The Testing Effect: An Empirical Examination of the Underlying Influence of Self-Efficacy

Daniel R. Hawthorne

University of Missouri – Saint Louis

A thesis submitted to the Graduate School at the University of Missouri – St. Louis in partial fulfillment of the requirements of the degree of Masters in Psychology

2010

Therese Macan, Ph.D.

Chair

Mark Tubbs, Ph.D.

Stephanie Merritt, Ph.D.

Abstract

This study examines the testing effect (a phenomenon wherein people who learn by testing show less forgetting over time) as a potential organizational training mechanism. Specifically, the impact of self-efficacy and information regarding the testing effect on early and later retention was investigated. Ninety participants completed a computer-based map learning task either through a study protocol or a test protocol, and then responded to measures of post-training self-efficacy and judgments of learning. Half of the participants read information about the testing effect. No support for the proposed hypotheses was found. However, individual difference variables and motivation may provide directions for future research.

The Testing Effect: An Empirical Examination of the Underlying Influence of Self-Efficacy

An overall goal of organizational training programs is the transfer of learning. That is, the skills or information presented in the training program should eventually be used on the job. But before transfer can occur, a trainee has to be able to retain information presented in the program (Goldstein & Ford, 2002). Therefore, strategies that can positively impact a trainee's retention of knowledge should improve the chances that trained behaviors will transfer to the workplace.

The focus of this study was to examine mechanisms that may improve trainee retention of information and learning. A compelling method for improving retention is a phenomenon from the area of cognitive psychology known as the *testing effect*. The testing effect is when "taking a test on material can have a greater positive effect on future retention of that material than spending an equivalent amount of time restudying the material, even when performance on the test is far from perfect and no feedback is given on missed information" (Roediger & Karpicke, 2006b). The present study sought to explore some of the intricacies of the testing effect applied to an organizationally relevant context as well as how individual differences can impact its effects.

Testing Effect

Cognitive psychologists have found—at numerous levels of ecological validity from research laboratories to classrooms—that testing is useful not only as an assessment tool, but also as a tool for learning (Leeming, 2002; Spitzer, 1939; Tulving, 1967). The improved information retention from the testing effect is theorized to operate because of the inherent practice of information retrieval that occurs during testing, and which is not a part of simply studying or reading information (Roediger & Karpicke, 2006a). Further research has provided support for this theory by showing that the testing effect is attributable to the act of retrieval, not just the reexposure of testing materials (Szpunar, McDermott, & Roediger, 2008; Toppino & Cohen, 2009).

Classic testing effect research design involves two major conditions (Roediger & Karpicke, 2006a). The first of these conditions is defined in the present study as the *study* condition. In this condition, participants are given a number of different timed study trials after an initial opportunity to study the material. In contrast, the second condition is known as *testing* and involves the participant having only the initial opportunity to study the material. Following the initial study opportunity, participants are given testing trials, wherein they are instructed to recall as much information as possible.

How the mind inherently processes information can be thought of similar to how fluid runs through a sieve. Information is taken in to the mind and eventually flows out of it, in the form of forgetting. While simplistic, this metaphor is useful for explaining the power behind the testing effect. An important concept is the *forgetting curve*, an exponential curve that shows the rate at which a person forgets information as a function of time (Ebbinghaus, 2003). Testing is theorized to have an effect via its influence on forgetting.

In most applications of the testing effect, the major difference between studying and testing conditions is in the slope of the forgetting curve (see Figure 1; Roediger & Karpicke, 2006b; Tulving, 1967). While studying tends to show stronger immediate retention of information, the rate at which a person forgets information is much steeper with studying. In contrast, testing shows less initial retention, but over time less forgetting of the information occurs, resulting in a shallower forgetting curve. Thus, to continue the previous metaphor of fluid through a sieve, testing increases the viscosity of the fluid, making it flow more slowly, allowing a person to utilize the information for a longer period of time.

