RICE UNIVERSITY

Is Retest Bias Biased? An Examination of Race, Sex, and Ability Differences in Retest Performance on the Wonderlic Personnel Test

by

Jason Gilbert Randall

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Master of Arts

APPROVED, THESIS COMMITTEE

Anton J. Villado, Chair
Assistant Professor of Psychology

Margaret E. Beier,
Associate Professor of Psychology

Frederick L. Oswald,
Associate Professor of Psychology

HOUSTON, TEXAS
NOVEMBER 2012

ABSTRACT

Is Retest Bias Biased? An Examination of Race, Sex, and Ability Differences in Retest Performance on the Wonderlic Personnel Test

by

Jason Gilbert Randall

Research suggests there may be race, sex, and ability differences in score improvement on different selection tests and methods when retested (Schleicher, Van Iddekinge, Morgeson, & Campion, 2010). However, it is uncertain what individual differences moderate retest performance on GMA assessments, and why. In this study, 243 participants were retested on the Wonderlic Personnel Test (WPT). There was no evidence that race, sex, emotional stability, or conscientiousness moderate retest performance on the WPT, although SAT scores did positively predict retest performance. Individuals within the interquartile range of the initial WPT scores gained more when retested than those with more extreme scores. Establishing artificial cut-off levels demonstrated that those below the cut-off gained more when retested than those above the cut-off. Therefore, average-scorers and in some cases lower-scorers who may have failed to meet a predetermined cut-off are encouraged to re-test as they have little to lose and much to gain.

Acknowledgements

I would like to thank my advisor, Dr. Anton Villado, for his mentorship in completing this thesis. I would also like to thank the committee members, Margaret Beier and Fred Oswald, for their valuable inputs and feedback which have improved this paper greatly. I would like to thank my colleague Christina Upchurch and other lab members for assistance with data collection on this project. Finally, I would like to thank my family for their constant support and encouragement, without whom I could not have completed this project.

Table of Contents

List of Tables

List of Figures

Is Retest Bias Biased? An Examination of Race, Sex, and Ability Differences in Retest Performance on the Wonderlic Personnel Test

High-stakes testing is pervasive in today's society and numerous important, potentially life-changing decisions are made based on the outcomes of such tests. Thus, it is inevitable that people will seek opportunities to retest as a way to re-prove or improve their scores. Additionally, because no test is error free and many factors may contribute to an individual underperforming, the practice of retesting is common in both academic and employment settings (Lievens, Reeve, & Heggestad, 2007; Hausknecht, Halpert, Di Paolo, & Gerrard, 2007). Re-taking cognitive ability tests in educational and organizational settings is very common, with as many as 25%-50% of applicants being retested (Hausknecht et al., 2007). Applicants retest for various reasons, which may reflect applicants' desire to improve their chances of qualifying for educational and occupational opportunities (Hausknecht et al., 2007), or alternatively an organization's suggestion or protocol for retesting (Tippins et al., 2006). The opportunity for retesting in selection and promotion decisions is sanctioned and even encouraged by both the Society for Industrial Organizational Psychology (2003) and the American Psychological Association (1999) as a way to control for measurement error and as a chance for individuals to improve their scores. Yet, despite the advantages and frequency of retesting, there is ample evidence to suggest that retesting inflates assessments of ability up to about one quarter of a standard deviation between the first and second administration (Hausknecht et al., 2007). Drawing on substantial evidence of adverse impact for cognitive ability assessments and preliminary evidence of subgroup differences for retest effects, in this study I investigated whether differences in retest

gains on ability assessments could be explained by various psychological phenomena associated with race and sex. I also examined how variation in ability-level affects the magnitude of retest gains and consider how these differences may impact selection procedures via the creation of artificial cut-off scores.

In an attempt to document and diagnose the existence and extent of retest effects, researchers have investigated a myriad of tests in educational and employment settings, as well as in research laboratories. A recent meta-analysis found no difference in retest effects between operational settings (i.e., educational and employment) and research or lab settings (Hausknecht et al., 2007). However, there are likely differences in the motivations for and allowances to retest between educational and employment settings, as well as differences in test content and the underlying constructs the tests are designed to assess. Interestingly, research suggests that despite its conceptual design (achievement vs. aptitude) or the setting in which a GMA assessment is used (employment, education, laboratory), cognitive ability assessments (e.g., SAT, GATB, Wonderlic, Raven's Progressive Matrices) are susceptible to coaching and retest bias (Reeve & Lam, 2005; Kulik, Kulik, & Bangert, 1984; Hausknecht et al., 2007). This study focuses on retest effects for an assessment of GMA that is often used in employment settings: the Wonderlic Personnel Test (WPT).

Retest Bias

The phenomenon of retest bias—consistent score increases across repeat administrations of the same test—is a well-established finding in many fields and for many types of assessments (Hausknecht et al., 2007; Kulik et al., 1984). For example, retest effects (also called testing effects) are extensively studied in the experimental and

educational psychology literatures. In these domains, most research on testing effects utilizes tests of free recall (or other memory tests) or multiple-choice tests to demonstrate how multiple administrations of a test affect subsequent performance and learning (Roediger & Karpicke, 2006). The common finding is that merely having taken a test previously contributes to improved test performance on subsequent assessments (Roediger & Karpicke, 2006). Such performance improvements may be attributable to various factors including the learning of test content or even familiarization of test format and strategy. Lievens, Reeve, and Heggestad (2007) succinctly classified the underlying rationale explaining the mechanisms of retest bias into three categories: 1) actual increases in the target ability, 2) a reduction in measurement error and debilitating factors, and 3) increases in test-specific, non-*g* skills. The first two categories are the reason the Society for Industrial Organizational Psychology (2003) and the American Psychological Association (1999) have cited as rationale for allowing applicants to re-take assessments in selection and promotion contexts, while the third is typically the rationale identified in the research as accounting for the retest phenomenon (Lievens et al., 2007; Hausknecht et al., 2007; Lievens et al., 2005; Kulik et al., 1984). This is especially true for retesting with cognitive ability assessments in which the target construct, GMA, is conceptualized as a stable psychological construct and thus extra practice or coaching designed to increase ability should do nothing to account for typical score increases seen at subsequent administrations (Reeve & Lam, 2005).

Retest Effects for Cognitive Ability Assessments

Cognitive ability is generally the most valid predictor of job performance across job categories (Schmidt & Hunter, 1998). Due to the predictive power and common use of

cognitive ability assessments, it is no surprise that there are also high rates of retesting with ability tests reported in studies examining employment retesting (Hausknecht et al., 2007; Lievens et al., 2005; Schleicher et al., 2010). Reeve and Lam (2005) addressed the paradox that retest effects for cognitive ability introduce. If it is possible and common to see score gains on measures of GMA, one must either question the conceptualization of ability as a stable construct, or the assumption of measurement invariance (that the relationship between GMA and test indicators remains stable), the latter of which would undoubtedly undermine the inferences drawn from test scores. After administering a general cognitive ability test several times to group of participants, Reeve and Lam (2005) found that mere retesting did not alter the latent construct (GMA) being measured as the reliability of factor scores and construct validity of the test remain the same when retested. Yet despite the stability of the construct and its psychometric properties in retest situations, these authors still saw sizable score increases typical of the retest phenomenon, with some scales approaching a one standard deviation increase. Thus, Reeve and Lam (2005) concluded that the retest phenomenon for assessments of cognitive ability are likely attributed to non-cognitive, test-specific or non-g factors associated with practice—a conclusion that has since been substantiated (Lievens et al., 2007; te Nijenhuis, van Vianen, & van der Flier, 2007).

In a recent meta-analysis, Hausknecht and colleagues (2007) evaluated 107 samples where participants took a cognitive ability assessment at least twice. These authors found an overall score inflation of about one quarter of a standard deviation ($\delta = 0.26$) between the first and second administration, with the increase dropping to 0.18 between second and third administrations. In their evaluations of moderators, Hausknecht et al. (2007)

found that retest score increases were positively related to time spent receiving formal test coaching, that the magnitude was greater for identical than for alternate forms, that larger increases were associated with shorter time periods between administration, and that the increase was attributable to factors other than regression to the mean alone. Contrary to their expectations, however, retest bias was not related to participant time in formal schooling or study context (operational selection vs. research setting), and did not differ by dimension of ability (analytical, quantitative, and verbal). These data present fairly conclusive evidence that retest bias is a real issue for ability assessments.

However, while many moderators were examined in this meta-analysis, there is an obvious emphasis on methodological factors as explanations for the retest phenomenon. Thus, while time between retest, test form, study context, coaching, and dimension of ability were all considered in this review, other individual differences such as race, sex, and ability level which may also moderate the magnitude of the retest phenomenon were not considered. This is especially problematic in selection contexts because in addition to issues of retest bias, one must also consider issues of adverse impact when using cognitive ability assessments. As a result, it is imperative to determine whether certain individuals gain more from retesting than others, or whether the mere act of re-administering a GMA assessment unfairly disadvantages certain subgroups. To address this gap in the literature, in the current study I addressed two areas of concern: subgroup (race and sex) differences, and ability differences.

Subgroup Differences for Cognitive Ability Assessments

Group differences on selection tests have long been an important concern in industrial/organizational psychology, and much of this preoccupation has to do with laws

and sanctions enacted to promote equal employment opportunity. Adverse impact is characterized by differential hiring rates for different groups (e.g., race, sex, age) because of mean group differences on selection tests (Hough, Oswald, & Ployhart, 2001). There is a wealth of evidence to show that group differences on cognitive ability tests do exist and that these tests demonstrate higher rates of adverse impact than do other selection tests such as personality measures (e.g., Hattrup, Rock, & Scalia, 1997; Hough et al., 2001; Sackett, Schmitt, Ellingson, & Kabin, 2001). Organizations are often torn between using selection tests that produce scores demonstrating the highest validity (cognitive ability assessments) but also result in adverse impact, or sacrificing validity for selection measures that are more likely to result in a diverse workforce (De Corte, Lievens, & Sackett, 2007; Pyburn, Ployhart, & Kravitz, 2008). This perplexing problem has been termed the "diversity-validity dilemma" (Pyburn et al., 2008, 144) and the "selection quality-adverse impact problem" (De Corte et al., 2007, 1380).

To assess the extent to which mean differences in cognitive ability tests differ by race, Roth, Bevier, Bobko, Switzer, and Tyler (2001) meta-analyzed differences between Black-Americans and White-Americans on cognitive ability tests and found that Whites outperformed Blacks overall by an average of one standard deviation for job applicants in employment settings on general intelligence measures. These authors also identified two important moderators: job complexity, such that the difference is more exaggerated for jobs of low complexity (.86) than those of high complexity (.63); and study design, such that the difference is more exaggerated for job applicants in cross-job study designs (1.0-1.23) than for within-job designs (.83). Research has determined that differences between Blacks and Whites on ability assessments are not due to changes in the underlying factor

structure (Jensen, 1980; Carretta & Ree, 1995), mirroring the more general finding that retesting does not change the factor structure of ability assessments (Reeve & Lam, 2005). Thus, it would appear that neither race differences nor retest bias can be explained by a distortion of the structure of the GMA construct.

Mean group score differences on general ability assessments by sex, however, are much lower than race differences. In fact, most GMA tests are designed so that there are no differences between male and female mean scores; thus if differences are found on GMA they are of small effect and are not in a consistent direction (Neisser et al., 1996).

Subgroup Differences in Retesting

Recognizing that adverse impact is a common and genuine problem when cognitive ability tests are used in employment selection contexts, one might ask whether retesting ameliorates or exaggerates pre-existing subgroup differences on such tests. Examinations of subgroup differences and retesting suggest that there are race, sex, and age differences on retest gains for varying selection tools. Examining retest scores of applicants for government jobs who re-took tests one year or more after an unsuccessful initial attempt, Schleicher et al. (2010) found significantly different average magnitudes of improvement upon retesting by race for only certain types of selection measures. The differential improvement of Whites over Hispanics and Blacks upon retesting was significant on what the authors classified as written tests: job knowledge, biodata, and verbal ability. No retest differences by race were found for performance-based ability tests: three types of interviews, a leaderless group discussion, and a case analysis exercise. In contrast, Blacks generally showed greater improvement than Whites when re-interviewing. Concerning sex and age, the authors found that women and applicants

under 40 showed larger improvements with retesting than did men and applicants over 40. Specifically, there were no differences by sex on the written tests, but women improved more than men on all of the performance tests. Van Iddekinge et al. (2011) found no significant differences by race (Black-White, Hispanic-White, or Asian-White) on retesting scores for a job knowledge test. The authors did, however, find that females improved more than males, and that younger candidates improved more than older candidates.

