

**INCREASED RETEST SCORES ON COGNITIVE TESTS:  
LEARNING OR MEMORY EFFECTS?**

A Dissertation

by

ANDREW MICHAEL NABER

Submitted to the Office of Graduate and Professional Studies of  
Texas A&M University  
in partial fulfillment of the requirements for the degree of

**DOCTOR OF PHILOSOPHY**

Chair of Committee, Committee Members,	Winfred Arthur, Jr. Stephanie C. Payne Steven Smith Ernest Goetz
Head of Department,	Douglas Woods

August 2015

Major Subject: Psychology

Copyright 2015 Andrew Michael Naber

## **ABSTRACT**

Ability and knowledge tests are often used in academic and personnel settings to evaluate individuals. For numerous reasons, providing individuals the opportunity to retest is recommended by scientific and professional guidelines. Retesting consistently results in increased test scores. Despite its pervasiveness, it is unclear whether this retesting effect reflects test-specific memory effects or functions as a learning intervention that increases the underlying construct.

This dissertation's objective was to competitively investigate whether learning or memory best explains the retesting effect using a 2 (same versus alternate form retest)  $\times$  2 (corrective feedback versus no feedback) repeated measures mixed factorial experimental design. Three hundred forty participants completed ability and knowledge retests. Additionally, as potential explanatory variables for the retesting effect, participants completed measures of working memory capacity, general mental ability, and test attitudes. Participants also reported their retest interval behaviors.

The results were more in accord with a memory instead of a learning explanation. Alternate retest forms attenuated the retesting effect and slowed item response times. Retest increases occurred at similar magnitudes across constructs (knowledge and ability). Greater working memory capacity facilitated retest increases only on same form retests. Ultimately, the retesting effect does not appear to result in increases in the underlying construct.

## **ACKNOWLEDGEMENTS**

First and foremost, I would like to thank my committee chair and advisor, Dr. Winfred Arthur, Jr., for his mentorship on my dissertation and throughout my time as his student. I hope to carry his wisdom with me for the rest of my career and hopefully pass on some small part to my own advisees one day. Additionally, my humblest thanks go to my committee members, Dr. Stephanie Payne, Dr. Steve Smith, and Dr. Ernest Goetz, for their guidance, thoughtful criticism, and support throughout this dissertation. I wish to extend my gratitude to Bryan Edwards of Oklahoma State University who provided me the advice, resources, and encouragement.

I must thank the many wonderful people that I have had the pleasure of working with, socializing with, and caring about since I came to Texas. Both the faculty and my fellow graduate students made my time at Texas A&M University a meaningful experience, both professionally and personally. My sincerest gratitude goes to my friends and family who kept me sane and committed throughout my time at A&M. Finally, I would never have gotten this far in my life if it were not for my mother and father's unconditional encouragement and love. Thank you most of all.

## TABLE OF CONTENTS

	Page
ABSTRACT .....	ii
ACKNOWLEDGEMENTS .....	iii
TABLE OF CONTENTS .....	iv
LIST OF FIGURES.....	vii
LIST OF TABLES .....	ix
1. INTRODUCTION: INCREASED SCORES ON RETEST: MEMORY OR LEARNING EFFECTS? .....	1
2. THE Pervasiveness of the Retesting Effect .....	5
3. MOTIVATIONS FOR IMPLEMENTING RETESTING .....	7
3.1 Increasing Perceptions of Organizational Fairness .....	8
3.2 Reducing Adverse Impact .....	8
3.3 Enhancing Validity.....	10
3.4 Reducing Measurement Error .....	12
4. EXPLANATIONS FOR THE RETESTING EFFECT: LEARNING OR MEMORY EFFECTS? .....	14
4.1 Learning Effects .....	16
4.1.1 Learning Effects in Laboratory Settings .....	18
4.1.2 Learning Effects in Operational Settings .....	20
4.2 Memory Effects .....	24
4.2.1 Memory Effects in Laboratory Settings .....	25
4.2.2 Memory Effects in Operational Settings .....	27
5. THE PRESENT STUDY .....	29
5.1 Investigating the Retesting Effect using Alternate Test Forms.....	29
5.2 Investigating the Retesting Effect using Item Response Time .....	32
5.3 Investigating the Retesting Effect by Construct Domain.....	34
5.3.1 Ability.....	37
5.3.2 Knowledge .....	39
5.4 The Role of Additional Explanatory Variables for the Retesting Effect .....	42

5.4.1 Working Memory Capacity and Memory Effects.....	43
5.4.2 GMA and Learning Effects .....	44
5.5 Examining the Retesting Effect using Corrective Feedback .....	48
5.5.1 Corrective Feedback by Construct Domain .....	53
5.5.2 Corrective Feedback for Memory Effects: The Role of Working Memory Capacity .....	55
5.5.3 Corrective Feedback for Learning Effects: The Role of GMA .....	56
5.6 Controlling for the Influence of Test Attitudes on the Retesting Effect .....	58
5.7 Summary .....	60
 6. METHOD .....	63
6.1 Sample .....	63
6.2 Research Design .....	64
6.3 Power Analysis to Detect Hypothesized Effects.....	65
6.4 Measures.....	68
6.4.1 Knowledge Test.....	68
6.4.2 General Mental Ability (GMA).....	68
6.4.3 Alternate Forms .....	69
6.4.4 Working Memory Capacity .....	71
6.4.5 Test-taking motivation .....	71
6.4.6 Test anxiety .....	72
6.4.7 Retest Interval Behavioral Inventory .....	72
6.5 Procedure.....	72
 7. RESULTS.....	77
7.1 Descriptive Statistics and Control Variables.....	77
7.1.1. Test Attitudes .....	78
7.1.2. Retest Interval Behaviors .....	78
7.2 Hypothesis Testing.....	84
7.2.1 Hypothesis 1 .....	84
7.2.2 Hypothesis 2 .....	85
7.2.3 Hypothesis 3 .....	86
7.2.4 Hypothesis 4 .....	87
7.2.5 Hypothesis 5 .....	89
7.2.6 Hypothesis 6 .....	91
7.2.7 Hypothesis 7 .....	93
7.2.8 Hypothesis 8 .....	95
 8. DISCUSSION .....	98
8.1 Implications.....	107
8.1.1 Scientific Implications.....	108
8.1.2 Practical Implications .....	112

8.2 Limitations .....	114
8.3 Future Directions.....	116
8.3.1 Negative Transfer.....	116
8.3.2 Meta-Cognition .....	120
8.3.3 Working Memory Training .....	121
8.4 Summary and Conclusions.....	122
 REFERENCES.....	124
 APPENDIX A: MEASURES .....	153
Sample items from General Psychology Competency Examination (GPCE) 2.0 .....	153
Test Attitudes and Perceptions Survey.....	154
Retest Interval Behavioral Inventory .....	156
 APPENDIX B: GENERAL PSYCHOLOGY COMPETENCY EXAM (GPCE) MEASURE REVISION, UPDATE, AND REVALIDATION (CONTENT-RELATED) .....	157
Appendix B.1 GPCE SME Content Distribution Rating Form .....	169
Appendix B.2 GPCE Measure Review Booklet for Subject Matter Experts .....	170
 APPENDIX C: HYPOTHEZIZED PATTERN OF RESULTS FOR MEMORY VERSUS LEARNING EXPLANATIONS.....	172
 APPENDIX D: RAW SCORE FIGURES .....	180
 APPENDIX E: RETEST INTERVALS BY CONDITIONS.....	185

## LIST OF FIGURES

	Page
Figure 1 Hypotheses 1 and 3: Total test score (percentage) by same form retest versus alternate form retest over time .....	85
Figure 2 Hypothesis 2: Mean item response time by same form retest versus alternate form retest over time. ....	86
Figure 3 Hypothesis 4: GMA and working memory predicting retest score increases by same form retest versus alternate form retest .....	89
Figure 4 Hypothesis 5: Total test score (percentage) by same form retest versus alternate form retest over time (collapsed across knowledge and ability).....	91
Figure 5 Hypothesis 6: Mean item response time by same form retest versus alternate form retest with corrective feedback or no corrective feedback over time (collapsed across knowledge and ability) .....	93
Figure 6 Hypothesis 7: Total test score (percentage) by same form retest versus alternate form retest with corrective feedback or no corrective feedback on knowledge and ability tests over time.....	95
Figure 7 Hypothesis 8: GMA and working memory predicting retest score increases by same form retest versus alternate form retest with corrective feedback and no feedback .....	97
Figure B.1 Sample item from GPCE Measure Review Booklet for Subject Matter Experts.....	160
Figure B.2 Initial GPCE content distribution goal (black) versus the current distribution of items rated as acceptable from SMEs (gray)....	165
Figure B.3 Final GPCE content distribution goal (black) versus the current distribution of items rated as acceptable from SMEs (gray)....	167
Figure C.1 Hypotheses 1a and 3a versus 1b and 3b: Total test score (percentage) by same versus alternate form retest over time.. ....	172

Figure C.2	Hypothesis 2a versus 2b: Mean item response time by same versus alternate form retest over time. ....	173
Figure C.3	Hypothesis 4: Hypothesis 4: GMA and working memory predicting retest score increases by same form retest versus alternate form retest. ....	174
Figure C.4	Hypothesis 5a versus 5b: Total test score (percentage) by same versus alternate form retest over time. ....	175
Figure C.5	Hypothesis 6a versus 6b: Mean item response time by same versus alternate form retest with corrective feedback or no corrective feedback over time. ....	176
Figure C.6	Hypothesis 7a versus 7b: Total test score (percentage) by same versus alternate form retest with corrective feedback or no corrective feedback over time .....	178
Figure C.7	Hypothesis 8a versus 8b: GMA and working memory predicting retest score increases by same form retest versus alternate form retest with corrective feedback and no feedback .....	179
Figure D.1	Hypotheses 1 and 3: Total test score by same versus alternate form retest over time. ....	180
Figure D.2	Hypothesis 2: Mean item response time by same versus alternate form retest over time. ....	181
Figure D.3	Hypothesis 5: Total test score by same versus alternate form retest over time. ....	182
Figure D.4	Hypothesis 6: Mean item response time by same versus alternate form retest with corrective feedback or no corrective feedback over time .....	183
Figure D.5	Hypothesis 7: Total test score (percentage) by same versus alternate form retest with corrective feedback or no corrective feedback over time. ....	184

## LIST OF TABLES

	Page
Table 1 List of Competitive Hypotheses .....	61
Table 2 Research Design and Measures.....	65
Table 3 Power Analysis for Detecting a One-Tailed Difference Between Two Independent Correlations .....	67
Table 4 Test-Retest Reliability by Test Forms.....	70
Table 5 Protocol for Time 1 and Time 2 Administration .....	76
Table 6 Mean Differences Between Participants Who Completed and Attrited from the Retest Protocol. ....	78
Table 7 Mean Differences Between Participants Who Covered Test Content During Retest Interval. ....	80
Table 8 Mean Differences Between Participants Who Sought Test Content During Retest Interval. ....	81
Table 9 Intercorrelations amongst Study Variables. ....	82
Table 10 Competitive Hypotheses supporting either Memory versus Learning Explanations .....	100
Table 11 Frequency Distribution of Mock Pass/Fail Retest Scores .....	117
Table 12 Frequency Distribution of Mock Pass/Fail Retest Scores by Same versus Alternate Retest Form .....	118
Table B.1 Subject Matter Expert Ratings of General Psychology Competency Exam .....	162
Table E.1 Descriptive Statistics for Retest Intervals Across Form Manipulations .....	185

Table E.2 Descriptive Statistics for Retest Intervals across Feedback Manipulation .....	185
---	-----

## **1. INTRODUCTION: INCREASED SCORES ON RETEST: MEMORY OR LEARNING EFFECTS?**

Standardized tests of cognitive ability and knowledge are commonly used in academic and personnel settings to evaluate test-takers, and hence, often operate as gatekeepers for educational and employment opportunities. Providing individuals the opportunity to retest is recommended in terms of both scientific and professional guidelines (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA/APA/NCME], 2014; Equal Employment Opportunity Commission [EEOC], 1978; Society for Industrial and Organizational Psychology [SIOP], 2003). Not surprisingly, many test-takers choose to retest in order to improve their chances of attaining these educational and employment opportunities, especially after failing previous attempts (Brounstein & Holahan, 1987; Cliffordson, 2004; Lievens, Buyse, & Sackett, 2005; Tuzinski, Laczo, & Sackett, 2005). A commonly observed effect of retesting is an increase in test scores. Despite the pervasiveness of this retesting effect, which is robust across both knowledge and ability construct domains, as well as academic and applied settings (Hausknecht, Halpert, Di Paolo, & Moriarty Gerrard, 2007; Kulik, Kulik, & Bangert, 1984), it is unclear whether these increases reflect test-specific memory effects or meaningful increases in the assessed construct domain.

Interestingly, explanations for the retesting effect are often embedded within particular disciplines examining particular content domains. Thus, cognitive psychology and education researchers who focus more on knowledge-based tests explain the

retesting effect in terms of learning. In contrast, personnel psychologists who are generally focused on ability tests, explain the retesting effect in terms of memory-related effects; primarily as a source of error variance. Due to this domain-specific embeddedness, little if any research simultaneously examines the competing learning and memory effect explanations for retest score increases across construct domains.

In the context of the retesting effect, this discussion is confined to *cognitive* abilities and cognitively-loaded constructs such as knowledge. Whereas there is a broad literature on retesting in noncognitive domains, such as personality, the mechanism for retest score increases on noncognitive constructs are quite different (i.e., self-presentation, faking) from those within cognitively-loaded domains (i.e., memory, learning). Thus, the use of the term ability will be subsequently confined to cognitive abilities, and the use of the term retesting effect will be confined to the explanations of memory or learning effects on cognitively-loaded constructs.

Comparatively investigating the memory and learning accounts of the retesting effect advances the understanding of this phenomenon and presents practical implications and recommendations as well. For instance, in educational contexts, repeated retesting has been advanced as a means of actually improving learning (Carpenter, 2012). Thus, it seems plausible that testing not only assesses test-takers but may actually act as an intervention to increase the test-takers' standing on the underlying construct (Roediger & Karpicke, 2006a). Hence, testing has been recommended as a proactive technique for educators to enhance learning (Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Pashler, Bain, Bottge, Graesser, Koedinger, McDaniel, &

Metcalfe, 2007). In contrast, in high-stakes testing contexts, personnel psychologists have argued that the retesting effect threatens the construct-related validity of tests, and subsequently, tests' criterion-related validity as well (Lievens, Reeve, & Heggestad, 2007). Thus, depending on whether one takes the perspective that retest score gains can best be explained by learning or memory effects, one would draw quite different inferences from retest scores for both assessment and educational purposes.

Learning occurs when the test-taker's initial test functions as an intervention to further develop the underlying construct. In contrast, memory effects occur when a test-taker increases his/her retest score due to remembering previous responses and test-specific information from initial testing with no concomitant development in the underlying construct. Clearly, learning and memory processes (and consequently, the assessments related to those processes) are inextricably interconnected. In terms of the present study, learning is operationally defined as retest score increases that *do* transfer across test forms; whereas memory effects are operationally defined as retest increases that do *not* transfer across test forms. Considering the pervasiveness of the retesting effect across settings and constructs, as well as the differing motivations for offering retesting opportunities (including, fairness, adverse impact, validity, and measurement error), the inferences drawn from retesting scores as reflecting either learning or memory has profound implications for test-takers, organizations, and society.

Furthering the understanding of the competing theoretical explanations for the retesting effect could potentially be a major scientific and applied contribution to the field. Consequently, this dissertation comparatively investigates these two competing

schools of thought—specifically, whether learning or memory best explains observed test score increases upon retesting.

## **2. THE Pervasiveness of the retesting effect**

Score increases upon retest have been observed across numerous cognitive construct domains and settings, including classroom achievement tests (Friedman, 1987), general mental ability (GMA) tests (Kaufman, 1994), college admissions exams (Donlon & Angoff, 1971), credentialing exams (Geving, Webb, & Davis, 2005), and employment tests of various types (Sackett, Burris, & Ryan, 1989; Wing, 1980). These retest score increases are generally in the small to medium effect size range for both knowledge and ability tests ( $d = 0.26$ , Hausknecht et al., 2007;  $d = 0.27$ , Lievens et al., 2005;  $d = 0.48$ , Raymond, Neustel, & Anderson, 2007;  $d = 0.15$ , Schleicher, Van Iddekinge, Morgeson, & Campion, 2010), however, some large retesting gains have been observed as well ( $d = 0.79$ ; Raymond et al., 2007;  $d = 0.93$ , Van Iddekinge, Morgeson, Schleicher, & Campion, 2011). These gains are also robust across numerous underlying cognitive abilities (visual perception, mechanical comprehension, and selective attention [Matton, Vautier, & Raufaste, 2009]; numerical ability and abstract reasoning [Matton, Vautier, & Raufaste, 2011]; and short-term memory [Watson, Pasteur, Healy, & Hughes, 1994]). Furthermore, final test performance exhibits score improvements even after *multiple* initial tests compared to following a single practice test in both knowledge (e.g., Karpicke & Roediger, 2007, 2010; Logan & Balota, 2008; Pavlik & Anderson, 2005) and ability content domains (Hausknecht et al., 2007), indicating that retest increases continue to occur (to some extent) across successive testing administrations.

The frequency and incidence of retesting is not trivial. In academic settings, as much as 50% of high school students retake the SAT I (Nathan & Camara, 1998), 30% of MCAT candidates retest (Koenig & Leger, 1997), and 40% of applicants retested for medical school in Belgium (Lievens et al., 2005). In organizational settings, studies have reported that 11% of retail manager candidates (Tuzinski et al., 2005), 32% of government employees (Van Iddekinge et al., 2011), and 25% of law enforcement candidates (Sin, Farr, Murphy, & Hausknecht, 2004) chose to retest. Not surprisingly, 70% of the candidates who initially failed a promotion test chose to retest (Van Iddekinge et al., 2011). Despite the extensive research investigating the retesting effect, organizations, educators, and policy-makers allow or forbid retesting for a variety of reasons, irrespective of the learning or memory effects explanation.

### **3. MOTIVATIONS FOR IMPLEMENTING RETESTING**

Whereas the incentives for applicants to retest are apparent, organizations and policy-makers allow retesting for a variety of motivations. Offering retesting opportunities within selection and promotion systems is commonplace among large private and public organizations (Muchinsky, 2004), and educational institutions (Wheeler, 2004). Despite the prevalence of retesting, and the professional advice to allow it (AERA/APA/NCME, 2014; EEOC, 1978; SIOP, 2003), there is a great deal of variability in the specific recommendations regarding retesting policies and procedures (Hausknecht et al., 2007; Raymond et al., 2007). For example, some major standardized test developers and publishers (e.g., SAT, GRE, LSAT, GMAT) allow retesting, albeit with restrictions on the minimum time intervals before retest. Other test publishers recommend specific procedures for retesting (e.g., the Miller Analogies Test notes that “if an examinee’s second [or most recent] test score is 25 points or greater than the first [or most recent previous] score, the second score is invalidated”, Harcourt Assessment, 2005). Some strictly prohibit retesting altogether (e.g., United States Air Force’s Basic Attributes Test, a test battery used to select military pilots, [Carretta, Zelenski, & Ree, 2000]). Differences between these retesting policies reflect a lack of consensus for how to address the retesting effect operationally.

Parallel to the competing learning and memory effect explanations for the retesting effect, retesting policy has received increased attention in recent years among both educators (Dunlosky et al., 2013) and personnel practitioners (e.g., Bourdeau, 2008; Wheeler, 2004). In the absence of integrated research across different disciplines,

organizations and policy-makers that allow retesting likely do so based on four fundamental motivations, which are: (1) increasing perceptions of organizational fairness, (2) reducing adverse impact, (3) enhancing validity, and (4) reducing measurement error.

### **3.1 Increasing Perceptions of Organizational Fairness**

Organizations seeking to attract and retain applicants prefer testing procedures that appear fair to applicants. During the selection process, test-takers desire adequate opportunities to demonstrate their job qualifications, which may be influenced by the retesting policy (Schleicher, Venkataramani, Morgeson, & Campion, 2006), that in turn may affect test-takers' perceptions of the organization's procedural justice (Arvey & Sackett, 1993; Gilliland, 1993; Gilliland & Steiner, 2001). Beyond the perceived fairness of selection procedures by test-takers, retesting policy may also affect test-takers' subsequent performance within an organization. Hausknecht, Trevor, and Farr (2002) found that applicants who retested, compared to applicants who were hired after the initial test, exhibited greater organizational commitment, greater training performance, and reduced turnover. Thus, organizations that wish to be perceived fairly and encourage committed test-takers may allow retesting for this reason alone, irrespective of whether retest increases reflect learning or memory effects.

### **3.2 Reducing Adverse Impact**

Hausknecht et al. (2002) noted that, "systematic score changes across repeated administrations can result in a qualitative change in the make-up of the workforce" (p. 244). This statement not only reflects fluctuations in test scores within a test-taker

across testing administrations, but also the possibility that the retesting effect exhibits systematic differences between subgroups of test-takers (e.g., by race, sex, and age). Regardless of the literature, organizations seeking employee diversity sometimes use retesting in the potentially erroneous belief that retesting opportunities will reduce subgroup differences and in turn, reduce the risk of adverse impact in final selection decisions (Ployhart & Holtz, 2008). However, in light of the numerous factors influencing selection at different stages of the process (e.g., differential decisions to retest, differential magnitude of the retesting effect, selection drop-out rates across subgroups), retesting appears to exacerbate subgroup differences for some groups of test-takers (i.e., older candidates, African-Americans), while reducing subgroup differences in others (i.e., females, Schleicher et al., 2010; Van Iddekinge et al., 2011).

Previous research indicates that Whites demonstrated significantly larger score improvements upon retest than African-Americans and Hispanics on selection tests of job knowledge and verbal ability; whereas females exhibit significantly larger score increases upon retesting than men on performance tests, yet not ability measures (Schleicher et al., 2010; Van Iddekinge et al., 2011). Additionally, test-takers under the age of 40 demonstrated significantly larger score gains compared with test-takers aged 40 and older (Schleicher et al., 2010; Van Iddekinge et al., 2011). Thus, research indicates that the retesting effect differentially affects subgroup differences in final test scores, and therefore may affect adverse impact heterogeneously across subgroups within organizations.

Opportunities to retest may also be disproportionately taken, consequently compounding these heterogeneous retest score increases (Schleicher et al., 2010; Van Iddekinge et al., 2011). Data show differential selection drop-out rates across subgroups following initial testing (Schmit & Ryan, 1997) and retesting (Sin et al., 2004). As Hausknecht et al. (2002) noted, (re)testing *is* likely to affect employment opportunities and therefore the diversity of entire organizations, but retesting may in fact exacerbate adverse impact due to the subgroup differences between retest scores and differential drop-out rates following (re)testing. Accordingly, Ployhart and Holtz (2008) concluded that the implementation of retesting was *not* effective in directly reducing subgroup differences, and therefore unlikely to reduce the adverse impact of a test.

### **3.3 Enhancing Validity**

Ability and knowledge tests are among the best predictors of job performance (Dye, Reck, & McDaniel, 1993; Schmidt & Hunter, 1998), and are consequently, widely used to facilitate selection and staffing decisions. Thus, implementing a retesting policy that enhances the operational validity of these measures should include distinguishing between retest score gains that are primarily reflecting learning versus memory effects, as these will result in more interpretable and therefore useful test scores.

Some researchers have posited that retesting enhances validity through two possible mechanisms: test familiarity and learning. That is, initial testing provides practice with the types of items encountered on the test, reducing error due to unfamiliarity with the test's format, and therefore enhancing validity. Additionally, testing may enhance validity by functioning as a learning intervention on the construct

itself (Anastasi, 1981; Anastasi & Urbina, 1997; Maurer, Salomon, & Troxtel, 1998).

Anastasi and Urbina (1997) further distinguish between these two mechanisms as retesting may enhance validity through *practice* (raising scores and validity through practice on specific operationalizations by reducing error) on one hand, and *teaching* on the other (learning effects, through raising the true score across operationalizations).

Anastasi (1981) posited that short test orientations (i.e., practice) would increase a test's validity, whereas long-term coaching would reduce validity. However, little if any research has drawn this nuanced distinction between practice and learning, and the present study does not draw this distinction either, but instead focuses on investigating the more fundamental learning and memory explanation for the retesting effect.

In either case, if the retesting effect meaningfully increases test-takers' standing on the construct of interest, then retesting allows an organization to make more informed decisions regarding the test-takers' true standing on the construct of interest (Lievens et al., 2005). Conversely, if retest scores are *not* as valid as initial test scores, but instead reflect memory of the previous test administration, then allowing retesting may result in the selection of less qualified workers and ultimately reduce organizational effectiveness (Van Iddekinge et al., 2011). Although it is clear that test users wish to make the best use of the inferences drawn from test scores, it is not necessarily apparent whether retest scores reflect enhanced validity (i.e., learning) or score increases confounded with previous test experience (i.e., memory).

### **3.4 Reducing Measurement Error**

Organizations may also permit retesting under the assumption that an individual's initial assessment was subject to heterogeneous and random measurement error; that is, some transient characteristics of the test-taker or the testing environment contaminated test scores but may or may not be present upon a secondary assessment (Lievens et al., 2005). For example, test-taker's mood, illness, or distraction, or the nuances in the test's administration may affect an individual test-taker's performance. Under the assumption that the operational validity of the *particular* test-taker's initial test score is somehow contaminated due to this random error, organizations may allow a retest where this random error is less likely to be present (true, or not). The primary influences of measurement error in interpreting retesting scores have been investigated in the context of regression to the mean.

In operational settings, typically only the test-takers who were *not* selected (i.e., test-takers who performed poorly) will retest because the test-takers who performed well are selected and do not retest. Thus, in operational contexts, disproportionately lower performing test-takers than the initial test-taking pool will retest while higher performing test-takers (who were selected) will not. Consequently, the assumed regression to the mean is upward (i.e., an increase in test scores) because it is the lower-performing test-takers who upon retesting, are improving their scores. Furthermore, Lievens et al. (2005) found some evidence that individuals who scored particularly low on an initial test were more likely to retest. Thus, retest applicant pools will invariably contain fewer test-takers than the initial test-taking pool and a greater proportion of lower-performing

test-takers, potentially leading to scores regressing to the mean. It is well documented that regression to the mean can affect the inferences that researchers and practitioners are able to draw from test scores across administrations (e.g., Bobko, 2001; Hausknecht et al., 2007; Hunter, Schmidt, & Le, 2006; Lievens et al., 2005; Roth, Van Iddekinge, Huffcutt, Eidson, & Bobko, 2002).

In summary, these four motivations for implementing a particular retesting policy rest on diverse assumptions and values reflecting organizational concerns. These motivations are important as retesting policy could result in awarding credentials, jobs, and promotions to unqualified individuals and therefore may have significant consequences for test-takers, organizations, and society at large (Millman, 1989). Although organizations and policy-makers are clearly not of any singular motivation, retesting policy decisions would be more scientifically defensible if made with a greater understanding as to whether the retesting effect is best explained by learning or memory (Chamorro-Premuzic & Furnham, 2010; Hausknecht et al., 2007).

#### **4. EXPLANATIONS FOR THE RETESTING EFFECT: LEARNING OR MEMORY EFFECTS?**

*“It is one thing, however, to remember, another to know.” Seneca* (p. 80, Campbell, 1969)

The finding that practicing one type of material improves performance on similar material extends back over a hundred years to psychology’s earliest pioneers (James, 1890; Thorndike & Woodworth, 1901). These findings were not interpreted merely as an effect of memory, but immediately recognized as a potentially useful tool for learning. As early as 1906, Edward Thorndike proposed that “the active recall of a fact from within is, as a rule, better than its impression from without” (p. 123, Thorndike, 1906). Since Abbott’s (1909) seminal study, several hundred experiments consistently demonstrated that testing enhances learning and retention (for recent reviews, see Rawson & Dunlosky, 2011; Roediger & Butler, 2011; Roediger, Putnam, & Smith, 2011). Nevertheless, mean score increases upon retest also occur across constructs by which learning and retention would *not* be expected (i.e., abilities), leading to concerns about precisely what process may underlie the retesting effect. More than 50 years ago, personnel researchers questioned the value of retesting applicants’ necessary skills and abilities in personnel selection (van der Reis, 1963). Similarly, researchers in educational settings have questioned the interpretation of retest scores since the early 1920s (e.g., Richardson & Robinson, 1921).

The two lines of inquiry investigating the retesting effect as reflecting either learning or memory have proceeded in parallel for some time. Concomitantly, personnel researchers have more frequently investigated abilities, whereas cognitive and

educational psychologists have more frequently investigated memory and knowledge, resulting in segregated streams of research. In the absence of integrated research across constructs and fields of study, knowledge retest increases are commonly assumed to reflect learning, whereas ability retest increases are assumed to reflect memory, and thus, a source of measurement error. However, there is no reason why these two theoretically competing explanations cannot be conceptually crossed. In an effort to bridge this gap in the literature, this dissertation focused on advancing the literature of the retesting effect by investigating both explanations simultaneously.

The strongest evidence for learning is the generalizability of score increases of the assessed construct across different settings (Jensen, 1998). In cognitive psychology, transfer is the term used to refer to score gains that occur after initial exposure to a stimulus that are subsequently reflected across novel operationalizations, settings, and tasks to which a test-taker has not been previously exposed (Carpenter, 2012; Jensen, 1998). The transfer of ability and knowledge to these novel settings reflects learning, as ultimately, the underlying ability and knowledge measured by these tests must be utilized in contexts different from those in which they were originally assessed or acquired (Carpenter, 2012). In terms of transfer, learning refers to an increase in test scores that reflect an increase in the *underlying construct* which transfers across operationalizations, settings, and tasks. Conversely, memory effects refer to an increase in the *specific test score* that does not transfer across operationalizations, settings, and tasks (Anastasi, 1981).

Transfer is a broad term and numerous taxonomies exist. For example, Barnett and Ceci (2002) propose a taxonomy of evidence of the retesting effect as reflecting learning through transfer across temporal contexts, test formats, and test construct domains. Transfer may also be conceptualized in terms of near or far transfer. Near transfer occurs when a rule or concept from a domain is applied to a new item from the same domain (Chen & Klahr, 1999) and far transfer occurs when a rule or concept is applied in a novel domain or context but requires a similar underlying solution (Gick & Holyoak, 1980).

Despite extensive research on the generality of the retesting effect, little research exists regarding the benefits of initial testing in transferring to novel contexts across time, test forms, construct domains, and external criteria (Carpenter, 2012); and there is also limited theory as to *why* retesting improves learning across both constructs and fields of study (Dunlosky et al., 2013). In the absence of integrated research across constructs and research disciplines, it is difficult to demonstrate whether retesting facilitates learning or memory because it is unclear whether these retesting effect explanations are confounded by the particular constructs investigated within these respective fields.

#### **4.1 Learning Effects**

In terms of the retesting effect, learning refers to an increase in test scores that reflect an underlying increase in the psychological construct of interest that is due to the intervention of testing itself (Roediger & Karpicke, 2006a), and is manifested through transfer across settings, time, and test operationalizations (Carpenter, 2012). Thus, the

learning explanation posits that score increases upon retest are equally, if not more valid assessments of the construct since they reflect a true construct increase due to the initial testing providing the opportunity to improve on the underlying construct domain.

The retesting effect on knowledge tests is a historically robust phenomenon in both the education and cognitive psychology literatures, consistently showing evidence that testing itself facilitates knowledge by modifying and organizing existing memories, and developing greater retrieval of memories (Abbott, 1909; Bjork & Bjork, 1992; Gates, 1917; Hunt, 2006; Karpicke & Roediger, 2008; Roediger & Karpicke, 2006a; Spitzer, 1939). Little work, however, has investigated the processes by which retesting may affect abilities and furthermore, whether these increases reflect learning or memory effects. Nevertheless, tests that require reasoning and the manipulation of knowledge appear to facilitate learning to a greater extent than memory recall tests (Chan, McDermott, & Roediger, 2006; Hamaker, 1986). Thus, it is not unreasonable to posit that the retesting effect may increase both knowledge *and* ability.

The principal cognitive explanation for the retesting effect is spreading activation theory (Bjork & Bjork, 1992; Collins & Loftus, 1975; Little, Bjork, Bjork, & Angello, 2012; Roediger & Karpicke, 2006a). Spreading activation theory models human cognition in terms of an interrelated network of concepts (Bjork & Bjork, 1992; Collins & Loftus, 1975). Concepts are commonly represented as nodes in a network, where the properties of all concepts are represented as relational links between the nodes of other concepts (Collins & Loftus, 1975). Nodes that are more closely related are more easily retrieved. Thus, when a stimulus is presented (i.e., primed), the concepts most closely

related to that stimulus are activated throughout the network, priming the recollection of related information within the network (Bjork & Bjork, 1992; Collins & Loftus, 1975). Furthermore, practice retrieving information from memory (e.g., through testing) improves the ability to recall both initially encoded information *and* related nodes in the future (Bjork & Bjork, 1992; Little et al., 2012). Spreading activation theory has primarily been used to explain the retesting effect in terms of learning; however, this explanation also covaries with the construct investigated (i.e., knowledge) and fields of study (i.e., cognitive psychology). Although spreading activation theory has been primarily used to explain the retesting effect under these boundary conditions, it nevertheless offers a viable alternate explanation to memory in ability domains as well.

Previous empirical research, which is discussed further below, presents diverse forms of evidence suggesting that the act of testing itself results in learning the underlying construct. Broadly, initial testing facilitates learning to a greater extent than traditional learning methods, transfers beyond test-specific information, transfers to increasingly complex conceptualizations of the construct domain, occurs (although at different magnitudes) across both knowledge and ability domains, and transfers to external criteria in both academic and applied settings.

**4.1.1 Learning Effects in Laboratory Settings.** Retesting consistently produces test score increases and enhances retention greater than traditional learning methods, such as additional studying (Chan et al., 2006; Karpicke & Blunt, 2011; Roediger & Karpicke, 2006a). In fact, the retesting effect produces retest score gains beyond that of even test-takers' own expectations, as it is consistently found that

initially, test-takers do not anticipate that retesting will improve performance greater than additional studying (Karpicke & Blunt, 2011; Roediger & Karpicke, 2006a).

The retesting effect is not confined to simply remembering the previous test experience and repeating responses. Test-takers do, in fact, change their responses following initial testing. Also, this response-changing is not merely confined to recalled guesses, confidence in previously incorrect responses, or even previously successful retrieval attempts (Kang, Pashler, Cepeda, Rohrer, Carpenter, & Mozer, 2011; Kornell, Hays, & Bjork, 2009).

The retesting effect is also not confined to the similarity or simplicity of particular response formats, but also occurs across response formats, increasingly abstract content, and complex skills. In fact, initial tests that are more difficult to remember or require higher-level comprehension of the construct, such as recall and constructed-response items, result in *greater* retesting effects than recognition tests (McDaniel, Roediger, & McDermott, 2007). Retest score increases also occur on higher-level comprehension items, including short-answer application and multiple-choice, inference-based items (e.g., Agarwal & Roediger, 2011; Butler, 2010; Johnson & Mayer, 2009). Retest score increases have also been shown on tests that require higher-level inferences and the application of previously learned information to new settings or problems (Agarwal & Roediger, 2011; Butler, 2010; Foos & Fisher, 1988; Johnson & Mayer, 2009; Karpicke & Blunt, 2011; McDaniel, Howard, & Einstein, 2009). These increases even occur when the retests included different items or test forms than initial testing (Karpicke & Blunt, 2011). Retesting increases even occur for abstract tasks, such

as inductive function learning tasks (Kang, McDaniel, & Pashler, 2011) as well as more complex skills, such as resuscitation procedures (Kromann, Jensen, & Ringsted, 2009).