Another key variable of interest in testing effect research has been information *retention* (Spitzer, 1939). Information retention is related to the slope of the forgetting curve. That is, as the slope of the curve decreases, participants exhibit more information retention. Another element of testing effect research regarding retention involves the points on the forgetting curve. *Immediate retention* is usually measured a few minutes after the final test trial to establish a base score for retention (Wheeler, Ewers, & Buonanno, 2003). That is, the initial amount of information retained is the maximum amount of information a person can recall in future test trials. Thus, it establishes the top point of the forgetting curve. Other measurements of information retention at future points in time are referred to as *delayed retention* (Wheeler et al., 2003). These assessments are points along the forgetting curve and establish the degree of information retention that a participant exhibits at a particular time.

In the case of the current study, my hypotheses and methods refer to *immediate* retention and *delayed* retention. I have chosen these terms to better represent the methods and procedures I am using.

The Testing Effect in Applied Settings

Decades of research into the testing effect have been compelling and stable (Roediger & Karpicke, 2006a). However, only recently have researchers begun to explore the application of testing effect procedures beyond basic cognitive functioning. Any scientific finding can be considered advancement to other fields, but potential concerns exist with the application of a finding in one field of psychology to another. As Kraiger (2003) advocated, it is important to consider the differences across fields and conduct research to test the effects in other domains.

One difference pertinent to this study is that cognitive psychologists tend to measure learning over a comparatively short time frame, sometimes as short as a few seconds or minutes, whereas I-O psychologists tend to measure learning over a longer time frame, such as several weeks. However, even though it comes from the field of cognitive psychology, the testing effect exhibits promising results that address Kraiger's warning about cross-field implementation. This is because the testing effect exhibits more knowledge retention than studying alone over a more delayed period of time (e.g. as much as 1-4 weeks; Butler, Karpicke, & Roediger, 2007; Roediger & Karpicke, 2006a). The present study sought to examine retention in this manner, establishing a forgetting curve by measuring immediate information retention and delayed information retention (after a 7-10 day period).

Another difference between I-O and cognitive psychology is the criterion of interest (Kraiger, 2003). In general, when learning is measured in cognitive psychology studies, the criterion of interest tends to be recall speed. However, in most I-O psychology research the criterion of interest is job performance or application of learned skills to the job. In a more recent study, the testing effect has been demonstrated using a more applied, performance-based criterion—retention of map information and use of that information in a task—producing results that are more generalizable to an applied setting (Carpenter & Pashler, 2007).

Carpenter and Pashler (2007) extended previous research by applying the testing effect to a visuospatial task using map-learning, instead of using a basic cognitive mechanism (e.g., word lists). The experiment measured participants' accuracy when reproducing maps that had been learned either by studying alone or by testing. Test exhibited improved accuracy of information recall above the study conditions. Although a true applied task—in the sense of I-O psychology—would have been to measure the application of retained knowledge instead of reproduction, their task moves research forward by using a task closer to an applied context.

Even though Carpenter and Pashler (2007) presented promising results, they measured retention after a short time period, testing after only 30 minutes. It is important to examine if these effects remain after a longer time interval. The findings of Roediger and Karpicke (2006b) present evidence demonstrating improved retention—in the form of less forgetting—after a week. The present study examined retention after a 7-10 day interval. I expected similarly that the amount of immediate information retention would be greatest in study conditions, where participants study the material only and are assessed soon after. However, over a longer period of time (7-10 days) participants in the test conditions were expected to show a greater amount of information retention.

Hypothesis 1a – Participants in study conditions will show more immediate information retention than participants in memory test conditions.

Hypothesis 1b – Participants in memory test conditions will show more delayed information retention than participants in study conditions.

Post-Training Self-Efficacy versus Judgments of Learning

Making a learning plan, monitoring one's own learning, and revising learning strategies is the natural process a trainee goes through while learning (Goldstein & Ford, 2002). A by-product of this process is self-efficacy. *Self-efficacy* is the self-referent thoughts that people have regarding their ability with a task or a given set of tasks (Bandura, 1985; Bandura & Locke, 2003; Ford, Kraiger, & Merritt, 2009; Gist & Mitchell, 1992). An application of self-efficacy in training research is *post-training self-efficacy* (PTSE). Post-training self-efficacy is the self-referent thoughts that a person has about a previous training experience. Thus, in this regard, PTSE is a self-reflective attitudinal judgment about how much knowledge an individual believes he or she has learned and reflects a performance orientation regarding knowledge (Cuevas, 2004).