In summary, there is evidence for subgroup differences when retesting in operational settings for different constructs (job knowledge, verbal ability) and methods (biodata, interviewing, assessment center exercises; Schleicher et al., 2010; Van Iddekinge et al., 2011; Lievens et al., 2005). However, the evidence of subgroup differences for retesting with assessments of general cognitive ability (GMA) is not well established.

Initial evidence suggests that Whites improve more in a retest setting for written assessments of verbal ability than do Blacks, Hispanics, or Asians. Schleicher and colleagues (2010) expected to see this pattern of results since they argued that adverse impact is most common when assessments of knowledge, skills, and abilities are far removed from actual performance (i.e., written or tested on paper instead of performed in person). Schleicher et al. (2010) found no difference on the verbal ability assessment between men and women, but again found a significant difference in the improvement of people under 40 when compared to re-testers over 40. However, their assessment of verbal ability was designed to measure grammar, spelling, punctuation, word use, and organization required for writing and editing (2010). Such a test is clearly different from

typical assessments of GMA (e.g., WPT, SAT, APM) which are designed to assess more than just writing and editing knowledge and capability. Thus, it remains to be seen whether there are subgroup differences on retest scores for general cognitive ability assessments.

Ability differences in Retesting

Investigations of differences among re-testers have also demonstrated that individuals retested on a number of ability assessments show larger improvements in those with higher initial ability than those with lower ability (Rapport et al., 1997; Salthouse & Tucker-Drob, 2008), including aptitude and achievement tests (Kulik et al., 1984). For example, in their meta-analysis, Kulik and colleagues (1984) categorized studies as comprising low-, middle-, or high-ability samples and computed the difference scores for each group to determine whether retest magnitude differed by ability group. These authors found effect sizes (Cohen's *d*) characterizing score increases from time one to time two of .17 for students of low ability, .40 for students of middle ability, and .82 for students of high ability. Rapport and colleagues (1997) similarly divided individuals into groups of low-, average-, and high-average intelligence based on initial scores on the WAIS-R and found that average- and high-average groups improved more from time one to time two than did the low-average group, though their sample was small (n = 12 in each group; N = 36). These two studies offer important evidence that cognitive ability moderates the increase of retest gains, with higher-scoring individuals gaining more from retesting than lower-scoring individuals. However, this evidence was accumulated by categorizing or segmenting individuals into discrete levels of ability and comparing means of these artificial groups. I hope to build on this previous work by considering

GMA as a continuous variable in order to make a more definitive conclusion concerning potential for ability-level to moderate retest effects.

Underlying Mechanisms

There are a number of possible mechanisms that may contribute to an explanation of why differences in cognitive ability, sex, and race arise when re-taking cognitive ability assessments. Most explanations for retest differences can be classified as 1) ability-based, 2) personality-based, 3) attitude-based, or 4) motivation-based. Examination of each of these factors is critical, yet beyond the scope of the current research. Thus, in the current study I focus only on ability- and personality-based mechanisms, recognizing that investigation of attitude- and motivation-based mechanisms are an important and necessary extension for future research.

There are several reasons why individuals of different races might improve differentially from a retest opportunity which, beyond factors such as stereotype threat and test-taking attitudes, are mostly ability-based. Ability-based mechanisms are not a far step from the finding that those who score higher on cognitive ability tests gain more from retesting than do those who score lower on such tests (Rapport et al., 1997; Kulik et al., 1984; Salthouse & Tucker-Drob, 2008). This connection hinges on earlier research showing that Whites generally outperform Black and Hispanic minorities on measures of GMA (Roth et al., 2001; Chan & Schmitt, 1997). Logically, if Whites score higher on cognitive ability tests and those who score higher on such tests the first time gain more from retesting than those who score lower, then it may be expected that Whites may gain more from retesting than Blacks or Hispanics.

The mechanisms I chose to focus on for sex differences are captured by two key personality variables: emotional stability and conscientiousness. First, research has found a consistent negative relationship between emotional stability and test anxiety (e.g., Schmidt & Riniolo, 1999), with test anxiety mediating the positive relationship between emotional stability and performance on intelligence tests (Moutafi, Furnham, & Tsaousis, 2006). Moreover, Dobson (2000) found that performance on cognitive tests is underestimated for individuals low on emotional stability, even in real selection situations. Thus, women, who typically score lower on emotional stability than men (Schmitt, Realo, Voracek, & Allik, 2008), may be more negatively affected by test anxiety and thus perform worse on the WPT at the initial assessment. However, to the extent that the first administration in a retest paradigm represents the most anxiety-inducing situation where the test and setting are new to the test-taker, the retest opportunity should be marked by lower levels of anxiety, thus the negative effects of test anxiety (i.e., the positive effects of higher emotional stability) should be attenuated upon retesting, resulting in larger gains for women. Second, people high in conscientiousness who are more detail-oriented and achievement-striving may be better at remembering specific test content and forming effective test-taking strategies, which would also aid retest performance. Evidence also shows that individuals high in conscientiousness are more likely to set their own performance goals and to engage in processes to maintain the motivation and effort necessary to attain these self-set goals (Barrick, Mount, & Strauss, 1993). Women also report higher mean levels of conscientiousness than men do (Schmitt et al., 2008), so this difference may also contribute to women's differential retest improvement.

Finally, there are at least two reasons why higher-ability individuals should gain more from retesting than lower-ability individuals: memory and test-wiseness. Lievens, Reeve, and Heggestad (2007) found that memory correlated significantly with retest gains, suggesting it may account in part for score increases upon retesting. There is also some preliminary evidence that certain types of memory, a lower-order dimension of cognitive ability, may differ by race and sex (Hough et al., 2001). Verive and McDaniel (1996) meta-analyzed race differences on memory-span tests and saw that Whites out-performed Blacks by about one half of a standard deviation. Concerning sex differences, Maccoby and Jacklin (1974) found that women out-performed men on memory tests containing verbal content, but that differences were much smaller on memory tests of objects and digits. Test-wiseness has typically been characterized as test-specific non-*g* skills which enable an individual to improve their score on a test upon re-administration of a test (Lievens et al., 2007; Te Nijenhuis, van Vianen, & van der Flier, 2007). Although the definition of test-wiseness precludes the impact of GMA as the source for the retest gains (i.e., non-*g* skills), it may be that, similar to the effect of cognitive ability on performance and learning, people who score higher on GMA assessments acquire test-specific non-*g* skills more quickly and adeptly than do those who score lower, and thus can utilize test strategies in order to see greater score improvements upon retesting. To the extent that higher levels of GMA enable an individual to acquire test-wiseness, higher-ability (majority) individuals should exhibit higher retest gains.

In summary, explanations for differential retest gains for race center on underlying ability (at least as captured by typical GMA assessments), explanations for

sex center on differences in emotional stability and conscientiousness, and those for ability focus on memory and test-wiseness.

Redemptive versus Non-redemptive Retesting

Investigations of any individual differences (ability, race, or sex) concerning the retest phenomenon are complicated by the fact that the extant literature does not always consider the entire range of cognitive ability when retesting. In order to understand who retests, and why, the population of re-testers in employment settings can be divided into two categories: redemptive, and non-redemptive. Redemptive retesting is characterized by individuals who fail an ability test, typically administered in a selection or promotion context, and then are given the option to retest. For example, Schleicher et al. (2010) examined retest effects using a sample of individuals applying for government positions where roughly 15,000 people apply, but applications only occur once a year. As a result, several thousand who fail one or more portions of the selection battery are allowed to re-apply, and thus retest, the next year.

Non-redemptive re-testers are individuals who are re-taking an ability test in a selection context for a parallel or more advanced job in the same or a different organization who did not fail an initial assessment. The practice of re-testing without having first failed captures several types of testers. First, those who were previously successful in applying for one job, but are either switching jobs or applying for several positions at once, and thus retest on one assessment several times. Although this may be more difficult to identify because it is cross-organizational, some test publishers keep a database of individual test-takers, suggesting that retesting does occur in cross-organizational situations. For example, Hogan, Barrett, and Hogan (2007) reported that

thousands of people taking the Hogan Personality Inventory (R. Hogan & Hogan, 1995) had taken the test multiple times over the course of their life. Additionally, non-redemptive re-testers may be individuals who scored high enough on an online version of the test to pass an initial screening, and then are subsequently called in for a second, proctored administration of the test. Tippins et al. (2006) presented this two-stage selection strategy as an attempt to utilize the advantages of unproctored internet testing (e.g., reduced administration costs, increased applicant reach, and enhanced company image), while simultaneously reducing their applicant pool to a more manageable number and attempting to control for cheating. In other words, the first, online administration of the test is used to screen out obvious poor performers, and the follow-up test is administered in-person to control for issues of malfeasance accompanying online testing.

Although there have been no direct investigations into the potential for redemptive and non-redemptive re-testers to differ, a handful of studies point to the potential for this distinction to be meaningful. With a sample of redemptive re-testers who failed on their initial attempt, Lievens and colleagues (2005) found that the validity coefficient of an ability test as a predictor of performance (operationalized as GPA) was higher for one-time test takers than for either initial or retest scores of re-testers. In other words, Lievens et al. (2005) found meaningful differences between people in an operational selection context (admission to medical school in Belgium) who passed the test on their initial try, and those who initially failed, and then had to retest in order to reach the cutoff score even after correcting for the restriction of range in the samples. This finding serves as direct support for the argument that there may be meaningful differences for different populations of re-testers, especially between typical scenarios

where people retest because of an initial failure (redemptive retesting), and those who are capable of passing a cutoff score on the first try (non-redemptive retesting).

Most studies in the current literature, and especially those in operational selection contexts focus on the redemptive category of re-testers—those who are re-trying after an initial failure (Lievens, Buyse, & Sackett, 2005; Lievens et al., 2007; Schleicher et al., 2010; Van Idekkinge et al., 2011; Hausknecht et al., 2007). As a result, most of the findings on retest effects in the extant literature are based on samples that do not represent the entire applicant pool. Those neglected in this population of re-testers are applicants who typically score higher on assessments of GMA who are more likely to be non-redemptive re-testers.

Current Study

Drawing on the research and arguments presented above, it is known that retesting with cognitive ability assessments is a common procedure in selection settings and that GMA tests are susceptible to the same score inflation seen in other assessments (Hausknecht et al., 2007). It is also known that there are mean group differences by race for ability assessments (Hough et al., 2001). Additional research in applied settings suggests that race and sex differences exist for different selection methods and assessments (Schleicher et al., 2010; Van Iddekinge et al., 2011). However, the majority of these findings for retest differences by subgroup were determined in a retest paradigm in which the only individuals retesting were those who failed the assessments the first time. Thus, it is unclear whether these results are limited by the restricted and incomplete sample of redemptive re-testers or those with a potentially lower range of cognitive ability.

Likewise, although both Schleicher et al. (2010) and Van Iddekinge et al. (2011) have made important advancements in the area of retesting using various selection measures in operational contexts, these authors reference the prevalence of retest studies that deal with cognitively-oriented tests as their rationale for not including typical GMA measures in their analyses. However, this literature (e.g., Hausknecht et al.., 2007; Kulik et al., 1984;) neglects to evaluate the entire population of re-testers, with a common over-emphasis on low-ability individuals. More seriously, previous research evaluated ability differences in retest performance by sub-categorizing levels of ability and testing for group differences (e.g., Rapport et al., 1997) rather than maintaining the continuous nature of the data. Thus, a gap in the literature remains to evaluate the potential for subgroup and ability differences when re-administering cognitive ability assessments for the entire range of ability characterizing the retesting population in a continuous non-categorical analysis.