Retest increases are also not bound to only test-specific information, but transfer even to related yet previously *untested* content (Chan et al., 2006; Karpicke & Blunt, 2011; Roediger & Karpicke, 2006a). That is, the retesting effect is stronger when test content is interrelated and retested as opposed to *unrelated* content (i.e., arbitrarily grouped test content) and retested. Furthermore, test-takers exhibit greater retest gains on related knowledge even when that particular knowledge was not directly tested on the initial test. Thus, retest performance increases do not appear to be the result of the mere exposure to test items (Chan et al., 2006). Instead, retesting appears to develop the underlying construct domain even when no direct memory effects are possible (i.e., retest items are not repeated from initial test). From the laboratory research presented above, it is clearly plausible that initial testing not only measures, but also develops the specified construct domain as reflected by increased retest scores.

**4.1.2 Learning Effects in Operational Settings.** The retesting effect extends beyond lab settings into operational settings in both academic and organizational employment settings. These learning effects also occur across both knowledge and ability domains in the prediction of external criteria. Of particular relevance to learning, the literature also demonstrates some preliminary evidence that the retesting effect not only *increases* retest performance, but may in fact *enhance* the criterion-related validity of retests. Although the retesting effect is often conceptualized as mean increases in test-scores that reflects error (i.e., memory effects), this increase in the criterion-related

validity of retests raises question as to whether this reflects the retest capturing some additional variance in the underlying construct (i.e., learning), or the reduction of some other confounding error that initially suppressed test scores. Nevertheless, this evidence is mixed and difficult to parse out from memory effects in the absence of experimental designs and criterion-related validity data.

Research has consistently demonstrated that the retesting effect on knowledge tests extends outside lab settings to long-term educational settings and classroom materials (Carpenter, Pashler, & Cepeda, 2009; Cranney, Ahn, McKinnon, Morris, & Watts, 2009; Glass & Sinha, 2013; Kromann et al., 2009; Metcalfe, Kornell, & Son, 2007; Rawson & Dunlosky, 2011; Rees, 1986; Vojdanaska, Cranney, & Newell, 2010). In fact, the retesting effect also exhibits significant gains even after summative course assessments (Balch, 1998; Daniel & Broida, 2004; Lyle & Crawford, 2011; McDaniel, Agarwal, Huelser, McDermott, & Roediger, 2011; McDaniel, Anderson, Derbish, & Morrisette, 2007; McDaniel, Wildman, & Anderson, 2012). While these studies primarily investigated the retesting effect using formal educational materials and younger samples, evidence indicates that the retesting effect occurs at a similar magnitude across age groups, including primary, secondary, and undergraduate education, and from mid- to late age (Meyer & Logan, 2013).

Retest score increases also occur in applied selection contexts on both knowledge and ability tests. Lievens et al. (2005) reported a retesting effect  $d$  of 0.27 for scores on a test of science knowledge used to assess medical school applicants. Similarly, candidates who chose to retest on a job knowledge test improved their scores by almost a

full standard deviation ( $d = 0.93$ , Van Iddekinge et al., 2011). Schleicher et al. (2010) reported a retest increase  $d$  of 0.15 on a job-knowledge test used to select applicants for professional jobs within a federal agency. Raymond et al. (2007) found even larger increases on two certification tests completed by medical imaging workers ( $d = 0.79$  and 0.48). Similarly, in the ability domain, Hausknecht et al.’s (2007) meta-analysis reported a retesting effect  $d$  of 0.26 across ability measures in selection settings.

Retest score increases occur consistently across diverse constructs and settings; however, it is less clear whether these retest score increases consistently reflect learning effects. So, mean score increases upon retest do not necessarily imply learning, as increases may still in fact reflect memory effects. There is mixed evidence in both academic and selection contexts, across both ability and knowledge tests, that second administrations of tests exhibit higher scores *in addition to* displaying equal if not greater criterion-related validity. Although some differences between validity coefficients were not significant, increasing validity coefficients are not a necessary condition to demonstrate evidence of learning. That is, mean retest score increases with *no* concomitant change in criterion-related validity would still provide evidence of learning. However, the opposite—a significant retest score increase coupled with a significant *decrease* in validity coefficients—would more clearly indicate the contaminating influence of memory on retest scores.

Research within academic settings has found that secondary test scores are sometimes *better* predictors of academic criteria than initial test scores, although differences between initial and retest validity are often small (e.g., Allalouf & Ben-

Shakhar, 1998; Coyle, 2006; Lievens et al., 2005, 2007; Reeve & Lam, 2005). Within academic settings, ability retest score increases are common while criterion-related validity increases or stays the same (Allalouf & Ben-Shakhar, 1998; Coyle, 2006). For example, ability retest scores correlated more highly with college matriculation exam scores than did initial test scores (Allalouf & Ben-Shakhar, 1998), and retest scores on the SAT correlated more highly with college grade point average than initial SAT scores correlated with college grade point average (GPA,  $r = .54$  versus  $r = .50$ , respectively; although these two validities were not significantly different; Coyle, 2006).

Higher validity coefficients for retest scores are not confined to academic settings. For example, in a sample of law enforcement applicants, Hausknecht et al. (2002) found evidence that the second administration of a GMA test showed greater, although not significantly greater, criterion-related validity coefficients with training performance than initial scores (e.g.,  $r = .31$  versus  $.27$ , respectively [limited only to applicants who chose to retest]). Similarly, Lievens et al. (2005) found that medical school applicants' retest scores on a science-knowledge test were significantly more predictive of subsequent grade point average than initial test scores ( $r = .21$  versus  $.11$ ; limited only to applicants who chose to retest). Similarly, Van Iddekinge et al. (2011) found that a second job knowledge-test exhibited considerably greater prediction of job performance ( $r = .38$  vs.  $.27$ ; limited only to applicants who chose to retest), and more generally, that second administration scores were comparable in criterion-related validity when evaluated against numerous alternate operationalizations (i.e., most recent test scores, highest test scores, or the mean of initial and retest scores).

Thus, research in both laboratory and operational settings demonstrate that retest increases occur across response formats, test forms, and that these retest scores predict relevant external criteria. Nevertheless, little research has directly pitted this learning explanation against memory in laboratory or operational settings.

#### **4.2 Memory Effects**

One cannot assume that retest score increases necessarily reflect learning, as numerous alternate explanations exist. Even when test-takers participate in construct-focused training, coaching, or practice, researchers agree that retest score increases are not necessarily indicative of increases in the underlying construct (Cole, 1982; Snow, 1982; Sternberg, Ketron, & Powell, 1982). Instead, researchers posit that score increases may reflect any number of heterogeneous influences interacting with the retesting process to increase observed test scores (Anastasi, 1981; Hausknecht et al., 2007; Shadish, Cook, & Campbell, 2001). Although varied influences exist (and will be discussed further later), the most likely alternate explanation for the retesting effect is memory.

Memory effects occur when the memory of the initial testing process influences *only* the observed test score upwards, rather than develops the underlying construct (Anastasi, 1981). In comparison to learning effects, memory effects increase the test-taker's observed test score with no concomitant change in the individual's true score or standing on the construct of interest. Thus, the memory effect explanation would posit that the observed score increases upon retesting do not reflect learning, but memory of the previous test administration. Support for the memory explanation of the retesting

effect includes reduction in the test's construct-related validity, lack of generalization to convergent measures, weaker criterion-related validity, and retest scores' correlations with memory-related constructs (e.g., working memory capacity) that are unrelated to a test's focal construct.

**4.2.1 Memory Effects in Laboratory Settings.** Some researchers present evidence that second administrations of ability tests do not assess the construct in the same way as initial testing (Matton et al., 2011; Reeve & Lam, 2005), and may instead reflect contamination by memory (Lievens et al., 2007). Memory effects may alternatively explain retest increases by a single general, homogenous factor reflecting memory of the initial testing experience (Anastasi, 1981; Lievens et al., 2007). Accordingly, Matton et al. (2009) found that a single factor could account for the retesting effect across a battery of diverse ability measures despite an absence of a relationship with the other factors. Matton et al. (2009) concluded that this single factor may reflect a reliable, yet independent, measurement artifact inherent to all retesting that consistently increases test scores across diverse constructs. Similar to this single factor explanation, Lievens et al. (2007) found that second administrations of a variety of ability tests were significantly correlated with a memory association test ( $r = .29$ ), whereas initial test scores were not ( $r = .03$ ).

If the retesting effect actually develops the underlying construct, as the proponents of using testing as a learning intervention posit, then it appears unlikely that these learning benefits for retested content would decline over time (see Roediger & Karpicke, 2006a, for a review). Conversely, if the retesting effect is due to memory,

increases are likely to dissipate over time as memories of previous test responses decay. Following this prediction, meta-analytic evidence shows the retesting effect for abilities diminish over greater time intervals between administrations (Hausknecht et al., 2007). These results follow the pattern of memory, not learning.

Initial testing may provide test-takers the opportunity to remember test-specific information; that is, test-takers may remember previous items and responses rather than actively responding to items, as would reflect the assessed construct (Kulik et al., 1984). Even within cognitively loaded reasoning tasks, sufficient practice results in overlearning which reduces the *g*-saturation of these tasks (Ackerman, 1987). Specifically, it appears that initial testing promotes the use of algorithms, heuristics, or rule-applying behavior that may be reproduced upon retest, rather than reflecting the underlying construct (te Nijenhuis, Voskuijl, & Schijve, 2001). That is, testing provides the opportunity for test-takers to develop and practice test-specific, short-cut strategies that reduce the cognitive efforts required and therefore, improve performance without a concomitant increase in the underlying construct (Neubauer & Freudenthaler, 1994; Roberts & Newton, 2003). One may also posit that the memory of these strategies would allow test-takers to quickly complete the elementary tasks of responding to items, gaining additional test-taking time to complete other items which were not previously successfully or correctly completed.

If the retesting effect ultimately reflects learning, then interventions designed to increase the underlying construct should be generalizable such that they are also reflected in score increases on alternate operationalizations of the same construct.

Although testing may act as an intervention to facilitate learning, and mean score increases consistently occur upon retesting, little evidence suggests that these retest increases transfer across tests assessing even the same construct (te Nijenhuis, van Vianen, & van der Flier, 2007). In fact, the magnitude of the retesting effect is greatest on tests with the lowest ability loadings at both the item- (te Nijenhuis et al., 2007) and the test-level (Jensen, 1998). For example, retests of ability exhibit less *g*-saturation and lower convergent validity with alternate ability operationalizations (Jensen, 1998; Lievens et al., 2007; Matton et al., 2011). Thus, this literature would suggest that the retesting effect leads to score improvement through the influence of unrelated constructs, such as memory of test-specific information and previous responses that contaminate and increase scores upon retest.

**4.2.2 Memory Effects in Operational Settings.** Although retest score increases are robust across constructs and settings, the evidence pertaining to criterion-related validity is less consistent. That is, some research in applied settings indicates that the criterion-related validity of retests is significantly *higher* than initial tests, including both ability (Embretson, 1987) and knowledge (e.g., Lievens et al., 2005; Van Iddekinge et al. 2011). Conversely, other research in applied settings indicates that the criterion-related validity of retests is significantly *lower* than initial tests (e.g., Hausknecht et al., 2002; Lievens et al., 2005; Lievens et al., 2007). The criterion-related validity of these tests rests on their capacity to measure the underlying construct (Ree, Earles, & Teachout 1994; Ree & Earles 1991), thus a decrease in the construct-related validity of the measure and an increase in contaminating constructs (i.e., memory) would reduce the

measure's criterion-related validity as well (te Nijenhuis et al., 2001). If retest scores increase with a concomitant decrease in criterion-related validity in operational settings, then it appears more likely that this reflects memory of initial testing rather than learning of the underlying construct.

## **5. THE PRESENT STUDY**

This dissertation undertakes a comparative investigation of the two explanatory processes for the retesting effect, that is, whether retest score gains reflect learning or memory. Consequently, the primary objective of this dissertation is a competitive test of learning or memory as the best explanation for the retesting effect. To disentangle these competing explanations, this study's design assesses and demonstrates the magnitude and pattern of relationships associated with the retesting effect as reflecting either learning or memory. Therefore, an initial goal is a replication of the retesting effect, which in turn, provides a baseline that permits the competitive test of these explanations. However, to investigate these two competing explanations, it is first necessary to control for the effects of memory of test-specific material while allowing test-takers the opportunity to learn the underlying construct. To accomplish this, the test forms were manipulated.

### **5.1 Investigating the Retesting Effect using Alternate Test Forms**

Explicating the role of learning versus memory is ultimately tested by the transfer and broad generalizability of score increases across operationalizations of the construct of interest (Carpenter, 2012). Accordingly, many researchers have called for investigations of the retesting effect using alternate test forms (Lievens et al., 2007). Memory of previous test information (e.g., previous responses) relies on test-specific information, whereas learning reflects true construct gains that extend beyond particular operationalizations. By extension, minimizing the test-specific information between test and retest by using alternate forms is often recommended as a method to reduce the test-

specific overlap between test administrations, and therefore reduce the confounding influence of memory on subsequent administrations (Shadish et al., 2001).

Although there are various practical reasons for developing alternate forms (e.g., changes in the underlying content domain, test security), educational institutions and organizations often develop and administer alternate test forms under the belief that this prevents the purposeful (i.e., a breach in test security) or inadvertent influence of memory of test-specific information from contaminating tests' construct- and criterion-related validity (Geving et al., 2005; Hausknecht et al., 2007; Wendt & Harris, 2004). Alternate test forms do significantly reduce the magnitude of retest score gains but do not entirely eliminate the retesting effect. For instance, meta-analytic research supports the fact that significant retest score gains still occur even with the use of alternate test forms. Specifically, Kulik et al. (1984) computed effect sizes from over 40 studies involving retest score gains on both ability and knowledge tests, finding that score increases are substantively greater on the same form ( $d = 0.42$ ) than alternate forms ( $d = 0.23$ ). In the absence of test-takers' ability to remember test-specific information, the remaining score increase on alternate form retests are presumed to reflect learning instead.

Retest increases occur across alternate forms and extend to increasingly complex test content. For example, testing facilitates memory classification and organization (e.g., bird types, families); even when new examples are introduced within the learned categories (Jacoby, Walheim, & Coane, 2010). Testing also facilitates transferring previously learned problems to novel, never before-seen material within similar domains

(e.g., testing versus restudying math rules/relationships, Kang et al., 2011). Initial testing was also found to be superior to additional studying in the ability to draw inferences from previously learned material, even when the inferences about the domain were quite different (e.g., the underlying structure of bird and bat wings compared to aircraft; Butler, 2010; see also Karpicke & Blunt, 2011).

Investigating the competing explanations of learning and memory requires the use of alternate forms, as these explanations are confounded when using only retest score increases from the same test forms. Therefore, the present study manipulates test forms to eliminate test-specific overlap and therefore controls for the influence of memory, which in turn, permits a test for learning as an explanation for the retesting effect. If learning underlies the retesting effect, then alternate forms should not greatly reduce score increases upon retest or significantly alter the pattern of results. However, if memory effects best explain almost all score gains, then alternate forms should substantially reduce if not eliminate the ability of test-takers to utilize test-specific information and therefore eliminate the score gains associated with the retesting effect. As such, alternate test forms were used to competitively test for memory and learning as best accounting for the retesting effect across both ability and knowledge tests. This was also accomplished in conjunction with a manipulation of corrective feedback.

In light of the extensive literature on the retesting effect previously reviewed, as prerequisite hypotheses, if memory best accounts for the retesting effect, then:

**Hypothesis 1a:** Retest scores will be higher than initial scores; however, the magnitude of this effect will be larger on the same form retest than the alternate form retest.

However, if learning best accounts for the retesting effect, then:

**Hypothesis 1b:** Retest scores will be higher than initial test scores and the magnitude of this effect will be similar for both the same and alternate form retests.

## 5.2 Investigating the Retesting Effect using Item Response Time

The cognitive process underlying learning and memory effects upon retest are likely to be reflected by differences in the speed of processing and response production between initial and retest performance. Item response time reflects the cognitive processing speed that test-takers take between the initial item's cue and their production of the item's response.

Memory effects likely reflect a simpler production process upon retest than learning. That is, test-takers do not require time to process, solve, and subsequently respond to the item, but are instead responding to the item's cue from memory. Initial testing provides test-takers practice to develop, remember, and use test-specific, short-cut strategies to quickly complete the elementary tasks of retest items faster, despite no concomitant increase in the underlying construct (Ackerman, 1987, 1988; Neubauer & Freudenthaler, 1994; Roberts & Newton, 2003). Test-takers may retain test-specific information from initial testing that allows a previous response to be identified (Webb, Pridemore, Stock, Kulhavy, & Henning, 1997). Accordingly, this is often reflected in

faster response times; test-takers spend *less* time processing items that they previously answered (e.g., Stock, Kulhavy, Pridemore, & Krug, 1992; Webb, Stock, Kulhavy, & White, 1990; Webb, Stock, & McCarthy, 1994). A number of researchers have found that speeded and psychomotor-loaded tests exhibit larger retest increases, potentially reflecting that remembering previous responses allows test-takers to quickly respond (Burke, 1997; Larson & Alderton, 1997). Thus, if retest score increases are the result of memory rather than learning, then test-takers recall previous responses and the strategies that they previously used to quickly produce a response.

However, following from the premise that response latency reflects the depth of cognitive processing that differentiates memory from learning effects, alternate test forms eliminate test-takers' ability to remember test-specific information. That is, test-takers completing an alternate form retest have no previous item-specific responses to remember, and subsequently, respond to items faster. Test-takers cannot quickly retrieve item- or test-information from memory as alternate test forms comprise novel test information and thus require original processing (Webb et al., 1997). As memory effects are controlled, and any retest score increases presumably reflect learning, then test-takers must still cognitively process an item before responding. Thus, test-takers with faster response times on a second test administration are likely remembering previous responses (reflecting memory effects), whereas test-takers with slower response times are likely taking additional time to cognitively process an item upon retest. Accordingly, Raymond et al. (2007) found shorter response times for the same-form knowledge test compared to alternate-form knowledge tests (although, not significant

differences in the magnitude of test scores). Similarly, Powers (1986) found that ability retest score increases were positively related to the response time allotted per item. If learning effects underlie the retesting effect, then test-takers must still process the novel, alternate form item before responding. Thus, alternate forms versus the same form will exhibit little difference between response times if learning underlies the retesting effect. On the other hand, if memory effects underlie the retesting effect, then test-takers *cannot* use test-specific information and will require greater time to process and respond to novel items. Thus, if memory best accounts for the retesting effect, then:

***Hypothesis 2a:*** Retest response times will be faster than initial test response times. However, the magnitude of the difference between the initial and retest response times will be larger on the same form retest than the alternate form retest.

However, if learning best accounts for the retesting effect, then:

***Hypothesis 2b:*** Retest response times will be faster than initial test response times. However, the magnitude of the difference between the initial and retest response time for the same and alternate form retest will be small (i.e., similar).

### **5.3 Investigating the Retesting Effect by Construct Domain**

Although the retesting effect occurs across ability and knowledge domains (Kulik et al., 1984), past explanations for the retesting effect have been embedded in research literatures that covary with these constructs. Thus, in the absence of competitive investigations using *both* ability and knowledge tests, it is unclear whether retest score increases are the result of learning or memory as these explanations are

confounded with the construct domain (i.e., explaining ability retest increases in terms of memory and knowledge tests in terms of learning). Furthermore, using both ability and knowledge tests permits the comparative examination of learning and memory as explanations for the retesting effect. Specifically, whereas it is true that knowledge tests are widely used in personnel selection and ability tests are widely used in cognitive psychology and education, selection researchers still invoke the error variance explanation for knowledge tests and education and cognitive psychology researchers are relatively silent on the interpretation of retest scores on ability tests. Thus, there remains the question of why a learning explanation would hold for knowledge but not ability tests, especially in light of research supporting the effectiveness of practice in skill and knowledge acquisition (Arthur, Day, Bennett, & Portney, 2013).

Researchers posit that test construct domain interacts with the retesting process in some way (Greene, 1941; Kingston & Turner, 1984; Raymond et al., 2007; Wing, 1980). Abilities, as fundamentally stable and enduring characteristics of individuals (Deary, Pattie, & Starr, 2013), are less likely to be affected by learning interventions (Skuy, Gewer, Osrin, Khunou, Fridjon, & Rushton, 2002). Conversely, knowledge is malleable and is therefore more likely to be affected by learning interventions. This is possibly due to the finding that the retesting effect is greater when test items can be answered by the systematic application of general problem solving skills, whereas the observed retesting effect is smaller for items that require application of previously acquired, domain-specific knowledge (Kingston & Turner, 1984; Wing, 1980). Item types that are “most subject to practice [learning] effects are those used to assess fluid as opposed to

“crystallized intelligence” (p. 153, Wing, 1980). Although both the learning and memory explanations may account for retest score increases across both construct domains, the magnitude of these influences are likely to differ by construct domain.

In accordance with this prediction, the criterion-related validity of knowledge tests appears to increase upon retesting (e.g., Van Iddekinge et al., 2011), whereas the criterion-related validity of ability tests often decreases (e.g., Lievens et al., 2005). Specifically, Lievens et al. (2005) found that while both knowledge and ability test scores increased, their criterion-related validity diverged, such that ability retest scores’ criterion-related validity decreased whereas the knowledge retest scores’ criterion-related validity stayed the same.

Drawing from this literature, researchers and practitioners commonly assume that score increases on ability retests over short retest intervals or without substantive training interventions are more likely to reflect memory, not learning. Thus, following the design of Lievens et al. (2005), ability and knowledge were contrasted in the present study. In comparison with knowledge, ability, specifically GMA, is less likely to meaningfully increase over short time periods simply due to the initial test acting as a learning intervention. Conversely, knowledge appears more likely to develop from testing. Thus, if testing functions as a learning intervention that develops the underlying construct, then knowledge tests are likely to be influenced to a greater extent than ability tests.

As previously detailed, alternate test forms offer the ability to parse out the influence of test-takers remembering test-specific information (memory effects) from

learning. Thus, manipulating alternate test forms in conjunction with varying the degree to which the construct domain is amenable to intervention (i.e., ability versus knowledge), offered a research paradigm that permitted a competitive test of whether learning or memory best accounts for the retesting effect.

**5.3.1 Ability.** The term ability is used to describe a relatively broad and enduring capability that an individual uses to learn and perform a task and that differs between individuals (Ackerman, 1987, 1988; Lubinski, 2000). For the purposes of the present study, GMA was used as an exemplar of ability testing in general. GMA is associated with the ability to process, understand, and learn information (Jensen, 1998; Schmidt & Hunter, 1998) and is commonly used to differentiate between test-takers in both educational and organizational contexts (Lubinski, 2000; Vandenberg & Lance, 2000). Extensive research supports the stability of GMA over time (Carroll, 1993; Gottfredson, 1986). Specifically, research shows that GMA is stable after early adolescence (Dixon, Kramer, & Baltes, 1985; Jensen, 1998) and into late adulthood (Deary et al., 2013). Furthermore, even deliberate, construct-focused, learning interventions aimed at increasing GMA do not generalize to alternate test forms assessing the same construct or external criteria (Skuy et al., 2002).

Nevertheless, the retesting effect on ability tests in general (and GMA specifically) are consistent, robust, and occur across organizational (e.g., Hausknecht et al., 2002), educational (e.g., Powers & Rock, 1999), clinical (e.g., Basso, Carona, Lowery, & Axelrod, 2002), and research settings (e.g., Woehlke & Wilder, 1963).

Furthermore, the retesting effect also occurs on alternate ability test forms, permitting the possibility that these retest score increases reflect learning, not only memory effects.

Meta-analytic evidence indicates that ability tests show evidence of the retesting effect across alternate forms, although at a reduced magnitude compared to same forms ( $d = 0.22$  compared to  $d = 0.40$ , respectively; Hausknecht et al., 2007). Hausknecht et al. (2007) also found that the magnitude of the retesting effect on *same form* ability tests diminished over time. However, it is interesting to note that there was no relationship between retest interval and the retesting effect increases when test-takers took *alternate* forms. That is, although retest increases dissipated over time on the same test form (as would be predicted by memory), the retest increases did not dissipate over time for the alternate test form (as would be predicted by learning). These findings suggest that retesting with the same form leads to test score increases due to memory effects, but does not refute the possibility of learning effects across same *and* alternate forms (albeit at a smaller magnitude). In contrast, some evidence indicates that test-takers with previous experience of an alternate form of an ability test do not perform any better than those who *initially* completed the alternate form (Matton et al., 2011). Thus, it is possible that some research utilizing alternate test forms reflect differences between alternate test forms that were not psychometrically equivalent and therefore confound comparison. So, in the proposed study, GMA was selected as a stable, ability construct that is *comparatively less* likely to be influenced by learning, as opposed to knowledge which is likely more amenable to learning interventions.

**5.3.2 Knowledge.** Knowledge reflects a specific body of information, be it factual information (declarative) or the steps, techniques, and organization for implementing information (procedural) in the successful performance of a task (Dye et al., 1993). Although knowledge may be influenced by the abilities necessary to learn, retain, and retrieve it (e.g., GMA, Schmidt, Hunter, & Outerbridge, 1986), knowledge is domain-specific and may reflect numerous and diverse bodies of information.

For knowledge tests, transfer occurs across varying degrees of overlap in test content. Although equivalent alternate forms for knowledge retests have not been commonly examined, the retesting effect on knowledge tests occurs across operationalizations assessing similar knowledge without test-specific overlap. Generally, it appears that initial testing influences score increases for later retests, yet changes in the subsequent retest form do not appear to entirely remove retest increases (Carpenter, 2009; 2011; Carpenter & DeLosh, 2006). For example, an increasing number of studies have shown that retesting increases occur on knowledge tests even when the specific test content was not initially tested, but related to content from previous tests (Chan, 2009, 2010; Chan et al., 2006; Cranney et al., 2009). Thus, testing appears to develop test-takers ability to draw inferences related to underlying test content, even if the specific test items were not initially assessed.

Although there is an extensive literature investigating the retesting effect on knowledge tests in education and cognitive psychology, some researchers (Carpenter, 2012; Karpicke & Blunt, 2011; McDaniel et al., 2007) point out that the majority of research on the retesting effect focuses on relatively simple memory of novel material

(e.g., word lists, word pairs, short passages) that may or may not generalize to more complex knowledge. Thus, this research has largely used relatively simple test content requiring the cued recall of target information from memory (Dunlosky et al., 2013), despite the clear value of more complex knowledge and its frequent use by personnel researchers (Dye et al., 1993; Raymond et al., 2007; Schmidt & Hunter, 1998).

Nevertheless, some research shows that more complex knowledge continues to exhibit retest increases in operational settings. For instance, Geving et al. (2005) tracked the performance of over 9,000 repeat examinees on a state real estate licensure examination comprising multiple alternate forms drawn from an item pool where alternate forms contained 9% of their items in common. While average retests scores increased by 0.62 standard deviations, there was no advantage to seeing the same items twice. That is, test-takers improved their scores upon retest, but when responding to the same items on retest they were just as likely to change their responses from correct to incorrect as from incorrect to correct. Similarly, Raymond et al. (2007) found no evidence of greater retest increases if test-takers received the same versus alternate forms for a radiography and computer tomography licensing exam. Thus, it appears that the retesting effect results in comparable score gains across same and alternate test forms even for complex knowledge within operational settings.

Consonant with the proposition that knowledge is more amenable to learning interventions than ability, comparing the magnitude of the retesting effect on alternate forms across construct domains permits an explication of whether memory or learning best accounts for the retesting effect. That is, if memory best explains the retesting

effect, then test-takers who remember previous test-specific information are unlikely to differ on whether they remember knowledge or ability test information. Furthermore, if memory best explains the retesting effect, then test-takers who complete an alternate form retest cannot exploit previous test-specific information, again, resulting in no difference in retest score gains between construct domains. Thus, if memory best accounts for the retesting effect, then:

***Hypothesis 3a:*** Retest score increases from initial test scores will be approximately equal for knowledge and ability; however, the magnitude of this retest increase will be larger on the same form retest compared to the alternate form retest (across construct domains).

Again, based on the premise that knowledge is more amenable to learning than ability, if learning best explains the retesting effect, then test-takers are likely to show greater retest score gains for knowledge compared to ability. Furthermore, if learning best explains the retesting effect, then test-takers who complete an alternate form retest should show retest score gains that are comparable in magnitude to test-takers who complete a same retest form. Consequently, if learning best accounts for the retesting effect, then:

***Hypothesis 3b:*** Retest score increases from initial test scores will be higher for knowledge than ability, irrespective of whether the same or alternate form retest is administered.

## **5.4 The Role of Additional Explanatory Variables for the Retesting Effect**

Whereas both primary research and meta-analyses have examined the magnitude and moderators of the retesting effect (Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; Hausknecht et al., 2007), little research has examined the cognitive process underlying retesting in terms of GMA or working memory capacity (Dunlosky et al., 2013).

Broadly, if learning underlies the retesting effect, then these score increases are likely to be influenced by GMA; whereas if memory underlies the retesting effect, then these score increases are likely to be influenced by working memory capacity.

To fully understand the influence of these explanatory variables on retest score increases, alternate forms are necessary. When alternate forms are used, test-takers cannot exploit previous test-specific information to improve their score without a concomitant increase in the underlying construct. Therefore, alternate test forms attenuate the influence of test-takers' working memory capacity to encode, recall, and consider previous test-specific content to improve their score. However, if the retesting effect is influenced by learning, then individuals are not bound by test-specific information and should improve their score across operationalizations of the same construct. That is, after controlling for the influence of memory by eliminating any test-specific overlap between initial and retest administrations, test-takers may still learn the underlying construct and exhibit retest score gains. Furthermore, if the retesting effect reflects memory, then test-taker's working memory capacity is an individual difference that likely influences the magnitude of his/her retest score increases, but working memory capacity's influence would be attenuated by alternate form retests. Conversely,

if the retesting effect reflects learning, then test-takers' GMA likely influences the magnitude of retest score increases, yet should generalize beyond same retest forms to alternate form retests.

**5.4.1 Working Memory Capacity and Memory Effects.** If memory explains the retest score increases, then working memory, the ability of test-takers to encode, manipulate, and recall information while simultaneously performing another cognitive task, is likely relevant. Working memory capacity has been defined as “a brain system that provides temporary storage and manipulation of the information necessary for . . . complex cognitive tasks” (Baddeley, 1992, p. 556). In the context of testing, test-takers use working memory capacity to solve cognitive problems by considering competing responses, drawing comparisons between alternatives, and then selecting their response by simultaneously matching the demands of the item, prior knowledge, and relevant information retrieved from memory (Hancock, Stock, & Kulhavy, 1992; Kulhavy & Stock, 1989; Webb, Stock, Kulhavy, Haygood, Zulu, & Robinson, 1990; Webb et al., 1997).

Working memory capacity is differentiated from short- and long-term memory as a general limitation on attentional capacity when maintaining multiple concepts in active memory while simultaneously performing distracting activities (Case, Kurland, & Goldberg, 1982; Daneman & Carpenter, 1980; Engle, 2002; Engle, Tuholski, Laughlin, & Conway, 1999). That is, working memory capacity reflects limits of an individual's ability to retrieve information from memory that has been lost from the focus of attention due to competing cognitive tasks (Baddeley, 1992; Case et al., 1982; Daneman

& Carpenter, 1980). Accordingly, individuals with greater working memory capacity are more able to inhibit distractions across cognitive tasks (Conway, Cowan, & Bunting, 2001; Kane, Bleckley, Conway, & Engle, 2001; Kane & Engle, 2003).

The relationship between working memory capacity's influence on ability and knowledge tests is due to an overlap in the underlying cognitive components of these tests. Typical tests of working memory capacity comprise interleaving a set of information to be remembered with a secondary processing task (e.g., math operations, reading), requiring individuals to actively maintain to-be-remembered information through focused attention (Daneman & Carpenter, 1980; Turner & Engle, 1989). Similarly, the complexity of items on cognitively-loaded tests requires test-takers to use working memory capacity to maintain and retrieve multiple elements from items concurrently while solving the problem (Shelton, Elliott, Matthews, Hill, & Gouvier, 2010). Thus, working memory capacity may impact retest scores with similar working memory capacity task requirements, and yet not transfer across operationalizations and constructs to explain learning.

**5.4.2 GMA and Learning Effects.** If learning effects occur because initial testing facilitates the development of the underlying construct, then the underlying cognitive process that applies to other learning interventions should similarly apply to retesting. Retest score increases are likely affected by the GMA of test-takers. Individuals higher in GMA benefit the most from learning interventions in general, as is reflected by the consistent relationship between GMA scores and training performance (Ackerman, 1987, 1988; Jensen, 1998; Schmidt & Hunter, 1998) and is the direct

precursor to learning in applied settings (Schmidt et al., 1986). The higher the cognitive complexity of a task, such as learning from a test rather than simply remembering previous responses, the higher the level of ability necessary to perform the task effectively (Gottfredson, 1997; Jensen, 1998). Accordingly, it is commonly found that the higher the cognitive complexity of a test, the more difficult it becomes for test-takers to improve their scores through practice (Jensen, 1998; te Nijenhuis et al., 2001). Therefore, the differences in trainability between tests can be attributed to the differences between different tests' *g* loading.

The learning explanation would posit that individuals with greater ability would be better equipped to process, integrate, and utilize information from initial testing to improve their scores upon retest. Therefore, if the retesting effect reflects learning, then GMA likely moderates this effect. As the learning effects explanation would predict, GMA appears to moderate the magnitude of retest gains on both ability and knowledge tests; however, the evidence is sparse and mixed. That is, individuals with greater ability also appear to gain more from initial testing and receive higher scores on retest than individuals with less ability. For example, Kulik et al. (1984) found that GMA moderated the magnitude of mean score increases on second administrations of ability tests. Using a sample of 3,605 sixth graders, Spitzer (1939) similarly found a larger retesting effect for higher-ability readers than for lower-ability readers on a text comprehension multiple-choice test.

GMA also moderates the magnitude of the retesting effect on knowledge tests. Marsh, Agarwal, and Roediger (2009) reported a proportional relationship between

initial successful performance and the magnitude of the benefits accrued from retesting. That is, if test-takers perform poorly on initial tests, poor initial memory of test material may limit the benefits of testing due to a lack of elaborative, cognitive processing that usually occurs from successful retrieval (Carpenter & DeLosh, 2006). Individuals who perform poorly on initial tests have gained less practice that might benefit learning, resulting in lower retest increases. In fact, after controlling for GMA, Meyer and Logan (2013) found that the differences in the magnitude of the retesting effect between initially low and high performing test-takers disappeared.

If the retesting effect is the result of memory and *not* learning, then GMA should display a comparatively weaker relationship with the magnitude of retest score increases. That is, retest increases that cannot be explained with GMA are unlikely to reflect learning. Consistent with this prediction, Coyle (2006) found that retest increases on scholastic aptitude tests were unrelated to GMA. Similarly, te Nijenhuis et al. (2001) found that retesting and coaching reduced the *g*-loadedness of ability test scores. Furthermore, te Nijenhuis et al. (2007) found that the *lowest* GMA test-takers actually exhibited the largest retest score gains after an intervention designed to develop the ability. As individuals greater in ability are consistently more likely to improve in training situations (Ackerman, 1987, 1988; Jensen, 1998; Schmidt & Hunter, 1998; Schmidt et al., 1986), these findings instead suggest that the retesting effect is *not* due to initial levels of GMA, but rather may include more narrow abilities such as memory (Jensen, 1998; Lievens et al., 2007; te Nijenhuis et al., 2007).

As such, test-takers who remember previous test-specific information (Kulik et al., 1984) through working memory capacity are able to meet the cognitive demands of the retest while simultaneously recalling, processing, and considering previous items, strategies, and responses from the initial test. However, working memory capacity cannot aid test-takers in improving their retest scores on alternate form retests as there is no test-specific overlap. Thus, if memory effects best account for the retesting effect, then working memory capacity is likely the ability that allows test-takers to improve their retest scores irrespective of the underlying construct assessed. However, test-takers with higher GMA are no more likely to remember previous test experiences than test-takers with lower GMA.

