage selection exercise. You will receive further instruction later.

When you finish the information exchange, you will spend about 20 minutes on the **Exercise 2**. This time, we ask you to circle the items if you think the information was exchanged to you. Again, please try your best to perform well and answer all of the questions.

If you finish the Exercise 2, you may proceed with a short **Questionnaire 2**, which will take less than 3 minutes to complete the last part of the experiment. The experiment will take no more than 50 minutes of your time, and you will receive a full hour's credit for your participation.

If you have any questions about what you will be doing, feel free to ask now or whenever they arise. For further information about the experiment, you may contact:

Ben-Roy Do at 333-9631, benroydo@s.psych.uiuc.edu, or

Dr. Fritz Drasgow at 333-2739, fdrasgow@s.psych.uiuc.edu.

E. Questionnaire 1

Please write a number next to each statement to indicate the extent to which you agree or disagree with that statement.

Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
1	2	3	4	5

Think about a recent team project you had to work on with your classmates or coworkers. In general, I see Myself as Someone Who...

___ 1. Am trusted to keep secrets.

___ 2. Keep my promises.

___ 3. Believe that honesty is the basis for trust.

___ 4. Can be trusted to keep my promises.

___ 5. Am true to my own values.

___ 6. Lie to get myself out of trouble.

___ 7. Am hard to understand.

___ 8. Feel like an imposter.

___ 9. Like to exaggerate my troubles.

Please continue to work on the Exercise 1.

F. Verbal Reasoning

Each item consists of a pair of words that are related to one another. From a set of five additional pairs of words, please **select the pair that best expresses the relationship conveyed by the original pair**. Note: The order of the words in the pairs is significant!

Q1. diving : water

a) surfing : swimming
b) fly : bird
c) snow : skiing
d) parachuting : air
e) sand : building

Q2. depressed : cheerful

a) sad : melancholy
b) doctor : pain
c) angry : calm
d) blue : sky
e) singing : smiling

Q3. wings : bees

a) dog : legs
b) driven : car
c) birds : wings
d) fins : fish
e) honey : sweet

Q4. surf : water

a) island : land
b) waves : boat
c) glide : air
d) ice : skate
e) storm : climate

Q5. fuel : bus

a) fire : wood
b) coal : plane
c) sun : tree
d) wind : kite
e) water : boat

Q6. silver : gold

a) diamond : glass
b) whiskey : alcohol
c) glass : crystal
d) coal : wood
e) concrete : stone

Q7. water : plant

a) body : food
b) engine : fuel
c) cake : diet
d) petrol : car
e) table : wood

Q8. sugar : caramel

a) water : ice
b) mould : bread
c) milk : cheese
d) butter : cream
e) vegetable : salad

Q9. oxygen : water

a) juice : sugar
b) water : ice
c) word : letter
d) forest : tree
e) alcohol : beer

Q10. wind : clock

a) weigh : kilogram
b) violin : tune
c) fly : bird
d) sharpen : pencil
e) cold : water

Q11. waitress : restaurant

a) clinic : doctor
b) teacher : class
c) receptionist : hotel
d) bank : clerk
e) farmer : ranch

Q12. clothing : skirt

a) smile : laugh
b) noodles : food
c) music : melody
d) mammal : monkey
e) fish : dolphin

Q13. sentence : word

a) text : book
b) snow : ice
c) atmosphere : gas
d) mist : cloud
e) liquid : oxygen

Q14. shovel : pitchfork

a) pen : paper
b) plate : knife
c) spoon : fork
d) broomstick : brush
e) butter : milk

Q15. warm : hot

a) storm : drizzle
b) snow : cold
c) frost : chill
d) hill : mountain
e) boulder : pebble

G. Numerical Reasoning

In each item of the assessment, you will find a series of numbers arranged according to a certain rule. From a set of 5 possible options, your task is to **choose the option whose corresponding number best continues the series**.