Also lacking in previous research is a direct assessment of the mechanisms which might explain subgroup differences in retesting. Fundamentally, race and sex are, in and of themselves, not psychologically meaningful variables. Instead, race and sex may serve as proxies for differences that could be better explained by underlying psychological phenomena that vary across different subgroups: cognitive ability, emotional stability, and conscientiousness. However, research studies documenting subgroup and ability differences in retest performance as of yet have only speculated concerning what mechanisms might explain differences in retest gains without actually measuring and testing them. Thus the proxies of race and sex remain, yet an understanding of why subgroup identification matters in predicting retest gains is still missing.

I redressed these limitations by examining whether retest bias differentially impacts protected subgroups and individuals of varying levels of ability on a GMA assessment used in employment: the Wonderlic Personnel Test (WPT). Such an investigation is necessary to enhance understanding of assessments of GMA and their re-administration in selection settings as well as the impact retesting may have on different race and sex subgroups which are protected by law during the selection process. The current study was conducted in a controlled, low-stakes lab setting where participants were not informed of their performance on the initial test or that they would be re-taking the same test upon their return in an attempt to control for motivation- and attitude-based explanations of retest differences of the phenomenon.

In addition to assessing the potential for retesting to influence subgroup and ability differences on GMA assessments, I also investigated the impact retesting may have in a selection setting by instituting a set of reasonable, yet artificial cutoff scores on the initial administration of the WPT without notifying participants of the cutoff or the results. This artificial division of individuals into initial-pass and initial-fail samples of re-testers allows comparison of the two groups to see if the magnitude of retest bias differs as a function of cut-off group.

Situating these purposes into the preceding review of the literature on both retesting and adverse impact, I address four very fundamental, yet important sets of research questions and hypotheses. First I want to assess whether retesting ameliorates, exaggerates, or does not affect race differences upon retesting, and whether GMA is more suited to account for these differences than race is.

Hypotheses 1a-b: Retesting exaggerates racial differences on the WPT such that Whites and Asians improve more upon retesting than Blacks and Hispanics on an (1a) identical or (1b) alternate form of the WPT.

Hypotheses 1c-d: GMA accounts for the racial differences in retest gains, such that when the variance attributed to GMA is accounted for, race does not contribute any additional variance in retest gains for an (1c) identical or (1d) alternate form of the WPT.

Second, I want to assess whether retesting introduces sex differences on the WPT, with the expectation that retesting aids women more than men, but that these differences may be more attributable to women's lower levels of emotional stability and higher levels of conscientiousness.

Hypothesis 2a-b: Women improve more than men when retested with an (2a) identical or (2b) alternate form of the WPT.

Hypotheses 2c-d: Personality differences account for the sex differences in retest gains, such that when the variance attributed to emotional stability and conscientiousness is accounted for, sex does not contribute any additional variance in retest gains for an (2c) identical or (2d) alternate form of the WPT.

The third area of investigation centers on the potential for individuals with varying levels of ability to differ in the degree of score increase they see upon retesting. Specifically, I expect that cognitive ability moderates the increase of retest gains such that the higher an individual scores on the initial WPT assessment, the greater the rate of change (retest gain) between initial and follow-up assessments.

Hypothesis 3a-b: Individuals who score higher on the initial WPT assessment see greater score gains at Session 2 than those who score lower on their initial assessment when retested with an (3a) identical or (3b) alternate form of the WPT.

Finally, I want to know whether there may be differences between re-testers who score high enough to initially pass a cutoff score and those who do not. Admittedly, such an investigation is limited because the reason for and the decision to retest is inextricably tied to retest performance. However, dividing the current sample at several reasonable cut-off scores allows preliminary investigation for considering whether redemptive and non-redemptive groups may be distinct.

Hypothesis 4: The magnitude of retest gains is significantly higher for re-testers whose initial score on the WPT is high enough to be considered passing than for those below the cut-off, with a more stringent cut-off resulting in a larger difference.

Method

Participants

Participants (N = 243, 67% female, mean age 20.37 [SD = 5.33]) were recruited from the campus of a small southern private university and surrounding community colleges. The demographics composition of the same was 39.5% White, 30.5% Asian, 10.5% African-American, and 19.5% Latino/Hispanic. Participants were awarded research credit or extra credit for their participation in the study.

Procedure

The entire study was comprised of two sessions separated by a six-week interval and is outlined in Figure 1.

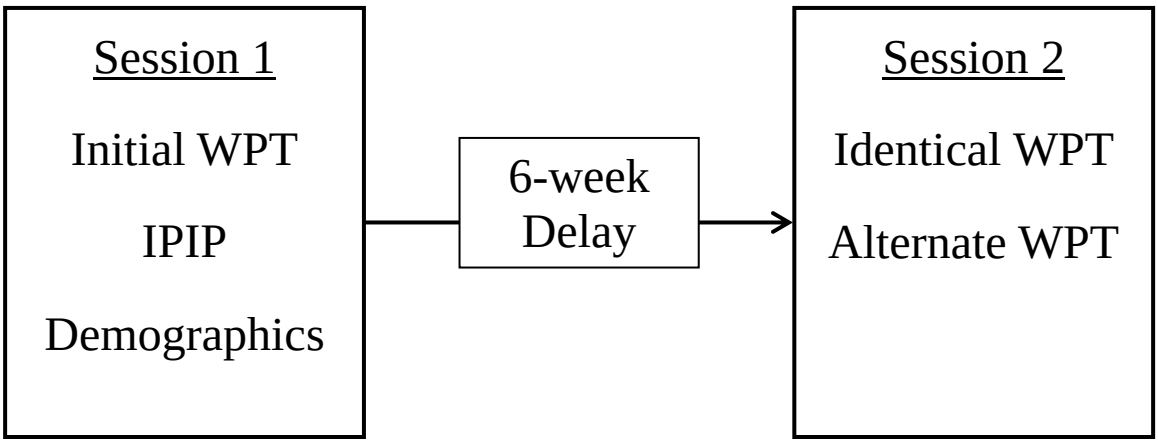

Figure 1. Outline of Study Protocol.

At Session 1, participants were administered one form of the Wonderlic Personnel Test (WPT), a personality assessment, and a demographics form. After a six-week interval, participants returned for Session 2 and were administered an identical and alternate form of the WPT. Form A and Form B of the WPT were designed to be equivalent in their assessment of GMA. Nevertheless, in order to control for potential order or test form effects, participants were randomly assigned to one of four conditions: A-AB, A-BA, B-AB, B-BA. Thus, at Session 1 participants were randomly assigned to complete either Form A or Form B. Upon their return to the laboratory, all participants completed two assessments of the WPT: one of each test form, and were randomly assigned such that they either took Form A before Form B, or vice versa. I found no evidence of test form or test order effects, thus conditions were collapsed into one group and for the remainder of this manuscript I refer to the three WPT with the following terminology: Initial, Identical, and Alternate. Initial represents the one test subjects took

during Session 1 (regardless of form). Identical represents the assessment taken during Session 2 that was the same form of the WPT taken at Session 1 (regardless of form or order). And alternate refers to the assessment taken during Session 2 that was not the same as the form of the WPT taken at Session 1 (regardless of form or order).

In order to approximate a retest interval that would more realistically reflect that seen in an organization's selection or promotion process, an interval of six weeks between test administrations was implemented. Such an interval is longer than the minimum interval suggested by the Wonderlic manual (2002) of at least one half hour (established to prevent cognitive fatigue). Other assessments of a similar nature suggest similar time-frames (e.g., GRE [60 days], GMAT [31 days]).

Measures

Wonderlic Personnel Test (WPT). The Wonderlic tests are used in employment settings to provide accurate, reliable measures of general cognitive ability (Wonderlic, Inc., 2002). The test contains 50 questions with a wide variety of problem types (e.g., disarranged sentences, number series, and story problems requiring mathematics or logic solutions) which are arranged in order of difficulty with the easiest first. Participants are allowed 12 minutes to complete the test. The average score reported in the test manual for all job applicants is approximately 21 and is equal to the average score for high school graduates; the mean reported in the manual for college graduates is 29 (2002). Meta-analytic results show the predictive validity of the WPT as .63 in its prediction of ability, and .33 in its prediction of college grades (Hunter & Hunter, 1984). Split-half (odd-even) reliability of the WPT yielded a correlation of .82 for Form A and .82 for Form B on Session 1 scores, and .84 for Form A and .82 for Form B on Session 2 scores. Other

estimates of reliability were similarly high, with test-retest reliability for identical forms at .86 and alternate form reliability at .83.

Personality. Participants completed one of the forms (Form A or Form B) of the 50-item version of the International Personality Item Pool NEO-PI (Goldberg, 1999), a Big-5 measure of personality, in order to assess participants' conscientiousness and emotional stability. Participants were provided with statements and asked to rate how accurately the statements describe them on a 5-point scale (1 = *very inaccurate* and 5 = *very accurate*). The scores on the conscientiousness (Form A α = .86; Form B α = .81) and emotional stability (Form A α = .87; Form B α = .80) subscales all demonstrated adequate levels of reliability.

Demographics and SAT. Participants completed a demographics form after taking the WPT in Session 1 in order to minimize triggering any potentially negative stereotypes that may have been induced by asking participants to identify their sex or race prior to taking the assessment. Participants were asked to identify their sex as either *Male* or *Female*, and were provided seven response options for race, including: *White/Caucasian; Black/African American; Asian; Latino/Hispanic; Native American; Middle Eastern*; and *Other* (with a write-in response space provided for *Other*). Participants provided consent for me to contact the University Office for Enrollment in order to obtain their official SAT scores.

Results

Table 1 reports the means, standard deviations, and intercorrelations of all study variables. Table 2 displays the means and standard deviations of all study variables by subgroup as well as the effect size (Cohen's *d*) comparing mean differences by subgroups.

Table 1

Intercorrelations, Means, and Standard Deviations of All Study Variables

Variable	*M*	*SD*	1	2	3	4	5	6	7	8	9	10
1. WPT Initial (Session 1)	30.09	7.10	-									
2. WPT Identical (Session 2)	33.59	7.21	.86	-								
3. WPT Alternate (Session 2)	32.45	7.18	.83	.85	-							
4. WPT Retest Difference[a] - Identical	3.50	3.75	-.23	.29	.07	-						
5. WPT Retest Difference[a] - Alternate	2.36	4.17	-.27	.00	.31	.52	-					
6. SAT Score	1420	113.21	.57	.58	.59	.00	.01	-				
7. Emotional Stability	3.16	0.72	.00	.02	-.03	.04	-.04	.02	-			
8. Conscientiousness	3.54	0.67	-.13	-.05	-.08	.14	.08	.09	.10	-		
9. Quadratic Term	50.15	64.27	-.44	-.48	-.45	-.09	-.02	-.07	.01	.02	-	
10. Sex	0.67	0.47	-.26	-.26	-.25	-.02	.01	-.14	-.27	.06	.06	-
11. Age	20.37	5.33	-.39	-.39	-.39	-.01	-.01	-.18	.01	.17	.30	.09

Note. $N = 243$ except for SAT $N = 199$. Sex is dummy-coded so that 0 = male and 1 = female. [a] Differences were computed so that positive values indicate Session 2 scores were greater than initial scores. For all relationships except those with SAT, correlations above .12 or below -.12 are significant, $p < .05$, two-tailed. For relationships with SAT, correlations above .13 or below -.13 are significant, $p < .05$, two-tailed.

Table 2
Means, Standard Deviations, and Mean-group Differences by Subgroup for All Study Variables

	Race					Sex				
	White/Asian		Black/Hispanic			Male		Female		
	M	SD	M	SD	*d*	M	SD	M	SD	*d*
Sample Size (*N*)[a]	170	-	72	-	-	81	-	162	-	-
Age	19.96	4.87	21.35	6.23	-0.26	19.69	3.56	20.71	6	-0.19
WPT Time 1	31.78	6.46	26.19	7.03	0.84**	32.67	6.23	28.80	7.17	0.56**
WPT Time 2 (Identical)	35.32	6.51	29.60	7.25	0.85**	36.28	6.57	32.24	7.16	0.58**
WPT Time 2 (Alternate)	34.16	6.51	28.50	7.15	0.84**	34.98	6.24	31.19	7.29	0.54**
WPT Retest Difference[b] —Identical	3.54	3.83	3.40	3.61	0.04	3.62	3.9	3.44	3.69	0.05
WPT Retest Difference[b] —Alternate	2.38	4.22	2.31	4.11	0.02	2.31	3.77	2.39	4.37	-0.02
SAT Score	1444	103.06	1356	106.54	0.84**	1439.6	110.3	1407.9	113.7	0.28
Emotional Stability	3.15	0.71	3.20	0.74	-0.06	3.44	0.61	3.02	0.73	0.60**
Conscientiousness	3.56	0.66	3.49	0.68	0.10	3.49	0.67	3.57	0.67	-0.12

Note. Mean group differences for each variable (Cohen's *d*) were computed such that a positive *d*-score indicates a higher score in favor of the majority (White/Asian and Male).
[a] Sample Size for the SAT variable only was *N* = 147 (White/Asian), *N* = 51 (Black/Hispanic), *N* = 72 (Male), and *N* = 125 (Female). [b] Differences were computed so that positive values indicate Session 2 scores were greater than initial scores.