Given that GMA is the “ability to learn”, an additional and separate GMA test is necessary to investigate the extent to which GMA might explain learning due to retesting on both the knowledge and another ability test. Specifically, the present study uses Raven’s Advanced Progressive Matrices (APM, Carpenter, Just, & Shell, 1990; Raven, 1989; Raven, Court, & Raven, 1985; Raven, Raven, & Court, 1994) to explain retest score gains, as it was originally constructed as an operationalization of Spearman’s *g* (Spearman & Jones, 1950) and has been found to have the highest *g* loading across a variety of GMA tests (te Nijenhuis et al., 2001). Thus, it is likely that this test would be most resistant to the training of, and transfer from the other ability tests in the present study (te Nijenhuis et al., 2001; te Nijenhuis et al., 2007), and therefore may be used to explain the retest score increases as reflecting learning on the retest measures. Thus, if memory best accounts for the retesting effect, then it is expected that:

**Hypothesis 4a:** The relationship between working memory capacity and retest score increases will be stronger than the relationship between GMA and retest scores increases and be of greater magnitude for the same form retest than the alternate form retests.

On the other hand, if learning best accounts for the retest score increases, then test-takers with higher GMA are more likely to experience greater learning from initial testing and therefore greater development of the underlying construct. Test-takers with greater working memory capacity, however, are unable to use their memory of test-specific information to improve their score on alternate form retests. Thus, under a learning explanation for the retesting effect, it is expected that:

**Hypothesis 4b:** The relationship between GMA and retest score increases will be stronger than the relationship between working memory capacity and retest scores increases and be of a similar magnitude between the same form retest and the alternate form retest.

## **5.5 Examining the Retesting Effect using Corrective Feedback**

Historically, providing corrective feedback has long been investigated and used to improve learning (Pressey, 1926; Thorndike, 1913; Trowbridge & Carson, 1932). Common theory on the positive effect of corrective feedback on performance is largely derived from the behaviorist law of effect (Thorndike, 1913), but the influence of corrective feedback is more complex (Kluger & DeNisi, 1996). Given the inconsistent benefit of providing corrective feedback for test performance, this review of the feedback literature focuses on the value of corrective feedback as a moderator of

learning interventions in general (Kluger & DeNisi, 1996) and the retesting effect specifically (Butler, Karpicke, & Roediger, 2007, 2008; Butler & Roediger, 2007).

Individuals use feedback (whether actively provided by an intervention or not) to evaluate their performance relative to their goals and focus attention on specific targets to achieve those goals (Kluger & Denisi, 1996). Specific to retesting, corrective feedback draws attention to the test-taker to learn task rules and recognize errors (e.g., Frese & Zapf, 1994). Kulhavy (1977) proposed that the primary benefit of corrective feedback for testing lies in decreasing response competition between incorrect and correct responses, thus guiding the test-taker to correct errors. However, corrective feedback after even initially correct responses also appears important to learning, as feedback allows individuals to confirm the accuracy of correct guesses that may not be maintained otherwise (Butler et al., 2007; 2008).

Across diverse tasks, Kluger and Denisi's (1996) meta-analysis demonstrated that corrective feedback generated greater performance increases when interventions provide cues that directly support learning and goal setting compared to feedback that provides cues directing attention towards the self (i.e., were not directly related to task performance and learning). The benefit of retesting with corrective feedback is that it consistently outperforms retesting without corrective feedback (Dunlosky et al., 2013). Not surprisingly, more extensive feedback interventions show greater score increases on ability tests compared to retesting alone ( $d = 0.51$  and  $0.25$ , respectively, Kulik et al., 1984). Despite the evidence that the retesting effect demonstrates score increases greater than common learning techniques, including additional studying, concept mapping, note-

taking, and imagery use (Fritz, Morris, Acton, Voelkel, & Etkind, 2007; Karpicke & Blunt, 2011; McDaniel et al., 2009; Neuschatz, Preston, Toglia, & Neuschatz, 2005), the magnitude of these retest increases is moderated by the extent to which initial testing was accompanied with corrective feedback. Specifically, Kulik et al. (1984) differentiate between the magnitude of retesting increases based on the degree to which an external intervention actively encourages learning rather than merely providing performance assessment (i.e., corrective feedback as to why a response is correct or incorrect rather than simply identifying responses as correct or incorrect). Furthermore, the learning benefits of corrective feedback are also moderated by initial test performance, in that corrective feedback provides greater score increases to test-takers who performed better on an initial test (Kang, McDermott, & Roediger, 2007).

Although corrective feedback may provide test-takers an additional means for facilitating learning from initial testing, test-takers may also use corrective feedback as a crutch with no concomitant learning (Anderson, 1987), and subsequently, reduce the likelihood that learning will transfer across operationalizations of the construct. In accordance with this proposition, corrective feedback was found to be detrimental to the performance of tasks that were somewhat different than the task on which corrective feedback was initially provided (Carroll & Kay, 1988). Furthermore, corrective feedback may improve test scores but this benefit may not generalize to alternate forms or external criteria. Some interventions, such as practice or coaching, may significantly increase ability scores (Hausknecht et al 2007); yet these gains are not always generalizable across tests (e.g., te Nijenhui et al., 2007). Indeed, even learning

interventions specifically aimed at increasing GMA do not generalize to alternate test forms assessing the same construct or the external criteria that GMA commonly predicts (Skuy et al., 2002). Thus, it appears possible that corrective feedback exacerbates memory effects on same retest forms and interferes with more elaborate learning that might transfer across alternate form retests (Kluger & Denisi, 1996). To investigate the competing learning and memory effects explanation for the retesting effect with corrective feedback, the use of alternate test forms are necessary.

Little research has directly examined the effect of corrective feedback on alternate test forms; however, there is substantial evidence that corrective feedback moderates the retesting effect across alternate response formats (e.g., initial multiple-choice testing followed by short answer response formats, Kang et al. 2007). If testing does in fact function as a learning intervention, providing test-takers with corrective feedback should intensify the retesting effect by further facilitating the learning value of initial testing. That is, corrective feedback from initial test performance has the potential to promote more elaborate understanding, conceptualization, and consolidation of the underlying construct domain. Accordingly, previous research demonstrates that corrective feedback provides cues that support and encourage goal setting, error correction, and learning from initial testing (Butler et al., 2007; 2008; Frese & Zapf, 1994; Kluger & Denisi, 1996; Kulik et al., 1984). If corrective feedback intensifies the retesting effect to develop a test-taker's underlying construct through learning, then any subsequent operationalization assessing that construct should reveal this increase.

As reviewed previously, corrective feedback provides test-takers with performance information that may be used to facilitate learning from the initial test, however, it will also provide test-takers the opportunity remember test-specific information. Thus, corrective feedback offers a means of investigating the cognitive process—memory or learning—that underlie the retesting effect, but is limited by the extent to which the same test form is used. Therefore, the use of alternate forms eliminates the test-takers' ability to use the memory of previous test-specific information that corrective feedback may provide. Thus, if memory best accounts for the retesting effect, then:

***Hypothesis 5a:*** Retest score increases will be higher for the corrective feedback condition compared to the no corrective feedback condition; however, the magnitude of this effect will be larger on the same form retest than the alternate form retest.

On the other hand, if learning best explains the retesting effect, then test-takers receiving corrective feedback followed by the alternate form retest should show comparable retesting increases to test-takers who received the same retest form. Thus, if learning best accounts for the retesting effect, then:

***Hypothesis 5b:*** Retest score increases will be higher for the corrective feedback condition compared to the no corrective feedback condition and the magnitude of this effect will be similar for both the same form retest and alternate form retest.

The preceding conceptual arguments and expected pattern of results engender analogous hypotheses in reference to response time such that if memory best accounts for the retesting effect, then it is expected that:

*Hypothesis 6a:* Retest response times will be faster than initial test response times for the corrective feedback condition compared to the no corrective feedback condition. However, the magnitude of the difference between the initial and retest response times will be larger on the same form retest than the alternate form retest.

On the other hand, if learning best accounts for the retesting effect, then:

*Hypothesis 6b:* Retest response times will be faster than initial test response times for the corrective feedback condition compared to the no corrective feedback condition and the magnitude of this effect will be similar for both the same form retest and alternate form retest.

**5.5.1 Corrective Feedback by Construct Domain.** Corrective feedback facilitates score improvements across both ability and knowledge assessments. Corrective feedback exhibits a significant, but small effect on knowledge testing ( $d = 0.26$ , Bangert-Drowns et al., 1991), and when corrective feedback is provided, these retest gains are greater than common learning interventions (Dunlosky et al., 2013; Kang et al., 2007). Learning increases following corrective feedback are not confined to knowledge tests, as meta-analytic evidence indicates that learning interventions (such as coaching) on ability tests can positively impact ability test scores upon retest at a magnitude greater than simply retesting alone ( $d = 0.70$  versus  $0.24$ , respectively;

Hausknecht et al., 2007). Specifically, providing corrective feedback after test-takers responded to each item improved tests of verbal (Betz & Weiss, 1976a, 1976b) and vocabulary ability at a greater rate than non-feedback conditions (Prestwood, 1979). Although no studies have directly compared the magnitude of the retesting effect with corrective feedback for both knowledge and ability, because knowledge is more amenable to intervention than ability, one would therefore expect that corrective feedback after initial testing to be a more robust intervention for knowledge than ability. Conversely, if the retesting effect is primarily influenced by the memory of previous test-specific information, then one would not expect differential memory effects for knowledge versus ability tests. Thus, if memory best accounts for the retesting effect, then:

***Hypothesis 7a:*** Retest score increases will be approximately equal for ability and knowledge, but exhibit an interactive effect between receiving corrective feedback conditions and retest form, such that receiving corrective feedback will increase retest scores at a greater magnitude for the same form retest than the alternate form retest.

However, if learning best accounts for the retesting effect, then:

***Hypothesis 7b:*** Retest score increases will be higher for knowledge than ability and exhibit an interactive effect with corrective feedback condition, such that receiving corrective feedback will increase retest scores at a greater magnitude for knowledge than ability, but the magnitude of this effect will be similar for the same form retest and alternate form retest.

### **5.5.2 Corrective Feedback for Memory Effects: The Role of Working**

**Memory Capacity.** Corrective feedback prior to retesting may facilitate learning, but may also provide test-takers with test-specific information and strategies that may be remembered and repeated to improve test scores despite no concomitant increase in the underlying construct (Powers, 1986). As would be predicted by memory, the retesting effect occurs even in the absence of corrective feedback (Roediger & Karpicke, 2006a). In fact, corrective feedback shows greater benefits for simple-task performance (e.g., memory) compared to more complex-task performance (e.g., learning; Kluger & Denisi, 1996). Thus, corrective feedback may actually exacerbate memory effects rather than facilitate learning.

Corrective feedback is not universally beneficial. Corrective feedback may promote learning if test-takers are receptive; however, corrective feedback may actually *inhibit* learning if test-takers do not actively consider or use the feedback to develop a greater understanding of the construct domain (Saloman & Globerson, 1987). Instead of promoting learning, corrective feedback may act as a crutch that reduces the need to learn to perform the task better (Anderson, 1987). For example, in situations where test-takers may simply copy correct responses without making any attempt at cognitive elaboration or integration with their previously developed knowledge, corrective feedback actually hinders learning (Anderson, Kulhavy, & Andre, 1971, 1972). Irrespective of the potential learning influence of testing, corrective feedback also provides test-takers with test-specific information that promotes memory effects, which

are likely to be retained and exploited using working memory capacity, rather than learned using GMA.

**5.5.3 Corrective Feedback for Learning Effects: The Role of GMA.** If learning underlies the retesting effect, then test-takers provided with specific and accurate performance information after an initial test will likely experience greater learning on the underlying construct and will therefore score higher upon retest than test-takers not provided with corrective feedback. Despite the differences in magnitude in the retesting effect with and without corrective feedback, it is not clear whether this reflects learning, rather than merely providing additional test-specific information and leading to memory effects. Test-takers' behaviors lend some insight into whether learning actually occurs.

When individuals receive corrective feedback, the rate of error correction appears proportional to one's confidence in the initial error (Nelson & Dunlosky, 1991; Stock et al., 1992). Previous research has shown that the retesting effect without corrective feedback is lower, as incorrectly selecting non-keyed responses on an initial test may cause test-takers to retain non-keyed responses as correct on a retest rather than recall that they did not know the answer and try again (Butler & Roediger, 2008). Yet this incorrect guessing does *not* impair the learning of the correct response, as long as corrective feedback is provided after these errors (Butler & Roediger, 2008; Kang et al., 2011). In fact, corrective feedback sometimes reduces the test-takers' retrieval of previous responses that were incorrect, even when test-takers are specifically instructed to remember their original (incorrect) responses (Webb et al., 1997).

The efficacy of corrective feedback in facilitating the learning of the underlying test content is largely determined by whether the corrective feedback includes the correct response (for meta-analyses, see Bangert-Drowns et al., 1991; Kluger & DeNisi, 1996). For example, corrective feedback that simply indicates that a response is right or wrong is less effective than the presentation of the correct response itself (e.g., Pashler, Cepeda, Wixted, & Rohrer, 2005). Accordingly, corrective feedback is primarily beneficial through error correction and prompting individuals to meta-cognitively re-evaluate their responses, not merely providing additional opportunities to remember previous responses (Bangert-Drowns et al., 1991). Test-takers are likely to exert greater effort to understand the tested content when corrective feedback is provided, as responses with greater discrepancy will promote greater error correction on subsequent tests (Kulhavy & Stock, 1989; Stock et al., 1992; Webb et al., 1994). Corrective feedback is particularly effective in protecting against these high-confidence and perseveration errors (i.e., when test-takers consistently respond incorrectly to the same items) to a greater effect than retesting alone (Butler & Roediger, 2008; Butterfield & Metcalfe, 2001). Test-takers with greater GMA are more able to reconcile this discrepancy and correct their responses on retest.

Nevertheless, learning can still take place without corrective feedback if test-takers self-assess (Meichenbaum, 1985; Wong, 1985). Without corrective feedback on test performance, test-takers who experience difficulty in responding to certain test items may subsequently identify content that requires further study or thought (Roediger & Karpicke, 2006a). Thus, corrective feedback is not *necessary* for the retesting effect to

reflect learning. In any case, the potential value of corrective feedback to facilitate testing as a learning intervention invariably provides test-takers the cues to recall previous test-specific responses and is best investigated in conjunction with alternating retest forms. The magnitude of retest score increases by corrective feedback condition are likely to be differentially influenced by GMA and working memory capacity depending on whether the retesting effect is best explained by learning or memory effects. Thus, if memory best accounts for the retesting effect, then:

*Hypothesis 8a:* A stronger relationship will exist between working memory capacity and retest score increases compared to GMA and retest score increases for the corrective feedback condition compared to the no corrective feedback condition, and the magnitude of this relationship will be larger on the same form retest than the alternate form retest.

However, if learning best accounts for the retesting effect, then:

*Hypothesis 8b:* A stronger relationship will exist between GMA and retest score increases compared to working memory capacity and retest score increases for the corrective feedback condition compared to the no corrective feedback condition, and the magnitude of this relationship will be similar for both the same form retest and alternate form retest.

## **5.6 Controlling for the Influence of Test Attitudes on the Retesting Effect**

Although most external influences on retest performance are unsystematic and comprise the random error present with all testing (e.g., test-taker mood, illness, distraction, nuances in test administration), fluctuations in error variance that

systematically *covary* with testing administrations present plausible alternate explanations for the retesting effect (Anastasi, 1981; Messick, 1989). Accordingly, researchers have often explained retest increases by citing a decrease in debilitating test attitudes (e.g., test anxiety; Cassady & Johnson, 2002; Hausknecht, Day, & Thomas, 2004; McCarthy & Goffin, 2005) or, relatedly, increases in facilitating test attitudes (e.g., test motivation, test familiarity; Chan, Schmitt, DeShon, Clause, & Delbridge, 1997; Lievens et al., 2005; Reeve & Lam, 2007). Thus, test attitudes are a source of error variance that suppress initial test scores, but exert less influence at retest, and therefore, increase retest scores and subsequently enhance validity as the error variance is removed.

A substantial body of literature exists investigating the influence of test-taker attitudes on initial test scores, thus it is not unreasonable to posit that test attitudes systematically vary upon retest and similarly affect retest performance (Reeve & Lam, 2007). Van Iddekinge et al. (2011) present evidence that test attitudes covarying with retesting would not only increase test scores but affect validity coefficients as well. Various test attitudes may affect test (and retest) performance, the present discussion is limited to the two test attitudes most likely to covary with retesting, reduced test anxiety (Cassady & Johnson, 2002; Messick & Jungeblut, 1981) and increased test motivation (Hausknecht et al., 2002). Although, it is certainly plausible that test attitudes systematically covary with retesting, few, if any, studies have directly examined the fluctuation of test-takers' attitudes at both initial and retest performance (Schleicher et al., 2006) nor how test-taker attitudes directly interact with the retesting effect to

influence test score' validity (Lievens et al., 2007). Thus, to explicate whether the retesting effect reflects either learning or memory requires controlling for the potential confounding influence of test attitudes (Reeve, Heggestad, & Lievens, 2008).

### **5.7 Summary**

The prevalence and influence of the retesting effect on the scientific community and society at large underscore the importance of understanding whether these retest increases reflect either memory effects or learning. It is clear that the retesting effect consistently influences the tests that organizations, educators, and policy-makers use to make decisions across work, educational, and social settings. Despite the divergent implications that a learning versus memory explanation of the retesting effect has for retesting policies and practices in both academic and applied domains, a lack of integrated research between cognitive, education, and personnel researchers limits advancements in this domain and the subsequent inability to make clear evidence-based recommendations. Consequently, examining whether learning or memory best accounts for the retesting effect is the major contribution of the present study to literatures in psychology, education, and management and its findings potentially have implications for test developers, test users, test-takers, policy makers, and society at large. A summary of the competing hypotheses for the posited memory and learning explanations are presented in Table 1.

Table 1

*List of Competing Hypotheses*

Hypotheses	If memory best explains the retesting effect:	If learning best explains the retesting effect:
1	(a) Retest scores will be higher than initial scores; however, the magnitude of this effect will be larger on the same form retest than the alternate form retest.	(b) Retest scores will be higher than initial test scores and the magnitude of this effect will be similar for both the same and alternate form retests.
2	(a) Retest response times will be faster than initial test response times. However, the magnitude of the difference between the initial and retest response times will be larger on the same form retest than the alternate form retest.	(b) Retest response times will be faster than initial test response times. However, the magnitude of the difference between the initial and retest response time for the same and alternate form retest will be small (i.e., similar).
3	(a) Retest score increases from initial test scores will be approximately equal for knowledge and ability; however, the magnitude of this retest increase will be larger on the same form retest compared to the alternate form retest (across construct domains).	(b) Retest score increases from initial test scores will be higher for knowledge than ability, irrespective of whether the same or alternate form retest is administered.
4	(a) The relationship between working memory capacity and retest score increases will be stronger than the relationship between GMA and retest scores increases and be of greater magnitude for the same form retest than the alternate form retests.	(b) The relationship between GMA and retest score increases will be stronger than the relationship between working memory capacity and retest scores increases and be of a similar magnitude between the same form retest and the alternate form retest.
5	(a) Retest score increases will be higher for the corrective feedback condition compared to the no corrective feedback condition; however, the magnitude of this effect will be larger on the same form retest than the alternate form retest.	(b) Retest score increases will be higher for the corrective feedback condition compared to the no corrective feedback condition and the magnitude of this effect will be similar for both the same form retest and alternate form retest.

Table 1

*List of Competing Hypotheses (Continued)*

Hypotheses	If memory best explains the retesting effect:	If learning best explains the retesting effect:
6	(a) Retest response times will be faster than initial test response times for the corrective feedback condition compared to the no corrective feedback condition. However, the magnitude of the difference between the initial and retest response times will be larger on the same form retest than the alternate form retest.	(b) Retest response times will be faster than initial test response times for the corrective feedback condition compared to the no corrective feedback condition and the magnitude of this effect will be similar for both the same form retest and alternate form retest.
7	(a) Retest score increases will be approximately equal for ability and knowledge, but exhibit an interactive effect between receiving corrective feedback conditions and retest form, such that receiving corrective feedback will increase retest scores at a greater magnitude for the same form retest than the alternate form retest.	(b) Retest score increases will be higher for knowledge than ability and exhibit an interactive effect with corrective feedback condition, such that receiving corrective feedback will increase retest scores at a greater magnitude for knowledge than ability, but the magnitude of this effect will be similar for the same form retest and alternate form retest.
8	(a) A stronger relationship will exist between working memory capacity and retest score increases compared to GMA and retest score increases for the corrective feedback condition compared to the no corrective feedback condition, and the magnitude of this relationship will be larger on the same form retest than the alternate form retest.	(b) A stronger relationship will exist between GMA and retest score increases compared to working memory capacity and retest score increases for the corrective feedback condition compared to the no corrective feedback condition, and the magnitude of this relationship will be similar for both the same form retest and alternate form retest.

## **6. METHOD**

### **6.1 Sample**

Participants comprised 374 individuals recruited from Texas A&M University who had completed an introductory psychology class. The sample was restricted to individuals who had completed an undergraduate introduction to psychology course because the knowledge test that was used in the study (i.e., the General Psychology Competency Exam; Arthur, Tubré, Paul, & Edens, 2003) is a walking-knowledge, basic psychology competency exam.

Participants were recruited using a number of different methods. Primarily, participants were recruited using academic advising listservs of departments where introductory to psychology courses were common or required (e.g., Psychology, Health and Kinesiology, Industrial Distribution). Additionally, participants were recruited through flyers, recruiting visits to upper-level psychology courses, and social media websites (e.g., Facebook). Participants recruited from the preceding sources were paid \$20 for their participation.

Finally, participants were also recruited from the present university's Psychology Department subject pool. Consequently, these particular participants completed this study in return for partial course credit, rather than compensation. These participants were only eligible to participate if they had already completed their introductory psychology course, but had not yet completed their required subject pool credit hours.

In addition to the \$20 participation compensation or subject pool credit, performance incentives were used to address the population and ecological validity

concerns associated with using a low-stakes student sample. Consequently, participants performing in the top 10% of the sample were awarded a \$50 performance reward based on the mean of *all* their knowledge and ability test scores (i.e., Time 1 and Time 2). Participants were informed of the compensation and performance rewards at recruitment, and the beginning of Time 1 and Time 2 sessions.

## **6.2 Research Design**

The study was a 2 (retest form; same form versus alternate form)  $\times$  2 (corrective feedback versus no feedback) repeated measures, mixed factorial, experimental design in which participants were tested on two separate occasions separated by a 7-10 day retest interval. That is, test format and corrective feedback were both manipulated as between-subjects conditions. Consequently, participants were randomly assigned to one of the four conditions, and commenced the study protocol by first completing the ability and knowledge tests. They then received item-level corrective feedback or not, and then were retested 7-10 days later on either the same or an alternate form of the ability and knowledge tests. Test type (ability versus knowledge) was a within-subjects variable in that all participants were administered both ability and knowledge tests at Time 1 and Time 2. The research design is illustrated in Table 2.

Table 2

*Research Design and Measures*

Time 1		Time 2
<b>Measures</b>	<b>Between-Subjects Feedback Manipulation</b>	<b>Between-Subjects Test Form Manipulation</b>
Working Memory Capacity		
General Mental Ability		
Test Anxiety		<b>7-10 Day Retest Interval</b>
Test-Taking Motivation		
Knowledge Test Form A or B	No Corrective Feedback	
Ability Test Form A or B		
	Corrective Feedback	

### 6.3 Power Analysis to Detect Hypothesized Effects

To estimate the number of participants necessary to detect the anticipated effects posited in the hypotheses, an a priori power analysis was conducted using G\*Power 3.1 (Erdfelder, Faul, & Buchner, 1996; Faul, Erdfelder, Buchner, & Lang, 2009). This power analysis was based on the most conservative hypothesis in the study. Consequently, if the study has sufficient power to detect this specified effect, then it would by inference have sufficient power to detect the less conservative effects as well.

The hypothesis that served as the basis for the power analysis pertains to a one-tailed test of the difference between two independent correlations (i.e., a  $q$  effect size). Specifically, this hypothesis consisted of the comparison of the effect size based on the extant meta-analytic literature of corrective feedback on knowledge retest score increases ( $d = .30$ , Bangert-Drowns et al., 1991) to the effect size of corrective feedback

on ability retest score increases ( $d = .11$ , Hausknecht et al., 2007; difference between these effect sizes,  $q = .20$ ). These parameter estimates served as the basis the effect sizes used for the power analysis, and alpha was set to .05 and power for .80. Table 3 presents these a priori parameters which estimated that the optimal sample size to detect the difference between these effects was 638 participants.

On completing the study, a post-hoc power analysis was also computed using the present study's effect sizes and  $N$ , which indicated a power level of .12. Implications of the differences between the a priori and post hoc power analyses are discussed in the Discussion section.

Table 3

*Power Analysis for Detecting a One-Tailed Difference between Two Independent Correlations*

<b>Parameters</b>	<b>Power Analysis Parameters</b>	<b>Theoretical Basis</b>	<b>Present Study's Parameters</b>
Corrective feedback effect size on retest ability performance	.11 (0.22)	Hausknecht et al.'s (2007) meta-analysis found a corrected retest increase of .22 standard deviations retesting score increase after test coaching following the initial general mental ability test.	.15 (.31)
Corrective feedback effect size on retest knowledge performance	.30 (0.63)	Bangert-Drowns et al.'s (1991) meta-analysis found a corrected retest increase of .63 standard deviations retesting score increase after test feedback following the initial knowledge test.	.20 (.41)
<i>q</i> effect size difference estimate	.20		.05
$\alpha$ Level	.05		.05
Power level	.80		.12 (actual power level).
Between-subjects sample (Alternate Test Form A)	319 participants required		171 participants
Between-subjects sample (Alternate Test Form B)	319 participants required		169 participants
Total Sample	638 participants required		340 participants

*Note.*  $d$  values are reported in parentheses and these were converted into  $r$  effect size estimates and  $q$  effect size differences for the power analysis.

## **6.4 Measures**

With exception of the Raven's Advanced Progressive Matrices, all measures were computer-administered via Inquisit.

**6.4.1 Knowledge Test.** The General Psychology Competency Examination (GPCE) 2.1 is a 40-item, three-alternative, content-valid, multiple-choice, walking-knowledge psychology competency exam that assesses mastery of basic concepts, principles, and facts covered in the standard introductory psychology course. This current measure is an updated, revised, and revalidated (content-related) version of the exam that was initially developed by and used in Arthur et al. (2003), Fehrman, Woehr, and Arthur (1991), and Woehr, Arthur, and Fehrman (1991). Sample items are reported in Appendix A. The procedures that were undertaken in the revision and revalidation effort are presented in Appendix B.

Two alternate forms of the GPCE 2.1, each consisting of 40 items were used and participants' scores were the summed number of items answered correctly. Woehr et al. (1991) reported an internal consistency reliability estimate of .51 for the previous GPCE (1.0) scores. Test-retest reliabilities for the knowledge test forms are reported in Table 4.

**6.4.2 General Mental Ability (GMA).** Two GMA tests were used. One GMA test was used as the comparative ability test to the knowledge test. This GMA test is a speeded 4-alternative, multiple-choice, 60-item (30 verbal and 30 numeric) test similar to that was developed by Arthur (2004; 2005). Participants were allotted 10 minutes to complete the test. Participants' test scores were calculated as the total number of items

answered correctly. Two alternate forms of this test were used, each consisting of 60 items. Arthur, Glaze, Villado, and Taylor (2010) reported a retest reliability coefficient of .78 (mean retest interval = 429.16 days,  $SD = 54.84$ ) for an equivalent form of this test, along with a convergent validity of .72 with the Thurstone Test of Mental Alertness. Test-retest reliabilities across ability test forms are reported in Table 4 below.

GMA was also measured using the short form of the Raven's Advanced Progressive Matrices (APM; Arthur & Day, 1994; Arthur, Tubré, Paul, & Sanchez-Ku, 1999) which consists of 2 practice and 12 test items. Participants were allotted 15 minutes to complete the test. Participants' scores were the total number of items answered correctly. Arthur et al. (1999) reported a 1-week test-retest reliability of .76. This assessment was used to investigate the relationships between GMA and retest gains on both the knowledge and ability tests, primarily as evidence of learning.

**6.4.3 Alternate Forms.** Alternate forms exist along a continuum. In the present study, alternate forms for both the knowledge and ability tests are conceptualized as reflecting the same item format and the underlying content at the item-level and are operationalized as such. For example, the first item for both Knowledge Test A and B requires knowledge of social psychology, whereas the first item for both Ability Test A and B requires division. For the knowledge test forms, the content and item order were based on the initial GPCE 1.0 and assessments by relevant subject matter experts (see Appendix B). For the ability tests, the content and order of the items mirrored the original test (Arthur, 2004, 2005).

Assessments of the reliability of alternate forms are inevitably conservative estimates, as the magnitude of the relationship between the tests exists across operationalizations (reflecting the degree of equivalency of the alternate forms) as well as time (reflecting temporal stability). This is visible in the magnitude of consistency between alternate form's reliability versus same form reliability over time. All things being equal for the same time interval, different test forms should exhibit lower than same form retest reliability. As seen in Table 4, these alternate forms do not exhibit a perfect correlation (indicating perfectly similarity), but are of a sufficiently high enough test-retest reliability as to indicate that these measures are not totally different.

Table 4

*Test-Retest Reliability by Test Forms*

		<b>Time 1</b>	
		<b>Form A</b>	<b>Form B</b>
<b>Knowledge Test</b>			
<b>Time 2</b>	<b>Form A</b>	.76 Same Form (A→A, N = 82)	.75 Alternate Form (B→A, N = 82)
	<b>Form B</b>	.71 Alternate Form (A→B, N = 89)	.60 Same Form (B→B, N = 87)
<b>Ability Test</b>			
<b>Time 2</b>	<b>Form A</b>	.76 Same Form (A→A, N = 82)	.58 Alternate Form (B→A, N = 82)
	<b>Form B</b>	.64 Alternate Form (A→B, N = 89)	.70 Same Form (B→B, N = 87)

*Note.* Total N = 340. See Appendix E for retest interval descriptive statistics by form and corrective feedback manipulations.

**6.4.4 Working Memory Capacity.** A computerized version of the N-back lag task was used to assess participants' working memory capacity (Shelton, Metzger, & Elliott, 2007). The N-back lag task consists of a list of items presented at a rate of one item (a letter) per second. At the end of each list, participants are asked to recall the last item in the list—the one presented 1-back, 2-back, or 3-back in the list—to operationalize working memory capacity function. Participants' scores are the average number of items correctly recalled in the list minus incorrect recalls (i.e., misses).

Like the Raven's APM, the inclusion of this measure permits an investigation of the relationships between working memory capacity and retest scores on both the knowledge and GMA tests. No test-retest reliability data are reported in the extant literature for Shelton et al.'s (2007) N-back lag task; however, in a convergent validation study, Geffen (2004) reported an average correlation of .51 between the subscales (0-, 1-, 2-, or 3-back trials) of the N-back lag task, reflecting some degree of internal consistency between the trials.

**6.4.5 Test-taking motivation.** Test-taking motivation was measured using eight items adopted from the Test Attitude Survey (Arvey, Strickland, Drauden, & Martin, 1990). Examples of items are “Doing well on this test is important to me” and “I will try my best on this test.” Ratings were made on a 5-point Likert scale (1 - strongly disagree; 5 - strongly agree). Arvey et al. (1990) reported an internal consistency reliability estimate of .85 for the test-taking motivation scores. The measure was scored by obtaining the mean of the eight items ( $T_1 \alpha = .72$  and  $T_2 \alpha = .74$ , test-retest reliability

coefficient = .65). Test-taking motivation was included in the study as a potential control variable.

**6.4.6 Test anxiety.** Test anxiety was measured using 10 items adopted from the Test Attitude Survey (Arvey et al., 1990). Examples of items are “I usually get very anxious about taking tests.” and “During a test, I get so nervous I can't do as well as I should have.” Ratings were made on a 5-point Likert scale (1 - strongly disagree; 5 - strongly agree). Arvey et al. (1990) report an internal consistency reliability estimate of .80 for the test-taking motivation scores. The measure was scored by obtaining the mean of the 10 items ( $T_1 \alpha = .64$  and  $T_2 \alpha = .66$ , test-retest reliability coefficient = .82). Like test-taking motivation, test anxiety was included as a potential control variable.

**6.4.7 Retest Interval Behavioral Inventory.** A number of possible participants' behaviors between the initial and retest measures could serve as a potential confound to the study's results (i.e., seeking test answers in some way between initial test and retest). Consequently, self-reports of how participants spent their time between the initial test and retest sessions were also collected. The full measure can be found in Appendix A.

## **6.5 Procedure**

Depending on the number of participants who were scheduled for and actually attended a session, participants were run in groups of up to 8 individuals with each participant at their own computer workstation. In terms of the details of the protocol, at Time 1, participants first read and signed the informed consent declaration, which explained the purpose and instructions for the study. Participants then completed the individual difference measures, which consisted of the working memory capacity

measure and the paper-and-pencil Raven's APM. Prior to completing the Raven's APM and initial test measures, participants then completed the test anxiety and test-taking motivation measures. Participants next went on to complete the initial knowledge (i.e., GPCE version 2.1) and ability tests. The order of the knowledge and ability tests was counterbalanced across conditions (test forms; corrective feedback versus no feedback condition).

Participants were randomly assigned to the between-subjects corrective feedback manipulation (i.e., either receiving corrective feedback or no feedback) prior to beginning the study. So, following the completion of the initial tests, participants in the corrective feedback condition received their total test performance as well as item-level corrective feedback for both the knowledge and ability tests. Corrective feedback was administered via Inquisit. Each item was presented on the computer screen individually, along with the participant's response and the keyed response. Participants could spend a maximum of thirty seconds reviewing each item or choose to advance to the next item at their own pace. Corrective feedback was only provided for the items that participants responded to. Participants could not take feedback with them. In contrast, participants in the control condition received no feedback at either the item- or test-level. Upon completion of the tests and/or feedback session, participants were thanked for their participation and scheduled to return for Time 2.

Research has consistently demonstrated that the length of the retest or retention interval is an important moderator of the retesting effect (Hausknecht et al. 2007; Roediger & Karpicke, 2006b). Thus, the choice of the length of the retest interval was

not trivial. Hausknecht et al.'s (2007) meta-analytic mean retest interval was 134.52 days ( $SD = 304.67$  days), with a median of 20 days. Considering this extreme variability of retest intervals in the extant literature and the study's goal of investigating the cognitive process underlying the retesting effect rather than the magnitude of said effect, the present study used a shorter, 7- to 10-day retest interval (based on the scheduling of participants). That is, the stronger the retesting effect this protocol could engender, the more amenable it was to investigating the research questions of interest. Furthermore, past research (e.g., Roediger & Karpickie, 2006a) has indicated that a 7-day retest interval is sufficient for obtaining the retesting effect.

Following this 7-10 day retest interval, participants returned to the lab to complete the second administrations of the knowledge and ability test (Time 2). Participants again completed the test anxiety and test-taking motivation measures prior to completing the retests. Participants were randomly assigned to either the same or alternate forms of the retests. Assignment of participants to the same or alternate forms occurred across test construct domains (i.e., knowledge and ability). That is, individuals assigned to the same form received the same forms of *both* the knowledge and ability tests, and individuals assigned to the alternate form received the alternate forms of *both* the knowledge and ability tests. Again, the order of the knowledge and ability tests was counterbalanced across conditions. Additionally, Test Forms A and B were counterbalanced across both conditions (same versus alternate test forms; corrective feedback versus no feedback) and time points, such that Test Form A was not always the initial test.

At the conclusion of the second study session, participants were debriefed as to the purpose of the study, paid for their participation, completed their payment receipt, and were thanked for their time and efforts. Participants were asked to refrain from telling anyone about the study until after January 1st, 2015 (the anticipated data collection completion date), so that future participants were not furnished with information that might bias their participation or performance. The first study (Time 1) session was approximately 1.5 hours in length. The second and final study session (Time 2) was approximately 1 hour long. Table 5 provides an overview of the study protocol.