Q1. 2, 5, 11, 20, 32, 47, 65 ...
a) 68
b) 71
c) 74
d) 75
e) 86

Q2. 32, 16, 19, 20, 10, 13, 14 ...
a) 7
b) 14
c) 15
d) 17
e) 28

Q3. 47, 38, 30, 23, 17, 12, 8 ...
a) 4
b) 5
c) 6
d) 7
e) 8

Q4. 75, 72, 24, 27, 9, 6, 2 ...
a) 2
b) 5
c) 9
d) 12
e) 18

Q5. 3, 4, 6, 10, 11, 13, 17 ...
a) 18
b) 20
c) 22
d) 24
e) 25

Q6. 1, 2, 5, 11, 12, 15, 21 ...

a) 22
b) 24
c) 25
d) 27
e) 30

Q7. 19, 91, 18, 81, 17, 71, 16 ...

a) 14
b) 15
c) 61
d) 71
e) 81

Q8. 2, 5, 15, 18, 54, 57, 171 ...

a) 173
b) 174
c) 176
d) 178
e) 180

Q9. 42, 45, 15, 18, 6, 9, 3 ...

a) 0
b) 3
c) 6
d) 9
e) 12

Q10. 36, 12, 16, 18, 6, 10, 12 ...

a) 3
b) 4
c) 10
d) 12
e) 14

Q11. 3, 6, 10, 15, 21, 28, 36 ...

a) 36
b) 37
c) 39
d) 42
e) 45

Q12. 44, 48, 12, 16, 4, 8, 2 ...

a) 0
b) 2
c) 4
d) 6
e) 8

Q13. 9, 1, 10, 2, 11, 3, 12 ...

a) 3
b) 4
c) 5
d) 14
e) 15

Q14. 10, 30, 32, 16, 48, 50, 25 ...

a) 27
b) 40
c) 50
d) 75
e) 99

Q15. 14, 11, 15, 16, 13, 17, 18 ...

a) 15
b) 19
c) 22
d) 23
e) 36

H. Personality Inventory

On the following page, there are phrases describing people's behaviors. Please use the rating scale below to describe how accurately each statement describes ***you***. Describe yourself as you generally are now, not as you wish to be in the future. Describe yourself as you honestly see yourself, in relation to other people you know of the same sex as you are, and roughly your same age. So that you can describe yourself in an honest manner, your responses will be kept in absolute confidence. Please read each statement carefully, and then write a number next to each statement to indicate the extent to which you agree or disagree with that statement.

Response Options

Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree strongly
1	2	3	4	5

___ 1. Am the life of the party.
___ 2. Feel little concern for others.
___ 3. Am always prepared.
___ 4. Get stressed out easily.
___ 5. Have a rich vocabulary.
___ 6. Don't talk a lot.
___ 7. Am interested in people.
___ 8. Leave my belongings around.
___ 9. Am relaxed most of the time.
___ 10. Have difficulty understanding abstract ideas.
___ 11. Feel comfortable around people.
___ 12. Insult people.
___ 13. Pay attention to details.
___ 14. Worry about things.
___ 15. Have a vivid imagination.
___ 16. Keep in the background.
___ 17. Sympathize with others' feelings
___ 18. Make a mess of things.
___ 19. Seldom feel blue.
___ 20. Am not interested in abstract ideas.
___ 21. Start conversations.
___ 22. Am not interested in other people's problems.
___ 23. Get chores done right away.
___ 24. Am easily disturbed.

___ 25. Have excellent ideas.
___ 26. Have little to say.
___ 27. Have a soft heart.
___ 28. Often forget to put things back in the proper place.
___ 29. Get upset easily.
___ 30. Do not have a good imagination.
___ 31. Talk to a lot of different people at parties
___ 32. Am not really interested in others.
___ 33. Like order.
___ 34. Change my mood a lot.
___ 35. Am quick to understand things.
___ 36. Don't like to draw attention to myself.
___ 37. Take time out for others.
___ 38. Shirk my duties.
___ 39. Have frequent mood swings.
___ 40. Use difficult words.
___ 41. Don't mind being the center of attention.
___ 42. Feel others' emotions.
___ 43. Follow a schedule.
___ 44. Get irritated easily.
___ 45. Spend time reflecting on things.
___ 46. Am quiet around strangers.
___ 47. Make people feel at ease.
___ 48. Am exacting in my work.
___ 49. Often feel blue.
___ 50. Am full of ideas.