**p < .001. *p < .05.

Hypotheses evaluating subgroup differences in retest gains (1a-b, 2a-b) were assessed with a mixed ANOVA using time (Session 1, Session 2) as the within-subjects independent variable and subgroup identification (race and sex) as the between-subjects independent variable. WPT scores served as the dependent variable.

Race Differences in Retest Performance

Hypothesis 1a. Hypothesis 1a predicted that Whites and Asians would improve more than Blacks and Hispanics when retested with an identical form of the WPT. The mixed ANOVA demonstrated a significant main effect for race, $F(1, 240) = 39.26$, $p < .001$, $\eta^2 = .14$, with means showing that the White-Asian group outperformed the Black-Hispanic group on the initial WPT assessment (White-Asian $M = 31.78$, Black-Hispanic $M = 26.19$) and for the identical form at session 2 (White-Asian $M = 35.32$, Black-Hispanic $M = 29.60$). There was also a main effect of session, $F(1, 240) = 171.67$, $p < .001$, $\eta^2 = .42$, with means showing that Session 2 scores on an identical form of the WPT ($M = 33.59$) were higher than initial scores ($M = 30.09$), indicating the presence of a retest effect. However, the interaction between race and session was not significant, $F(1, 240) = 0.07$, $p = .794$, $\eta^2 = .00$, indicating that the magnitude of retest gains on an identical form of the WPT did not differ for the two racial subgroups. As there was no difference in the rate of improvement by racial subgroup, Hypothesis 1a was not supported.

Hypothesis 1b. Hypothesis 1b predicted that Whites and Asians would improve more than Blacks and Hispanics when retested with an alternate form of the WPT. A similar mixed ANOVA showed a significant main effect for race, $F(1, 240) = 39.83$, $p < .001$, $\eta^2 = .14$, with the White-Asian group outperforming the Black-Hispanic group on

the initial WPT assessment (see above) and on the alternate form at Session 2 (White-Asian M = 34.16, Black-Hispanic M = 28.50). There was a main effect of session, $F(1, 240) = 63.31$, $p < .001$, $\eta^2 = .21$, with means showing that session 2 scores on an alternate form of the WPT (M = 32.45) were higher than initial scores (M = 30.09), indicating the presence of a retest effect. However, the interaction between race and session was not significant $F(1, 240) = 0.02$, $p = .896$, $\eta^2 = .00$, indicating that the magnitude of retest gains on an alternate form of the WPT did not differ for the two racial subgroups. Thus, there was no support for Hypothesis 1b, which predicted that retesting on an alternate form of the WPT would exaggerate racial differences.

Hypotheses 1c and 1d were based on the assumption that there would be racial differences in retest gains on the WPT (H1a and H1b) and predicted that GMA, as approximated by subjects' SAT scores, would account for more variance in the prediction of retest gains than race. Despite the lack of support for Hypotheses 1a and 1b, I conducted hierarchical regression analyses to determine if racial category (White-Asian and Black-Hispanic) accounted for additional variance above and beyond initial WPT scores, and SAT scores. The dependent variable in this equation was Session 2 scores (i.e., retest scores) on the WPT, with the form of the test differing for the two hypotheses (1c: identical, 1d: alternate). Table 3 displays the unstandardized regression coefficients (B) and their standard error ($SE\ B$), the standardized regression coefficients (β), R, R^2, adjusted R^2, and the change in R^2 for each step for the prediction of retest performance on an identical form of the WPT (H1c). Table 4 displays the same information for an alternate form of the WPT (H1d).

Table 3
Hierarchical Regression of Initial WPT Scores, SAT scores, and Race on Identical-form Retest Scores

Models and variables	*B*	*SE B*	β	*R*	R^2	Adjusted R^2	ΔR^2
Step 1				.74	.55	.55	.55*
WPT Initial (Session 1)	.73	.06	.74*				
Step 2				.76	.58	.58	.04*
WPT Initial (Session 1)	.60	.06	.61*				
SAT Score	.01	.00	.23*				
Step 3				.77	.59	.58	.00
WPT Initial (Session 1)	.59	.06	.60*				
SAT Score	.01	.00	.21*				
Race	-.79	.60	-.07				

Note. N = 198. Race is dummy-coded so that 0 = White and Asian and 1 = Black and Hispanic.
*$p < .001$.

Hypothesis 1c. The DV for Hypothesis 1c was the WPT identical form retest score. The IV in step 1 was WPT initial score, $R^2 = .55$, $F(1, 196) = 236.80$, $p < .001$. SAT was added as the second IV in step 2, $R^2 = .58$, $F(2, 195) = 136.38$, $p < .001$. Addition of SAT to the initial WPT scores resulted in a significant increment in R^2, $\Delta R^2 = .04$, $F(1, 195) = 16.83$, $p < .001$, accounting for an additional 4% of the variance in retest performance beyond the initial WPT. Racial category (White-Asian and Black-Hispanic) was introduced as the third IV in step 3, $R^2 = .59$, $F(3, 194) = 91.85$, $p < .001$. Addition of racial category to the model resulted in a non-significant increment in R^2, $\Delta R^2 = .00$, $F(1, 194) = 1.75$, $p = .187$, explaining no variance in retest performance beyond initial WPT scores and SAT. Thus, even after partialing out variance due to initial WPT scores, SAT—a measure of GMA distinct from WPT performance—explained significant amounts of variance in retest performance on an identical form of the WPT and racial category did not. Hypothesis 1c was supported.

Table 4
Hierarchical Regression of Initial WPT Scores, SAT scores, and Race on Alternate-form Retest Scores

Models and variables	*B*	*SE B*	β	*R*	R^2	Adjusted R^2	ΔR^2
Step 1				.67	.45	.45	.45*
WPT Initial (Session 1)	.66	.05	.67*				
Step 2				.72	.51	.51	.06*
WPT Initial (Session 1)	.50	.06	.50*				
SAT Score	.01	.00	.30*				
Step 3				.72	.52	.51	.00
WPT Initial (Session 1)	.49	.06	.40*				
SAT Score	.01	.00	.23*				
Race	-.61	.65	-.05*				

Note. N = 198. Race is dummy-coded so that 0 = White and Asian and 1 = Black and Hispanic.
*p < .001.

Hypothesis 1d. The DV for Hypothesis 1c was WPT alternate form retest score. The IV in step 1 was WPT initial scores, R^2 = .45, $F(1, 196) = 161.02$, $p < .001$. SAT was added as the second IV in step 2, R^2 = .51, $F(2, 195) = 102.52$, $p < .001$. Addition of SAT to the initial WPT scores resulted in a significant increment in R^2, ΔR^2 = .06, $F(1, 195) = 24.62$, $p < .001$, accounting for 6% of the variance in retest performance. Racial category (White-Asian and Black-Hispanic) was introduced as the third IV in step 3, R^2 = .52, $F(3, 194) = 68.60$, $p < .001$. Addition of racial category to the model resulted in a non-significant increment in R^2, ΔR^2 = .00, $F(1, 194) = 0.88$, p = .349, explaining no variance in retest performance beyond initial WPT scores and SAT. Thus, after partialing out variance due to initial WPT scores, SAT scores explained significant amounts of variance in retest performance on an alternate form of the WPT and racial category did not. Hypothesis 1d was supported.

Sex Differences in Retest Performance

Hypothesis 2a. The second set of hypotheses dealt with sex-based differences in retest gains. Hypothesis 2a predicted that women would improve more than men when retested with an identical form of the WPT. A mixed ANOVA with session as the within-subjects independent variable and sex as the between-subjects variable demonstrated a significant main effect of sex, $F(1,237) = 18.89$, $p < .001$, $\eta^2 = .07$, with means showing that men outperformed women on the initial WPT assessment (Male $M = 32.67$, Female $M = 28.80$) and on the identical form completed at session 2 (Male $M = 36.28$, Female $M = 32.24$). The main effect of session demonstrating the presence of a retest effect was significant $F(1, 240) = 171.67$, $p < .001$, $\eta^2 = .42$, with session 2 scores on an identical form of the WPT ($M = 33.59$) higher than initial scores ($M = 30.09$). However, the interaction between sex and session was not significant, $F(1, 237) = 0.41$, $p = .525$, $\eta^2 = .00$, suggesting that the magnitude of retest gains on an identical form of the WPT did not differ by sex. Hypothesis 2a was not supported.

Hypothesis 2b. Hypothesis 2b predicted that women would improve more than men when retested with an alternate form of the WPT. A similar mixed ANOVA showed a significant main effect for sex, $F(1, 237) = 17.82$, $p < .001$, $\eta^2 = .07$, with men outperforming women on the initial WPT assessment (Male $M = 32.67$, Female $M = 28.80$) and on the alternate form completed at Session 2 (Male $M = 34.98$, Female $M = 31.19$). The main effect of session indicative of retest effects for the alternate form of the WPT was significant, $F(1, 240) = 63.31$, $p < .001$, $\eta^2 = .21$, with means showing that session 2 scores on an alternate form of the WPT ($M = 32.45$) were higher than initial scores ($M = 30.09$)). The interaction between sex and session was not significant, $F(1,$

237) = 0.00, p = .959, η^2 = .00, suggesting that the magnitude of retest gains on an alternate form of the WPT did not differ by sex. Hypothesis 2b was not supported.

Hypotheses 2c and 2d were based on the assumption that there would be sex differences in retest gains on the WPT (H2a and H2b) and predicted that emotional stability and conscientiousness would account for more variability in the prediction of retest gains than the proxy of sex would. Despite the lack of support for Hypotheses 2a and 2b, I conducted hierarchical regression analyses to determine if sex accounted for additional variance above and beyond initial WPT scores, and personality variables (emotional stability and conscientiousness). The dependent variable in this model was Session 2 scores (i.e., retest scores) on the WPT, with the form of the test differing for the two hypotheses (2c: identical, 2d: alternate). Table 5 displays the unstandardized regression coefficients (*B*) and their standard error (*SE B*), the standardized regression coefficients (β), R, R^2, adjusted R^2, and the change in R^2 for each step for the prediction of retest performance on an identical form of the WPT (H2c). Table 6 displays the same information for an alternate form of the WPT (H2d).

Table 5

Hierarchical Regression of Initial WPT Scores, Emotional Stability, Conscientiousness, and Sex on Identical-form Retest Scores

Models and variables	*B*	*SE B*	β	*R*	R^2	Adjusted R^2	ΔR^2
Step 1				.86	.74	.74	.74*
WPT Initial (Session 1)	.88	.03	.86*				
Step 2				.86	.74	.74	.00
WPT Initial (Session 1)	.88	.03	.86*				
Emotional Stability	.21	.33	.02				
Step 3				.86	.75	.74	.00
WPT Initial (Session 1)	.88	.03	.87*				
Emotional Stability	.16	.33	.02				
Conscientiousness	.59	.36	.06				
Step 4				.87	.75	.75	.00
WPT Initial (Session 1)	.87	.03	.86*				
Emotional Stability	.03	.34	.00				
Conscientiousness	.62	.36	.06				
Sex	-.71	.54	-.05				

Note. N = 243. Sex is dummy-coded so that 0 = Male and 1 = Female.

*p < .001.