Table 5

*Protocol for Time 1 and Time 2 Administration*

<b>Time 1 Protocol</b>				
<b>Session</b>	<b>Scheduled Activity</b>	<b>Activity Length</b>	<b>Session Time</b>	<b>Cumulative Time Elapsed</b>
0	<i>Administrative</i> Informed Consent	1 Minutes	1 Minutes	1 Minutes
1	<i>Individual Difference Measure</i> Working Memory	15 Minutes	15 Minutes	16 Minutes
2	<i>Test Attitudes</i> Test Anxiety (T <sub>1</sub> ) Test-Taking Motivation (T <sub>1</sub> )	8 Minutes 8 Minutes	16 Minutes	32 Minutes
3	<i>Individual Difference Measure</i> General Mental Ability	15 Minutes	15 Minutes	47 Minutes
4	<i>Initial Test Measures</i> Knowledge Test Form A or B (T <sub>1</sub> ) Ability Form A or B (T <sub>1</sub> )	15 Minutes 10 Minutes	25 Minutes	1 Hour 12 Minutes
5	<i>Demographics</i>	2 Minutes	2 Minutes	1 Hour 14 Minutes
6	<i>Between-Subjects Manipulation:</i> Item-Level Feedback versus Control	10 Minutes	10 Minutes	1 Hour 24 Minutes
<b>Time 2 Protocol</b>				
<b>Session</b>	<b>Scheduled Activity</b>	<b>Activity Length</b>	<b>Session Time</b>	<b>Cumulative Time Elapsed</b>
1	<i>Retest Interval Behavioral Inventory</i>	2 Minutes	2 Minutes	2 Minutes
2	<i>Test Attitudes</i> Test Anxiety (T <sub>2</sub> ) Test-Taking Motivation (T <sub>2</sub> )	8 Minutes 8 Minutes	16 Minutes	18 Minutes
3	<i>Retest Measures</i> Knowledge Test Form A or B (T <sub>2</sub> ) Ability Form A or B (T <sub>2</sub> )	15 Minutes 10 Minutes	25 Minutes	43 Minutes
4	<i>Debrief</i>	2 Minutes	2 Minutes	45 Minutes

## **7. RESULTS**

Mirroring the design of the present study, the analyses and results are organized such that the (a) memory and (b) learning effects explanations for the retesting effect are pitted against each other such that either one or the other could be supported. All analyses were conducted using SAS 9.2 PROC GLM. This SAS procedure allows testing a multivariate generalized linear model containing both within and between subject variables, as well as using continuous variables (working memory and Raven's APM).

### **7.1 Descriptive Statistics and Control Variables**

Three-hundred seventy-four participants participated in the Day 1 session of the protocol. Of those, three-hundred forty participants (mean age = 21.04 years old,  $SD = 4.43$ , 238 females [70.00%], 76 Psychology majors [22.69%]) completed both Day 1 and Day 2 sessions of the retest protocol and comprised the final sample. Thus, thirty-four participants were dropped from the analyses due to attrition. For a comparison between the participants who completed versus attrited from the study protocol, see Table 6. Based on these analyses, only Time 1 ability scores significantly differed between participants who completed versus attrited. However, participants who completed the study exhibited significantly *lower* ability scores at Time 1, opposite of what would be expected if participants who performed poorly did not anticipate they would receive the performance award. Thus, there appears no reason to believe that there was any systematic attrition.

Table 6

*Mean Differences Between Participants Who Completed and Attrited from the Retest Protocol*

Variable	Completed Participants <sup>a</sup>		Attrited Participants <sup>b</sup>		<i>d</i>
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	
Age	21.04	4.43	20.59	1.31	0.14
Raven's APM	8.64	2.03	8.47	2.11	0.08
Working Memory	3.74	.97	3.59	1.14	0.14
Ability Score T <sub>1</sub>	34.67	7.48	37.82	8.29	0.40*
Knowledge Score T <sub>1</sub>	20.21	5.21	19.91	5.11	0.06
Test Motivation T <sub>1</sub>	3.84	.35	3.79	.32	0.06
Test Anxiety T <sub>1</sub>	2.84	.56	2.97	.60	0.22

*Note.* <sup>a</sup> *N* = 340. <sup>b</sup> *N* = 34. \* *p* < .05. All significance tests are two-tailed Satterhwaite *t*-tests, which do not assume equal variances between groups. (There was no difference between the results using the traditional pooled variance *t*-tests between groups).

### 7.1.1 Test Attitudes

To test whether test motivation and test anxiety offered alternate explanations for the effects, and therefore whether they would be included as covariates, paired sample two-tailed *t*-tests were conducted. Test motivation significantly *decreased* from Time 1 to Time 2, opposite the anticipated direction ( $t_{(339)} = 4.05, p < .05, d = 0.20$ ). Test anxiety increased; however, this increase was not significant ( $t_{(339)} = -.77, p > .05, d = -0.02$ ). Thus, test motivation and test anxiety were not included as covariates in the following analyses.

### 7.1.2 Retest Interval Behaviors

It is possible that participants engaged in activities and behaviors that may have improved their scores on the retest measures, regardless of the manipulations or the retesting intervention itself. To assess whether participants' behaviors during the retest interval could present an alternate explanation for the retest increases, mean differences

on ability and knowledge retest scores between the individuals who reported engaging in test-content related retest interval behaviors (or did not) were examined (see Table 7 and Table 8). Retest interval behaviors were investigated across all of the study's conditions.

One-hundred fourteen participants reported that an instructor covered similar psychology content in a class after they completed the Day 1 protocol session (33.53%). Forty-two participants (12.35%) reported that they completed psychology course work (including reading, studying) that was relevant to the initial test's content during the retest interval. Participants who reported that they either had an instructor cover similar content in class or that they had completed related psychology course work ( $N = 73$ , 21.47% [not all participants provided hours if they had indicated they covered similar course content]), reported that they spent on average 1.64 hours ( $SD = 2.24$ ) on this coursework. Thirty participants (8.82%) reported that they sought the answers of the tests' items in some way between the initial and retest measures and that they spent on average .80 hours ( $SD = .47$ ) seeking answers. A single participant (.29%) reported that she/he had spent 3 hours engaged in test preparation activities for the GRE Psychology Test between the two study sessions.

Table 7

*Mean Test Score Increases Between Participants Who Covered Test Content During Retest Interval*

		Covered in Class		Did Not Cover in Class		<i>d</i>
		Participants <sup>a</sup>		Participants <sup>b</sup>		
Variable		<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	
Same Form	Mean Ability Score Increase	8.20	5.37	8.11	6.07	0.02
	Mean Knowledge Score Increase	2.15	4.55	3.35	5.61	-0.23

		Covered in Class		Did not Cover in Class		<i>d</i>
		Participants <sup>c</sup>		Participants <sup>d</sup>		
Variable		<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	
Alternate Form	Mean Ability Score Increase	5.98	6.68	5.65	5.68	0.06
	Mean Knowledge Score Increase	0.92	3.83	1.34	3.65	-0.11

*Note.* <sup>a</sup> *N* = 60. <sup>b</sup> *N* = 109. <sup>c</sup> *N* = 60. <sup>d</sup> *N* = 111. All significance tests are one-tailed Satterhwaite *t*-tests, which do not assume equal variances between groups. (There was no difference between the results using the traditional pooled variance *t*-tests between groups).

There were no significant mean differences on knowledge and ability retest increases across same versus alternate retest forms between participants who covered relevant test material from their instructor, psychology course content, or sought answers during the retest interval. In fact, many of these differences between the groups were in the *opposite* direction of what one would expect if these activities served as a plausible alternate explanation for the retest increases. That is, participants who covered relevant test content or sought test answers during the retest interval performed *worse* on the retest than those who did not. Furthermore, few participants reported engaging in these behaviors at all. These results do not indicate that retest interval behaviors were associated with greater retest increases, and therefore do not offer a plausible alternate explanation for the retesting effect in the present study.

Table 8

*Mean Test Score Increases Between Participants Who Sought Test Content During Retest Interval*

		Sought Test Answers		Did Not Seek Test Answers		<i>d</i>	
		Participants <sup>a</sup>		Participants <sup>b</sup>			
Variable		<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>		
Same Form	Mean Ability Score Increase	9.87	5.36	7.96	5.87	0.33	
	Mean Knowledge Score Increase	-0.07	4.42	3.42	5.33	-0.66	

		Sought Test Answers		Did not Seek Test Answers		<i>d</i>	
		Participants <sup>c</sup>		Participants <sup>d</sup>			
Variable		<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>		
Alternate Form	Mean Ability Score Increase	5.60	6.98	5.72	5.92	-0.02	
	Mean Knowledge Score Increase	-0.47	4.12	1.24	3.68	-0.46	

*Note.* <sup>a</sup> *N* = 15. <sup>b</sup> *N* = 136. <sup>c</sup> *N* = 15. <sup>d</sup> *N* = 148. Twenty-six participants did not report whether or not they sought test answers during the retest interval. All significance tests are one-tailed Satterwaite *t*-tests, which do not assume equal variances between groups. (There was no difference between the results using the traditional pooled variance *t*-tests between groups).

Table 9 shows the descriptive statistics and zero-order correlations among the study variables. The relationships between variables in the present study were in the anticipated direction. All analyses were based on the raw scores. However, for the purposes of interpretation, all test scores were standardized to total percent correct for the presented figures. For hypothesized memory effects versus learning explanation figures, refer to Appendix C. For figures reporting the raw scores, refer to Appendix D.

Table 9

*Intercorrelations amongst Study Variables*

	<i>M</i>	<i>SD</i>	1	2	3	4	5	6	7	8
<i>Individual Differences</i>										
1. Raven's APM	8.64	2.03	-							
2. Working Memory	3.74	0.97	0.35*	-						
<i>Time 1</i>										
3. Ability Score T <sub>1</sub>	34.67	7.48	0.30*	0.26*	-					
4. Knowledge Score T <sub>1</sub>	20.21	5.20	0.21*	0.22*	0.35*	-				
5. Ability RT T <sub>1</sub> (ms)	12,991	3,126	-0.17*	-0.18*	-0.80*	-0.24*	-			
6. Knowledge RT T <sub>1</sub> (ms)	19,040	4,881	0.02	-0.02	-0.37*	-0.35*	0.42*	-		
7. Test Motivation T <sub>1</sub>	3.84	0.35	0.12*	0.24*	0.11*	-0.05	-0.04	0.00	(.72)	
8. Test Anxiety T <sub>1</sub>	2.84	0.56	-0.24*	-0.23*	-0.22*	-0.25*	0.09	0.03	-0.15*	(.64)
<i>Time 2</i>										
9. Ability Score T <sub>2</sub>	41.62	7.02	0.27*	0.32*	0.65*	0.38*	-0.54*	-0.34*	0.14*	-0.30*
10. Knowledge Score T <sub>2</sub>	22.26	6.15	0.25*	0.27*	0.25*	0.68*	-0.09	-0.06	0.02	-0.28*
11. Ability RT T <sub>2</sub> (ms)	10,220	2,466	-0.10*	-0.19*	-0.54*	-0.27*	0.58*	0.48*	-0.02	0.13*
12. Knowledge RT T <sub>2</sub> (ms)	14,732	4,160	-0.10*	-0.10*	-0.32*	-0.30*	0.33*	0.51*	0.01	0.09*
13. Test Motivation T <sub>2</sub>	3.77	0.36	0.12*	0.10*	0.18*	0.06	-0.11*	-0.09	0.65*	-0.14
14. Test Anxiety T <sub>2</sub>	2.85	0.52	-0.25*	-0.26*	-0.22*	-0.34*	0.09	0.06	-0.13*	0.82*
15. Retest Interval Length (Days)	7.61	1.19	-.06	.04	-.01	-.08	.00	.03	.08	-.04

Note. *N* = 340. Raven's APM = Raven's Advanced Progressive Matrices. RT = Response Time. ms = Milliseconds. T<sub>1</sub> = Time 1. T<sub>2</sub> = Time 2.

Standardized coefficient alphas are reported along the diagonal in parentheses (where appropriate). \* *p* < .05, one-tailed.

Table 9

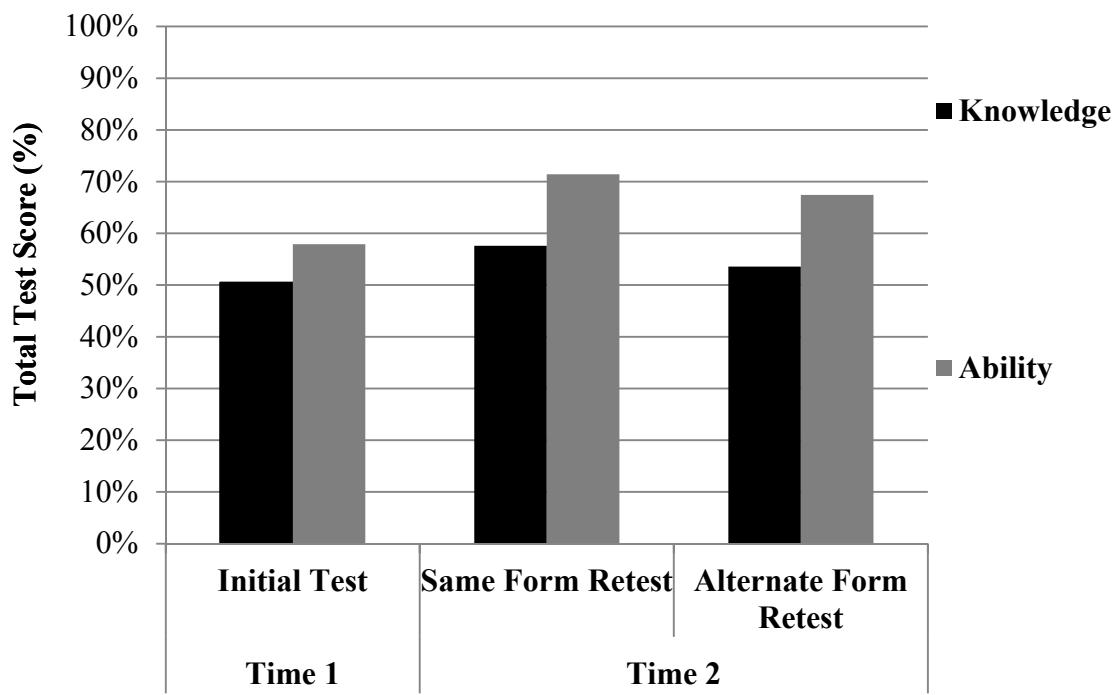
*Intercorrelations amongst Study Variables (Continued)*

	9	10	11	12	13	14
<i>Time 2</i>						
9. Ability Score T <sub>2</sub>	-					
10. Knowledge Score T <sub>2</sub>	0.42*	-				
11. Ability RT T <sub>2</sub> (ms)	-0.69*	-0.16*	-			
12. Knowledge RT T <sub>2</sub> (ms)	-0.39*	-0.21*	0.47*	-		
13. Test Motivation T <sub>2</sub>	0.13*	0.00	-0.05	-0.01	(.74)	
14. Test Anxiety T <sub>2</sub>	-0.30*	-0.35*	0.13*	0.05	-0.15*	(.66)
15. Retest Interval Length (Days)	-.05	-.02	.08	.07	.00	-.04

Note. N = 340. Raven's APM = Raven's Advanced Progressive Matrices. RT = Response Time. ms = Milliseconds. T<sub>1</sub> = Time 1. T<sub>2</sub> = Time 2. Standardized coefficient alphas are reported along the diagonal in parentheses (where appropriate). \* p < .05, one-tailed.

## 7.2 Hypothesis Testing

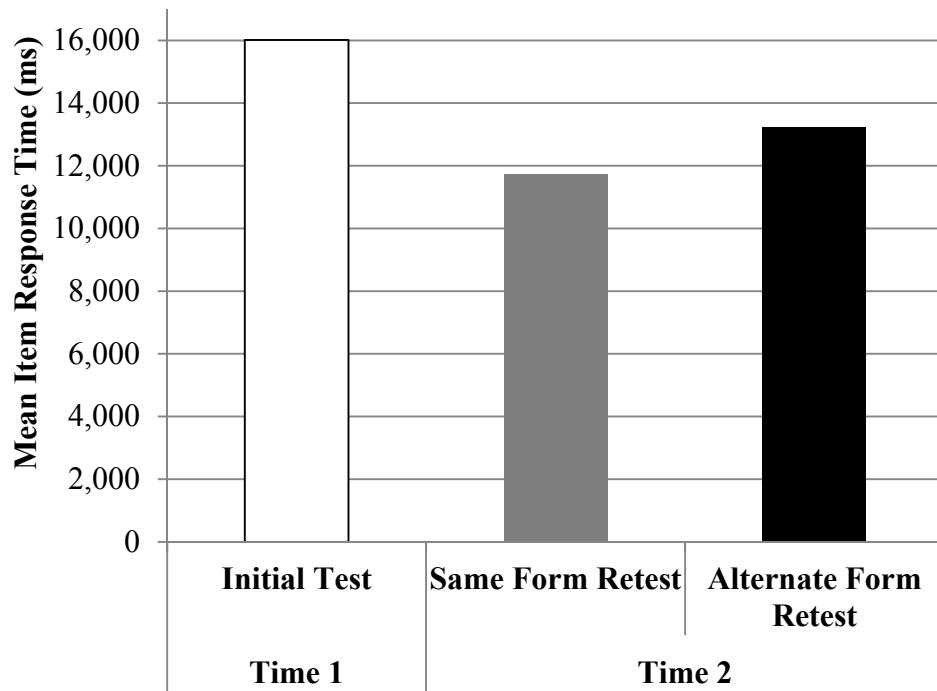
**7.2.1 Hypothesis 1.** To competitively test Hypotheses 1a versus 1b, a two-way (retest form; same form versus alternate form) repeated measures mixed factorial ANOVA was used to test differences in total test scores (collapsing across construct domains [knowledge and ability]). There was a significant interaction of retest form (same versus alternate form) by time ( $F_{(2, 337)} = 11.86, p < .05, \eta^2 = .03$ ), such that the magnitude of the retest increase was greater for same retest forms compared to alternate form retests. There were also significant main effects for time ( $F_{(2, 337)} = 6179.15, p < .05, \eta^2 = .53$ ) and test form ( $F_{(1, 338)} = 3.29, p < .05, \eta^2 = .17$ ). Thus, the results were supportive of Hypothesis 1a; in accordance with a memory explanation, the observed retest score increases were higher on the same form compared to the alternate form (see Figure 1).



*Figure 1.* Hypotheses 1 and 3: Total test score (percentage) by same form retest versus alternate form retest over time.

**7.2.2 Hypothesis 2.** To competitively test Hypotheses 2a versus 2b, a two-way (retest form; same form versus alternate form) repeated measures mixed factorial ANOVA testing differences in mean item response times (collapsing across construct domains [knowledge and ability]) was used. There was a significant interaction of retest form (same form versus alternate form) by time ( $F_{(2, 337)} = 24.10, p < .05, \eta^2 = .03$ ), such that mean item response times were faster for same form retest compared to alternate form retests. There were also significant main effects for time ( $F_{(2, 337)} = 4301.68, p < .05, \eta^2 = .63$ ) and test form ( $F_{(1, 338)} = 5.39, p < .05, \eta^2 = .02$ ) on mean item response times. Thus, the results were supportive of Hypothesis 2a; in accordance with

a memory explanation rather than learning, the observed retest mean item response times were faster on the same form retest compared to the alternate form retest (see Figure 2).



*Figure 2. Hypothesis 2: Mean item response time by same form retest versus alternate form retest over time.*

**7.2.3 Hypothesis 3.** To competitively test Hypotheses 3a and 3b, a three-way (retest form, same form versus alternate form; by construct type, ability and knowledge) repeated measures mixed factorial ANOVA testing differences in total test scores was used. Construct domain was included as a within subjects variable. Support for Hypothesis 3a, that memory best explains the retest score increases, would be found if there is a significant interaction of same versus alternate form retest by time, but no interaction with construct domain, such that the magnitude of the retest score increase is

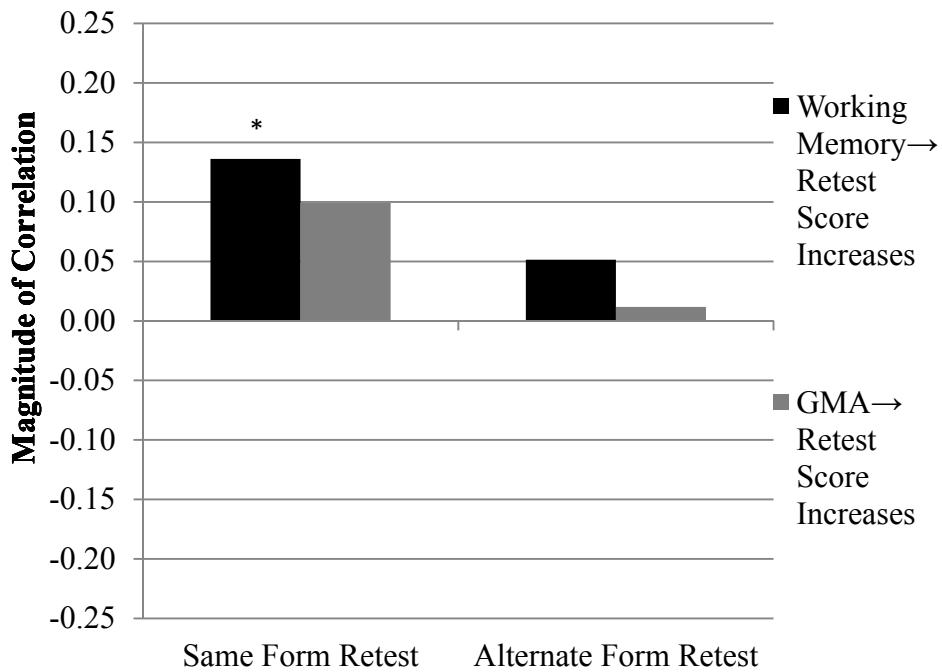
larger on the same form retest compared to the alternate form retest (irrespective of the construct domain). Support for Hypothesis 3b, that learning best explains the retest score increases, would be found if a significant three-way interaction is observed between construct domain and total test scores increases from Time 1 to Time 2, such that the magnitude of the retest score increase is greater for knowledge compared to ability, yet there is no significantly greater retest increase for same form retest compared to alternate form retest.

There was a significant interaction of retest form by time (as reported in Hypothesis 1 above); however, there was *no* significant three-way interaction by construct domain ( $F_{(2, 337)} = .54, p > .05, \eta^2 = .00$ ). Furthermore, there was a significant two-way interaction of construct domain over time, which was the *opposite* of the hypothesized direction ( $F_{(2, 337)} = 1281.47, p < .05, \eta^2 = .35$ ). That is, the ability test showed greater retest score increases compared to the knowledge test (irrespective of test form changes). Thus, the pattern of results were supportive of Hypothesis 3a; in accordance with a memory explanation rather than a learning explanation, the observed retest score increases were higher on the same form compared to the alternate form, but the magnitude was no different between knowledge and ability constructs (see Figure 1).

**7.2.4 Hypothesis 4.** Hypotheses 4a versus 4b was tested by comparing the influence of GMA and the magnitude of the retest score increase for the same and alternate form retests, and the influence of working memory capacity and the magnitude of retest score increase for the same and alternate form retests. Support for Hypothesis 4a, that memory best explains the retest score increases, would be found if there was a

significantly stronger correlation between working memory capacity and retest score increases compared to GMA and retest score increases, and that the magnitude of working memory capacity-retest scores increases is of a greater magnitude for same form retest rather than alternate form retest. Conversely, support for Hypothesis 4b, that learning best explains the retest score increases, would be found if there is a significantly stronger influence of GMA and retest score increases compared to the relationship between working memory capacity and retest score increases, irrespective of test-takers completing the same form retest or alternate form retest.

Hypotheses 4 was tested using the generalized linear model, including the retest form manipulation and both the individual differences (GMA and working memory) as predictors of retest score increases. There was a significant main effect of GMA over time, indicating that GMA was related to the magnitude of retest increases ( $F_{(18, 360)} = 1.99, p < .05, \eta^2 = .03$ ), thus, demonstrating some evidence for learning. There was a significant interaction of working memory by retest form on the magnitude of retest increases ( $F_{(54, 360)} = 1.32, p < .05, \eta^2 = .08$ ). There was no significant main effect of working memory on retesting effects ( $F_{(94, 360)} = 1.22, p > .05, \eta^2 = .12$ ). Thus, the pattern of results was supportive of Hypothesis 4a; in accordance with a memory explanation, the observed retest score increases were higher on the same form compared to the alternate form, and this effect was stronger for individuals with higher working memory capacity but not for individuals with higher GMA (see Figure 3).



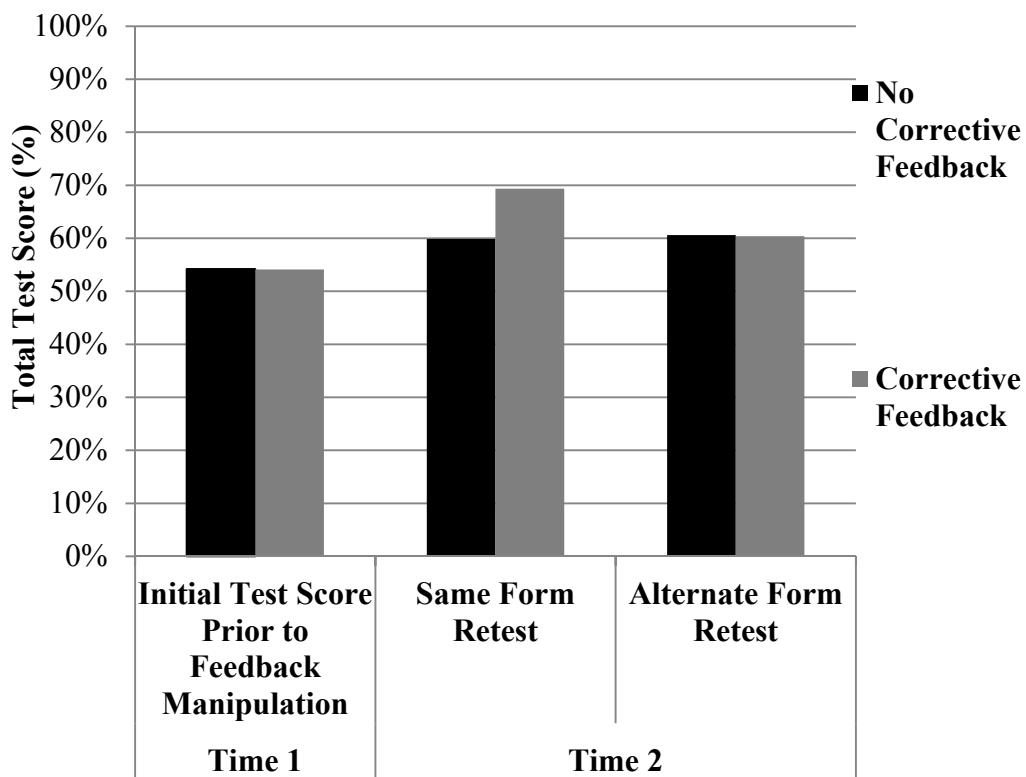
*Figure 3.* Hypothesis 4: GMA and working memory predicting retest score increases by same form retest versus alternate form retest.

*Note.* Effect sizes are  $r$ 's. \*  $p < .05$ .

**7.2.5 Hypothesis 5.** Hypothesis 5a versus 5b were competitively tested using a 2 (retest form; same form versus alternate form)  $\times$  2 (corrective feedback versus no feedback) repeated measures mixed factorial ANOVA (collapsing across construct domains [knowledge and ability]). Support for Hypothesis 5a, that memory best explains the retest score increases, would be found if there was a three-way interaction of retest form and corrective feedback on retest score increases, such that the magnitude of the retest score increase is greatest for the corrective feedback condition on the same form retest compared to the alternate form retest. Support for Hypothesis 5b, that learning best explains the retest score increases, would be found if there is a two-way

interactive effect of corrective feedback on retest score increase, but of a similar magnitude whether test-takers completed the same form retest or alternate form retest.

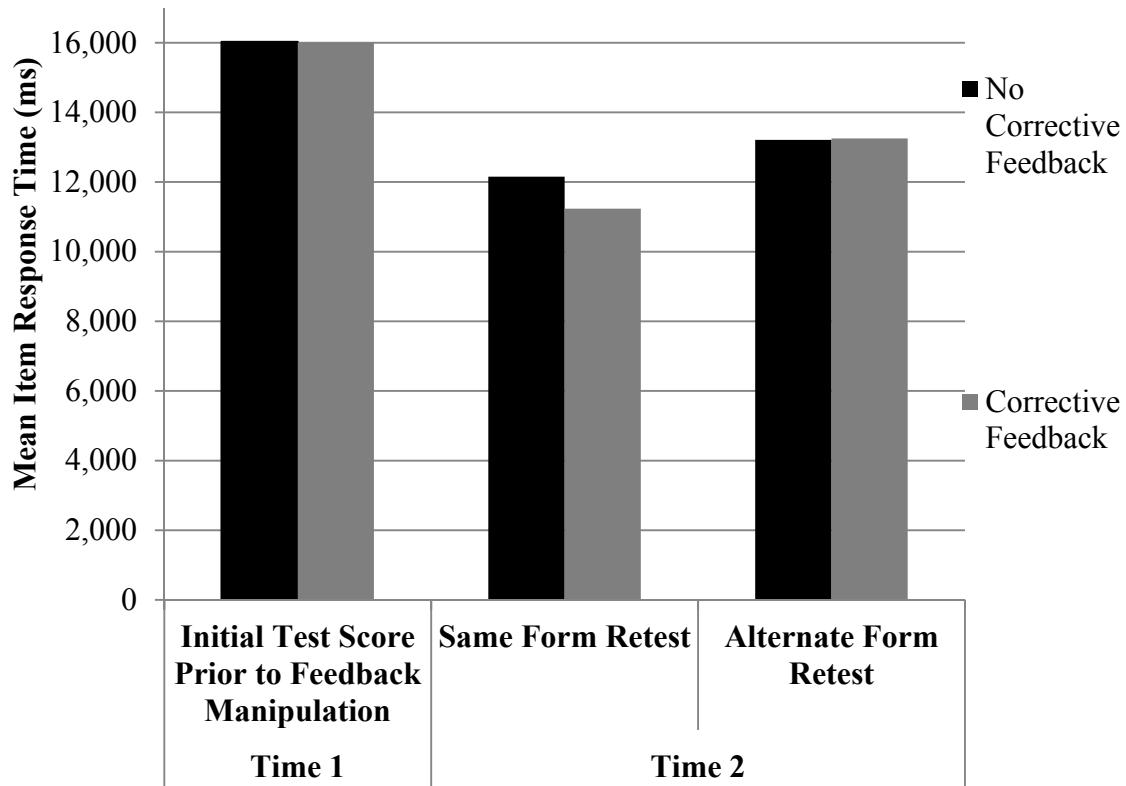
There was a significant three-way interaction of retest form by feedback manipulation by time ( $F_{(2, 335)} = 9.95, p < .05, \eta^2 = .01$ ); such that there was a greater retesting effect from corrective feedback on same form rather than alternate form. There were significant main effects of retest form over time ( $F_{(2, 335)} = 13.90, p < .05, \eta^2 = .03$ ) and corrective feedback over time ( $F_{(2, 335)} = 15.64, p < .05, \eta^2 = .03$ ). Thus, the patterns of results were supportive of Hypothesis 5a; in accordance with a memory explanation rather than learning, the observed retest score increases were higher on the same retest form compared to the alternate retest form, and this effect was stronger when participants were provided corrective feedback (see Figure 4).



*Figure 4.* Hypothesis 5: Total test score (percentage) by same form retest versus alternate form retest over time (collapsed across knowledge and ability).

**7.2.6 Hypothesis 6.** Hypothesis 6a versus 6b were tested using a 2 (retest form; same form retest versus alternate form retest)  $\times$  2 (corrective feedback versus no feedback) repeated measures mixed factorial ANOVA examining the differences in mean item response times (collapsing across construct domains [knowledge and ability]). Support for Hypothesis 6a, that memory best explains the decreased retest item response time, would be found if there was a significant three-way interactive effect of corrective feedback condition with retest form on mean retest item response times, such that faster mean item response times are observed with corrective feedback but of a

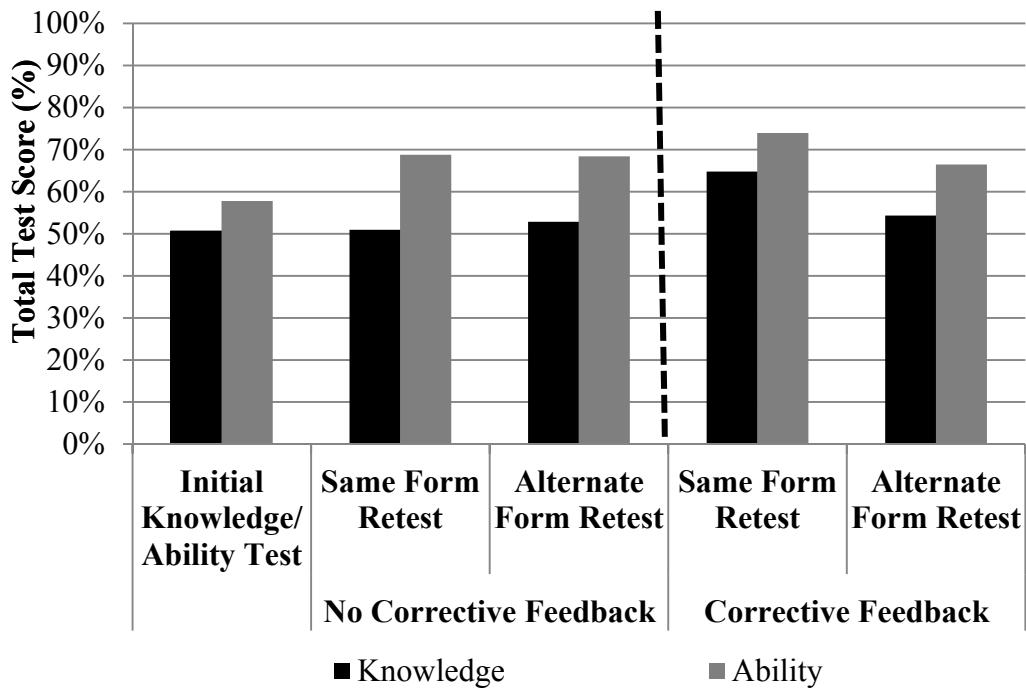
significantly larger magnitude on same form retest compared to alternate form retest. Support for Hypothesis 6b, that learning best explains the decreased retest item response time, would be found if there was a significant effect of corrective feedback on mean item response time, such that faster mean item response times scores are observed with corrective feedback, but of a similar magnitude whether test-takers completed the same form retest or alternate form retest. There was *no* significant three-way interaction of time by retest form by feedback manipulation ( $F_{(2, 335)} = 1.52, p > .05, \eta^2 = .00$ ). There were significant main effects of the retest form manipulation over time ( $F_{(2, 335)} = 24.62, p < .05, \eta^2 = .03$ ), but *no* effect of the feedback manipulation over time ( $F_{(2, 335)} = 1.88, p > .05, \eta^2 = .00$ ). Thus, there was no conclusive support for *either* Hypothesis 6a or 6b; in general, the pattern of results did not display clear support for either a memory or learning explanation (see Figure 5).



*Figure 5.* Hypothesis 6: Mean item response time by same versus alternate form retest with corrective feedback or no corrective feedback over time (collapsed across knowledge and ability).