In cognitive psychology literature, the feeling of confidence people report about their ability to retrieve retained information from memory described in the Roediger et al. (2006b) study is referred to as a *judgment of learning* (JOL). A JOL is generally defined as a metacognitive judgment about how capable a person feels about being able to recall previously presented information (Arbuckle & Cuddy, 1969). The two major types of JOLs are immediate JOLs (i.e.: measured immediately after learning) and delayed JOLs (i.e.: measured at some extended period of time after learning; Dunlosky & Nelson, 1992; Nelson & Dunlosky, 1991). There are some differences in how these judgments are made, but both are still a metacognitive test of retrieval fluency (Koriat, Sheffer, & Ma'ayan, 2002; Metcalfe & Finn, 2008). *Retrieval fluency* is defined by two characteristics: the speed with which information is retrieved and the completeness with which information is retrieved (Dunlosky & Metcalfe, 2009). Thus, retrieval fluency is high when a piece of information can be retrieved from memory quickly and is retrieved as completely as possible from the original encoding of the information. This contrasts PTSE, as the former is an attitudinal judgment of what has been learned, whereas JOLs are based more in the accuracy with which a person metacognitively recalls learned information.

Roediger and Karpicke (2006b) reported on how confident participants appeared to be in their ability to recall the information presented within their experiment. This report was presented anecdotally in the discussion section of the article, because of JOL/PTSE was not a focus of the study, nor was assessed directly. They reported that while participants in the test condition of the study felt less confidence in their ability to recall information than participants in the study condition, they were actually able to recall more information. An empirical examination of these variables therefore seems warranted.

In addition, a comparison of post-training self-efficacy and JOLs should be conducted. On the surface, the two operational definitions appear to be similar, separated perhaps only by field of study. In general, JOLs are measured by asking participants to indicate the percentage of information they feel capable of recalling at some future point in time (Koriat, 2007). Post-Training Self-Efficacy is instead measured using a self-report scale which participants indicate the degree of confidence they have in their ability to recall information at a future point in time (Colquitt, LePine, & Noe, 2000; Schwoerer, May, Hollensbe, & Mencl, 2005; Sitzmann, Brown, Casper, Ely, & Zimmerman, 2008; Tziner, Fisher, Senior, & Weisberg, 2007).

Thus, each measure may be examining the same construct, but tapping it differently. Previous research has found significant correlations between metacognitive measures and measures of self-efficacy ($r = .31$, $p < .05$; Ford, Smith, Weissbein, Gully, & Salas, 1998). Research has also explored how metacognition relates to pre-training and post-training self-efficacy (Cuevas, 2004). This experiment found relationships for metacognitive measures with pre-training measures ($r = .48$, $p < .0005$) and post-training measures ($r = .39$, $p < .005$). Thus, because of the apparent similarity of constructs, the present study measured both PTSE and JOLs and examined both as outcomes of participant self-efficacy.

Illusory Self-Confidence and Knowledge of the Testing Effect

One of the major findings from a recent meta-analysis by Sitzmann et al. (2008) on trainee reactions provides a compelling reason to study the accuracy of JOLs or PTSE when testing is used as a learning method. They found that PTSE was a significant predictor of post-training procedural knowledge transfer ($\beta=.37$), and delayed procedural knowledge ($\beta=.48$) even after controlling for declarative knowledge possessed before training. However, Sitzmann et al. did not address if PTSE provided incremental validity above post-training declarative knowledge.