Hypothesis 2c. The DV for hypothesis 2c was WPT Identical form retest scores. The IV in step 1 was WPT initial score. This step accounted for 74% of the variance in retest performance, R^2 = .74, F(1, 241) = 699.82, p < .001. Emotional stability was added as the second IV in step 2, R^2 = .74, F(2, 240) = 349.26, p < .001. Addition of emotional stability to the initial WPT scores did not result in a significant increment in R^2, ΔR^2 = .00, F(1, 240) = 0.41, p = .524, accounting for no variance in retest performance beyond initial WPT scores. Conscientiousness was introduced as the third IV in step 3, R^2 = .75, F(3, 239) = 235.47, p < .001. Addition of conscientiousness to the equation with emotional stability also did not result in a significant increment in R^2, ΔR^2 = .00, F(1, 239) = 2.76, p = .098, and accounted for none of the variance in retest performance beyond initial WPT scores and emotional stability. Sex was introduced as the fourth IV in

step 4, $R^2 = .75$, $F(4, 238) = 177.59$, $p < .001$. Addition of sex to the model resulted in a non-significant increment in R^2, $\Delta R^2 = .00$, $F(1, 238) = 1.75$, $p = .187$, explaining no variance in retest performance beyond initial WPT scores, emotional stability, and conscientiousness. This pattern of results suggests that none of the hypothesized predictors: emotional stability, conscientiousness, or sex predicted variability in retest performance on an identical form of the WPT beyond the initial WPT scores. Thus, hypothesis 2c was not supported.

Table 6
Hierarchical Regression of Initial WPT Scores, Emotional Stability, Conscientiousness, and Sex on Alternate-form Retest Scores

Models and variables	*B*	*SE B*	β	*R*	R^2	Adjusted R^2	ΔR^2
Step 1				.83	.69	.69	.69*
WPT Initial (Session 1)	.84	.04	.83*				
Step 2				.83	.69	.69	.00
WPT Initial (Session 1)	.84	.04	.83*				
Emotional Stability	-.26	.36	-.03				
Step 3				.83	.69	.69	.00
WPT Initial (Session 1)	.84	.04	.83*				
Emotional Stability	-.29	.36	-.03				
Conscientiousness	.30	.39	.03				
Step 4				.83	.69	.69	.00
WPT Initial (Session 1)	.83	.04	.82*				
Emotional Stability	-.43	.38	-.04				
Conscientiousness	.33	.39	.03				
Sex	-.79	.59	-.05				

Note. $N = 243$. Sex is dummy-coded so that 0 = Male and 1 = Female.
*$p < .001$.

Hypothesis 2d. The DV for hypothesis 2d was WPT alternate form retest scores. The IV in step 1 was WPT initial score. This step accounted for 69% of the variance in retest performance, $R^2 = .69$, $F(1, 241) = 529.92$, $p < .001$. Emotional stability was added as the second IV in step 2, $R^2 = .69$, $F(2, 240) = 264.70$, $p < .001$. Addition of emotional

stability to predict retest performance did not result in a significant increment in R^2, ΔR^2 = .00, $F(1, 240) = 0.52$, $p = .470$, accounting for no variance in retest performance beyond initial WPT scores. Conscientiousness was introduced as the third IV in step 3, R^2 = .69, $F(3, 239) = 176.35$, $p < .001$. Addition of conscientiousness to the equation with emotional stability did not result in a significant increment in R^2, $\Delta R^2 = .00$, $F(1, 239) = 0.58$, $p = .446$, and accounted for none of the variance in retest performance beyond initial WPT scores and emotional stability. Sex was introduced as the fourth IV in step 4, $R^2 = .69$, $F(4, 238) = 133.15$, $p < .001$. Addition of sex to the equation with emotional stability and conscientiousness resulted in a non-significant increment in R^2, $\Delta R^2 = .00$, $F(1, 238) = 1.79$, $p = .183$, explaining no variance in retest performance beyond initial WPT scores, emotional stability, and conscientiousness. This pattern of results suggests that none of the hypothesized predictors: emotional stability, conscientiousness, or sex predicted variability in retest performance on an alternate form of the WPT. Hypothesis 2d was not supported.

Ability Differences in Retest Performance

Hypothesis 3 predicted that individuals higher in cognitive ability would exhibit larger score gains than those with lower ability when retested with an identical (3a) or alternate (3b) form of the WPT. Hierarchical regression was employed to determine if the addition of a curvilinear term for initial scores improved prediction of retest scores on the WPT beyond that afforded by the linear prediction of retest scores alone. The first IV in these equations was individuals' initial scores on the WPT, and the second IV was the squared product of initial scores on the WPT (quadratic term), with the DV depending on the test form of the Session 2 scores (identical: 3a, alternate: 3b). Table 7 displays the

unstandardized regression coefficients (*B*) and their standard error (*SE B*), the standardized regression coefficients (β), *R*, R^2, adjusted R^2, and the change in R^2 for each step for the prediction of retest performance on an identical form of the WPT (H3a). Table 8 displays the same for an alternate form of the WPT (H3b).

Table 7

Hierarchical Regression of the Linear and Quadratic Terms of Initial WPT Scores on Identical-form Retest Scores

Models and variables	*B*	*SE B*	β	*R*	R^2	Adjusted R^2	ΔR^2
Step 1				.86	.74	.74	.74*
WPT Initial (Session 1) Linear Term	.88	.03	.86*				
Step 2				.87	.76	.75	.01*
WPT Initial (Session 1) Linear Term	.82	.04	.81*				
WPT Initial (Session 1) Quadratic Term	-.01	.00	-.12*				

Note. $N = 243$.

*$p < .001$.

Hypothesis 3a. The DV for hypothesis 3a was WPT identical form retest scores. The IV in step 1 was WPT initial score, $R^2 = .74$, $F(1, 241) = 699.82$, $p < .001$. The quadratic term for initial WPT scores was added as the second IV in step 2, $R^2 = .76$, $F(2, 240) = 372.08$, $p < .001$. Addition of the quadratic term for initial scores to the initial scores resulted in a significant increment in R^2, $\Delta R^2 = .01$, $F(1, 240) = 12.10$, $p = .001$, accounting for 1% of additional variance in retest scores beyond the linear term. The adjusted R^2 value of .75 of the full model indicates that roughly three-quarters of the variability in retest scores on an identical form of the WPT is predicted by the linear and quadratic functions of initial scores.

Table 8

Hierarchical Regression of the Linear and Quadratic Terms of Initial WPT Scores on Alternate-form Retest Scores

Models and variables	*B*	*SE B*	β	*R*	R^2	Adjusted R^2	ΔR^2
Step 1				.83	.69	.69	.69**
WPT Initial (Session 1) Linear Term	.84	.04	.83**				
Step 2				.83	.70	.69	.01*
WPT Initial (Session 1) Linear Term	.79	.04	.78**				
WPT Initial (Session 1) Quadratic Term	-.01	.00	-.10*				

Note. $N = 243$.

*$p < .05$, **$p < .001$.

Hypothesis 3b. The DV for Hypothesis 3b was WPT alternate form retest scores. The IV in step 1 was WPT initial score, $R^2 = .69$, $F(1, 241) = 529.92$, $p < .001$. The quadratic term for initial WPT scores was added as the second IV in step 2, $R^2 = .70$, $F(2, 240) = 274.71$, $p < .001$. Addition of the quadratic term for initial scores to the equation with initial scores resulted in a significant increment in R^2, $\Delta R^2 = .01$, $F(1, 240) = 6.79$, $p = .010$, accounting for 1% of additional variance in retest scores beyond the linear term. The adjusted R^2 value of .69 of the full model indicates that just under three-quarters of the variability in retest scores on an alternate form of the WPT is predicted by the linear and quadratic functions of initial scores.

However, evidence of a quadratic trend alone does not answer the question whether higher- or lower-ability demonstrated a greater rate of increase between initial and retest scores: the direction of the trend is also important. Figure 2 displays the quadratic trend found in the prediction of retest scores on an identical form of the WPT, and Figure 3 displays the same trend for an alternate form of the WPT.

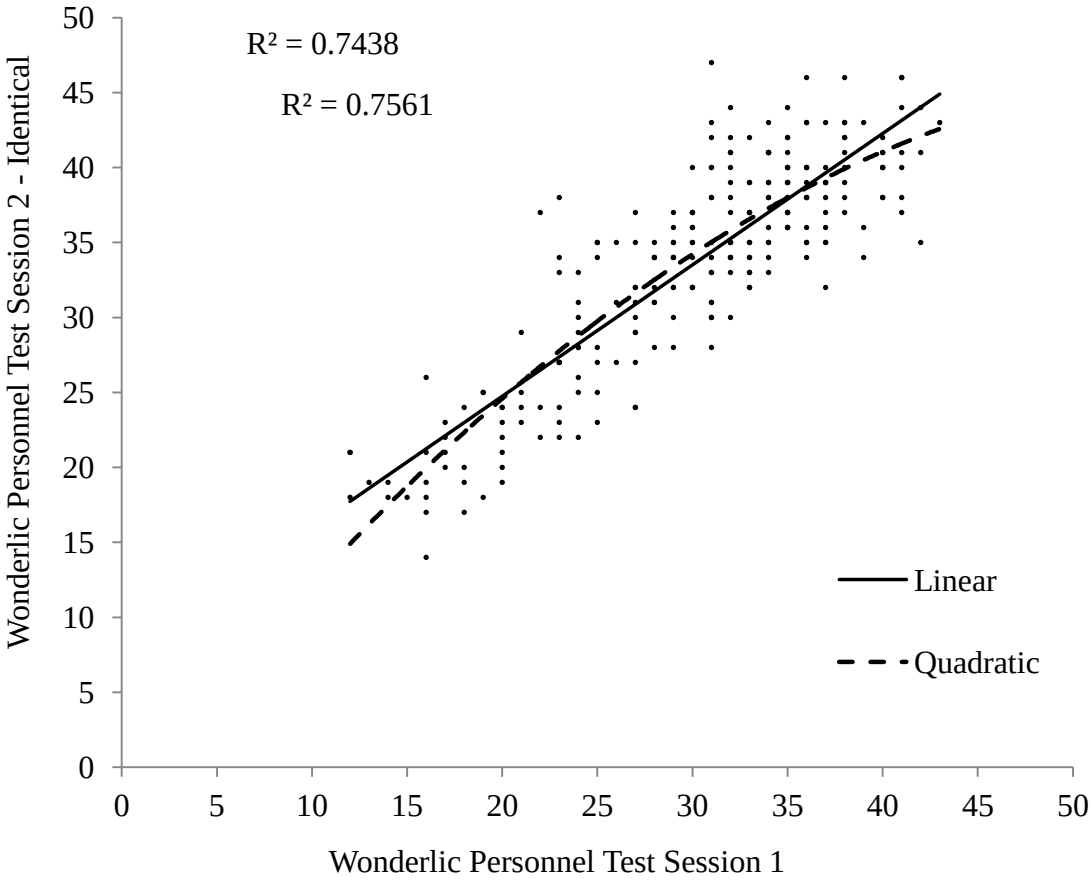

Figure 2. Prediction of Retest Performance on an Identical Form of the WPT by Linear and Quadratic Functions of Session 1 Scores.