**7.2.7 Hypothesis 7.** Hypotheses 7a versus 7b was tested using a 2 (retest form; same form versus alternate form)  $\times$  2 (corrective feedback versus no feedback)  $\times$  2 (construct domain, ability or knowledge, within-subjects) repeated measures mixed factorial ANOVA examining differences on total test scores. Support for Hypothesis 7a, that memory best explains the retest score increases, would be found if there was a three-way interaction of retest form and corrective feedback on retest score increases, such that the magnitude of the retest score increase is greater for the corrective feedback condition on the same form retest compared to the alternate form retest, yet be of a

similar magnitude across construct domains. Support for Hypothesis 7b, that learning best explains the retest score increases, would be found if there was an interactive effect of corrective feedback and construct domain on retest score increase, such that the magnitude of the retest score increase is greatest for the corrective feedback condition on knowledge compared to ability, but of a similar magnitude whether test-takers completed the same form retest or alternate form retest. However, there was *no* significant 4-way interaction indicating that this differed by construct domain ( $F_{(2, 335)} = .69, p > .05, \eta^2 = .00$ ). Thus, the pattern of results was supportive of Hypothesis 7a; in accordance with a memory explanation, the observed retest score increases were higher on the same retest form compared to the alternate retest form, and this effect was stronger when participants were provided corrective feedback, but the magnitude was no different between knowledge and ability constructs (see Figure 6).

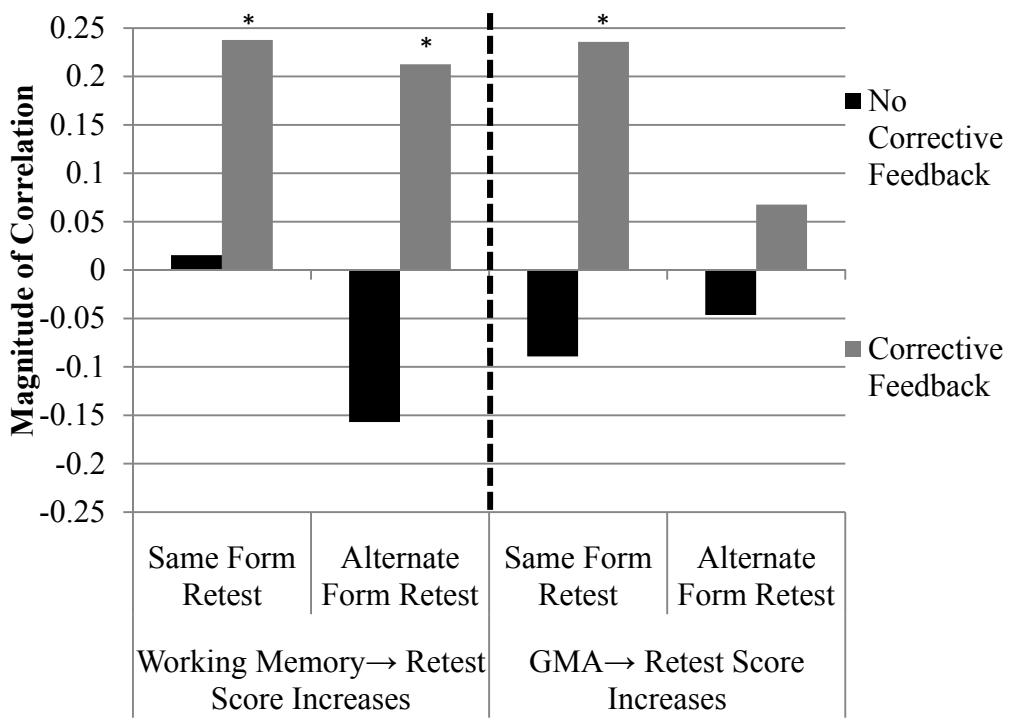


*Figure 6.* Hypothesis 7. Total test score (percentage) by same versus alternate form retest with corrective feedback or no corrective feedback over time.

**7.2.8 Hypothesis 8.** Support for Hypothesis 8a, that memory best explains the retest score increases, would be found if working memory capacity has a stronger influence on retest score increases compared to the influence of GMA on retest score increases in the corrective feedback condition compared to the no feedback condition, and that the magnitude of working memory capacity-retest score increases correlation is of greater magnitude for same form retest than alternate form retest. Conversely, support for Hypothesis 8b, that learning best explains the retest score increases, would be found if there was a significantly stronger influence of GMA on retest score increases compared to the influence of working memory capacity on retest score increases in the

corrective feedback condition compared to the control condition, but that there is no difference between same form retest versus alternate form retest.

There was no significant three-way interaction for GMA by corrective feedback over time ( $F_{(16, 360)} = 1.44, p > .05, \eta^2 = .02$ ), irrespective of retest form; thus, demonstrating no evidence of learning effects. However, there was *no* significant four-way interaction for working memory by retest form manipulation by corrective feedback over time ( $F_{(32, 360)} = .72, p > .05, \eta^2 = .02$ ); thus, demonstrating no evidence of memory effects either. In general, the pattern of results did not display conclusive support for either a memory (Hypothesis 8a) or learning explanation (Hypothesis 8b). Specifically, the observed retest score increases were higher on the same form compared to the alternate form, and this effect was again stronger for individuals receiving corrective feedback, but these effects were not moderated by either working memory capacity or GMA (see Figure 7).



*Figure 7. Hypothesis 8: GMA and working memory predicting retest score increases by same form retest versus alternate form retest with corrective feedback and no feedback*  
*Note.* Effect sizes are  $r'$ s. \*  $p < .05$ .

## **8. DISCUSSION**

It is clear that the retesting effect consistently influences the test scores that organizations, educators, and policy-makers use to make decisions across work, educational, and social settings. Despite the divergent and significant implications for retesting practices that a learning versus a memory explanation holds, a lack of integrated research between cognitive, education, and personnel researchers limits advancements in this domain and consequentially, the ability to make clear evidence-based recommendations. Theoretically, delineating whether learning or memory best accounts for the retesting effect is the major contribution of the present study to the psychology, education, and management literatures. Practically, these findings offer implications for test developers, test users, test-takers, policy makers, and society at large.

Table 10 summarizes the results of this study's competing hypotheses.

Generally, these results supported the explanation that the retesting effect reflects memory, rather than learning. Overall, there was a consistent and significant interaction of retesting by same versus alternate retest form, regardless of manipulations (corrective feedback or no feedback) or constructs (knowledge versus ability). That is, retest increases were significantly attenuated by alternate retest forms and these effects were stronger when corrective feedback was provided; however, the magnitude of the retest increases did not differ by constructs (knowledge versus ability). Furthermore, retesting resulted in significantly faster item response times, but again, these response times were attenuated (slowed) by alternate retest forms. This pattern of results appears unlikely to reflect meaningful learning across retest forms, but instead the memory of previous responses that are in fact dependent upon test-specific information.

Table 10

*Competitive Hypotheses supporting Memory Effects versus Learning Explanations*

Hypotheses	If memory effects best explains the retesting effect	If learning best explains the retesting effect	Results	Supported?
1	(a) Retest scores will be higher than initial scores; however, the magnitude of this effect will be larger on the same form retest than the alternate form retest.	(b) Retest scores will be higher than initial test scores and the magnitude of this effect will be similar for both the same and alternate form retests.	Retest scores were higher than initial test scores, but of a greater on same forms ( $F_{(2, 337)} = 11.86, p < .05, \eta^2 = .03$ ).	Memory Explanation
2	(a) Retest response times will be faster than initial test response times. However, the magnitude of the difference between the initial and retest response times will be larger on the same form retest than the alternate form retest.	(b) Retest response times will be faster than initial test response times. However, the magnitude of the difference between the initial and retest response time for the same and alternate form retest will be small (i.e., similar).	Retest response times were faster than initial test response times, such that mean item response times were faster for same forms compared to alternate form retests ( $F_{(2, 337)} = 24.10, p < .05, \eta^2 = .03$ ).	Memory Explanation
3	(a) Retest score increases from initial test scores will be approximately equal for knowledge and ability; however, the magnitude of this retest increase will be larger on the same form retest compared to the alternate form retest (across construct domains).	(b) Retest score increases from initial test scores will be higher for knowledge than ability, irrespective of whether the same or alternate form retest is administered.	There was <i>no</i> significant three-way interaction of retest score gains by construct domain ( $F_{(2, 337)} = .54, p > .05, \eta^2 = .00$ ).	Memory Explanation

Table 10

*Competitive Hypotheses supporting Memory Effects versus Learning Explanations (Continued)*

Hypotheses	If memory effects best explains the retesting effect	If learning best explains the retesting effect	Results	Supported?
4	(a) The relationship between working memory capacity and retest score increases will be stronger than the relationship between GMA and retest scores increases and be of greater magnitude for the same form retest than the alternate form retests.	(b) The relationship between GMA and retest score increases will be stronger than the relationship between working memory capacity and retest scores increases and be of a similar magnitude between the same form retest and the alternate form retest.	There was a significant interaction of working memory by retest form on the magnitude of retest increases, such that working memory had a greater effect on test scores for same versus alternate forms ( $F_{(54, 360)} = 1.32, p < .05, \eta^2 = .08$ ).	Memory Explanation
5	(a) Retest score increases will be higher for the corrective feedback condition compared to the no corrective feedback condition; however, the magnitude of this effect will be larger on the same form retest than the alternate form retest.	(b) Retest score increases will be higher for the corrective feedback condition compared to the no corrective feedback condition and the magnitude of this effect will be similar for both the same form retest and alternate form retest.	Retest score increases were greatest with corrective feedback on same form rather than alternate form ( $F_{(2, 335)} = 9.95, p < .05, \eta^2 = .01$ ).	Memory Explanation
6	(a) Retest response times will be faster than initial test response times for the corrective feedback condition compared to the no corrective feedback condition. However, the magnitude of the difference between the initial and retest response times will be larger on the same form retest than the alternate form retest.	(b) Retest response times will be faster than initial test response times for the corrective feedback condition compared to the no corrective feedback condition and the magnitude of this effect will be similar for both the same form retest and alternate form retest.	There was no significant three-way interaction of retest form by feedback manipulation on retest response times ( $F_{(2, 335)} = 1.52, p > .05, \eta^2 = .00$ ).	Inconclusive

Table 10

*Competitive Hypotheses supporting Memory Effects versus Learning Explanations (Continued)*

Hypotheses	If memory effects best explains the retesting effect	If learning best explains the retesting effect	Results	Supported?
7	(a) Retest score increases will be approximately equal for ability and knowledge, but exhibit an interactive effect between receiving corrective feedback conditions and retest form, such that receiving corrective feedback will increase retest scores at a greater magnitude for the same form retest than the alternate form retest.	(b) Retest score increases will be higher for knowledge than ability and exhibit an interactive effect with corrective feedback condition, such that receiving corrective feedback will increase retest scores at a greater magnitude for knowledge than ability, but the magnitude of this effect will be similar for the same form retest and alternate form retest.	Construct domain did <i>not</i> interact with corrective feedback conditions and retest form to increase the magnitude of retest increases ( $F_{(2, 335)} = .69, p > .05, \eta^2 = .00$ ).	Memory Explanation
8	(a) A stronger relationship will exist between working memory capacity and retest score increases compared to GMA and retest score increases for the corrective feedback condition compared to the no corrective feedback condition, and the magnitude of this relationship will be larger on the same form retest than the alternate form retest.	(b) A stronger relationship will exist between GMA and retest score increases compared to working memory capacity and retest score increases for the corrective feedback condition compared to the no corrective feedback condition, and the magnitude of this relationship will be similar for both the same form retest and alternate form retest.	Neither GMA across forms ( $F_{(16, 360)} = 1.44, p > .05, \eta^2 = .02$ ) nor working memory within same forms ( $F_{(32, 360)} = .72, p > .05, \eta^2 = .02$ ) interacted with corrective feedback to increase retest scores.	Inconclusive

Results related to corrective feedback were less conclusive. The lack of main effects of corrective feedback on retest increases for Hypotheses 6A versus 6B and Hypotheses 8A versus 8B made it difficult to draw support for *either* memory or learning explanations. Hypothesis 2A was supported: test-takers completed items significantly faster upon retest, but this was attenuated by alternate retest forms. However, neither Hypothesis 6A nor 6B were conclusively supported: corrective feedback did not accelerate mean item response times regardless of retest forms (i.e., evidence of learning) nor did corrective feedback accelerate mean item response times to a greater extent on same versus alternate form retests (i.e., evidence of memory). Hypotheses 4 and 8 showed a similar pattern. Hypothesis 4A supported memory effects: greater working memory capacity was related to greater retest increases on same, but not alternate retest forms. However, neither Hypothesis 8A nor 8B were conclusively supported: corrective feedback did not facilitate the magnitude of this effect. Working memory capacity did not interact with corrective feedback to produce greater score gains for the same form retest.

As indicated by the nonsignificant effects of corrective feedback on retest increases across hypotheses, it appears that corrective feedback was a less powerful manipulation than intended, potentially limiting the ability to effectively examine the learning versus memory explanations for these hypotheses. Despite a substantial sample ( $N = 340$ ), the *post hoc* power analysis indicated what appeared to be quite low power ( $\beta$  power level = .12); in fact, what appeared to be too low to detect almost any significant differences between corrective feedback on knowledge versus ability tests. Given the

effect sizes found, 8,771 participants would be necessary to achieve a power of .80 to detect this difference.

The goal of selecting this particular hypothesis for the power analysis was to establish a conservative test as to whether learning occurred compared to the assumption that memory effects occurred regardless of intervention. In this power analysis, evidence for the memory explanation was the null hypothesis: that there would be *no* difference between the effect of corrective feedback on knowledge retest increases compared to ability retest increases. The power analysis estimated the sample size required to detect evidence for the learning explanation, that corrective feedback facilitated greater retest increases on knowledge tests compared to ability tests. As indicated in Table 3, the present study's effect size of corrective feedback on ability and knowledge retest increases were of comparable size to meta-analytic estimates, but quite similar to one another in magnitude. In fact, the present study showed a *greater* effect of corrective feedback on the ability retest than previously reported in the meta-analytic literature ( $d = 0.31$  compared to 0.22, Hausknecht et al., 2007).

This similarity in effect sizes of corrective feedback on ability and knowledge retest increases likely reflects meaningful similarities between the influence of corrective feedback on retest increases across constructs. Under the memory explanation, which was predominantly supported, hypothesized differences between the effect sizes of corrective feedback for ability versus knowledge tests would be zero. That is, if individuals were simply remembering and repeating prior responses, rather than learning from testing, and that corrective feedback facilitated remembering those responses, this

process should not differ between ability and knowledge tests. Thus, the post hoc power analysis *appears* to indicate a low power level, but may actually indicate a meaningful similarity between these effects.

Additionally, this study accounted for three common methodological challenges when investigating the retesting effect, specifically, test attitudes, length of the retest interval, and attrition. Various test attitudes have been proffered as explanations for the resting effect and a potential boundary condition on the value of retesting as a means for improving learning (Reeve et al., 2009); yet no studies have examined test attitudes over time or how test attitudes may covary with retest increases. The present study's results indicated no significant effect of either test motivation or test anxiety over time. Test attitudes are further discussed in the Limitations section.

The extant literature has proposed that retesting is a viable learning intervention across quite variable retest intervals, from months to minutes. For example, research has shown that the benefits of retesting occur after retest interval delays of only minutes (0–20 minutes) to almost a year (Carpenter et al., 2009). Some studies have demonstrated that longer retest intervals may exhibit *greater* retesting effects (e.g., J. C. K. Chan, 2009; C. I. Johnson & Mayer, 2009; Kornell, Bjork, & Garcia, 2011; Roediger & Karpicke, 2006b); whereas others indicate that the retesting effect diminishes over time (Hausknecht et al., 2007). The majority of research has used retest intervals of at least 1 day with a modal retention interval of 1 week (Carretta et al., 2000), which consequently served as the basis for the present study's retest interval of 7-10 days. Considering the variability of retest intervals used in the literature, and the possibility that the length of

retest interval may moderate the magnitude of the retesting effect, the effect of the variability in the present study's retest interval of 7 and 10 day (4 days) was examined. These results indicated that the present study's retest interval length did not have any relationship with the variables of interest, nor did the length of the retest interval significantly predict either ability or knowledge test score increases ( $r = -.04$ ,  $r = .06$ , respectively). So on the basis of these results, it was deemed unwarranted to statistically control for the variability of the retest interval.

Finally, attrition was examined as a potentially systematic selection bias threat that could influence the study's results. The results of this examination indicated the absence of any meaningful differences between participants who completed versus attrited after the initial test scores or any individual difference variables that were collected at Time 1 (see Table 6). Participants who completed the protocol were not higher on any relevant individual differences (working memory capacity, GMA), nor were they more motivated or less anxious. Participants who attrited did not perform worse on either initial knowledge or ability tests. Thus, it does not appear that a particular kind of participant (e.g., more test anxious or with lower GMA) systematically attrited from the study, or that they left the study based on their initial test performance.

Systematic attrition has the potential to dramatically confound retesting research in operational settings. For example, particular test-takers may systematically self-select out before retesting or even be selected, precluding the need for retesting. In either case, operational test settings often compare quite different samples between the initial test and retest, obscuring the retesting effect. Therefore, a strength of the present design is

that the full, rather than truncated, sample was used, which allows for a more robust examination of the retesting effect that is unrelated to the confounding effect of attrition.

### **8.1 Implications**

The prevalence and influence of the retesting effect on the scientific community and society at large underscores the importance of understanding whether these retest increases reflect either learning or memory. The extant literature provides support for both explanations of the retesting effect; however, a memory versus a learning explanation would offer quite divergent scientific and practical implications. These divergent explanations fundamentally alter the interpretation of retest scores for both knowledge and ability constructs in science and its implications for practice across research domains. Consequently, the primary contribution of this work lies in providing support for the conclusion that memory, rather than learning, appears to better explain the retesting effect, and therefore offers insight into the processes that affect the inferences drawn from retest scores and the benefits of testing as a learning intervention.

By demonstrating that retest score increases best reflect memory, rather than learning, researchers and practitioners should exercise caution when assuming that retest increases reflect learning when the alternate memory explanation is also possible (e.g., same form retests are used, corrective feedback is provided, a short retest interval is used). Relatedly, researchers and practitioners should consider the purpose of testing when implementing regular testing programs and retesting policies: Is retesting for assessment and decision-making (e.g., academic and employment selection)? Or for providing feedback and formative assessment under contexts where learning is expected

and encouraged (e.g., education settings, during training programs)? Although the present study provides evidence supporting the memory explanation, understanding the effects of repeated testing on future test performance within its context and purpose offers greater utility to practitioners, rather than assuming that testing inherently facilitates *either* learning or memory effects. Nevertheless, it is important to note that the present study did not examine retest increases in the context of criterion-related validity relationships; thus, future research using academic and employment settings and criteria are likely to add additional insights.

**8.1.1 Scientific Implications.** The retesting effect is a robust empirical finding observed since the earliest studies in psychology and has also been observed across construct domains, academic and applied settings, and lab and operational settings. Yet much less is known about the benefits of testing on the application and transfer of this performance increase and the cognitive processes underlying this effect (Bjork & Bjork, 1992; Carpenter, 2012). Specifically, the major scientific implication of these results are to answer the call to investigate whether tests can promote the transfer of learned constructs across time, test forms, and construct domains or if these effects are bound to test-specific information (Carpenter, 2012). Greater understanding as to whether the retesting effect is best explained by memory rather than learning has implications for both the measurement of psychological constructs and the possibility that tests assess as well as teach (Carpenter, 2012; Dunlosky et al., 2013; Pashler et al., 2007; Roediger & Karpicke, 2006a). The extant literature makes it apparent that testing influences retest score gains and offers insight into the cognitive processes underlying memory (Bjork &

Bjork, 1992; Carpenter, 2012). However, on the basis of the present study's results, memory appears to better account for retest increases which suggests that retesting is unlikely to be viable learning intervention.

Research consistently supports the spreading activation theory of retesting, whereby the cognitive processes of learning and memory comprise both an encoding and retrieval process. That is, individuals first make connections between cognitions (encoding) but then must access these previously encoded connections and produce a response (retrieval; Bjork & Bjork, 1992; Karpicke & Blunt, 2011). These processes are not necessarily distinct or directly observable when retest increases occur; thus, it is difficult to separate whether retest increases facilitate encoding or retrieval as both are necessary to produce a response. Furthermore, spreading activation theory appears more applicable as an explanation for the retesting effect on knowledge tests, yet this is less conceptually transparent in the context of the robust and consistent effect of retesting on ability tests.

The present study showed no differences in the pattern or magnitude of retest increases across the knowledge and ability construct domains. That is, spreading activation theory would posit different processes for the retesting effect on knowledge versus ability constructs (which were not observed), and if spreading activation reflected encoding new information rather than merely retrieving previously encoded information, then there should be no differences between same versus alternate retest forms (which were observed). Thus, spreading activation theory's relevance to the retesting effect's

benefit appears bound to facilitating the retrieval of test-specific information, rather than developing the underlying construct.

The elucidation of these causal mechanisms is not trivial. The retesting effect shapes the inferences drawn from retest scores and has been used to justify the use of testing not only for evaluation, but as a learning intervention. Ultimately, the use and interpretation of repeated testing shapes the scientific community's understanding of how and why knowledge and abilities change over time. The present study was designed such that memory versus learning explanations were competitively examined as mutually exclusive processes for the retesting effect. However, the retesting effect may reflect numerous and complex underlying processes (Reeve & Lam, 2005). In accordance with this proposition, the retesting effect has been shown to not only increase mean retest scores, but also reduce the construct loading of a retest or even fundamentally alter which construct is assessed (Lievens et al., 2007). Thus, it seems possible that both memory *and* learning effects occur upon retesting.

Within the present research design, the data indicate that it is plausible that both learning and memory effects occurred, but that comparatively, the effects for memory were *stronger* than those for learning. Even when memory effects were experimentally controlled using alternate retest forms, retest increases still occurred. Similarly, retest mean item response time decreased on alternate forms, even when memory effects were experimentally controlled. Furthermore, one may posit that the present study's inconclusive results in terms of the relationship between individual differences and retest increases may be indicative of a learning explanation. As detailed previously,

individuals with greater GMA benefit more from learning interventions (Ackerman, 1987, 1988; Jensen, 1998; Schmidt & Hunter, 1998). Previous research has reported somewhat contradictory evidence for the relationship between GMA and retest increases. Test-takers with greater abilities sometimes achieve greater retest gains (e.g., Kulik et al., 1984; Kulik et al., 1984), yet other researchers have found that retest increases are unrelated to GMA (e.g., Coyle, 2006) or even that the *lowest* GMA test-takers exhibit the greatest retest increases (te Nijenhuis et al., 2007). Although Hypothesis 4A was supported, in that working memory capacity exhibited a stronger relationship with retest score increases for same retest forms, there was a main effect of GMA on the magnitude of retest score increases.

The objective of the present study was to examine which explanation, memory versus learning, was *best* supported, not necessarily to compare the magnitude of these explanations. Thus, an alternate interpretation of the present study's results may be that, after memory effects were experimentally controlled, *any* remaining retest increases reflected some degree of learning and test-takers with greater GMA benefited more from this learning intervention. Considering this possibility, future research may lend additional insights into this particular issue by (1) using more precise measures to detect these evidently smaller effects, (2) employing more robust retesting interventions (e.g., multiple retests, more elaborate feedback), and (3) examining the test-taking behaviors that are more likely related to learning versus memory explanations (e.g., response-changing, meta-cognition).

**8.1.2 Practical Implications.** Despite a lack of integrated or conclusive research regarding explanations for the retesting effect, retesting and retesting policies exert a profound impact on individuals and organizations across applied, educational, and policy domains by affecting the inferences that test users draw from test scores and score changes. The interpretation of the retesting effect as best explained by memory or learning affects the inferences drawn from test scores, and therefore, affects both retesting policy for assessment and the potential use of testing as a learning intervention. Based on the present evidence, the retesting effect appears best explained by memory, rather than learning, suggesting two primary implications for practitioners: (1) drawing conclusions based purely on retest scores likely leads to poorer and less-informed decisions and (2) implementing retesting for learning purposes is unlikely to meaningfully generalize to external learning, training, and performance criteria.

In the workplace, tests of knowledge and ability are used to make a multitude of human resource management decisions. In educational settings, tests are used to decide who receives education and to determine who has been successful in that education. In public settings, tests are used to decide who will be permitted to provide healthcare, manage finances, and make legal decisions. Incorrectly interpreting the influence of the retesting effect on test scores risks significant consequences to both individuals and society by potentially rewarding (denying) credentials, jobs, opportunities, and promotions to (un)qualified test-takers (Millman, 1989). Nevertheless, it is important to note that although the results did not support testing as a learning intervention, this does not indicate that a retest can *never* show evidence that learning has occurred. Instead,

interpreting retest scores requires careful consideration of other relevant learning mechanisms that may have occurred between an initial test and retest (e.g., additional study, training, education, and experiences).

Researchers present an extensive literature supporting the position that testing functions as a learning intervention for even complex concepts across educational materials (McDaniel et al., 2007; Pashler et al., 2007; Roediger & Karpicke, 2006a). These proponents also point out the *indirect* benefits of regular testing, such as encouraging self-evaluation and consequently, more focused learning (e.g., dynamic testing, formative assessment, Roediger and Karpicke 2006a). Indeed, many studies have shown that integrating regular testing into educational curricula shows beneficial results, regardless of the retesting effect (Pennebaker et al., 2013). So whereas the present study failed to provide support for the learning explanation and the viability of retesting as a learning intervention, these results do not preclude the potential benefit of regular testing as a complement to more familiar teaching practices.

Common teaching methods (e.g., additional study time, concept mapping) focus heavily on elaborative study which relies on memory encoding and strengthening of links between ideas, rather than practicing the retrieval of previously learned information. However, in terms of effective retrieval, many of these learning activities have been shown to be less effective than testing alone (Karpicke & Blunt, 2011). This distinction between encoding and retrieval mirrors the memory and learning explanations for the retesting effect. Perhaps due to this lack of translation across research domains and multiple pathways of causality, the evidence that testing is

beneficial through retrieval practice is often overlooked by education and personnel practitioners and therefore, rarely incorporated into educational practice and policy (Pashler et al., 2007). Considering the present study's results, it is difficult to justify implementing testing as a learning intervention per se; however, this does not preclude indirect benefits of retesting in learning settings.

## **8.2 Limitations**

Two potential limitations associated with the design of the present study are acknowledged. Specifically, the nature of a low-stakes test setting and multiple-choice item format may have attenuated the magnitude of these retest increases; yet these design characteristics seem unlikely to have affected the fundamental pattern of these results.

The study's design was selected to exploit the experimental control of a laboratory setting and capitalize on an initial level of content domain knowledge (general psychology) from a population that was readily available. This made it possible to examine the research questions, but was, inherently, low-stakes. Compared to high-stakes tests used to make decisions about an individual's academic standing or employment, it is possible that the relatively low-stakes nature of these tests' consequences impacted the magnitude of retest increases. Considering the present study's recruiting methods, it also possible that a particular subset of individuals systematically self-selected into the study (greater comfort with tests, desire to learn more). Nevertheless, participants reported mid to high test motivation ( $M = 3.84$ ,  $SD = 0.35$ , at Time 1;  $M = 3.77$ ,  $SD = 0.36$ , at Time 2) and low to mid test anxiety ( $M = 2.84$ ,

$SD = 0.56$ , at Time 1;  $M = 2.85$ ,  $SD = 0.52$ , at Time 2) and this did not affect the magnitude of retest score gains. Extensive previous research shows that the retesting effect generalizes across diverse populations of learners as well as laboratory, occupational, and educational contexts. Nevertheless, future research may offer additional insights by investigating learning versus memory effects using a similar design and measures in high-stakes testing environments such as operational education or employment settings.

Research investigating the retesting effect consistently demonstrates that using multiple-choice items on initial tests exhibit smaller retest scores gains compared to constructed response item formats (e.g., Carpenter & DeLosh, 2006; Foos & Fisher, 1988; Hamaker, 1986; McDaniel et al., 2007). Two primary explanations for these smaller retest increases have been offered: (1) multiple-choice items present the keyed response which allows test-takers to simply recognize the keyed response upon retest, or that (2) multiple-choice items present the keyed responses among alternatives which provides additional cues for test-takers to recognize the correct response (Roediger & Marsh, 2005). Regardless of the explanation, this suggests that using multiple-choice items may have reduced the overall magnitude of this study to detect evidence of either learning or memory effects.

Multiple-choice tests do, however, still facilitate retest score gains greater than simply restudying the same test content (McDaniel et al., 2007) and in some cases, actually generate greater retest gains than other item formats. For example, when initial tests contain plausible nonkeyed alternatives, test-takers must retrieve why the correct

alternatives were correct and why the incorrect alternatives were incorrect, thus developing more effective retrieval processes (Little et al., 2012). Item format clearly moderates the magnitude of the retesting effect; however, it seems unlikely that item format alters the underlying causal mechanism of the retesting effect. It is possible that test-takers could learn information, from either the same or alternate form, that allows them to eliminate multiple-choice alternative options in retest items. Nevertheless, this could reflect either learning versus memory effects, and therefore alternate item formats offer researchers an additional method to control for the effect of merely recognizing alternatives as cues. Relatedly, items that assess more complex knowledge domains compared to simpler memory of individual facts could be used to further explore learning versus memory effects. Does the magnitude of retest increases differ for declarative knowledge measures versus procedural knowledge measures? Are these retest increases more or less attenuated by alternate test forms?

### **8.3 Future Directions**

The results of the present study offer three avenues for future research: (1) negative transfer and response-changing upon retesting, (2) the effect of meta-cognition on retesting, and (3) investigating other theoretical explanations for facilitating learning using retesting (e.g., working memory training).

**8.3.1 Negative Transfer.** It is curious to note the relatively low test-retest reliability coefficients between initial tests and retests on alternate forms (e.g., as low as .58 and .64 for alternate ability test forms, compared to .76 and .70 for same forms). However, even same form knowledge retests showed a low test-retest reliability of only

.60, below commonly accepted standards for operational tests. Although retest *decreases* were not the focus of the present study, these relatively low test-retest reliabilities imply that retest scores fluctuated both up and down over time, regardless of same versus alternate form tests and construct domain (knowledge versus ability). Clearly, some participants exhibited negative transfer.

Table 11

*Frequency Distribution of Mock Pass/Fail Retest Scores*

	Time 1	Time 2	
		Fail	Pass
Fail	164	68	96
Pass	176	6	170

*Note.* N = 340 participants.

For illustrative purposes, the distribution of mock pass/fail scores was computed using the standardized, unit-weighted composite of initial knowledge and ability test scores (see Table 11). Participants who performed above the mean on the composite test score percentage ( $M = 54.15\%$ ) on the initial tests “passed” and those who performed below the mean “failed”. These initial test score means were then used to compute the retest score passes and fails; participants who performed above the mean initial test score on the retest “passed” and those who performed below the mean initial test score on the retest “failed”. Negative transfer occurs when those who initially passed subsequently fail upon retest. This appears relatively rare: only 6 participants who initially passed failed upon retest (1.76% of the total sample of 340 test-takers, 3.41% of the initially

passing test-takers). In contrast, 96 participants who initially failed passed upon retest (28.24% of the total sample of 340 test-takers, 58.54% of the initially failing test-takers).

Policies allowing retesting often assume that retest scores are given more weight as the final operational test score, otherwise retesting would not be considered at all. Considering the present study's results supporting the effects of memory, rather than learning, and the uneven distribution between test-takers who initially passed then failed compared to the distribution of test-takers who initially failed then passed, it appears quite likely that allowing retesting risks a greater ratio of false positives at the cost of relatively few false negatives. It is also interesting to note that these pass/fail rates do not appear substantially different between same versus alternate retest forms (see Table 12). Implementing alternate retest forms may eliminate the potential for memory effects and serve the purposes of the present study's research questions, yet be of a relatively small magnitude to exert an effect on operational decision-making.

Table 12

*Frequency Distribution of Mock Pass/Fail Retest Scores by Same versus Alternate Retest Form*

		Time 2			
Time 1		Fail		Pass	
		Same Retest Form	Alternate Retest Form	Same Retest Form	Alternate Retest Form
Fail	164	33	35	51	45
Pass	176	2	4	83	87

*Note.* N = 340 participants.

Negative transfer occurs at both the test- and item-level. Systematic retest increases (or even decreases) necessitate systematic response-changing between test administrations. Previous research shows that test scores increase when applicants change both incorrect and correct responses between administrations (Bors & Vigneau, 2003). Test-takers who respond or even guess incorrectly do not necessarily commit to the same wrong response on retest (Kang et al., 2011). For example, Bors and Vigneau (2002) found that although an overall test may exhibit acceptable reliability between administrations, it may nevertheless show low intra-item reliability. Relatedly, te Nijenhuis et al. (2007) found that items with greater *g*-saturation exhibited lower endorsement rates upon retest, whereas items with lower *g*-saturation show greater endorsement rates. This pattern of results implies that item-level response-changing may offer an additional avenue for competitively examining memory versus learning explanations at the item-level.

Consistently responding to the same item across initial test and retest administrations reflects a degree of confidence in that response, potentially indicating confidence in one's memory of the previous response or confidence in one's understanding of the item (i.e., learning). A test-taker may change responses from wrong to right (indicating learning) or right to wrong (indicating negative transfer). Thus, a taxonomy of response-changing may be constructed, whereby a pattern of results may best illustrate whether the retesting effect best reflects learning or memory at the item-level. Examining only total test scores blurs this distinction, assuming that test-takers merely repeat previous responses exactly and answer additional items correctly.

Response-changing offers another metric for investigating whether retest increases are more reflective of either learning or memory effects, the boundary conditions of retesting that may instead lead to negative transfer, and the particular behaviors underlying retest increases. As noted previously, memory effects may simply be stronger or more prevalent than learning, rather than mutually exclusive. It appears plausible that item-level response-changing offers a more finely-tuned method for examining that retesting acts as a learning intervention. Although the present study inevitably collected item-level response data, the purpose of the design was to investigate whether memory versus learning better explained the retesting effect overall, not compare the magnitude of memory *and* learning simultaneously. Thus, speaking to item-level issues and response-changing is beyond the scope of this study.

**8.3.2 Meta-Cognition.** Although substantial prior research has investigated meta-cognition through test-takers' self-reported confidence in responses over time, examining item-level response-changing offers a viable behavioral measure for assessing the underlying cognitive processes of the retesting effect. Testing reduces the otherwise ubiquitous tendency for test-takers to experience overconfidence in their own learning (e.g., Carpenter & Olson, 2012), leading some researchers to hypothesize that testing might improve meta-cognitive awareness or encourage the adoption of more effective encoding strategies (Carpenter, 2012; Pyc & Rawson, 2010).

The present study investigated two individual differences, GMA and working memory capacity, as potential explanatory variables for learning versus memory effects, respectively. Although these individual differences covary with meta-cognition, it

appears plausible that meta-cognitive behaviors and strategies are more directly related to retest increases than individual differences, and worth investigating further. Research shows that corrective feedback facilitates the retesting effect (Butler et al., 2007; Kang et al., 2007) and that participants are aware of the potentially beneficial effects of testing (Jacoby et al., 2010). Furthermore, particular meta-cognitive behaviors are likely to be more or less relevant to either the memory or learning explanation. For example, a behavioral goal of remembering previous responses is more likely to result in memory effects, whereas a behavioral goal of self-diagnosing comprehension is more likely to result in learning effects. Multiple methods with a wealth of prior research, such as meta-cognitive awareness inventories and think aloud protocols, do exist that would allow researchers to investigate test-taking behaviors, response changing, and response confidence. Based on previous research and the present data, examining meta-cognitive behaviors may offer additional evidence as to whether retest increases reflect learning versus memory explanations, and if so, provide a more direct, behavioral mechanism for how these increases occur.

**8.3.3 Working Memory Training.** Some research posits that working memory training programs, fundamentally a retesting intervention, may be used to improve one's standing on both ability and knowledge constructs. A body of work by Jaeggi, Buschkuhl, and colleagues have examined the influence of working memory training on both ability (e.g., fluid intelligence, Jaeggi, Studer-Luethi, Buschkuhl, Su, Jonides, & Perrig, 2010) and knowledge constructs (e.g., reading performance, Loosli, Buschkuhl, Perrig, & Jaeggi, 2012). Based on this work, it appears that working

memory capacity may be improved through retesting, which mediates the learning of ability and knowledge constructs.