Based on these findings, Sitzmann et al. (2008) recommended that practitioners should measure PTSE as a part of trainee reactions, because it can be considered a "cost-effective predictor of post-training knowledge transfer." However, this relationship can be influenced by factors such as the context of the training and the climate of the workplace (Budworth & Sookhai, 2008). Thus, even though self-efficacy may be useful as a predictor of transfer in certain situations, this wide-sweeping recommendation could prove to be problematic. For example, use of the testing effect during a training program in which PTSE is measured could provide an inaccurate assessment of later knowledge transfer. That is, a negative relationship between self-efficacy and performance may result (Roediger & Karpicke, 2006b). The effect of this proposed outcome on the testing effect will be examined in the present study.

Because self-efficacy is a measure of peoples' perception of their own abilities, this is a critical consideration of the present study. The level of perceived self-efficacy a person possesses in a specific domain has been thought to be a determinant of performance regardless of how much skill is possessed in the domain in question (Bandura & Locke, 2003). For example, a study exploring this effect used a manipulation of self-efficacy on student participants, and then had them solve problems with sentences containing nonsense words (Bouffard-Bouchard, 1990). The results of this study found a main effect of manipulated self-efficacy ($F(1,62) = 24.99$, $p < .001$) on performance. However, the definition of *domain* becomes important when considering Roediger et al.'s (2006b) study design. A domain can be thought of as an activity in which a person can possess a skill, such as playing basketball or chess (Bandura, 1985). Even though the study trials and the test trials are a method in the study, it is possible that these methods are perceived as a domain wherein they believe they possess skill and some degree of control over rather than a method that taps a cognitive process. After all, we all know how to study and may believe we are good at it. In fact, when people find themselves in novel situations—such as an experimental study—for which they have no context, JOLs are inflated and less accurate (Koriat, 2007). This inaccuracy is because they base JOLs on past experiences that may seem similar, although the connection may be illusory (Finn & Metcalfe, 2008).

Furthermore, participants—students in many research studies—likely have implicit associations with the act of "studying" and "testing." Academic learning situations tend to stress a connection between studying as a method of learning information, and testing as a method to evaluate learning (Bangert-Drowns, Kulick, & Kulick, 1991). However, testing in the context of the testing effect is theorized to create stronger connections in the brain through retrieval practice (Roediger & Karpicke, 2006a). In fact, recent studies have presented evidence that testing is a learning method because it taps the brain's natural method of information retrieval and storage (Leeming, 2002; Lyle & Crawford, in press).

If previous findings about illusory self-confidence operate as expected, then giving people information about the testing effect may help to dispel illusory self-confidence in this context (Finn & Metcalfe, 2008; Koriat, 2007; Metcalfe, 2000). Research has found that when people are given information that points out contamination of their subjective experiences, they tend to change judgments associated with those experiences in line with objective information (Strack, 1992). That is, a person moves from a cognitive stance of "what I feel" to "what I know (Koriat, 2007)." Thus, I expected that individuals who were given information about the testing effect would exhibit differences in self-efficacy and information retention than those who did not have information about the testing effect. That is, people are suspected to hold illusions of competency about how much control they may have regarding the ability to learn by studying as opposed to testing. If participants have information about the testing effect, they should exhibit a pattern of self-efficacy more in line with how much information can actually be retained.

Hypothesis 2a - Because information about the testing effect increases one's immediate map task self-efficacy in the study condition, there will be a significant interaction between testing effect information provided and experimental condition, such that the difference in immediate self efficacy between participants in the study condition will be greater when testing effect information is provided.

Hypothesis 2b - Because information about the testing effect increases one's delayed map task self-efficacy in the memory test condition, there will be a significant interaction between testing effect information provided and experimental condition, such that the difference in delayed self efficacy between participants in the testing condition will be greater when testing effect information is provided.

Further, the lack of self-efficacy felt by participants in the test condition reported by Roediger and Karpicke (2006b) in their research could be from this faulty perception. Information about the testing effect could represent a cognitive shift from a belief that testing in this context is a skill domain, to a belief that it is a training method. We already know that altering the context under which training occurs (remedial or advanced) can positively or negatively affect the performance of trainees (Quiñones, 1995, 1997). Because of this reframing, one may be able to improve the performance of participants in a testing effect study by simply changing their contextual perceptions of testing.