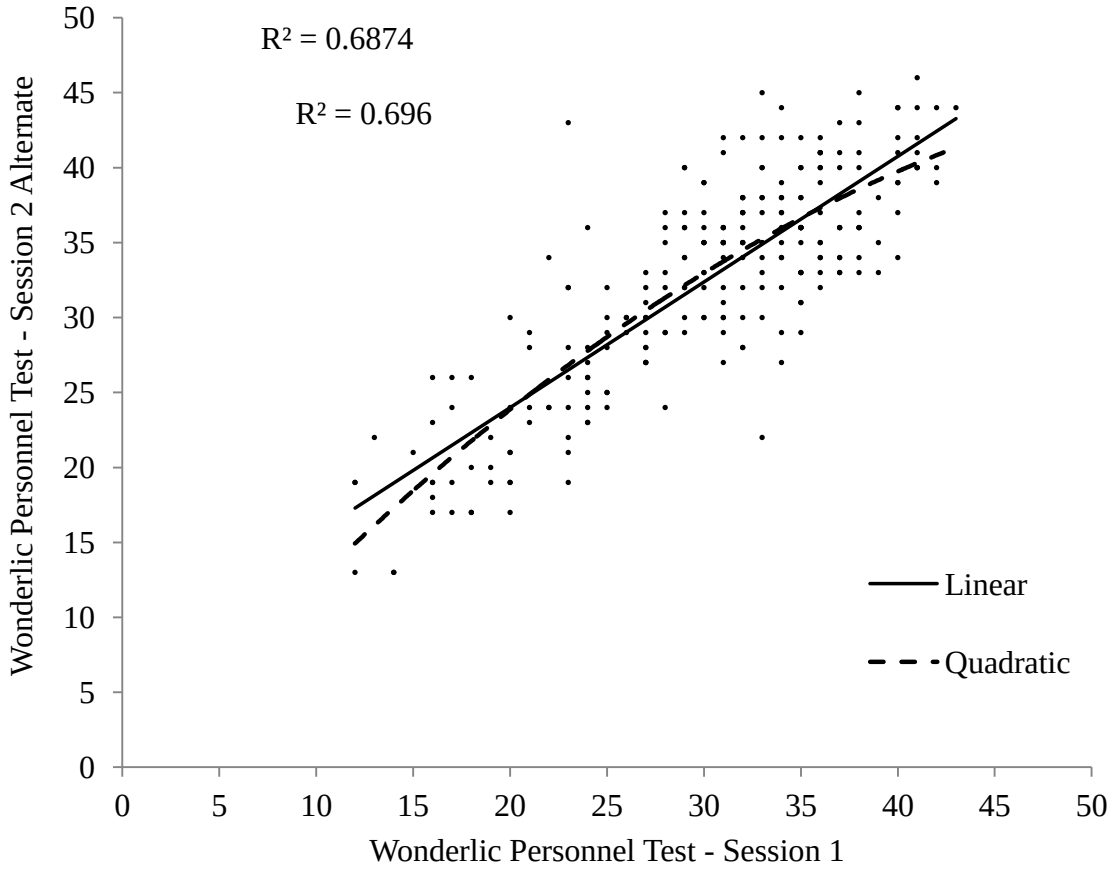

Figure 3. Prediction of Retest Performance on an Alternate Form of the WPT by Linear and Quadratic Functions of Session 1 Scores.

The trend is in fact a decelerating curve, rather than accelerating, and in comparison to the linear trend in the quadratic line predicts that people between roughly the 25th and 75th percentile (i.e., those scoring between 21 and 36 on the WPT at Session 1) would gain more on the retest administration than the linear trend predicts. Thus, the retest scores for those below about the 25th percentile and above the 75th percentile (i.e., below 21 and above 36 on the WPT at Session 1) are overestimated with the linear trend. This observation holds for both identical and alternate forms of the WPT (see Figure 2 and Figure 3, respectively). Thus, inconsistent with hypotheses 3a and 3b, the evidence suggests that individuals on both ends of the ability continuum (very high- and very low-

ability) actually improved less when retested than did those in the mid-ability range whether retest occurred with an identical or alternate form of the WPT.

Artificial Cut-off Score Analyses

Hypothesis 4 predicted that individuals who met or exceeded a predetermined cut-off level on their initial WPT assessment would gain more from retesting than would those who scored below the cut-off level. I considered three different cut-offs to assess this prediction. Tables 9, 10, and 11 present all relevant statistics for these data by Cut-off Score 1, 2, and 3, respectively. Figures 4, 5, and 6 also present the mean performance of the "pass" and "fail" groups on all WPT assessments for Cut-off Score 1, 2, and 3, respectively.

Table 9
Sample Size, Means, Standard Deviations and Mean-group Differences by Cut-off Score 1: A Score of 29 on the Initial WPT

	"Pass"			"Fail"			
	N	*M*	*SD*	*N*	*M*	*SD*	*d*
WPT Initial (Session 1)	158	34.42	3.54	85	22.02	4.52	3.17**
WPT Identical (Session 2)	158	37.63	3.91	85	26.07	5.77	2.49**
WPT Alternate (Session 2)	158	36.30	4.27	85	25.29	5.88	2.25**
WPT Retest Difference[a] — Identical	158	3.21	3.72	85	4.05	3.77	-0.22
WPT Retest Difference[a] — Alternate	158	1.87	4.16	85	3.27	4.07	-0.34*

Note. d-scores were calculated such that a positive value indicates an advantage for the "Pass" group.
[a] Retest differences were computed so that positive values indicate Session 2 scores were greater than Session 1 scores.

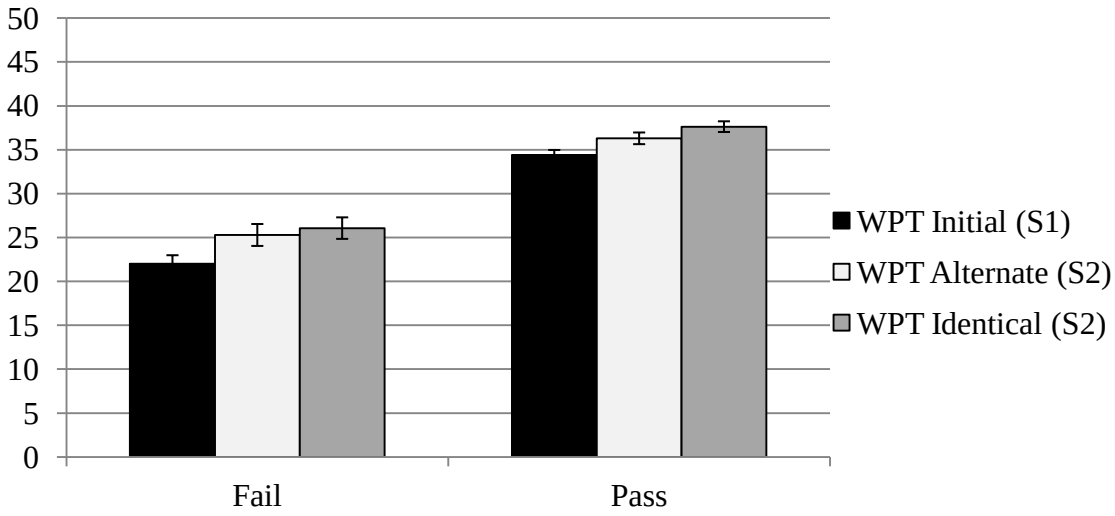

Figure 4. Mean Group Differences on the Three WPT Assessments for Cut-off Score 1: 29 on the Initial WPT. "Fail" represents those below the cut-off: $n = 85$. "Pass" represents those exceeding the cut-off: $n = 158$. Error bars represent the 95% confidence interval around the mean.

Cut-off Score 1. The first artificial cut-off was set at a score of 29 (the mean score for college graduates reported in the WPT manual) on the initial WPT assessment. 158 participants exceeded this cut-off level ($M = 34.42$) and 85 did not ($M = 22.02$). Those who exceeded cut-off level 1 gained an average of 3.21 points when retested with an identical form of the WPT and those below the cut-off gained an average of 4.05 points. However, as indicated by the failure to find an interaction between the two groups' retest scores on the identical form of the WPT, $F(1, 241) = 2.78$, $p = .097$, $d = -0.22$, these differences were not significant (a negative *d*-score indicates that those below the cut-off gained more by retesting than those above the cut-off). In contrast, for the alternate form retest, those exceeding the cut-off at Session 1 gained 1.87 points and those below the cut-off gained 3.27 points which was a significant, albeit relatively small

difference, $F(1, 241) = 6.33$, $p = .013$, $d = -0.34$, demonstrating that those below the cut-off gained more from retesting with an alternate form.

Table 10
Sample Size, Means, Standard Deviations and Mean-group Differences by Cut-off Score 2: The 70th Percentile on the Initial WPT

	"Pass"			"Fail"			
	N	*M*	*SD*	*N*	*M*	*SD*	*d*
WPT Initial (Session 1)	75	37.52	2.27	168	26.77	5.90	2.12**
WPT Identical (Session 2)	75	39.44	3.11	168	30.98	6.99	1.39**
WPT Alternate (Session 2)	75	37.79	3.87	168	30.07	7.03	1.24**
WPT Retest Difference[a] — Identical	75	1.92	3.34	168	4.21	3.72	-0.63**
WPT Retest Difference[a] — Alternate	75	0.27	3.36	168	3.30	4.17	-0.77**

Note. d-scores were calculated such that a positive value indicates an advantage for the "Pass" group.
[a] Retest differences were computed so that positive values indicate Session 2 scores were greater than Session 1 scores.

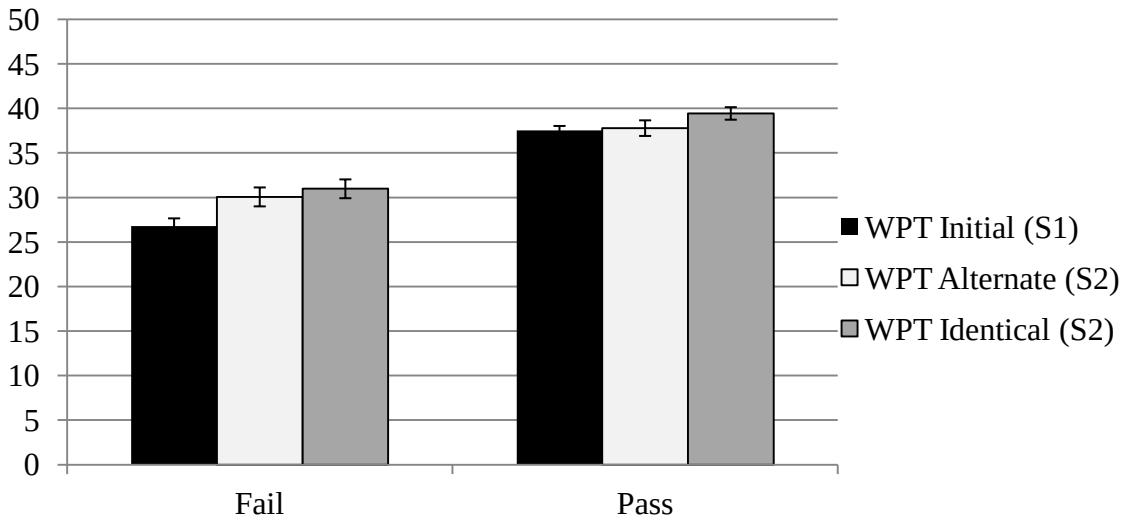

Figure 5. Mean Group Differences on the Three WPT Assessments for Cut-off Score 2: 70th Percentile on the Initial WPT. "Fail" represents those below the cut-off: $n = 168$. "Pass" represents those exceeding the cut-off: $n = 75$. Error bars represent the 95% confidence interval around the mean.

Cut-off Score 2. The second artificial cut-off was set at the 70th percentile for performance on the initial WPT assessment, meaning that only 30% of participants were in the pass group. This cut-off resulted in 75 individuals above the cut-off (M = 37.52) and 168 individuals below the cut-off (M = 26.77). Those scoring above the cut-off at Session 1 gained an average of 1.92 points when retested with an identical form of the WPT and those below the cut-off gained 4.21 points. The difference was significant, $F(1, 241) = 20.84$, $p < .001$, $d = -0.63$, providing evidence that those below the 70th percentile cut-off gained more from retesting than did those above the cut-off. Individuals above the cut-off also gained less (M = 0.27) when retested with an alternate form than those below the cut-off (M = 3.30) which was also a large and significant effect, $F(1, 241) = 30.71$, $p < .001$, $d = -0.77$.

Table 11
Sample Size, Means, Standard Deviations and Mean-group Differences by Cut-off Score 3: The Top 40 Scorers on the Initial WPT

	"Pass"			"Fail"			
	N	*M*	*SD*	*N*	*M*	*SD*	*d*
WPT Initial (Session 1)	43	39.07	1.78	200	28.16	6.28	1.90**
WPT Identical (Session 2)	43	39.81	3.3	200	32.25	7.13	1.14**
WPT Alternate (Session 2)	43	38.86	3.80	200	31.07	6.98	1.19**
WPT Retest Difference[a] — Identical	43	0.74	3.21	200	4.10	3.60	-0.95**
WPT Retest Difference[a] — Alternate	43	-0.21	3.26	200	2.92	4.15	-0.78**

Note. d-scores were calculated such that a positive value indicates an advantage for the "Pass" group.
[a]Retest differences were computed so that positive values indicate Session 2 scores were greater than Session 1 scores.