Previous meta-analytic evidence indicates that these programs can produce short-term improvements in working memory skills, but these improvements are not consistent, appear test-specific, and do not transfer to improve the external criteria (Melby-Lervåg & Hulme, 2013). Thus, working memory training may be yet another example of the retesting effect as best explained by memory, rather than learning. Perhaps relatedly, the present results indicated that working memory capacity exhibited small to moderate intercorrelations with the initial and retest scores for both knowledge and ability tests (see Table 9 for intercorrelations), but were poor predictors of the magnitude of actual retest increases (see Figure 6).

#### **8.4 Summary and Conclusions**

The retesting effect influences fundamental issues of measurement across the psychological, educational, and management literatures, and therefore, concerns the operational use of tests as well. Greater insight into the retesting effect directly affects research questions that are foundational to psychology, including learning, retrieval processes, and the valid interpretation of psychological measurements over time. The present study's results are more in alignment with the memory than a learning explanation for the retesting effect. Retest increases were attenuated by alternate retest forms (when memory effects were not possible), participants responded to retest items faster but this was attenuated by alternate test forms, and there were no differences between the magnitude of these retest increases regardless of the construct domain

(ability versus knowledge). Researchers have proposed that retesting offers a viable, even underused technique for facilitating learning, yet the present study demonstrated that the benefits of retesting do not appear to meaningfully reflect learning. Although no study is perfect, the limitations presented for the current study (e.g., choice of setting, item format) appear more likely to influence the *magnitude* of retest increases, rather than the pattern of results as reflecting either memory or learning. Future research investigating the learning versus memory explanations for the retesting effect could lend additional insights by examining factors affecting negative retesting transfer (at both the test- and item-levels), additional explanatory variables for memory versus learning explanations for the retesting effect (e.g., meta-cognitive skills), and retesting in the context of working memory training interventions.

## REFERENCES

- Abbott, E. E. (1909). On the analysis of the factors of recall in the learning process. *Psychological Monographs, 11*, 159–177.
- Ackerman, P. L. (1987). Individual differences in skill learning: An integration of psychometric and information processing perspectives. *Psychological Bulletin, 102*, 3–27.
- Ackerman, P. L. (1988). Determinants of individual differences during skill acquisition: Cognitive abilities and information processing. *Journal of Experimental Psychology: General, 117*, 288–318.
- Agarwal, P. K., & Roediger, H. L., III. (2011). Expectancy of an open-book test decreases performance on a delayed closed-book test. *Memory, 19*, 836–852.
- Allalouf, A., & Ben-Shakhar, G. (1998). The effect of coaching on the predictive validity of scholastic aptitude tests. *Journal of Educational Measurement, 35*, 31–47.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA, APA, & NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Anastasi, A. (1981). Coaching, test sophistication, and developed abilities. *American Psychologist, 36*, 1086–1093.
- Anastasi, A. & Urbina, S. (1997). *Psychological testing* (7<sup>th</sup> ed.). Upper Saddle River, NJ: Prentice-Hall.

- Anderson, J. R. (1987). Skill acquisition: Compilation of weak-method problem solutions. *Psychological Review*, 94, 192–210.
- Anderson, R. C., Kulhavy, R. W., & Andre, T. (1971). Feedback procedures in programmed instruction. *Journal of Educational Psychology*, 62, 148–156.
- Anderson, R. C., Kulhavy, R. W., & Andre, T. (1972). Conditions under which feedback facilitates learning from programmed lessons. *Journal of Educational Psychology*, 63, 186–188.
- Arthur, W., Jr. (2004). *Report on the development of an internet-administered general mental ability test (with verbal and numeric sub-scales)*. College Station, TX: Winfred Arthur, Jr. Consulting.
- Arthur, W., Jr. (2005). *ATMA test-retest reliability study*. College Station, TX: Winfred Arthur, Jr. Consulting.
- Arthur, W., Jr., & Day, D. V. (1994). Development of a short form for the Raven Advanced Progressive Matrices Test. *Educational and Psychological Measurement*, 54, 394–403.
- Arthur, W., Jr., Day, D. V., Bennett, W., Jr., & Portry, A. (Eds.) (2013). *Individual and team skill decay: The science and implications for practice*. New York: Taylor Francis/Psychology Press.
- Arthur, W., Jr., Glaze, R. M., Villado, A. J., & Taylor, J. E. (2010). The magnitude and extent of cheating and response distortion effects on unproctored internet-based tests of cognitive ability and personality. *International Journal of Selection and Assessment*, 18, 1–16.

- Arthur, W., Jr., Tubré, T., Paul, D. S., & Edens, P. S. (2003). Teaching effectiveness: The relationship between reaction and learning evaluation criteria. *Educational Psychology*, 23, 275–285.
- Arthur, W., Jr., Tubré, T. C., Paul, D. S., & Sanchez-Ku, M. L. (1999). College-sample psychometric and normative data on a short form of the Raven Advanced Progressive Matrices test. *Journal of Psychoeducational Assessment*, 17, 354–361.
- Arvey, R.D., & Sackett, P.R. (1993). Fairness in selection: Current developments and perspectives. In N. Schmitt & W. Borman (Eds.), *Personnel selection in organizations*, pp. 171–202. San Francisco: Jossey-Bass.
- Arvey, R. D., Strickland, W., Drauden, G., & Martin, C. (1990). Motivational components of test taking. *Personnel Psychology*, 43, 695–716.
- Baddeley, A. (1992). Working memory. *Science*, 255, 556–559.
- Balch, W. R. (1998). Practice versus review exams and final exam performance. *Teaching of Psychology*, 25, 181–185.
- Bangert-Drowns, R. L., Kulik, C. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research*, 61, 213–238.
- Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? *Psychological Bulletin*, 128, 612–637.

- Basso, M. R., Carona, F. D., Lowery, N., & Axelrod, B. N. (2002). Practice effects on the WAIS-III across 3- and 6-month intervals. *The Clinical Neuropsychologist*, 16, 57–63.
- Betz, N. E., & Weiss, D. J. (1976a). *Effects of immediate knowledge of results and adaptive testing on ability test performance (Research Report 76-3)*. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, June 1976. (NTIS No. AD A027147) (a).
- Betz, N. E., & Weiss, D. J. (1976b). *Psychological effects of immediate knowledge of results and adaptive testing (Research Report 76-4)*. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, June 1976. (NTIS No. AD A027170) (b).
- Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes*, (Vol. 2, pp. 35–67). Hillsdale, NJ: Erlbaum.
- Bobko, P. (2001). *Correlation and regression: Applications for industrial and organizational psychology and management* (2<sup>nd</sup> ed.). Thousand Oaks, CA: Sage.
- Brounstein, P. J., & Holahan, W. (1987). Patterns of change in scholastic aptitude test performance among academically talented adolescents. *Roepers Review*, 10, 110–116.
- Burke, E. F. (1997). A short note on the persistence of retest effects on aptitude scores. *Journal of Occupational and Organizational Psychology*, 70, 295–301.

- Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 36, 1118–1133.
- Butler, A. C., Karpicke, J. D., & Roediger, H. L., III. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied*, 13, 273–281.
- Butler, A. C., Karpicke, J. D., & Roediger, H. L., III. (2008). Correcting a metacognitive error: Feedback increases retention of low-confidence correct responses. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 34, 918–928.
- Butler, A. C., & Roediger, H. L., III (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology*, 19, 514–527.
- Butler, A. C., & Roediger, H. L., III. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple choice testing. *Memory & Cognition*, 36, 604–616.
- Butterfield, B., & Metcalfe, J. (2001). Errors committed with high confidence are hypercorrected. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 27, 1491–1494.
- Campbell, R. (1969). *Seneca, Letters from a Stoic*. Baltimore, MD: Penguin Books, Ltd.
- Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices test. *Psychological Review*, 97, 404–431.

- Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 35, 1563–1569.
- Carpenter, S. K. (2011). Semantic information activated during retrieval contributes to later retention: Support for the mediator effectiveness hypothesis of the testing effect. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 37, 1547–1552.
- Carpenter, S. K. (2012). Testing enhances the transfer of learning. *Current Directions in Psychological Science*, 21, 279–283.
- Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition*, 34, 268–276.
- Carpenter, S. K., & Olson, K. M. (2012). Are pictures good for learning new vocabulary in a foreign language? Only if you think they are not. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 38, 92–101.
- Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of U.S. history facts. *Applied Cognitive Psychology*, 23, 760–771.
- Carretta, T. R., Zelenski, W. E., & Ree, M. J. (2000). Basic Attributes Test (BAT) retest performance. *Military Psychology*, 12, 221–232.
- Carroll, J. B. (1993). *Human cognitive abilities*. Cambridge: Cambridge University Press.

- Carroll, J. M., & Kay, D. S. (1988). Prompting, feedback and error correction in the design of a scenario machine. *International Journal of Man-Machine Studies*, 28, 11–27.
- Case, R. D., Kurland, D. M., & Goldberg, J. (1982). Operational efficiency and the growth of short-term memory span. *Journal of Experimental Child Psychology*, 33, 386–404.
- Cassady, J. C., & Johnson, R. E. (2002). Cognitive test anxiety and academic performance. *Contemporary Educational Psychology*, 27, 270–295.
- Chamorro-Premuzic, T., & Furnham, A. (2010). *The psychology of personnel selection*. New York: Cambridge University Press.
- Chan, J. C. K. (2009). When does retrieval induce forgetting and when does it induce facilitation? Implications for retrieval inhibition, testing effect, and text processing. *Journal of Memory and Language*, 61, 153–170.
- Chan, J. C. K. (2010). Long-term effects of testing on the recall of nontested materials. *Memory*, 18, 49–57.
- Chan, J. C. K., McDermott, K. B., & Roediger, H. L. III. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General*, 135, 553–571.
- Chan, D., Schmitt, N., DeShon, R. P., Clause, C., & Delbridge, K. (1997). Reactions to cognitive ability tests: The relationships between race, test performance, face validity perceptions, and test-taking motivation. *Journal of Applied Psychology*, 82, 300–310.

- Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development*, 70, 1098–1120.
- Cliffordson, C. (2004). Effects of practice and intellectual growth on performance on the Swedish Scholastic Aptitude test (SweSAT). *European Journal of Psychological Assessment*, 20, 192–204.
- Cole, N. (1982). The implication of coaching for ability testing. In A. K. Wigdor & W. R. Garner (Eds.), *Ability testing: Uses, consequences, and controversies (part ii)*. Washington, DC: National Academy Press.
- Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, 82, 407–428.
- Conway, A. R. A., Cowan, N., & Bunting, M. F. (2001). The cocktail party phenomenon revisited: The importance of working memory capacity. *Psychonomic Bulletin and Review*, 8, 331–335.
- Coyle, T. R. (2006). Test-retest changes on scholastic aptitude tests are not related to  $g$ . *Intelligence*, 34, 15–27.
- Cranney, J., Ahn, M., McKinnon, R., Morris, S., & Watts, K. (2009). The testing effect, collaborative learning, and retrieval-induced facilitation in a classroom setting. *European Journal of Cognitive Psychology*, 21, 919–940.
- Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, 19, 450–466.
- Daniel, D. B., & Broida, J. (2004). Using web-based quizzing to improve exam performance: Lessons learned. *Teaching of Psychology*, 31, 207–208.

- Deary, I. J., Pattie, A., & Starr, J. M. (2013). The stability of intelligence from age 11 to age 90 years: The Lothian birth cohort of 1921. *Psychological Science*, 24, 2361–2368.
- Dixon, R., Kramer, D., & Baltes, P. (1985). Intelligence: A life-span developmental perspective. In B. Wolman (Ed.), *Handbook of intelligence: Theories, measurements and applications* (pp. 301–350). New-York: Wiley.
- Donlon, T. F., & Angoff, W. H. (1971). The Scholastic Aptitude Test. In W. H. Angoff (Ed.), *The College Board admissions testing program: A technical report on research and development activities relating to the Scholastic Aptitude Test and Achievement Test*. New York: College Entrance Examination Board.
- Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Sciences in the Public Interest*, 14, 4–58.
- Dye, D. A., Reck, M., & McDaniel, M. A. (1993). The validity of job knowledge measures. *International Journal of Selection and Assessment*, 1, 153–157.
- Edwards, B. D., Arthur, W., Jr., & Bruce, L. L. (2012). The three-option format for knowledge and ability multiple-choice tests: A case for why it should be more commonly used in personnel testing. *International Journal of Selection and Assessment*, 20, 65–81.
- Embretson, S. E. (1987). Improving the measurement of spatial aptitude by dynamic testing. *Intelligence*, 11, 333–358.

- Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science*, 11, 19–23.
- Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999). Working memory, short-term memory, and general fluid intelligence: A latent variable approach. *Journal of Experimental Psychology: General*, 128, 309–331.
- Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers*, 28, 1–11.
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G\*Power 3.1: Tests for correlation and regression analyses. *Behavioral Research Methods*, 41, 1149–1160.
- Fehrman, M. L., Woehr, D. J., & Arthur, W., Jr. (1991). The Angoff cutoff score method: The impact of frame-of-reference rater training. *Educational and Psychological Measurement*, 51, 857–872.
- Foos, P. W., & Fisher, R. P. (1988). Using tests as learning opportunities. *Journal of Educational Psychology*, 80, 179–183.
- Frese, M., & Zapf, D. (1994). Action as the core of work psychology: A German approach. In H. C. Triandis, M. D. Dunnette, & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2<sup>nd</sup> ed., Vol. 4, pp. 271 –340). Palo Alto, CA: Consulting Psychologists Press.
- Friedman, H. (1987). Repeat examinations in introductory statistics courses. *Teaching of Psychology*, 14, 20–23.

- Fritz, C. O., Morris, P. E., Acton, M., Voelkel, A. R., & Etkind, R. (2007). Comparing and combining retrieval practice and the keyword mnemonic for foreign vocabulary learning. *Applied Cognitive Psychology*, 21, 499–526.
- Gates, A. I. (1917). Recitation as a factor in memorizing. *Archives of Psychology*, 6, 104.
- Geffen, A. H. (2004). The concurrent validity and test-retest reliability of a visuospatial working memory task. *Intelligence*, 32, 591–605.
- Geving, A. M., Webb S., & Davis B. (2005). Opportunities for repeat testing: Practice doesn't always make perfect. *Applied Human Resource Management Research*, 10, 47–55.
- Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology*, 12, 306–355.
- Gilliland, S. W. (1993). The perceived fairness of selection systems: An organizational justice perspective. *Academy of Management Review*, 18, 694–734.
- Gilliland, S. W., & Steiner, D. D. (2001). Causes and consequences of applicant fairness. In R. Cropanzano (Ed.), *Justice in the workplace* (Vol. 2, pp. 175–195). Mahwah, NJ: Erlbaum.
- Glass, A. L. & Sinha, N. (2013). Multiple-choice questioning is an efficient instructional methodology that may be widely implemented in academic courses to improve exam performance. *Current Directions in Psychological Science*, 22, 471–477
- Gottfredson, L. S. (1986). Societal consequences of the *g* factor in employment. *Journal of Vocational Behavior*, 29, 379–410.

- Gottfredson, L. S. (1997) Why *g* matters: The complexity of everyday life. *Intelligence*, 24, 79–132.
- Greene E. B. (1941). *Measurement of human behavior*. New York, NY: Odyssey Press.
- Hamaker, C. (1986). The effects of adjunct questions on prose learning. *Review of Educational Research*, 56, 212–242.
- Hancock, T. E., Stock, W. A., & Kulhavy, R. W. (1992). Predicting feedback effects from response certitude estimates. *Bulletin of the Psychonomic Society*, 30, 173–176.
- Harcourt Assessment, Inc. (2013). *Miller Analogies Test: Scoring and score reporting*. Retrieved January 1<sup>st</sup>, 2013, from <http://www.milleranalogies.com>
- Hausknecht, J. P., Day, D. V., & Thomas, S. C. (2004). Applicant reactions to selection procedures: An updated model and meta-analysis. *Personnel Psychology*, 57, 639–683.
- Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Moriarty Gerrard, M. O. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology*, 92, 373–385.
- Hausknecht, J. P., Trevor, C. O., & Farr, J. L. (2002). Retaking ability tests in a selection setting: Implications for practice effects, training performance, and turnover. *Journal of Applied Psychology*, 87, 243–254.
- Hunt, R. R. (2006). The concept of distinctiveness in memory research. In R. R. Hunt & J. B. Worthen (Eds.), *Distinctiveness and memory* (pp. 3–25). New York, NY: Oxford University Press.

- Hunter, J. E., Schmidt, F. L., & Le, H. (2006). Implications for direct and indirect range restriction for meta-analysis methods and findings. *Journal of Applied Psychology*, 91, 594–612.
- Jacoby, L. L., Wahlheim, C. N., & Coane, J. H. (2010). Test enhanced learning of natural concepts: Effects on recognition memory, classification, and metacognition. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 36, 1441–1451.
- Jaeggi, S. M., Studer-Luethi, B., Buschkuhl, M., Su, Y., Jonides, J., & Perrig, W. J. (2010). The relationship between n-back performance and matrix reasoning—Implications for training and transfer. *Intelligence*, 38, 625–635.
- James, W. (1890). *Principles of psychology* (Vol. 1). New York, NY: Dover.
- Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT: Praeger.
- Johnson, C. I., & Mayer, R. E. (2009). A testing effect with multimedia learning. *Journal of Educational Psychology*, 101, 621–629.
- Kane, M. J., Bleckley, M. K., Conway, A. R. A., & Engle, R. W. (2001). A controlled-attention view of working memory capacity: Individual differences in memory span and the control of visual orienting. *Journal of Experimental Psychology: General*, 130, 169–183.
- Kane, M. J., & Engle, R. W. (2003). Working memory capacity and the control of attention: The contributions of goal neglect, response competition, and task set to Stroop interference. *Journal of Experimental Psychology: General*, 132, 47–70.

- Kang, S. H. K., McDaniel, M. A., & Pashler, H. (2011). Effects of testing on learning of functions. *Psychonomic Bulletin & Review*, 18, 998–1005.
- Kang, S. H. K., McDermott, K. B., & Roediger, H. L., III. (2007). Test format and corrective feedback modulate the effect of testing on memory retention. *European Journal of Cognitive Psychology*, 19, 528–558.
- Kang, S. H. K., Pashler, H., Cepeda, N. J., Rohrer, D., Carpenter, S. K. & Mozer M. C. (2011). Does incorrect guessing impair fact learning? *Journal of Educational Psychology*, 103, 48–59.
- Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, 331, 772–775.
- Karpicke, J. D., & Roediger, H. L., III. (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 33, 704–719.
- Karpicke, J. D., & Roediger, H. L., III. (2008). The critical importance of retrieval for learning. *Science*, 319, 966.
- Karpicke, J. D., & Roediger, H. L., III. (2010). Is expanding retrieval a superior method for learning text materials? *Memory & Cognition*, 38, 116–124.
- Kaufman A. L. (1994). Practice effects. In R. Sternberg (Ed.), *Encyclopedia of human intelligence* (vol. 2, pp. 828–833). New York, NY: Macmillan.
- Kingston, N., & Turner, N. (1984). *Analysis of score change patterns of examinees repeating the Graduate Record Examinations General Test*. GRE Board

Professional Report GREB No. 83-5P Princeton, NJ: Educational Testing Service.

- Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*, 254–284.
- Koenig, J. A., & Leger, K. F. (1997). Test-taking behaviors and their impact on performance. *Academic Medicine, 72*, S100–S103.
- Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language, 65*, 85–97.
- Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 35*, 989–998.
- Kromann, C. B., Jensen, M. L., & Ringsted, C. (2009). The effects of testing on skills learning. *Medical Education, 43*, 21–27.
- Kulhavy, R. W. (1977). Feedback in written instruction. *Review of Educational Research, 47*, 211–232.
- Kulhavy, R. W., & Stock, W. A. (1989). Feedback in written instruction: The place of response certitude. *Educational Psychology Review, 1*, 279–308.
- 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.

- Larson, G. E., & Alderton, D. L. (1997). Test-retest results for the ECAT battery. *Military Psychology*, 9, 39–47.
- 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.
- 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.
- Little, J. L., Bjork, E. L., Bjork, R. A., & Angello, G. (2012). Multiple-choice tests exonerated, at least of some charges: Fostering test-induced learning and avoiding test-induced forgetting. *Psychological Science*, 23, 1337–1344.
- Logan, J. M., & Balota, D. A. (2008). Expanded vs. equal interval spaced retrieval practice: Exploring different schedules of spacing and retention interval in younger and older adults. *Aging, Neuropsychology, and Cognition*, 15, 257–280.
- Loosli, S. V., Buschkuhl, M., Perrig, W. J., & Jaeggi, S. M. (2012). Working memory training improves reading processes in typically developing children. *Child Neuropsychology*, 18, 62–78.
- Lubinski, D. (2000). Scientific and social significance of assessing individual differences: Sinking shafts at a few critical points. *Annual Review of Psychology*, 51, 405–444.

- Lyle, K. B., & Crawford, N. A. (2011). Retrieving essential material at the end of lectures improves performance on statistics exams. *Teaching of Psychology*, 38, 94–97.
- Marsh, E. J., Agarwal, P. K., & Roediger, H. L., III. (2009). Memorial consequences of answering SAT II questions. *Journal of Experimental Psychology: Applied*, 15, 1–11.
- Matton, N., Vautier, S., & Raufaste, E. (2009). Situational effects may account for gain scores in cognitive ability testing: A longitudinal SEM approach. *Intelligence*, 37, 412–421.
- Matton, N., Vautier, S., & Raufaste, E. (2011). Test-specificity of the advantage of retaking cognitive ability tests. *International Journal of Selection and Assessment*, 19, 11–17.
- Maurer, T., Salomon, J. & Troxtel, D. (1998). Relationship of coaching with performance in situational employment interviews. *Journal of Applied Psychology*, 83, 128–136.
- McCarthy, J. M., & Goffin, R. D. (2005). Selection test anxiety: Exploring tension and fear of failure across the sexes in simulated selection scenarios. *International Journal of Selection and Assessment*, 13, 282–295.
- McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L., III. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology*, 103, 399–414.

- McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morissette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, 19, 494–513.
- McDaniel, M. A., Howard, D. C., & Einstein, G. O. (2009). The read-recite-review study strategy: Effective and portable. *Psychological Science*, 20, 516–522.
- McDaniel, M. A., Roediger, H.L. III, & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin and Review*, 14, 200–206.
- McDaniel, M. A., Wildman, K. M., & Anderson, J. L. (2012). Using quizzes to enhance summative-assessment performance in a web-based class: An experimental study. *Journal of Applied Research in Memory and Cognition*, 1, 18–26.
- Meichenbaum, D. (1985). Teaching thinking: A cognitive-behavioral perspective. In S. F. Chipman, J. W. Segal, & R. Glaser (Eds.), *Thinking and learning skills, Vol. 2: Research and open questions* (pp. 407–426). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Melby-Lervåg, M., & Hulme, C. (2013). Is working memory training effective? A meta-analytic review. *Developmental Psychology*, 49, 270–291.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.) *Educational measurement* (pp. 13–103). New York, NY: American Council on Education and Macmillan.
- Messick, S., & Jungeblut, A. (1981). Time and method in coaching for the SAT. *Psychological Bulletin*, 89, 191–216.

Metcalfe, J., Kornell, N., & Son, L. K. (2007). A cognitive-science based programme to enhance study efficacy in a high and low risk setting. *European Journal of Cognitive Psychology*, 19, 743–768.

Meyer, A. N. D., & Logan, J. M. (2013). Taking the testing effect beyond the college freshman: Benefits for lifelong learning. *Psychology and Aging*, 28, 142–147.

Millman J. (1989). If at first you don't succeed: Setting passing scores when more than one attempt is permitted. *Educational Researcher*, 18, 5–9.

Muchinsky, P. M. (2004). When the psychometrics of test development meets organizational realities: A conceptual framework for organizational change, examples, and recommendations. *Personnel Psychology*, 57, 175–209.

Nathan, J. S., & Camara, W. J. (1998, September). *Score change when retaking the SAT I: Reasoning Test* (Research Note No. RN-05). New York, NY: The College Board.

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOL) are extremely accurate at predicting subsequent recall: The delayed-JOL effect. *Psychological Science*, 2, 267–270.

Neubauer, A. C. & Freudenthaler, H. H. (1994) Reaction time in a sentence-picture verification test and intelligence: Individual strategies and effects of extended practice. *Intelligence*, 19, 193–218.

Neuschatz, J. S., Preston, E. L., Toglia, M. P., & Neuschatz, J. S. (2005). Comparison of the efficacy of two name-learning techniques: Expanding rehearsal and name-face imagery. *American Journal of Psychology*, 118, 79–102.

- Pashler, H., Bain, P., Bottge, B., Graesser, A., Koedinger, K., McDaniel, M., & Metcalfe, J. (2007). *Organizing instruction and study to improve student learning (NCER 2007-2004)*. Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education. Available from <http://ncer.ed.gov>
- Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, & Cognition, 31*, 3–8.
- Pavlik, P. I., Jr., & Anderson, J. R. (2005). Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect. *Cognitive Science, 29*, 559–586.
- Ployhart, R. E., & Holtz, B. C. (2008). The diversity-validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology, 61*, 153–172.
- Powers, D. E. (1986). Relations of test item characteristics to test preparation/test practice effects: A quantitative summary. *Psychological Bulletin, 100*, 67–77.
- Powers, D. E., & Rock, D. A. (1999). Effects of coaching on SAT I: Reasoning test scores. *Journal of Educational Measurement, 36*, 93–118.
- Pressey, S. L. (1926). A simple device which gives tests and scores- and teaches. *School and Society, 23*, 373–376.
- Prestwood, J. S. (1979). Knowledge of results and the proportion of positive feedback on tests of ability. *Applied Psychological Measurement, 3*, 155–160.

- Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science*, 330, 335.
- Raven, J. (1989). The Raven Progressive Matrices: A review of national norming studies and ethnic and socioeconomic variation with the United States. *Journal of Educational Measurement*, 26, 1–16.
- Raven, J. C., Court, J. H., & Raven, J. (1985). *A manual for Raven's Progressive Matrices and Vocabulary Scales*. London, UK: H. K. Lewis.
- Raven, J. C., Raven, J., & Court, J. H. (1994). *A manual for Raven's Progressive Matrices and Vocabulary Scales*. London, UK: H. K. Lewis.
- Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? *Journal of Experimental Psychology: General*, 140, 283–302.
- Raymond, M. R., Neustel, S., & Anderson, D. (2007). Retest effects on identical and parallel forms in certification and licensure testing. *Personnel Psychology*, 60, 367–396.
- Ree, M. J., & Earles, J. A. (1991). Predicting training success: Not much more than *g*. *Personnel Psychology*, 44, 321–332.
- Ree, M. J., Earles, J. A., & Teachout, M. S. (1994). Predicting job performance: Not much more than *g*. *Journal of Applied Psychology*, 79, 518–524.
- Rees, P. J. (1986). Do medical students learn from multiple choice examinations? *Medical Education*, 20, 123–125.

- Reeve, C. L., Heggestad, E. D., & Lievens, F. (2009). Modeling the impact of test anxiety and test familiarity on the criterion-related validity of cognitive ability tests. *Intelligence*, 37, 34–41.
- 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.
- Richardson, F., & Robinson, E. S. (1921). Effects of practice upon the scores and predictive value of the alpha intelligence examination. *Journal of Experimental Psychology*, 4, 300–317.
- Roberts, M. J., & Newton, E. J. (2003). Individual differences in the development of reasoning strategies. In D. Hardman & L. Macci (Eds.), *Thinking: Psychological perspectives on reasoning, judgment, and decision making*. Chichester, UK: John Wiley.
- Roediger, H. L., III, & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, 15, 20–27.
- Roediger, H. L., III, & Karpicke, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, 1, 181–210.
- Roediger, H. L., III, & Karpicke, J. D. (2006b). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, 17, 249–255.

- Roediger, H. L., III, Putnam, A. L., & Smith, M. A. (2011). Ten benefits of testing and their applications to educational practice. *Psychology of Learning and Motivation*, 44, 1–36.
- Roth, P. L., Van Iddekinge, C. H., Huffcutt, A. I., Eidson, C. E., Jr., & Bobko, P. (2002). Correcting for range restriction in structured interview ethnic group differences: The values may be larger than researchers thought. *Journal of Applied Psychology*, 87, 369–376.
- Sackett, P. R., Burris, L. R., & Ryan, A. M. (1989). Coaching and practice effects in personnel selection. In C. Cooper & I. Robertson (Eds.), *International review of industrial and organizational psychology* (pp. 145–183). New York, NY: Wiley.
- Saloman, G., & Globerson, T. (1987). Skill may not be enough: The role of mindfulness in learning and transfer. *International Journal of Educational Research*, 11, 623–637.
- 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, age, and gender differences in retesting improvement. *Journal of Applied Psychology*, 95, 603–617.
- Schleicher, D. J., Venkataramani, V., Morgeson, F. P., & Campion, M. A. (2006). So you didn't get the job . . . Now what do you think? Examining opportunity-to-perform fairness perceptions. *Personnel Psychology*, 59, 559–590.

- 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.
- Schmidt, F. L., Hunter, J. E., & Outerbridge, A. N. (1986). Impact of job experience and ability on job knowledge, work sample performance, and supervisory ratings of job performance. *Journal of Applied Psychology, 71*, 432–439.
- Schmit, M. J., & Ryan, A. M. (1997). Applicant withdrawal: The role of test-taking attitudes and racial differences. *Personnel Psychology, 50*, 855–876.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference* (2<sup>nd</sup> Ed.). Wadsworth Publishing: Florence, KY.
- Shelton, J. T., Elliott, E. M., Matthews, R. A., Hill, B. D., & Gouvier, W. D. (2010). The relationships of working memory, secondary memory, and general fluid intelligence: Working memory is special. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 36*, 813–820.
- Shelton, J. T., Metzger, R. L., & Elliott, E. M. (2007). A group-administered lag task as a measure of working memory. *Behavior Research Methods, 39*, 482–493.
- Sin, H. P., Farr, J. L., Murphy, K. R., & Hausknecht, J. P. (2004, August). *An investigation of Black-White differences in self-selection and performance in repeated testing*. Paper presented at the meeting of the Academy of Management, New Orleans, LA.

- Skuy, M., Gewer, A., Osrin, Y., Khunou, D., Fridjon, P., & Rushton, J. P. (2002). Effects of mediated learning experience on Raven's Matrices scores of African and non-African university students in South Africa. *Intelligence*, 30, 221–232.
- Snow, R. E. (1982). The training of intellectual aptitude. In D. K. Detterman & R. J. Sternberg (Eds.), *How and how much can intelligence be increased?* Norwood, NJ: Ablex.
- Society for Industrial and Organizational Psychology. (2003). *Principles for the validation and use of personnel selection procedures* (4<sup>th</sup> Ed.). College Park, MD: Author.
- Spearman, C. & Jones, L. L. (1950). *Human ability*. London, UK: Macmillan.
- Spitzer, H. F. (1939). Studies in retention. *Journal of Educational Psychology*, 30, 641–656.
- Spurlock, D. G., Goss, S., & Bernstein, D. A. (2000). *Psychology 5<sup>th</sup> Edition's Test Bank*. Boston: Houghton-Mifflin Company.
- Sternberg, R. J., Ketron, J. L., & Powell, J. S. (1982). Componential approaches to the training of intelligence performance. In D. K. Detterman & R. J. Sternberg (Eds.), *How and how much can intelligence be increased?* (pp. 155-172). Norwood, NJ: Ablex.
- Stock, W. A., Kulhavy, R. W., Pridemore, D. R., & Krug, D. (1992). Responding to feedback after multiple-choice answers: The influence of response confidence. *Quarterly Journal of Experimental Psychology*, 45A, 649–667.

- te Nijenhuis, J., Voskuijl, O. F., & Schijve, N. B. (2001). Practice and coaching on IQ tests: Quite a lot of g. *International Journal of Selection and Assessment*, 9, 302–308.
- te Nijenhuis, J., van Vianen, A. E. M., & van der Flier, H. (2007). Score gains on g-loaded tests: No g. *Intelligence*, 35, 283–300.
- Thorndike, E. L. (1906). *The principles of teaching based on psychology*. New York, NY: A. G. Seiler.
- Thorndike, E. L. (1913). *Educational Psychology, Vol 1: The psychology of learning*. New York: Teachers College, Columbia University.
- Thorndike, E. L., & Woodworth, R. S. (1901). The influence of improvement in one mental function upon the efficiency of other functions. *Psychological Review*, 8, 247–261.
- Trowbridge, M. H. & Carson, H. (1932). An experimental study of Thorndike's theory of learning. *Journal of General Psychology*, 7, 245–258.
- Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language*, 28, 127–154.
- Tuzinski, K. A., Laczo, R. M., & Sackett, P. R. (2005, April). *Impact of response distortion on retaking of cognitive and personality tests*. Paper presented at the meeting of the Society for Industrial and Organizational Psychology, Los Angeles, CA.
- Uniform guidelines on employee selection procedures. (1978). *Federal Register*, 43, 38290–38315.

- U. S. Department of Education, National Center for Education Statistics. (2012). *Digest of education statistics 2011* (NCES Report 2012-011). Retrieved from <http://nces.ed.gov/pubs2012/2012001.pdf>
- U.S. Department of Labor. (1999). *Testing and assessment: An employer's guide to good practices*. Washington, DC: Author, Employment and Training Administration.
- Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3, 4–70.
- van der Reis, A. P. (1963). Is retesting justified in personnel selection? *Psychologia Africana*, 10, 19–30.
- Van Iddekinge, C. H., Morgeson, F. P., Schleicher, D. J., & Campion, M. A. (2011). Can I retake it? Exploring subgroup differences and criterion-related validity in promotion retesting. *Journal of Applied Psychology*, 96, 941–955.
- Vojdanoska, M., Cranney, J., & Newell, B. R. (2010). The testing effect: The role of feedback and collaboration in a tertiary classroom setting. *Applied Cognitive Psychology*, 24, 1183–1195.
- Watson, F. L., Pasteur, M. L., Healy, D. T., & Hughes, E. A. (1994). Nine parallel versions of four memory tests: An assessment of form equivalence and the effects of practice on performance. *Human Psychopharmacology*, 9, 51–61.
- Webb, J. M., Pridemore, D. R., Stock W. A., Kulhavy R. W., & Henning J. E. (1997). Remembering responses and cognitive estimates of knowing: The effects of

- instructions, retrieval sequences, and feedback. *Contemporary Educational Psychology*, 34, 918–928.
- Webb, J. M., Stock, W. A., Kulhavy, R. W., Haygood, R. C., Zulu, D. N. D., & Robinson, D. H. (1990). Directed forgetting and feedback in written instruction. *Bulletin of the Psychonomic Society*, 28, 543–546.
- Webb, J. M., Stock, W. A., Kulhavy, R. W., & White, M. C. (1990). *Feeling-of-knowing ratings for 240 general information facts*. (Tech. Rep. ZN-15). Tempe, AZ: Arizona State University, Learning and Instructional Technology Program.
- Webb, J. M., Stock, W. A., & McCarthy, M. T. (1994). The effects of feedback timing on learning facts: The role of response confidence. *Contemporary Educational Psychology*, 19, 251–265.
- Wendt, A., & Harris, L. (2004). *National Council licensure examination for registered nurses*. Chicago, IL: National Council of State Boards of Nursing.
- Wheeler, J. K. (2004, April). *Practical implications of selection retesters on testing development and policy*. Practitioner forum presented at the meeting of the Society for Industrial and Organizational Psychology, Chicago, IL.
- Wing, H. (1980). Practice effects with traditional mental test items. *Applied Psychological Measurement*, 4, 141–155.
- Woehlke, A. B., & Wilder, D. H. (1963). Differences in difficulty of Forms A and B of the Otis Self-Administering Test of Mental Ability. *Personnel Psychology*, 16, 395–398.