I. Direction at the End of the Exercise 1

Now that you have finished the **Exercise 1**, we would like you to talk about the test content to your fellow classmates. **Please notify the researcher and s/he will assist you to locate your team member**.

J. Direction for Information Exchange

Now you have one or two fellow students in your group, we ask you to **try your best to recall any information** that may help other students to do well on the test you just took.

The information exchange will take about 5 minutes. During the process, you will:

1) Talk about **questions and answers** appeared on the **Exercise 1** you just took.

2) Listen to your fellow students talk about information on the test they just took, which will become your **Exercise 2.**

Please do not ask your fellow students to clarify what items were provided on the test. If every member in your group has exchanged test information, Please notify the researcher, and s/he will give you the packet for the second-stage selection exercise.

K. Direction for Exercise 2

Now you will spend about 20 minutes on the **Exercise 2**. This time, we ask you to **circle the items if you think the information was exchanged to you**. Again, please try your best to perform well and answer all of the questions.

If you finish the Exercise 2, you may proceed with a short **Questionnaire 2**, which will take less than 3 minutes to complete the last part of the experiment.

L. Questionnaire 2

Thank you for completing the Exercise 2. As a final part of the study, we would like you to answer the following questions about yourself and your approach to the study. Please answer honestly and candid.

		Strongly Disagree				Strongly Agree
1	I was very motivated to disclose the information in the first stage exercise.	1	2	3	4	5
2	I actively participated in the information exchange.	1	2	3	4	5
3	I believe that I did as well as most people who took this test.	1	2	3	4	5
4	I am good at taking this kind of test.	1	2	3	4	5
5	I think that this test was difficult.	1	2	3	4	5
6	I think the information exchange is helpful.	1	2	3	4	5
7	I think the information exchange provide accurate answer to the question.	1	2	3	4	5
8	I am more confident about my performance after the information exchange.	1	2	3	4	5
9	The quality of information exchange was high.	1	2	3	4	5
10	The information exchanged was very helpful for the second stage exercise.	1	2	3	4	5

11	What is your undergraduate GPA (estimation is fine)?		
12	What was your high school GPA?		
13	What was your overall ACT score?		
14	What was your SAT verbal score?		
15	What was your SAT math score?		
16	What is your gender?	Male	Female
17	What is your ethnic background?		
	a. White (Caucasian)	d. Asian/Pacific Islander	
	b. Black (African American)	e. Native American (Aleutian/American Indian)	
	c. Hispanic	f. Other	

Thank you for your participation in our study. Please return all materials to the researcher. Thanks!

M. Debriefing

Debriefing

Thank you for your participation in our study on Test Coaching. As you know, testing can have a large and significant impact on the lives of many people. For example, many of you may have already taken the SAT or ACT as a pre-requisite for admissions into college. An important concern in testing has been to preserve the validity of such tests. As such, researchers have been concerned how test coaching may influence test results.

In the experiment you just participated in, you first took Exercise 1, exchanged information about the test, and then took Exercise 2. Your performance will be used to evaluate what types of items are most likely to be exchanged during test coaching process, and how likely your score may be influenced by the process.

There has not been any published research regarding this issue, consequently we are unable to direct you to any publications regarding this issue. This is an exploratory research project and therefore we do not know what types of items are more likely to be remembered. We hope that your participation will help us better understand the types of items that are most memorable. If you have any questions or comments about this work, feel free to contact Ben-Roy Do at benroydo@s.psych.uiuc.edu.

Again, thank you very much for your participation!