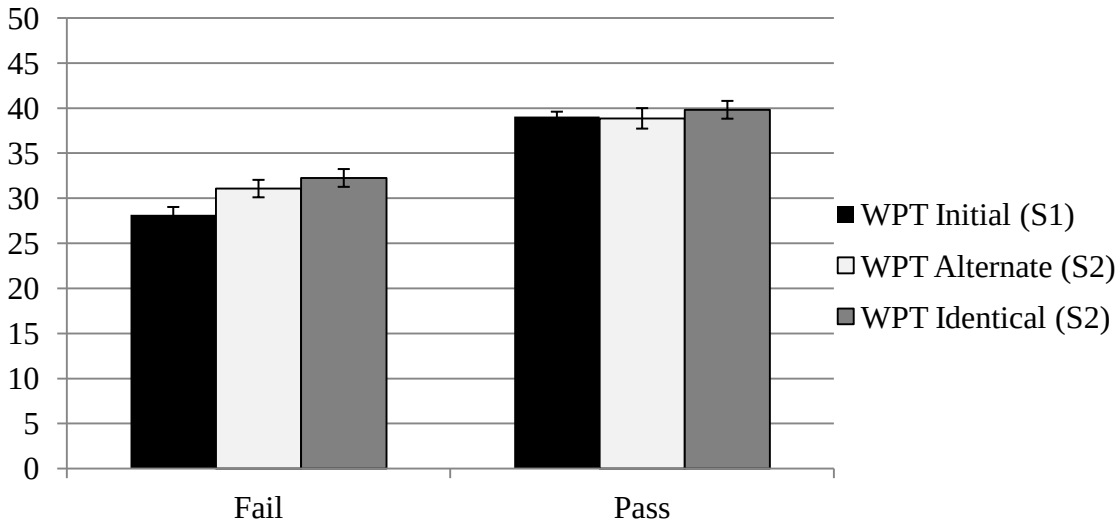

Figure 6. Mean Group Differences on the Three WPT Assessments for Cut-off Score 3: Top 40 Highest Scorers on the Initial WPT. "Fail" represents those below the cut-off: n = 200. "Pass" represents those exceeding the cut-off: n = 43. Error bars represent the 95% confidence interval around the mean.

Cut-off Score 3. The third and final artificial cut-off was a top-down selection approach where the top 40 scorers on the initial WPT assessment were classified as passing and all else were classified as failed. Due to scores being tied, 43 individuals were actually included in the pass group (M = 39.07) and 200 were in the fail group (M = 28.16). Those individuals in the pass group gained an average of 0.74 points when retested with an identical form of the WPT and those in the fail group gained an average of 4.10 points. The difference was significant and large, $F(1, 241) = 31.78$, $p < .001$, d = -0.95, suggesting those in the pass group gained less from retesting with an identical form than those in the fail group did. Similarly, the top group actually decreased their score by an average of 0.21 when retested with an alternate form of the WPT and those in the fail group gained an average of 2.92 points, $F(1, 241) = 21.52$, $p < .001$, d = -0.78. Thus, when the top 43 individuals were compared with the bottom 200, those in the bottom

group saw larger retest score increases on an alternate form of the WPT than those in the pass group.

Discussion

The primary objective of this study was to investigate the extent to which the practice of retesting disadvantages individuals based on race, sex, and ability-level. Additionally, I considered several mechanisms which might explain these differences. To further extend the investigation of ability-differences to a workplace setting, this study also considered how those passing or failing several artificial cut-offs differed in their retest performance.

Race and Sex Differences in Retest Performance

Although there was clear evidence of score gains when retaking both identical (t[242] = 14.54, dependent-d = 0.49) and alternate (t[242] = 8.82, dependent-d = 0.33) forms of the WPT, the results failed to provide any evidence that retesting advantaged or disadvantaged individuals based on their race or sex (see Dunlap, Cortina, Vaslow, & Burke, 1996, Equation 3, for the calculation of dependent d-scores). When retested, Blacks and Hispanics raised their scores on identical and alternate forms of the WPT as much as Whites and Asians: identical d = 0.04, alternate d = 0.02 (d-scores represent Session 2 performance – Session 1 performance, see Table 2). Males also gained as much as females when retested on identical and alternate forms of the WPT: identical d = 0.05, alternate d = -0.02. In line with previous research (Roth et al., 2001), I found evidence of mean differences by race on the initial WPT assessment in favor of the White/Asian racial subgroup (d = 0.84). In contrast to previous research (Neisser et al., 1996), I also found evidence of mean sex differences on the WPT in favor of Males (d = 0.56).

Nevertheless, the magnitude of these differences favoring Males and the White/Asian group remained unchanged at the retest administration: sex $d = 0.55$, race $d = 0.84$. Thus, failure to find race and sex differences in retest gains was not due to the absence of race and sex differences in test performance. Contrasted with evidence that race and sex differences in retest improvement exist for different selection tests and methods (Schleicher et al., 2010; Van Iddekinge et al., 2011), these null findings for subgroup differences on GMA retest performance should be comforting for organizations utilizing re-testing policies for ability assessments in the sense that retesting might not contribute to adverse impact beyond mean group differences on the test itself.

Underlying Mechanisms for Differences in Retest Performance

Corroborating the idea that retest differences do not depend on race and sex identification, only one of the three hypothesized mechanisms to explain differences in retest performance accounted for variance in retest gains. SAT scores predicted a significant amount of variance in retest performance on both an identical ($\Delta R^2 = .04$, $\beta = .21$, $p < .001$) and alternate ($\Delta R^2 = .06$, $\beta = .29$, $p < .001$) form of the WPT. The beta-weights are relatively large for the prediction of retest performance, especially given that initial scores on the WPT were included as the first step in the regression model. The positive direction of the beta weights indicates that SAT scores positively predict scores on the WPT at a retest administration, suggesting that high-scorers on the SAT also performed well when retested on the WPT. The finding that SAT scores can predict retest performance on an identical or alternate form of the WPT *after* partialing out initial WPT scores reinforces the idea that there is variance unaccounted for in retest scores that is not predicted by initial test performance, and that this variance can be explained by

individual differences. This is true even when the initial and additional predictors are assessments of the same construct. Thus, perhaps SAT scores capture components of ability or motivation that are not assessed by the WPT (which is itself a well-established test of cognitive ability, Hunter & Hunter, 1984; Wonderlic Inc., 2002), and display the importance of these unique aspects above those captured by any one assessment's measure of ability.

Similar to the inability of race and sex to predict retest scores, emotional stability and conscientiousness also did not contribute to the prediction of retest scores for either identical or alternate forms of the WPT. Failure to find effects may explain why the role of individual differences and especially non-ability differences in retest settings is largely ignored in the selection and testing literature (e.g., Reeve & Lam, 2007 is one noteworthy exception). This study highlights the difficulty in finding evidence for theoretically relevant individual differences (e.g., conscientiousness and emotional stability) to predict retest performance, while simultaneously highlighting the need for researchers to account for other individual differences (e.g., GMA). Moreover, it is still possible that personality variables will exert more of an influence in a more high-stakes setting or one in which feedback concerning test performance is provided: issues which shall be addressed in the limitations section.

Ability Differences in Retest Performance

The next part of this investigation tested whether individuals with higher ability-levels would gain more from retesting than those with lower ability. A curvilinear regression analysis was conducted to consider whether individuals who score higher on the initial WPT gain exponentially more when retested than those who score lower

initially. Although there was evidence of a curvilinear relationship in retest performance, contrary to my hypothesis the curve was decelerated, suggesting that moderate-ability individuals gain more from retesting than do either low- or high-ability individuals. A close examination of Figures 2 and 3, which displays the fit of the linear and quadratic function of Session 1 scores in the prediction of Session 2 scores on identical and alternate forms of the WPT (respectively) illustrates that people between roughly the 25th and 75th percentile (i.e., those scoring between 21 and 36 on the WPT at Session 1) would gain more on the retest administration than the linear trend predicts. Alternatively, retest scores for those below the 25th percentile and above the 75th percentile (i.e., below 21 and above 36 on the WPT at Session 1) are overestimated with the simple linear regression line. Thus, the evidence suggests that regardless of test form (i.e., identical or alternate) individuals on both ends of the ability continuum (very high- and very low-ability) improved less when retested than did those in the mid-ability range.

This finding challenges earlier investigations which categorized individuals into varying levels of ability and found that higher-ability individuals obtained larger score gains when retested (Kulik et al., 1984; Rapport et al., 1997). However, curvilinear regression analysis allows for a more nuanced test of ability-related differences in retest performance as it maintains the continuous nature of the data. Thus, the current findings extend previous research by making important qualifications suggesting that "the rich" (i.e., those high in GMA) may not always "get richer," but the middle-class might (Rapport et al., 375). This evidence, coupled with the large sample size (current study: N = 243, Rapport et al., 1997: N = 36) also lends more confidence to these results than those of earlier investigations. This finding suggests that average-scoring test-takers with

the option to retest may obtain larger score improvements than both relatively low-scoring and relatively high-scoring testers. Thus, it seems that those individuals within the interquartile range of the ability distribution may be the most successful at improving their score on a GMA assessment when they are retested relative to those in the lower quartile who may lack the ability to find ways to increase their score and to those in the upper quartile who may lack the room (in terms of the test scale) to increase.

There are several possible reasons that could explain why average-ability individuals may improve more than lower- and higher-ability individuals on GMA assessments. I will present two possible explanations for the under-improvement of high- and low-ability individuals. First, in accordance with Kanfer and Ackerman's (1989) resource allocation model, it may be that individuals with moderate levels of ability might have spent more of their Session 1 assessment in an earlier phase of skill acquisition requiring more cognitive effort due to inefficiencies in directing their attentional resources (Kanfer & Ackerman, 1989). Thus, preoccupation with understanding and following the test instructions and format, acquiring WPT-specific skills, or attending to the time limit or any test-irrelevant stimuli could have drawn away the more limited resources of these average-ability individuals, resulting in a lower score. In contrast, the highest-ability individuals may have more quickly acquired the test-specific skills (e.g., test-wiseness) and reached their plateau during Session 1 such that their retest performance remained relatively stable. On the other hand, low-ability individuals may have been so distracted or inefficient at skill acquisition and self-regulation during Session 1 to have not benefitted much from this initial assessment. Thus, by providing an opportunity to retest, average-ability individuals may have moved

past the more cognitively-taxing phase of skill acquisition after acquiring and executing any WPT-specific skills in order to better allocate their resources to test performance, resulting in higher scores. Future research should investigate these and other explanations for retest gains as moderated by ability-level in order to better understand this effect.

Second, an alternative explanation is that there could simply be a ceiling effect, such that individuals who scored higher on the WPT at Session 1 may not have much more room to improve before reaching the maximum score, whereas those who scored lower have much more room to improve. Although there was no evidence of a ceiling effect in the present study (WPT maximum score is 50 and the highest score obtained at Session 1 was 43 and at Session 2 was 46), there could still have been a partial ceiling effect. Nonetheless, this explanation is deficient in that it only explains the smaller retest score increases for people with higher levels of GMA, not those with lower levels.

Cut-off Scores and Retesting

The final goal was to demonstrate the consequences that retest effects may have on a common selection procedure: discrimination of test-takers on the basis of cut-off scores (Dwyer, 1996). In accordance with the ability-related differences in retest gains which challenged my predictions and the existing literature (Kulik et al., 1984; Rapport et al., 1997), analyses comparing retest gains for individuals above and below artificial cut-off levels on Session 1 scores support the idea that lower-scoring individuals gain more from retesting than do higher-scoring individuals. The first cut-off, a score of 29 on the WPT, was based on the mean reported score in the WPT manual for college graduates (Wonderlic, Inc., 2002) which was lower than the mean in this sample (M = 30.09, SD = 7.10) and thus resulted in a larger pass group (i.e., above the cut-off; N = 158) than fail

group (i.e., those below the cut-off; $N = 85$). At this cut-off there were no significant differences in retest gains between the groups on the identical form of the WPT ($d = -0.22$), however, there was a small advantage for those below the cut-off when retested on an alternate form of the WPT ($d = -0.34$; a negative d-score indicates that those below the cut-off gained more by retesting than those above the cut-off). When the cut-off was more stringent and set at the 70th percentile (i.e., the top 30th percentile passed), those above the cut-off level ($N = 75$) did not gain as much from retesting with an identical ($d = -0.63$) or alternate ($d = -0.77$) form of the WPT as those below the cut-off level ($N = 168$). These large, significant differences were even more pronounced for the final cut-off simulation where the top-scoring 43 individuals were selected to be in the pass group and the bottom-scoring 200 were in the fail group. Those above this cut-off were severely disadvantaged, gaining much less on identical ($d = -0.95$) and alternate ($d = -0.78$) forms of the WPT at retest administrations compared to the fail group.