Woehr, D. J., Arthur, W., Jr., & Fehrman, M. L. (1991). An empirical comparison of cutoff score methods for content-related and criterion-related validity settings. *Educational and Psychological Measurement*, 51, 1029–1039.

Wong, B. Y. L. (1985). Self-questioning instructional research. *Review of Educational Research*, 55, 227–268.

## **APPENDIX A**

### **MEASURES**

#### **Sample items from General Introductory Psychology Exam 2.0**

**This exam is designed to assess your knowledge of fundamental introductory psychology issues, concepts, and principles. Although this exam is being administered for research purposes only, please answer each item to the best of your ability.**

**There are 20 items on this exam. For each item, select the alternative that you think is the BEST answer. There is no penalty for getting an item wrong or guessing. Hence, it is in your best interest to guess if you do not know the answer to an item.**

A person is asleep, but her electroencephalograph (EEG) shows activity resembling that of the waking state. Her heart rate, blood pressure, and breathing are also similar to the waking state. Based on these physiological responses, this person is most likely in which stage of sleep?

- A. Stage 1
- B. Microsleep
- \*C. REM

In Pavlov's experiments, meat powder was placed on a dog's tongue to serve as the:

- A. conditioned stimulus.
- \*B. unconditioned stimulus.
- C. conditioned response.

To be stored in memory, information must first be:

- \*A. encoded.
- B. retrieved.
- C. decoded.

## Test Attitudes and Perceptions Survey

### DIRECTIONS

The following questions ask you to rate your general attitudes and perceptions about the test you are about to take. Your attitudes should be based on the information in the instruction set read aloud to you and the sample items you just completed. Read each of the questions below and check or bubble the circle on the accompanying scale that corresponds to your answer.

Strongly Disagree	Disagree	Neither Disagree or Agree	Agree	Strongly Agree
1	2	3	4	5

<i>Motivation</i>				
1. Doing well on this test is important to me.	<input type="radio"/> ① <input type="radio"/> ② <input type="radio"/> ③ <input type="radio"/> ④ <input type="radio"/> ⑤			
2. I want to do well on this test.	<input type="radio"/> ① <input type="radio"/> ② <input type="radio"/> ③ <input type="radio"/> ④ <input type="radio"/> ⑤			
3. I will try my best on this test.	<input type="radio"/> ① <input type="radio"/> ② <input type="radio"/> ③ <input type="radio"/> ④ <input type="radio"/> ⑤			
4. I will try to do the very best I can do on this test.	<input type="radio"/> ① <input type="radio"/> ② <input type="radio"/> ③ <input type="radio"/> ④ <input type="radio"/> ⑤			
5. While taking this test, I will concentrate and try to do well.	<input type="radio"/> ① <input type="radio"/> ② <input type="radio"/> ③ <input type="radio"/> ④ <input type="radio"/> ⑤			
6. I want to be among the top scorers on this test.	<input type="radio"/> ① <input type="radio"/> ② <input type="radio"/> ③ <input type="radio"/> ④ <input type="radio"/> ⑤			
7. I will push myself to work hard on this test.	<input type="radio"/> ① <input type="radio"/> ② <input type="radio"/> ③ <input type="radio"/> ④ <input type="radio"/> ⑤			
8. I am extremely motivated to do well on this test.	<input type="radio"/> ① <input type="radio"/> ② <input type="radio"/> ③ <input type="radio"/> ④ <input type="radio"/> ⑤			
9. I just don't care how I do on this test.	<input type="radio"/> ① <input type="radio"/> ② <input type="radio"/> ③ <input type="radio"/> ④ <input type="radio"/> ⑤			
10. I won't put much effort into this test.	<input type="radio"/> ① <input type="radio"/> ② <input type="radio"/> ③ <input type="radio"/> ④ <input type="radio"/> ⑤			
<i>Anxiety</i>				
11. I probably won't do as well as most of the other people who take this test.	<input type="radio"/> ① <input type="radio"/> ② <input type="radio"/> ③ <input type="radio"/> ④ <input type="radio"/> ⑤			
12. I am not good at taking tests.	<input type="radio"/> ① <input type="radio"/> ② <input type="radio"/> ③ <input type="radio"/> ④ <input type="radio"/> ⑤			

13. During a test, I often think about how poorly I am doing.	(1)   (2)   (3)   (4)   (5)
14. I usually get very anxious about taking tests.	(1)   (2)   (3)   (4)   (5)
15. I usually do pretty well on tests.	(1)   (2)   (3)   (4)   (5)
16. I expect to be among the people who score really well on this test.	(1)   (2)   (3)   (4)   (5)
17. My test scores don't usually reflect my true abilities.	(1)   (2)   (3)   (4)   (5)
18. I very much dislike taking tests of this type.	(1)   (2)   (3)   (4)   (5)
19. For this test, I find myself thinking of the consequences of failing.	(1)   (2)   (3)   (4)   (5)
20. During a test, I get so nervous I can't do as well as I should have.	(1)   (2)   (3)   (4)   (5)

## **Retest Interval Behavioral Inventory**

**DIRECTIONS:** For the following items, think about the time between the initial protocol session that you attended (i.e., the last time you participated in the study) and now. Respond to each of these questions to the best of your recollection with as much information as you can provide.

1. After you completed the tests from your last session, was any similar psychology content covered by a course instructor in class?	<input type="checkbox"/> Yes	<input type="checkbox"/> No
2. After you completed the tests from the last session, did you seek the answers of the tests' items in any way?	<input type="checkbox"/> Yes	<input type="checkbox"/> No
3. If so, provide an estimate of how long you spent seeking answers to the tests' items: _____ hours.		
4. Since the last time you were here, did you complete any psychology coursework (e.g., reading, studying) that was relevant to the previous tests' material?	<input type="checkbox"/> Yes	<input type="checkbox"/> No
5. If so, provide an estimate of how long you spent on psychology coursework (e.g., reading, studying) that was relevant to the previous tests' material: _____ hours.		
6. Since the last time you were here, did you complete any GRE Psychology Test preparation?	<input type="checkbox"/> Yes	<input type="checkbox"/> No
7. If so, provide an estimate of how many hours you spent preparing for the GRE Psychology Test: _____ hours.		

## **APPENDIX B**

### **GENERAL PSYCHOLOGY COMPETENCY EXAM MEASURE REVISION,**

#### **UPDATE, AND REVALIDATION (CONTENT-RELATED)**

To revise and update the General Psychology Competency Exam (GPCE), a content-related validation approach was undertaken. The goals of the GPCE's content-related revalidation were threefold: (1) to reassess whether the distribution of GPCE items adequately represented the introductory psychology content domain, (2) to ensure the GPCE reflected current knowledge within psychology as taught in an introductory psychology course, and (3) to expand the GPCE from 35 items to a total of 40 items which could be split into two alternate test forms of 20 items each, resulting in the GPCE 2.0. Due to marginal retest increases and concerns over a ceiling effect after pilot testing the measure (i.e., the GPCE 2.0), this test was subsequently further expanded to comprise 80 items which were split into two alternate test forms of 40 items each to create the GPCE 2.1. These two 40 item alternate test forms were the final measure used for the present study.

In updating the GPCE 2.0 with additional items, the first step consisted of determining core introductory psychology content and identifying subject matter experts (SMEs) to assess the extent to which the distribution of GPCE items adequately covered the introductory psychology content domain. First, the population of SMEs was identified as Texas A&M University Psychology Department instructors and faculty who had previously taught at least one introductory psychology course in the past five years. A sample of SMEs ( $n = 10$ ) were recruited and used throughout the content-

related validation procedures. That is, the same 10 SMEs participated in establishing the GPCE 2.0's content distribution and determining that the individual items of the GPCE 2.0 were adequate (which will be described in detail below). Next, copies of all available introductory psychology syllabi from the psychology department of Texas A&M University were obtained. Based on these syllabi, the most commonly used introductory psychology textbooks were identified. Using these introductory psychology textbooks' chapter headings, common groupings of introductory psychology content were identified. A proposed distribution of GPCE items reflecting core introductory psychology content domains was proposed based on the introductory psychology textbook content and the original GPCE's content distribution (Arthur et al., 2003).

To establish the content-related validity of the measure, a three-step process was undertaken. First, to establish that the GPCE 2.0's content distribution adequately reflected the content distribution of core introductory psychology knowledge, the 10 SMEs were individually interviewed to determine if the proposed GPCE test content distribution was representative of the breadth of introductory psychology content and, if not, what revisions they would recommend. Second, based on the SMEs' assessment, additional items were written as needed to meet the SMEs' proposed content distribution. Third, to establish that the individual items of the GPCE 2.0 were adequate, the SMEs completed ratings to assess the representativeness, correct labeling of content domain, and the correctness of the keyed responses of the GPCE's individual items. Steps two and three were repeated as necessary if there were not enough items

considered to be satisfactory, and therefore, an inadequate number of items to meet the proposed content distribution. Details of this process were as follows.

First, the proposed GPCE introductory psychology content distribution and a rating form was provided to the SMEs to report their recommended distributions (see Appendix B.1 below, *GPCE SME Content Distribution Rating Form*, for this proposed distribution). SMEs were informed that the purpose of these procedures was to update the GPCE to permit the creation of two alternate forms of a knowledge test assessing upper-level undergraduate psychology students' mastery of basic core psychology principles, concepts, and facts. Specifically, SMEs were asked whether they agreed with the proposed distribution, and if not, how they would change the distribution to better reflect the distribution of psychology content as they would teach the course. These content distribution ratings were collected with the author and SME individually.

After the SMEs' distributions of introductory psychology content were completed, the mean of SMEs' proposed content distributions was computed. The SMEs' distribution of core introductory psychology content was the same as that initially proposed on the *GPCE SME Content Distribution Rating Form*. However, additional items were required to develop two alternate forms of 20 items each.

<p><i>Item 1.</i> Research has shown that in situations where several potential helpers fail to help or provide assistance to someone in need, it is typically because they:</p> <ul style="list-style-type: none"> <li>*A. fail to take personal responsibility for helping.</li> <li>B. fail to notice the event.</li> <li>C. experience or have difficulty thinking of how to help.</li> </ul>		
<b>Introductory Psychology Content Domain: <i>Social Psychology</i></b>		
<p><b>1. Does this item correspond to the content domain as labeled?</b></p> <p><input type="checkbox"/> No      <input type="checkbox"/> Yes</p>	<p>If no, then select a more representative content domain from the drop down menu below:</p> <p>Content domain</p>	<p><b>2. Is this content covered in Introductory Psychology?</b></p> <p><input type="checkbox"/> No      <input type="checkbox"/> Yes</p>
<p><b>3. Does this item reflect “walking knowledge” following completion of Introductory Psychology?</b></p> <p><input type="checkbox"/> No      <input type="checkbox"/> Yes</p>	<p><b>4. Is the keyed response clearly the best one?</b></p> <p><input type="checkbox"/> No      <input type="checkbox"/> Yes</p>	<p><b>5. What is your estimate of the probability (0 to 100) that a minimally competent (i.e., C) student will answer the item correctly?</b></p> <p>—</p>

*Figure B.1.* Sample item from GPCE Measure Review Booklet for Subject Matter Experts.

Beginning with the initial pool of 35 items from the original GPCE, an additional 15 items were written based on standard item writing practices to generate an item pool that comprised 50 items. Next, these items were provided to the same 10 SMEs who were initially contacted, who were then asked to review the items and provide feedback pertaining to the quality of the items. Specifically, the SMEs were asked to review the items and provide information as to whether the items were representative and within the scope of the curriculum they taught, adequately represented the content domain as labeled (e.g., methods, social psychology, neuroscience), reflected “walking knowledge” following the completion of an introductory psychology course, and were keyed

correctly. SMEs also provided an Angoff-type rating estimating the probability that a minimally competent (i.e., C) student could answer the item correctly. (See Appendix B.2 for the *GPCE Measure Review Booklet for Subject Matter Experts* and a sample item rating in Figure B.1.)

Table B.1

*Subject Matter Expert Ratings of General Psychology Competency Exam*

Item	Appropriately Labeled Content (Yes)	Alternate Content Domain Proposed	Content Covered in Introductory Psychology (Yes)	Considered Walking Knowledge (Yes)	Keyed Response is Best (Yes)	Mean Probability a Minimally Competent Student Answers Correctly
1	100%	Consciousness	100%	90%	80%	70.00
2	90%		100%	80%	70%	71.50
3	100%		100%	100%	100%	80.40
4	100%		100% <sup>a</sup>	90%	70%	72.00
5	100%		100% <sup>a</sup>	90%	100%	64.00
6	100%	Learning	100%	90%	88.89% <sup>a</sup>	67.00
7	100%		90%	80%	100%	71.00
8	100%		100%	100%	100%	75.50
9	100%		100%	80%	100%	68.50
10	100%		100%	66.67% <sup>a</sup>	88.89% <sup>a</sup>	62.00 <sup>a</sup>
11	90%	Emotion and Motivation	90%	70%	80%	57.50
12	100%		70%	60%	100%	61.50
13	100%		90%	66.67% <sup>a</sup>	50%	49.30
14	100%		100%	100%	100%	74.20
15	100% <sup>a</sup>		100% <sup>a</sup>	88.89% <sup>a</sup>	100% <sup>a</sup>	81.11 <sup>a</sup>
16	100%	Personality Emotion & Motivation	90%	80%	66.67% <sup>a</sup>	55.50
17	100%		100%	100%	100%	79.30
18	100%		60%	50%	70%	52.30
19	100%		100%	100%	100%	74.00
20	90%		80%	70%	80%	63.80
21	70%	Personality Emotion & Motivation	100%	100%	100%	75.00
22	90%		50%	50%	88.89% <sup>a</sup>	56.30
23	100%		80%	70%	90%	67.90
24	100%		100%	90%	90%	62.30
25	100%		90%	100%	100%	74.00

Note. N = 10 SME Raters. <sup>a</sup>N = 9 raters. If alternate item content was proposed for the item by multiple raters, all raters proposed the same alternate item content category.

Table B.1 (*Continued*)*Subject Matter Expert Ratings of General Psychology Competency Exam*

Item	Appropriately Labeled Content (Yes)	Alternate Content Domain Proposed	Content Covered in Introductory Psychology (Yes)	Considered Walking Knowledge (Yes)	Keyed Response is Best (Yes)	Mean Probability a Minimally Competent Student Answers Correctly
26	100%		90%	90%	100%	65.00
27	100%		100%	90%	100%	67.50
28	100%		100%	90%	100%	83.50
29	100% <sup>a</sup>		88.89% <sup>a</sup>	66.67% <sup>a</sup>	100% <sup>a</sup>	60.00 <sup>a</sup>
30	100% <sup>a</sup>		66.67% <sup>a</sup>	66.67% <sup>a</sup>	100% <sup>a</sup>	58.67 <sup>a</sup>
31	100% <sup>a</sup>		88.89% <sup>a</sup>	66.67% <sup>a</sup>	100% <sup>a</sup>	61.67 <sup>a</sup>
32	100% <sup>a</sup>		88.89% <sup>a</sup>	88.89% <sup>a</sup>	100% <sup>a</sup>	78.67 <sup>a</sup>
33	90%	Human Development	100%	100%	100%	79.50
34	100%		100%	100%	100%	79.00
35	100%		90%	90%	100%	66.50
36	100%		100%	90%	100%	64.00
37	100%		100%	100%	100%	77.50
38	100%		90%	80%	100%	66.50
39	100%		40%	30%	80%	50.00
40	90%	Missing	80%	50%	70%	58.50
41	100%		100%	100%	90%	68.00
42	70%	Social Psychology	90%	90%	100%	71.50
43	100%		100%	100%	100%	79.50
44	100%		90%	80%	100%	69.30
45	100%		66.67% <sup>a</sup>	80%	100%	73.50
46	100%		100%	100%	100%	67.00
47	100%		60%	50%	80%	50.50
48	100%		100%	100%	100%	78.00
49	90%	Psychological Disorders	80%	70%	80%	65.50
50	90%		80%	60%	80%	59.50

Note. N = 10 SME Raters. <sup>a</sup> N = 9 raters. If alternate item content was proposed for the item by multiple raters, all raters proposed the same alternate item content category.

Next, based on the SMEs' feedback, items were either eliminated or modified as necessary as described below. Table B.1 presents the descriptive statistics for the SMEs'

ratings for the initial item pool. Items with 70% or less SME agreement (i.e., 7 out of the 10) that item's content was representative or within the scope of the curriculum they taught or reflected “walking knowledge” following completion of an introductory psychology course were eliminated. Seven items were eliminated on the basis of this decision rule. After these items were dropped, only 35 items remained. However, there was some disagreement among the SMEs as to whether some of these remaining items were correctly keyed. Specifically, eight of these 35 items were considered adequate, but 80% or less of the SMEs agreed that the item's keyed response was clearly the best. Dropping these eight items would substantially reduce the items available for the two test forms and drastically alter the content distribution of the measure (see Figure B.2). Thus, these items were provisionally retained pending additional review and editing by the SMEs.

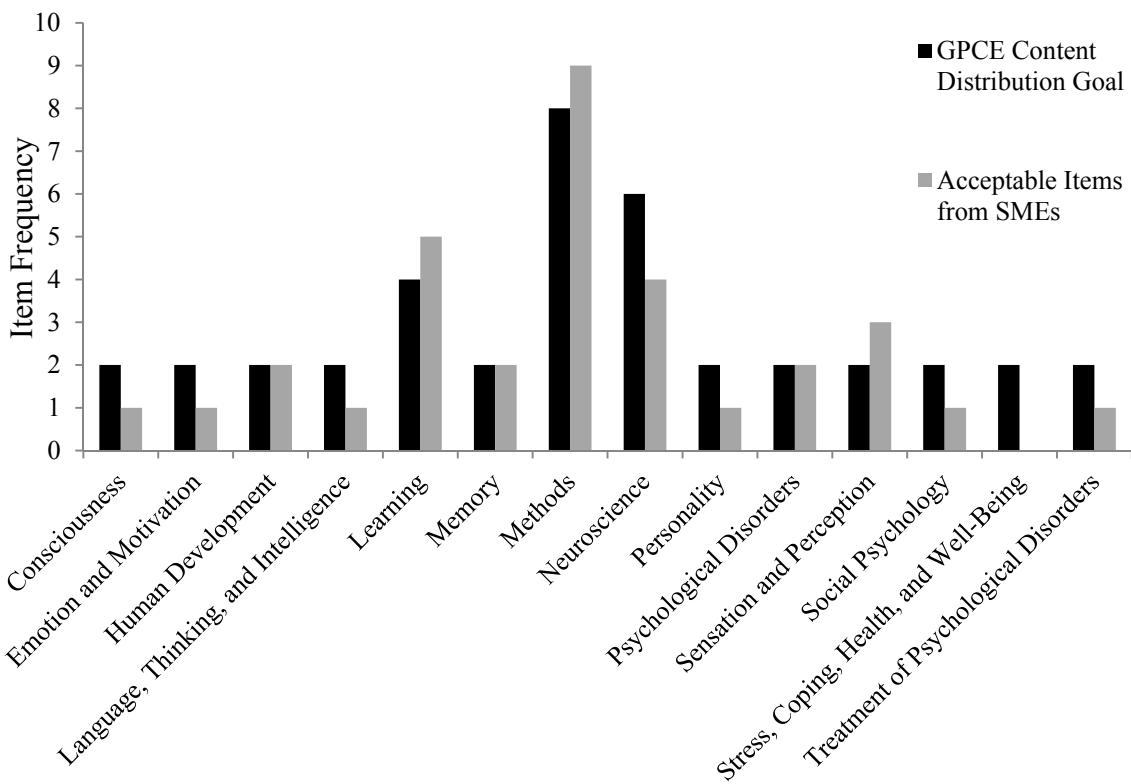


Figure B.2. Initial GPCE (2.0) content distribution goal (black) versus the current distribution of items rated as acceptable from SMEs (gray).

To meet the necessary number of items for the construction of two alternate test forms of 20 items each and the proposed GPCE content distribution, 12 additional items were written to ensure adequate content representation across the two test forms. The 8 discrepantly keyed items were also retained as potential items. Next, an additional SME meeting was conducted with a panel of four SMEs from the original sample of SMEs to review (1) previous items with adequate item ratings but discrepancies on the keyed response, and (2) the 12 new items. Specifically, the SMEs were instructed to identify, clarify, or modify the discrepantly keyed items by determining whether the keyed responses were in fact the best response, whether another alternative was the best response, whether writing an additional response was necessary as the best response, or

whether the item was inadequate and an entirely new item was needed. This process was repeated with the 12 new items. Items were not retained unless all SMEs unanimously agreed that the item was adequate (through rewriting the stem, response options, or both). Items were dropped if the SMEs could not reach consensus or believed that the item could not be rewritten to be adequate.

Upon completion of this process, 49 items remained. Because only 40 items were needed for the final version, items were retained for the final version by selecting the best 40 (in terms of covering introductory psychology content that was considered suitably broad, walking knowledge by the SME consensus panel) of the 49 items, coupled with selecting pairs of items with matching content for the alternate forms. (For clarity, only items that all SMEs unanimously approved were retained.) Finally, these items were split into two alternate test forms (Test Forms A and B) to match the distribution of introductory psychology content as developed by SMEs (see Figure B.3). Items were arranged in the same order of introductory psychology content across test forms.

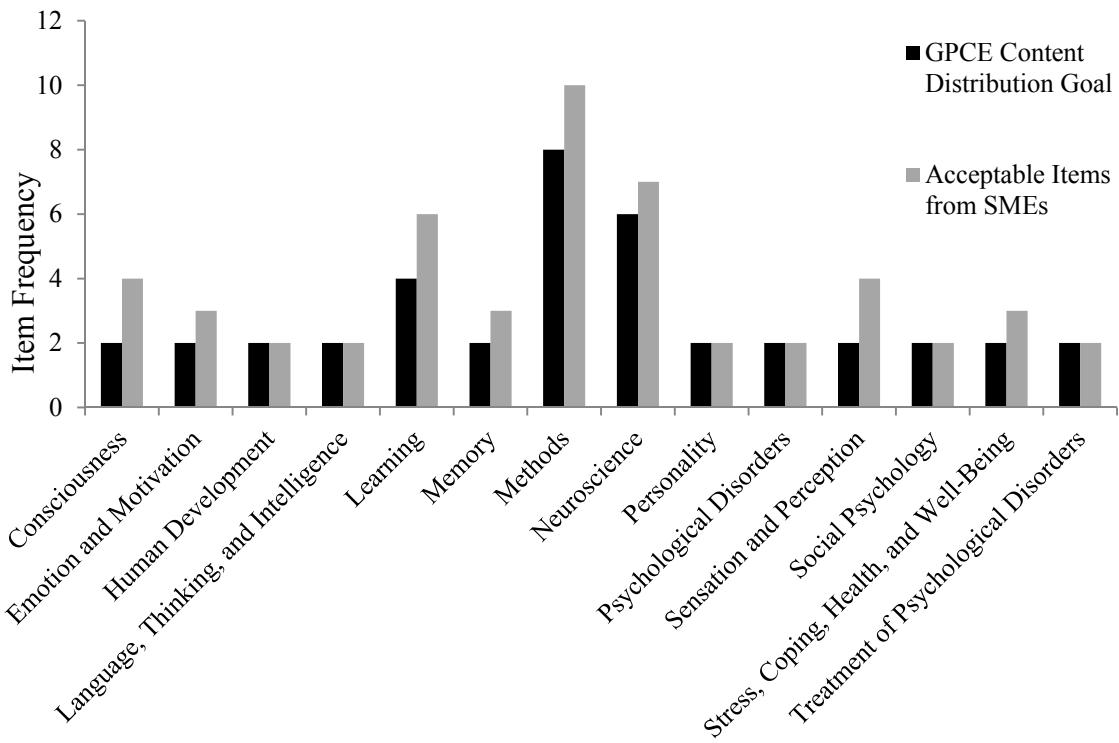


Figure B.3. Final GPCE (2.1) content distribution goal (black) versus the current distribution of items rated as acceptable from SMEs (gray).

After pilot testing this version of the GPCE, no retest increases were observed across test forms, causing concern that the short test (20 items per form) exhibited a ceiling effect. Thus, the GPCE 2.0 was expanded further to comprise 80 items total, which were split into two alternate test forms of 40 items to create the GPCE 2.1. The goal was to double the number of test items while maintaining the same distribution of psychology content across both test forms. Test administration time was kept constant (15 minutes).

Rather than replicating the previous process, additional items were obtained from *Psychology 5<sup>th</sup> Edition's Test Bank* (Spurlock, Goss, & Bernstein, 2000), which reported

item parameters from a normative sample. These additional items were not assessed by SMEs using same process as the GPCE's previous items; instead, their inclusion relied exclusively on psychometric item statistics reported by Spurlock et al. (2000).

Only items for which item statistics were reported were considered for inclusion. To be retained, items had to meet three criteria. First, items were selected if they were labeled as “conceptual/applied” rather than “factual”. Second, items were retained if the difficulty was less than or equal to .70 (i.e., 30% or less of the test-takers responded to the item correctly). Third, items were retained if the item discriminability was greater than or equal to .30. Items exhibiting these desired parameters were identified from each of the test bank’s content domains that matched the pre-established distribution of introductory psychology content.

Once this initial pool of items within each content domain was identified, items were retained as follows. First, items assessing overlapping content to the current GPCE alternate test forms’ items were excluded. Next, items reflecting similar topics within a particular content domain that could be easily separated to create the two alternate test forms were retained. Finally, the test bank items were modified from their initial 4-option multiple choice response format to match the GPCE’s 3-option multiple choice response format. To generate 3-option versions of the test bank items, the unkeyed alternative response with the lowest endorsement percentage was eliminated (Edwards, Arthur, & Bruce, 2012). The additional items were arranged were placed immediately after the original items, in the same order of content domain from the original test.

## Appendix B.1

### GPCE SME Content Distribution Rating Form

Name: \_\_\_\_\_ Date: \_\_\_\_\_

Area: \_\_\_\_\_

**Instructions:**

Do you consider the current GPCE Content Domain Breakdown to be representative?

Yes or  No?

If no, then how would you redistribute these items?

Psychology 107 Topic	Proposed GPCE Content Domain Item Breakdown	How would you redistribute these items?
Methods	4	
Neuroscience	3	
Sensation & Perception	1	
Consciousness	1	
Memory	1	
Learning	2	
Emotion and Motivation	1	
Language, Thinking, and Intelligence	1	
Human Development	1	
Personality	1	
Stress, Coping, Health, and Well-Being	1	
Psychological Disorders	1	
Treatment of Psychological Disorders	1	
Social Psychology	1	
Additional Domain?		
Additional Domain?		
Additional Domain?		
<b>Total</b>	<b>20</b>	

## Appendix B.2

### GPCE Measure Review Booklet for Subject Matter Experts

Name: \_\_\_\_\_

Area: Choose your program area:

As part of my dissertation, I am in the process of updating and revising a knowledge exam (i.e., the General Psychology Competency Exam [GPCE]) that assesses upper-level undergraduate psychology students' mastery of some basic, core psychology principles, concepts, and facts. Consequently, because of your background as an introductory psychology instructor, I am soliciting your input in reviewing and rating the items for this exam. Your judgments as a subject matter expert are valuable to this item review process as the results will help us determine which items should be retained, revised, or dropped, and subsequently, allow us to develop an exam that can be used to investigate the specified issues in my dissertation.

#### ITEM RATING INSTRUCTIONS

Please carefully read and follow all the instructions contained in this electronic form. If you do not understand an item or the instructions as written, feel free to email me at [AndrewMNaber@gmail.com](mailto:AndrewMNaber@gmail.com) or call me at 410-279-7125.

When you have finished your subject matter expert (SME) ratings of all 40 items, save the document as a word .docx file with your initials at the end of the file name. For example, I would save the document as GPCE SME Item Rating Form AMN.docx. Then, please email the document back to me at [AndrewMNaber@gmail.com](mailto:AndrewMNaber@gmail.com)

This electronic rating form contains five ratings for each test item. Please read each item carefully and answer each question according to the instructions below:

#### 1. Does this item correspond to the content domain as labeled?

- Click **Yes**, if this item represents the underlying content domain as it is labeled.
- Click **No**, if this item does not represent the underlying content domain as it is labeled.

If the item does **NOT** correspond to the content domain as labeled, please select from the PSYC 107 content domains listed below that you would consider the item as assessing. You may select from these domains on the rating form by using the drop down menu for each item.

**PSYC 107 Content Domains:**

Methods	Language, Thinking, and Intelligence
Neuroscience	Human Development
Sensation and Perception	Personality
Consciousness	Stress, Coping, Health, and Well-Being
Memory	Psychological Disorders
Learning	Treatment of Psychological Disorders
Emotion and Motivation	Social Psychology

**2. Is this content covered in Introductory Psychology as you teach or have taught it?**

- Click Yes, if this item was covered in your Introductory Psychology course.
- Click No, if this item was not covered in your Introductory Psychology course.

**3. Does this item reflect “walking knowledge” following completion of Introductory Psychology?**

- Click Yes, if this item reflects knowledge that minimally competent students (i.e., C students) would maintain and understand after completing Introductory Psychology and without needing to review reference materials.
- Click No, if this item does not reflect knowledge that minimally competent students (i.e., C students) would be expected to maintain and understand after completing Introductory Psychology without needing to review reference materials.

**4. Is the keyed response clearly the best one? Note that keyed responses are marked with an \*.**

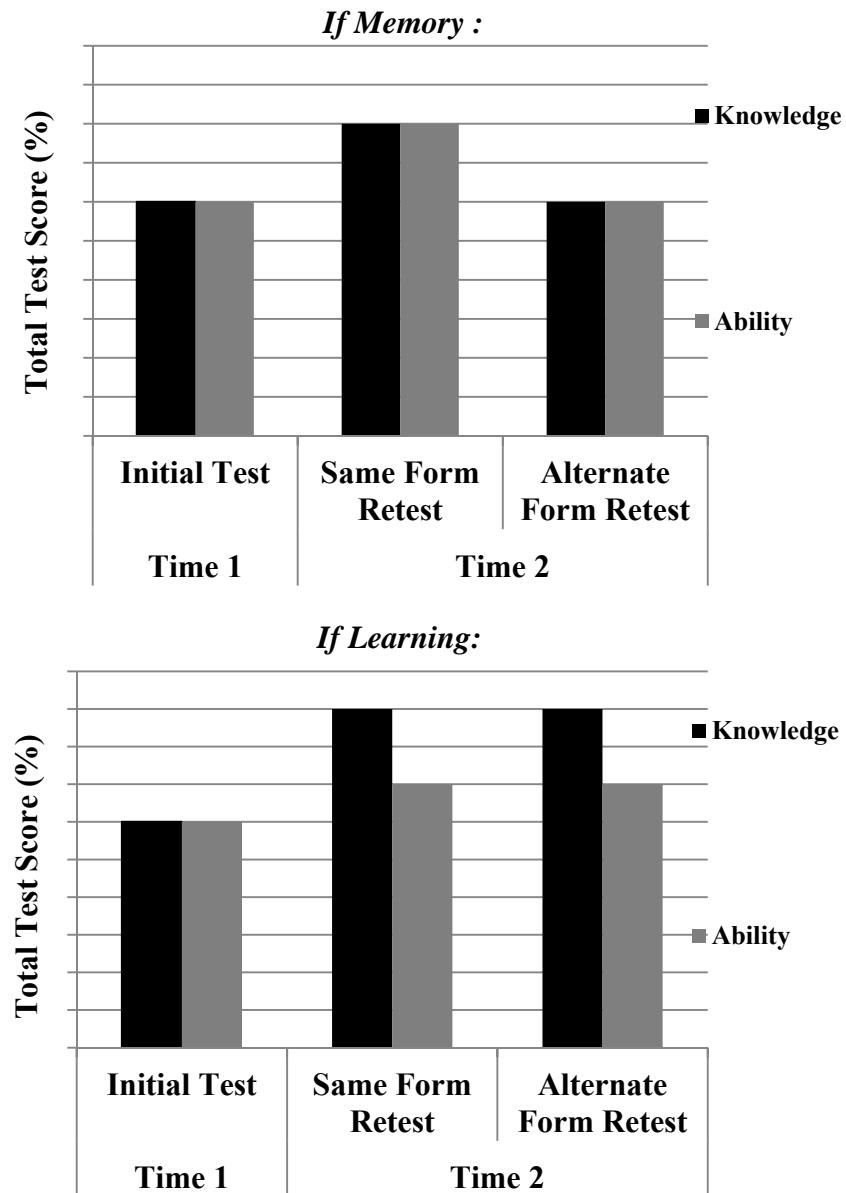
- Click Yes, if this item’s keyed response is clearly the best option.
- Click No, if this item’s keyed response is not clearly the best option.

**5. What is the likelihood or probability that a minimally competent student would correctly answer the item?**

A minimally competent student means that the student possesses enough knowledge to obtain a “C” in an introductory psychology course. If the student were any *less* knowledgeable, he or she could *not* obtain a “C” average upon completion of the course.

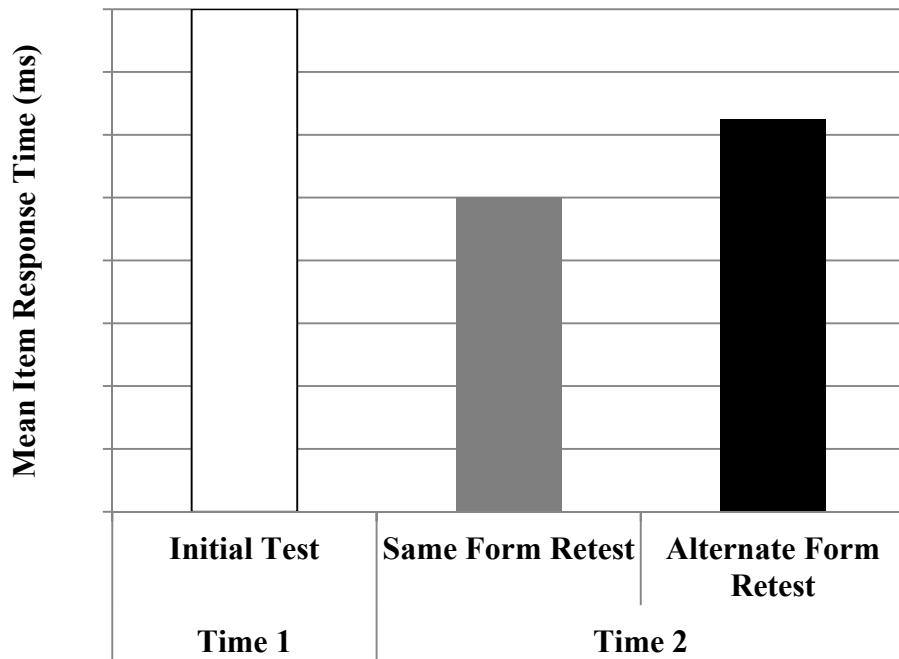
- Click the blank line to type the numerical likelihood or probability.

**APPENDIX C**  
**HYPOTHEZIZED PATTERN OF RESULTS FOR MEMORY VERSUS**  
**LEARNING EXPLANATIONS**



*Figure C.1.* Hypotheses 1a and 3a versus 1b and 3b: Total test score (percentage) by same versus alternate form retest over time.

*If Memory :*



*If Learning:*

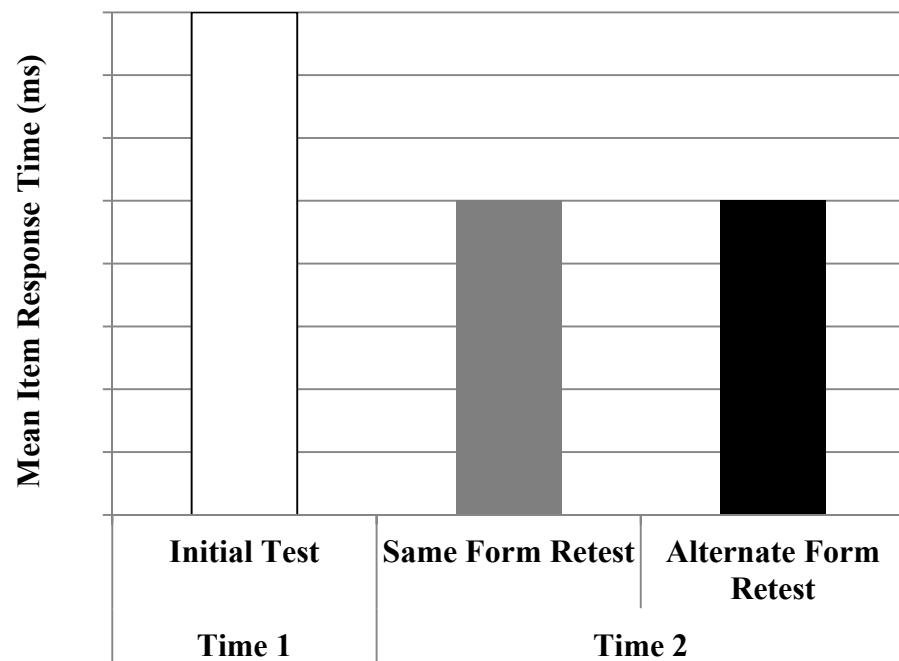
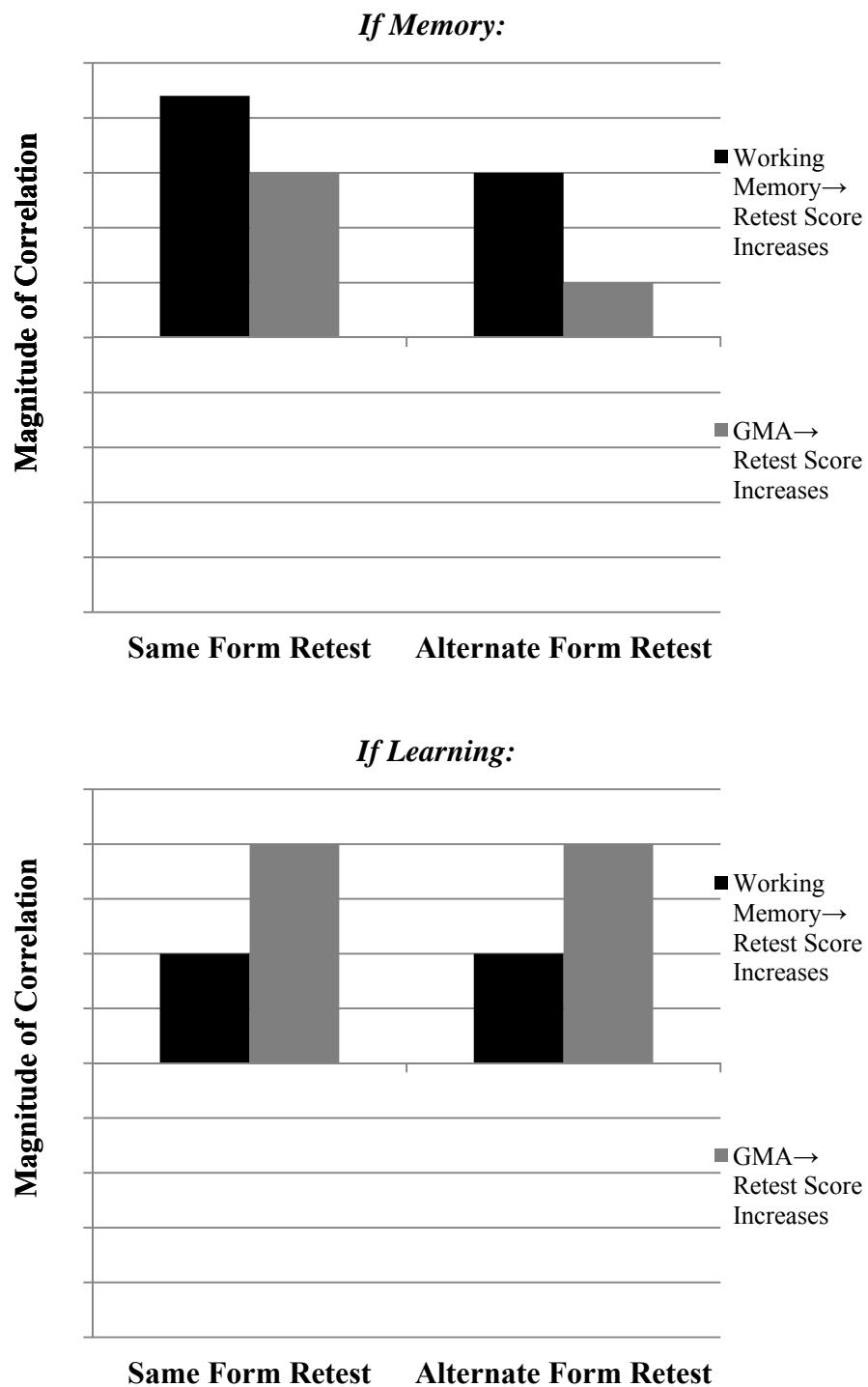
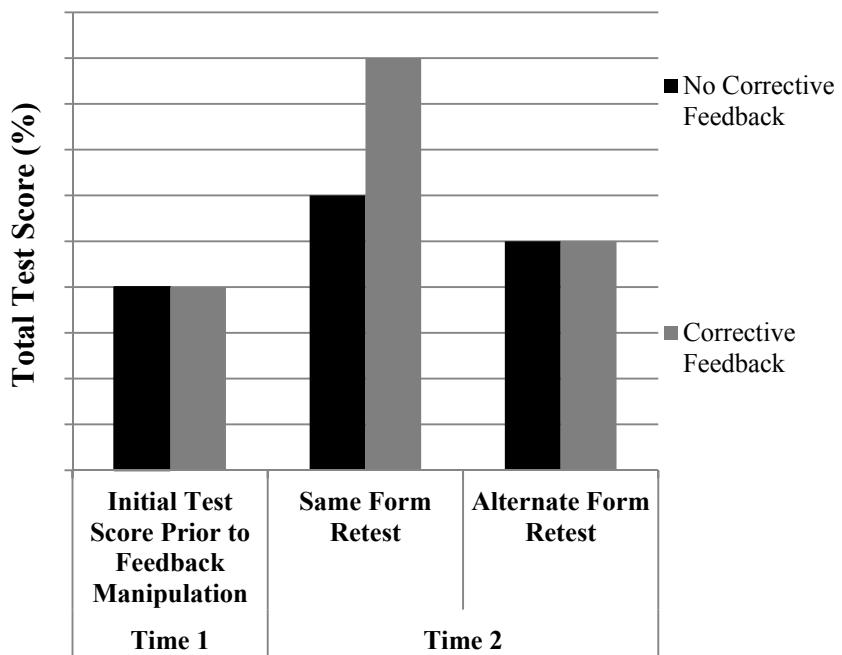


Figure C.2. Hypothesis 2a versus 2b: Mean item response time by same versus alternate form retest over time (irrespective of construct domain).



*Figure C.3.* Hypothesis 4a versus 4b: The relationship between working memory and retest score increases and the relationship between GMA and retest scores increases for the same form retest and alternate retest forms.

*If Memory:*



*If Learning:*

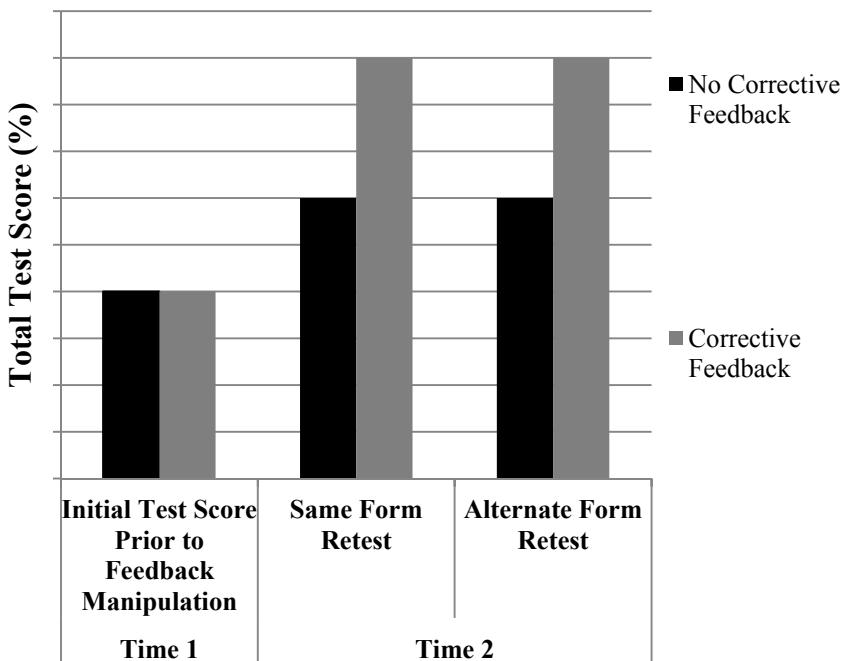
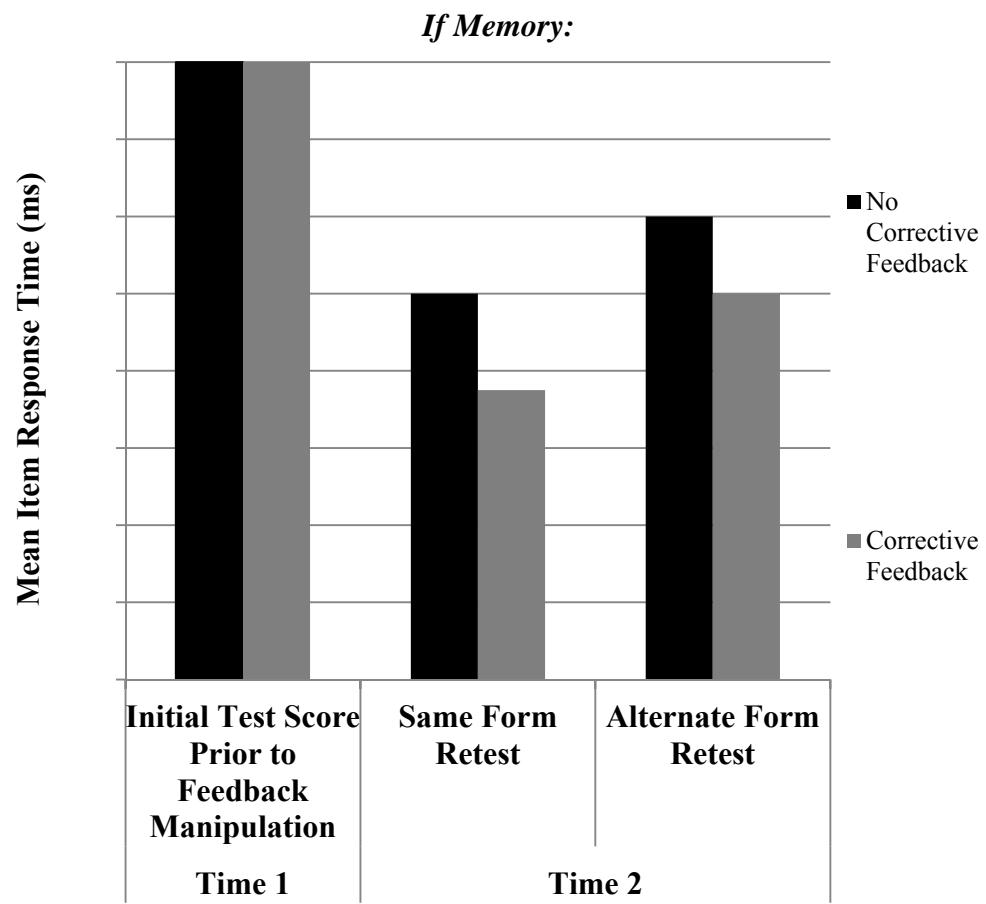
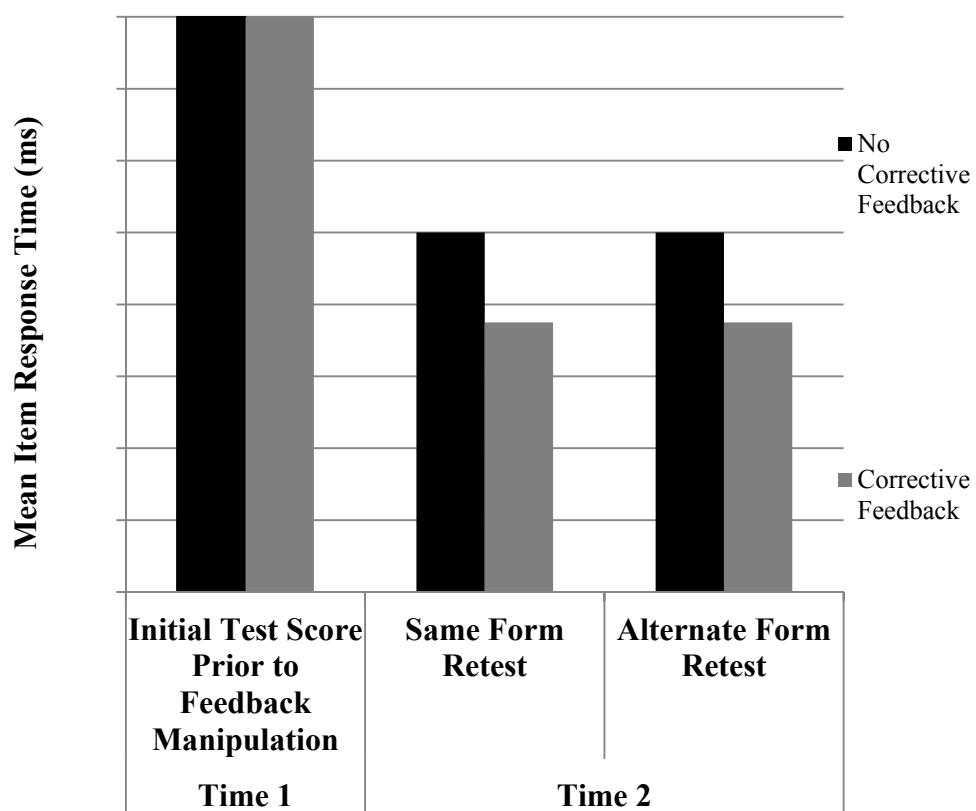


Figure C.4. Hypothesis 5a versus 5b: Total test score (percentage) by same versus alternate form retest over time (irrespective of construct domain).

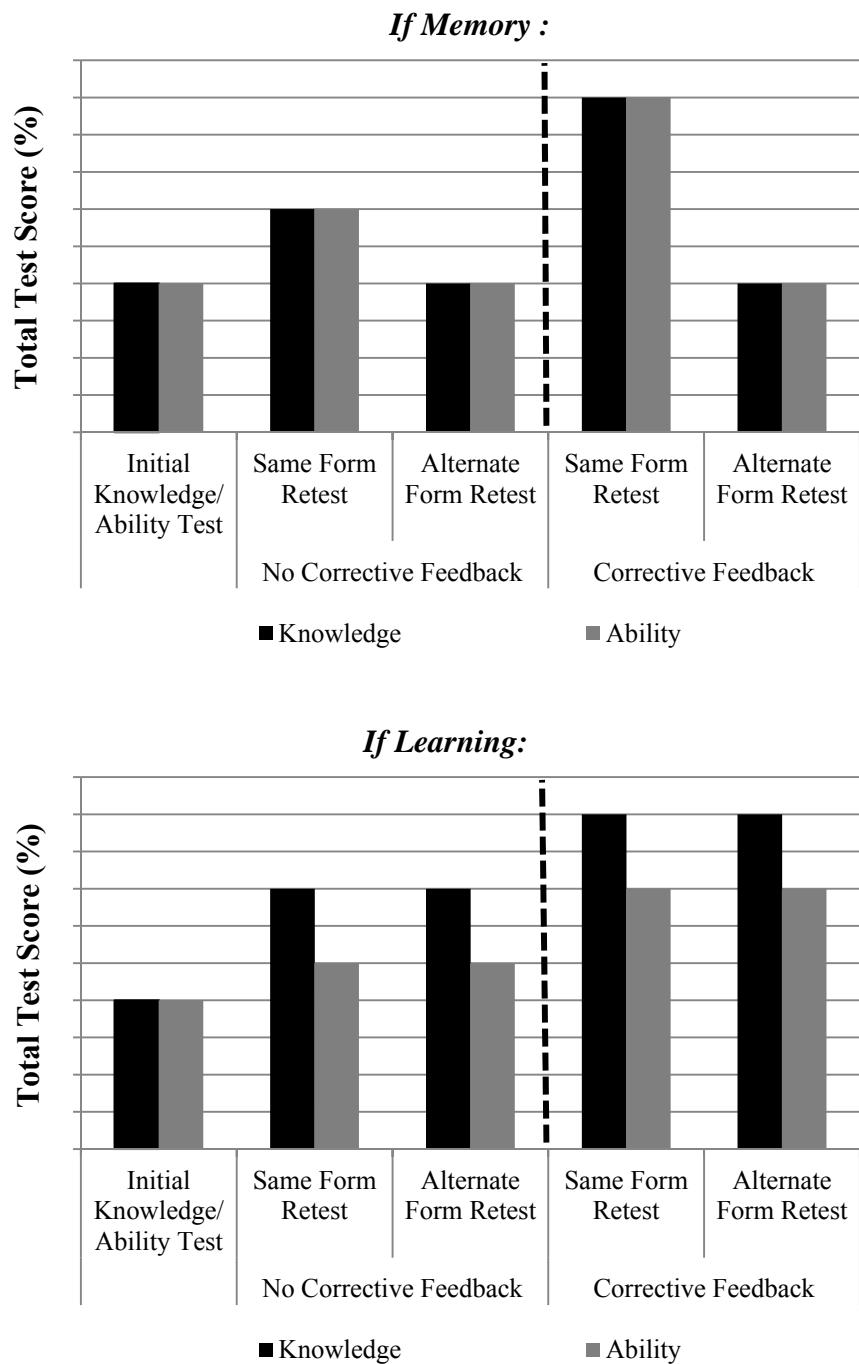


*Figure C.5.* Hypothesis 6a versus 6b: Mean item response time by same versus alternate form retest with corrective feedback or no corrective feedback over time (irrespective of construct domain).

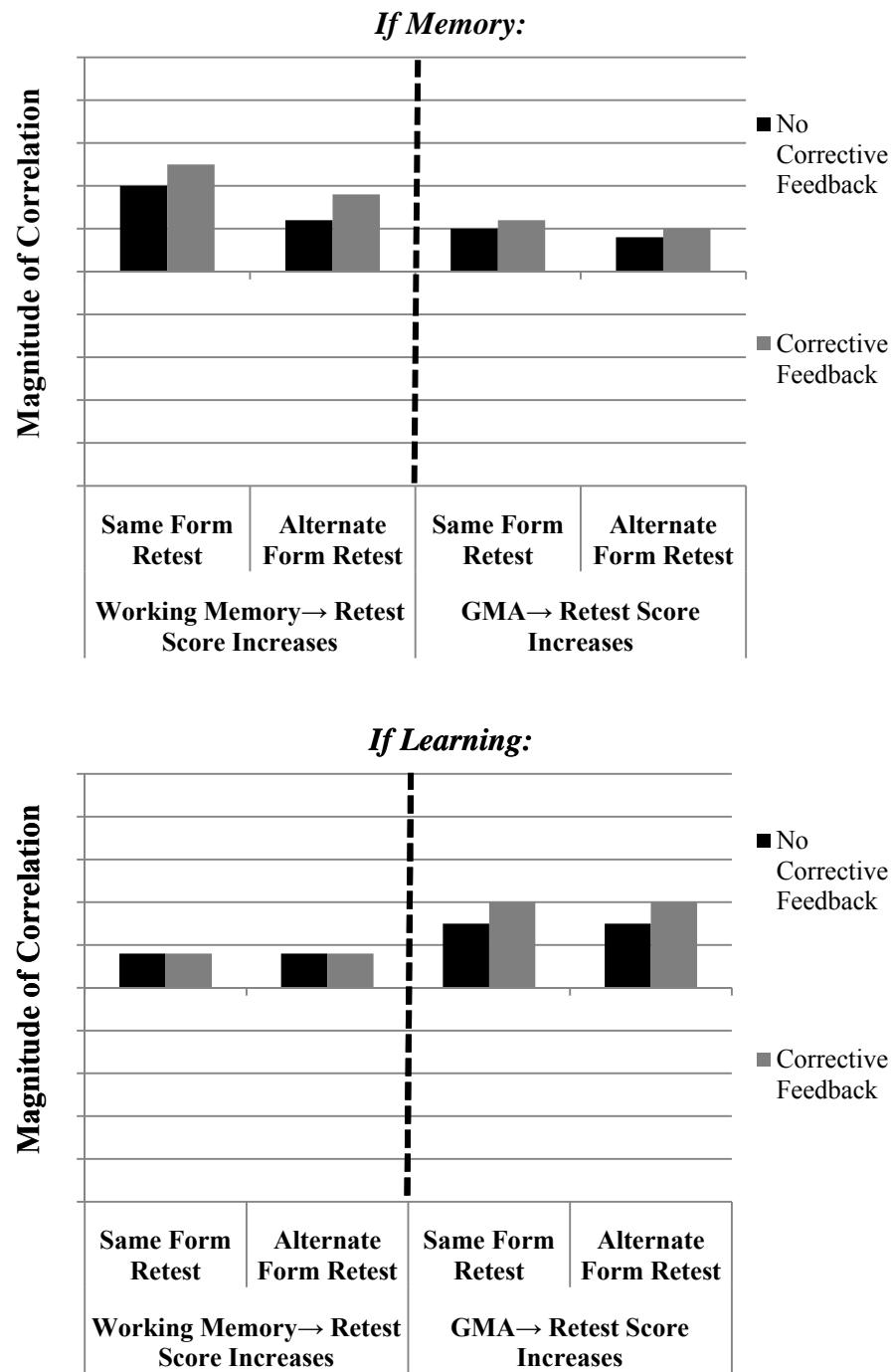
*If Learning:*



*Figure C.5 (Continued).* Hypothesis 6a versus 6b: Mean item response time by same versus alternate form retest with corrective feedback or no corrective feedback over time (irrespective of construct domain).



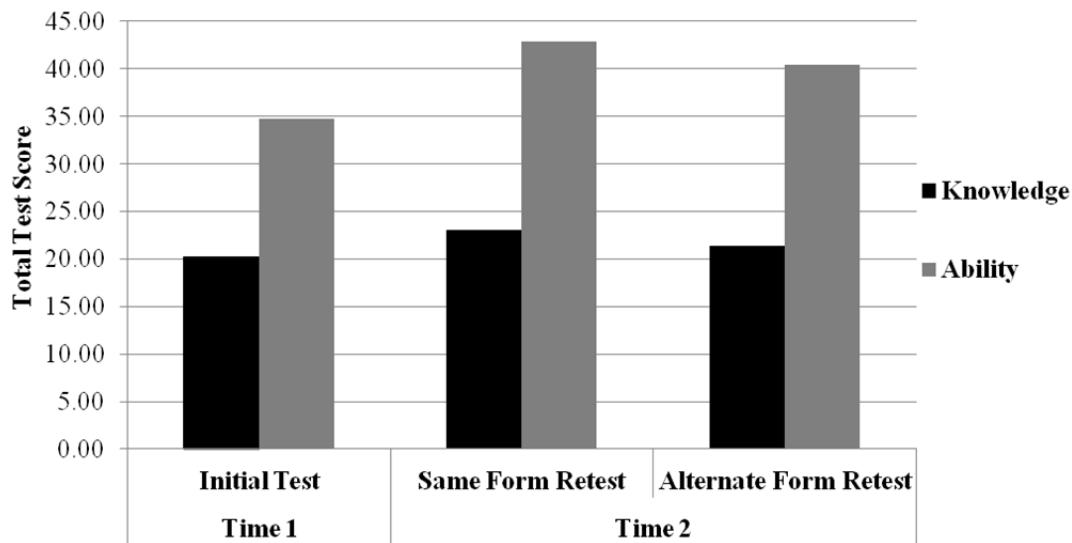
*Figure C.6.* Hypothesis 7a versus 7b: Total test score (percentage) by same versus alternate form retest with corrective feedback or no corrective feedback over time.



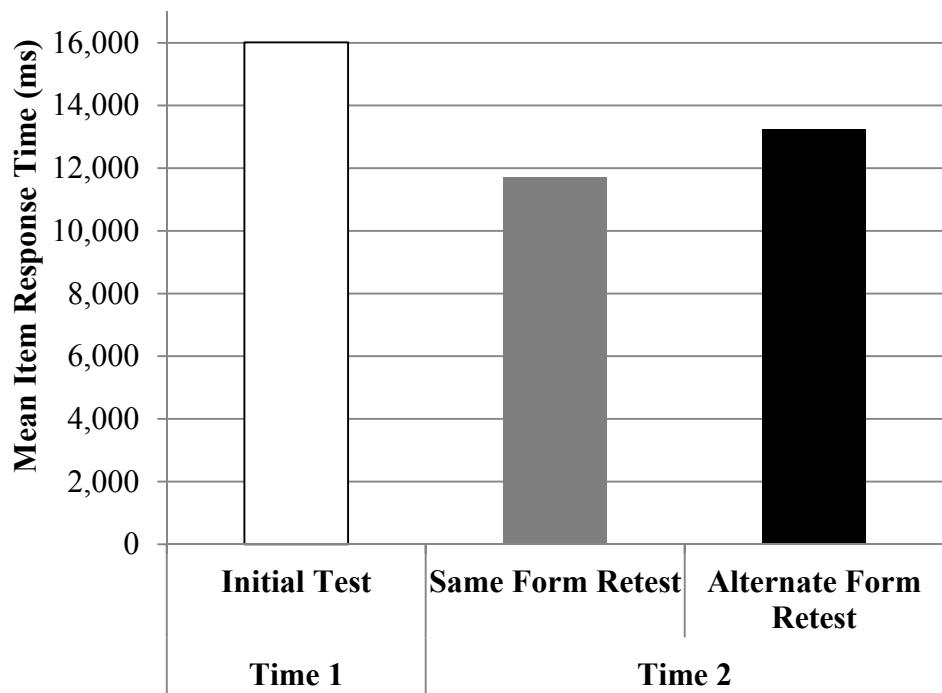
*Figure C.7.* Hypothesis 8a versus 8b: The relationship between working memory and retest score increases and the relationship between GMA and retest scores increases for the same form retest and alternate retest forms by corrective feedback versus no feedback conditions.

## APPENDIX D

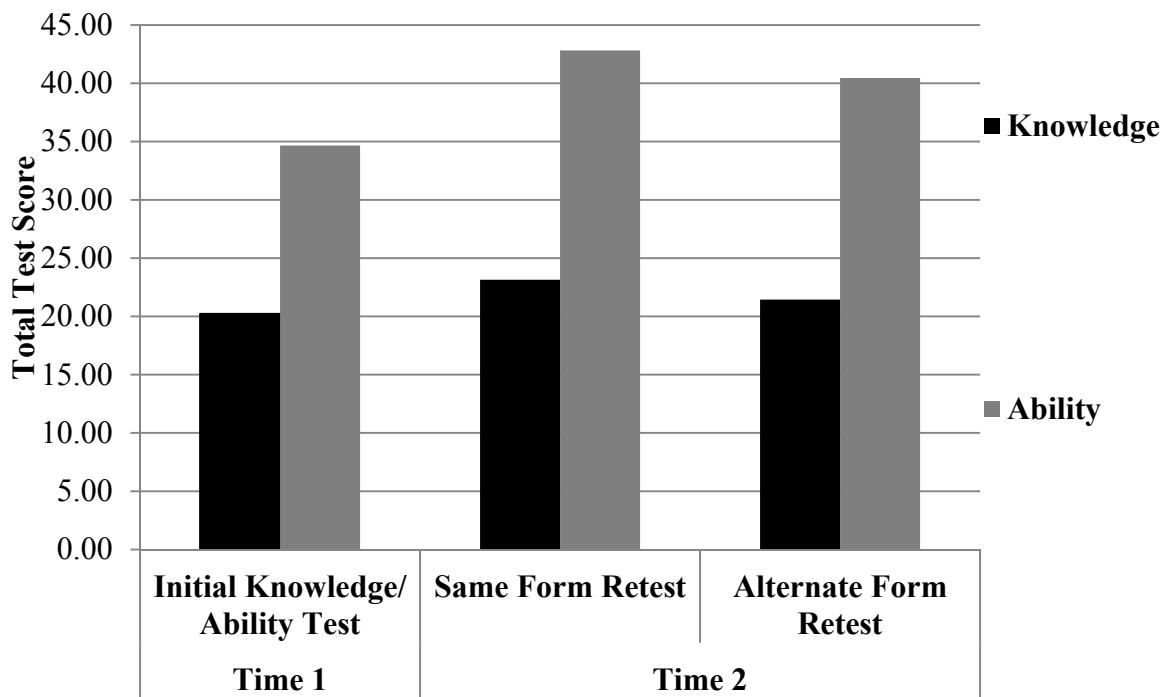
### RAW SCORE FIGURES



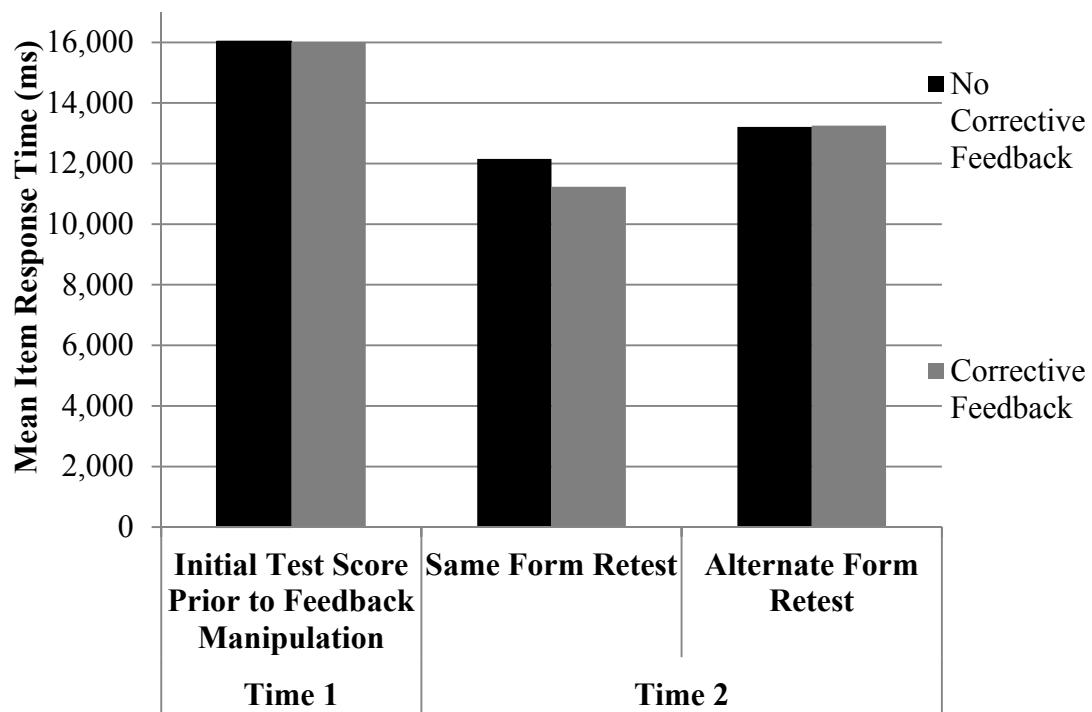
*Figure D.1.* Hypotheses 1 and 3: Total form retest versus alternate form retest over time.



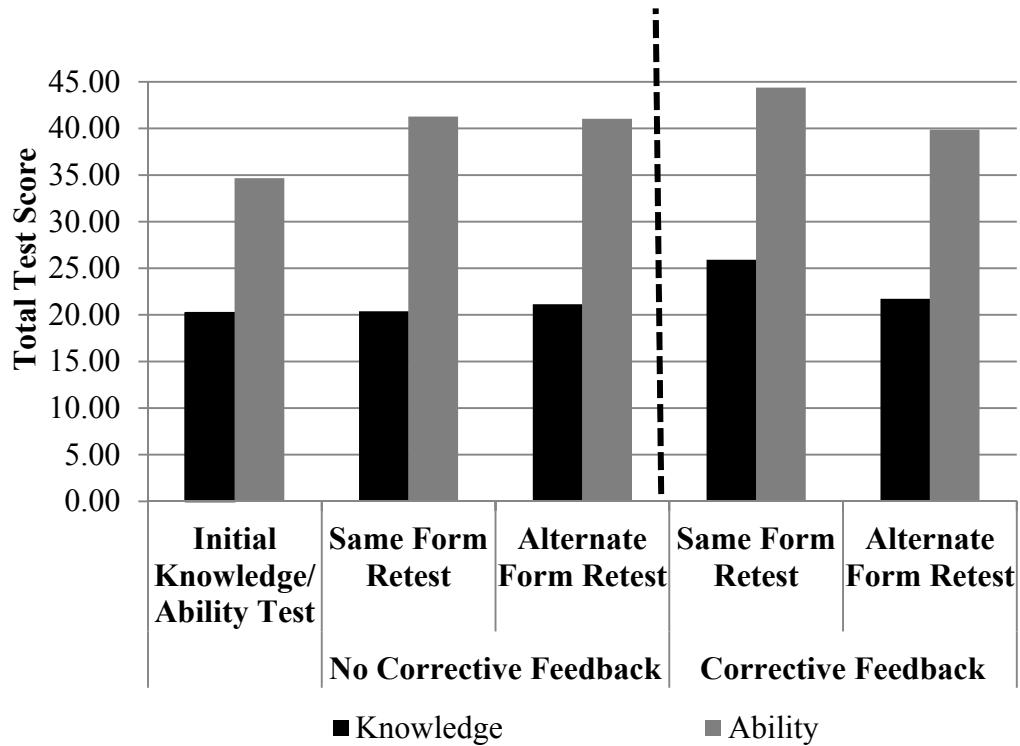
*Figure D.2.* Hypothesis 2: Mean item response time by same form retest versus alternate form retest over time.



*Figure D.3.* Hypothesis 5: Total test score by same versus alternate form retest over time.



*Figure D.4.* Hypothesis 6: Mean item response time by same versus alternate form retest with corrective feedback or no corrective feedback over time.



*Figure D.5.* Hypothesis 7: Total test score by same versus alternate form retest with corrective feedback or no corrective feedback over time.

## APPENDIX E

### RETEST INTERVALS BY CONDITIONS

Table E.1

*Descriptive Statistics for Retest Intervals across Form Manipulations*

		<b>Time 1</b>	
		<b>Form A</b>	<b>Form B</b>
Time 2	<b>Form A</b>	<i>M</i> = 7.52, <i>SD</i> = .88	<i>M</i> = 7.44, <i>SD</i> = .93
	Same Forms (A → A, <i>N</i> = 82)	Alternate Forms (B → A, <i>N</i> = 82)	
		<i>M</i> = 7.79, <i>SD</i> = 1.17	<i>M</i> = 7.62, <i>SD</i> = 1.07
	<b>Form B</b>	Alternate Forms (A → B, <i>N</i> = 89)	Same Forms (B → B, <i>N</i> = 87)

*Note.* Total *N* = 340. Retest interval minimum was 7 days and maximum was 10 days across all conditions.

Table E.2

*Descriptive Statistics for Retest Intervals across Feedback Manipulation*

	<i>M</i>	<i>SD</i>	<i>N</i>
No Feedback	7.60	1.05	172
Corrective Feedback	7.59	1.00	168

*Note.* Total *N* = 340. Retest interval minimum was 7 days and maximum was 10 days across both conditions.