Taken together, the pattern of cut-off score analyses suggests that individuals below cut-off levels may obtain larger score increases relative to individuals who exceed the cut-off level. Thus, to the extent that there are no other restraints or repercussions for retesting, it seems that lower-ability individuals who may have failed to meet a predetermined cut-off score have little to lose and much to gain when they re-take an assessment. An additional implication of the differences between the retest gains of those above and below these cut-off scores, and one reinforced by other analyses in this paper is that the retest phenomenon is not functionally equivalent across all levels of ability. Thus, it may be that the over-emphasis in the current literature on low-ability individuals,

or those more likely to be in a redemptive retesting situation (based on ability alone), prevents a comprehensive understanding of retest effects.

Limitations & Future Directions

There are several reasons why race and sex—the focal interest of the current study—may not have contributed to the explanation of variance in retest gains. One explanation could be the lack of variance due to the high correlations between the Session 1 and Session 2 scores on the WPT (Identical r = .86, Alternate r = .83). Nonetheless, the finding that SAT scores (which might be considered a less robust measure of cognitive ability compared to the WPT), predict retest performance on the WPT beyond initial scores provides evidence that there is enough variance in the pool of retest gains unaccounted for by initial scores to be predicted by other variables.

Additional limitations include characteristics of this study's testing conditions and the lack of feedback which typically drives the need for retesting in the first place. As this study was conducted in a lab where individuals were rewarded for merely completing the assessments, motivation to perform well on the assessment may have been less intense than motivation to perform well at initial or retest sessions in a high-stakes settings where important rewards depend on successful test performance (Arvey, Strickland, Drauden, & Martin, 1990). Moreover, all participants in this study were required to retest, so the choice to retest and the reasons underlying that choice were not represented in this paradigm, though they are certainly important to consider (Lievens et al., 2005; Messick & Jungeblut, 1981). Beyond direct influences of increased motivation on retest performance, higher stakes and higher motivation levels are likely to encourage individuals to seek out various forms of test-coaching to help prepare them for follow-up

assessments (Hausknecht et al., 2007; Messick & Jungeblut, 1981). In contrast, participants in this study were not provided feedback concerning their performance on the initial WPT assessment and also were not informed that they would be re-taking the assessment at their follow-up session. Thus, issues of redemptive versus non-redemptive retesting could not be accurately simulated in the current study as participants were unaware whether they exceeded a self- or other-set cut-off at their initial test session. Taking all of these limitations into consideration, it is safe to say that the findings provided here are likely lower-bound estimates of retest effects and differences.

Yet, despite the limitations of the current study, there are also a number of benefits to the approach taken. First, I was able to sample a wider range of ability as opposed to a restricted sample based on initial test performance (i.e., everyone in the sample retested). Second, the removal of explicit feedback and selection or reward decisions based on test performance was intended to minimize salient characteristics of motivation and attitudes in order to focus more directly on the goals of the current study: ability- and personality-based explanations of retest performance. Nevertheless, future research is needed to address retesting in a high-stakes setting, to evaluate the effects of test-taker motivation and test-taking attitudes, and to evaluate the effects of volition and reward. Even in a laboratory setting, valuable future research could be conducted on the effects of motivation on retest performance as variables such as initial performance feedback, opportunity to retest, and reward can be manipulated.

Finally, the failure to find differences for legally protected subgroups is itself evidence that race and sex differences in retest gains on ability assessments may not be particularly problematic. Instead, these findings may lend confidence to the utilization of

retesting in selection systems because the practice of retesting itself does not introduce or exaggerate subgroup differences. Moreover, individuals who are anxious that retest opportunities on GMA assessments will be disadvantageous to them because of their minority status, lower levels of emotional stability, or lower conscientiousness should be reassured that the practice of retesting alone does not exaggerate any subgroup differences—though it does preserve them. These implications are, of course, tempered by the fact that I only have retest data for one assessment of GMA. Future research should test for subgroup differences on various GMA assessments and should also examine mechanisms which might theoretically explain why certain subgroups would benefit from retesting more than others.

Conclusion

The purpose of this study was to examine the potential for score gains due to retesting on the WPT to differ depending on individuals' protected class (race and sex) and ability level, and to test what these findings might mean in a selection context. I found no evidence that retest gains are moderated by race or sex, implying that organizations may not be placing themselves at risk of manipulating subgroup differences on GMA test performance merely by asking employees or applicants to complete a follow-up assessment. There was also no indication that individuals' emotional stability or conscientiousness contributed to retest performance. However, SAT scores predicted variance in retest scores after controlling for initial test performance, suggesting that alternate GMA assessments besides the one used for retesting can explain variability in retest performance. Thus, the warning that retest performance might differ by race remains in effect as there remain mean group differences by race on the SAT, WPT, and

virtually all GMA assessments (Roth et al., 2001). There was also evidence of a curvilinear component in the prediction of retest scores, suggesting that individuals with moderate initial scores (i.e., within the interquartile range) on the WPT gain more from retesting than those with either low (i.e., lower quartile) or higher (i.e., upper quartile) initial scores. Furthermore, the establishment of artificial cut-off levels based on Session 1 scores demonstrated that those below the cut-off gained more when retested than those above the cut-off. Therefore, it may be most beneficial to encourage average-scorers and in some cases lower-scorers who may have failed to meet an important selection cut-off level to re-test as they have little to lose and much to gain when re-taking an ability assessment.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (1999). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

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

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.

Caretta, T. R., & Ree, M. J. (1995). Near identity of cognitive structure in sex and ethnic groups. *Personality and Individual Differences, 19*, 149-155.

Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology, 82*, 143-159.

De Corte, W., Lievens, F., & Sackett, P. R. (2007). Combining predictors to achieve optimal trade-offs between selection quality and adverse impact. *Journal of Applied Psychology, 92*, 1380-1393.

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.

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods, 1*, 170-177.

Dwyer, C. A. (1996). Cut scores and testing: Statistics, judgment, truth, and error. *Psychological Assessment, 8*, 360-362.

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality Psychology in Europe*, Vol. 7, 7-28. Tilburg, The Netherlands: Tilburg University Press.

Hattrup, K., Rock, J., & Scalia, C. (1997). The effects of varying conceptualizations of job performance on adverse impact, minority hiring, and predicted performance. *Journal of Applied Psychology, 82*, 656-664.

Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Gerrard, M. O. M. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology, 92,* 373-385. doi: 10.1037/0021-9010.92.2.373

Hogan, J., Barrett, P., & Hogan, R. (2007). Personality measurement, faking, and employment selection. *Journal of Applied Psychology, 92,* 1270–1285. doi:10.1037/0021-9010.92.5.1270

Hogan, R., & Hogan, J. (1995). *Hogan Personality Inventory manual*. Tulsa, OK: Hogan Assessment Systems.

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection, and amelioration of adverse impact in personnel selection procedures: Issues,

evidence, and lessons learned. *International Journal of Selection and Assessment, 9*, 152-194.

Hunter, J. E. & Hunter, R. F. (1984). Validity and utility of alternate predictors of performance. *Psychological Bulletin, 96*, 72-98.

Jensen, A. R. (1980). *Bias in mental testing*. New York: Free Press.

Kanfer, R. & Ackerman, P. L. (1989). Motivation and cognitive abilities: An integrative/aptitude-treatment interaction approach to skill acquisition. *Journal of Applied Psychology—Monograph, 74*, 657-690.

Kulik, J. A., Kulik, C. C., & Bangert, R. L. (1984). Effects of practice on aptitude and achievement test scores. *American Educational Research Journal, 21,* 435-447.

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

Lievens, F., Reeve, C. L., & Heggestad, E. D. (2007). An examination of psychometric bias due to retesting on cognitive ability tests in selection settings. *Journal of Applied Psychology, 92,* 1672-1682. doi: 10.1037/0021-9010.92.6.1672

Maccoby, E., & Jacklin, C. (1974). *The psychology of sex differences.* Stanford, CA: Stanford University Press.

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

Moutafi, J., Furnham, A., & Tsaousis, I. (2006). Is the relationship between intelligence and trait neuroticism mediated by test anxiety? *Personality and Individual Differences, 40*, 587-597.

Neisser, U., Boodoo, G., Bouchard, T. J., Jr., Boykin, A. W., Brody, N., Ceci, S. J., Halpern, D. F., Loehlin, J. C., Perloff, R., Sternberg, R. J., & Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51*, 77-101.

Pyburn, K. M., Jr., Ployhart, R. E., & Kravitz, D. A. (2008). The diversity-validity dilemma: Overview and legal context. *Personnel Psychology, 61*, 143-151.

Rapport, L. J., Brines, D. B., Axelrod, B. N., & Theisen, M. E. (1997). Full scale IQ as mediator of practice effects: The rich get richer. *Clinical Neuropsychologist, 11*, 375-380. doi:10.1080/13854049708400466

Reeve, C. L., & Lam, H. (2005). The psychometric paradox of practice effects due to retesting: Measurement invariance and stable ability estimates in the face of observed score changes. *Intelligence, 33*, 535-549. doi: 10.1016/j.intell.2005.05.003

Reeve, C. L., & Lam, H. (2007). The relation between practice effects, test-taker characteristics and degree of g-saturation. *International Journal of Testing, 7*, 225-242.

Roediger, H. L., III, & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*, 181-210.

Roth, P. L., BeVier, C. A., Bobko, P., Switzer, F. S., III, & Tyler, P. (2001). Racial differences in cognitive abilities: A meta-analysis. Personnel Psychology, 54, 297-330. doi: 10.1111/j.1744-6570.2001.tb00094.x

Sackett, P. R., Schmit, N., Ellingson, J. E., & Kabin, M. B. (2001). High-stakes testing in employment, credentialing and higher education. *American Psychologist, 56,* 302-318.

Salthouse, T. A., & Tucker-Drob, E. M. (2008). Implications of short-term retest effects for the interpretation of longitudinal change. *Neuropsychology, 22,* 800-811.

Schleicher, D. J., Van Iddekinge, C. H., Morgeson, F. P., & Campion, M. A. (2010). If at first you don't succeed, try, try again: Understanding race, sex, and age differences in retesting test score improvement. *Journal of Applied Psychology, 95,* 603-617. doi: 10.1037/a0018920

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. doi: 10.1037/0033-2909.124.2.262

Schmidt, L. A., & Riniolo, T. C. (1999). The role of neuroticism in test and social anxiety. *The Journal of Social Psychology, 139,* 394-395.

Schmitt, D. P., Realo, A., Voracek, M., & Allik, J. (2008). Why can't a man be more like a woman? Sex differences in Big Five personality traits across 55 cultures. *Journal of Personality and Social Psychology, 94,* 168-182.

Society for Industrial and Organizational Psychology. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green, OH: Author.

te Nijenhuis, J., van Vianen, A. E., & van der Flier, H. (2007). Score gains on *g*-loaded tests: No *g*. *Intelligence, 35,* 283-300.

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. doi: 10.1111/j.1744-6570.2006.00909.x

Van Iddekinge, C. H., Morgeson, F. P., Schleicher, D. J., & Campion, M. A. (2011). (2011, April 25). Can I retake it? Exploring subgroup differences and criterion-related validity in promotion retesting. *Journal of Applied Psychology*. Advance online publication. doi: 10.1037/a0023562

Verive, J. M., & McDaniel, M. A. (1996). Short-term memory tests in personnel selection: Low adverse impact and high validity. *Intelligence, 23,* 15-32.

Wonderlic, Inc. (2002). *Wonderlic Personnel Test & Scholastic Level Exam User's Manual*. Vernon Hills, IL: Wonderlic, Inc.