Thus, a person's beliefs about the outcomes of behavior or task performance can have an effect on actual task performance (Bandura, 1985; Bandura & Locke, 2003). An *outcome expectation* is defined as "a cognitive judgment of the likely consequence that behavior performance will produce" (Bandura, 1985). Several studies have found a positive relationship between outcome expectations and job/task performance (Geringer & Frayne, 1993; Latham & Frayne, 1989; Lee, 1984). Additionally, of particular interest to the proposed study, a positive relationship between outcome expectations and task performance was reported in a training context (Frayne & Geringer, 2000).

Because the information about the testing effect states that those in study conditions will exhibit less delayed-information retention, it is possible that participants in study conditions could exhibit decreased retention if given information about the testing effect. This could be because they internalize the message of being in a condition where they will exhibit reduced information retention, much like trainees who exhibit lower task performance because they have been given contextual information of being required to be in a remedial training program (Quiñones, 1995, 1997). Thus, even though they would normally have high performance, performance may be deflated because of this information.

Finally, the testing effect states that immediate information retention by those in test conditions is initially lower than those in study conditions (Roediger & Karpicke, 2006a). Because of this, participants in test conditions with information about the testing effect may suffer from similar effects to those described above (Quiñones, 1995, 1997). Given these possibilities, I expected that in the present study the effect of participants being given information about the testing effect would depend on which condition of the experiment (test or study) they were in as well as whether or not information retention is measured immediately or delayed.

Hypothesis 3a –Because information provided about the testing effect increases one's delayed information retention in the memory test condition, there will be a significant interaction between testing effect information provided and experimental condition, such that the difference in delayed information retention between participants in the testing condition will be greater when testing effect information is provided.

Hypothesis 3b –Because of the positive relationship between self-efficacy and task performance, it is expected that the relationship between learning condition and delayed retention will be mediated by self-efficacy.

Individual Differences in Map-Reading and Sketching

Any task will have some component that is addressed by some characteristic of a person better than another person. The same is true for visuospatial tasks. For instance, men tend to perform better at route-finding tasks (Halpern & Collaer, 2005), and those with artistic ability tend to perform better at map-sketching (Taylor, 2005). Additionally, people who have held jobs that require significant knowledge of map-reading and usage, such as taxi drivers tend to exhibit better memory in route-knowledge tasks (Kalakoski & Saariluoma, 2001).

However, many of these variables were not measured or accounted for in previous studies. Since the current study used a visuospatial task, these individual difference variables could be important to examine for influence on information retention or participant self-efficacy. Thus, the current study collected information about possible individual difference variables that have been identified in previous research as having a potential impact on studies of visuospatial information (Halpern & Collaer, 2005; Taylor, 2005).

Method

Participants

Participants were 114 undergraduate students at a mid-sized Midwestern university. Participants were recruited through the psychology department subject pool and through classes in the college of business. If participants were enrolled in participating classes, they were compensated for participation for their involvement with extra credit or psychology pool credit. Of the total 114 participants, 10 did not return for the second experiment session, creating a 9% attrition rate (study-no information = 4, study-information = 2, testing- no information = 0, testing-information = 4). In addition to the 10 participants who did not return, 5 participants were identified as outliers, and 9 failed manipulation checks, yielding a final participant total of 90 participants distributed randomly across the four conditions: study-no information (*n* =18), study-information (*n* =27), testing-no information (*n* =24), or testing-information (*n* =21). For the statistics required in the present study 80% power requires approximately 15 participants per group, thus I attained sufficient power.

Of the final 90 participants, all were undergraduates, of which 59 majored in business (65.6%), 11 majored in psychology (12%), and 20 were from another field of study (22.4%). The average participant was 24.48 years of age (*SD* = 6.12), with the youngest being 18 years of age and the oldest being 54 years old. With regard to year in college, 31 participants were seniors (34.4%), 48 were juniors (53.3%), 6 were sophomores (6.7%), and 5 were freshmen (5.6%). There were 44 males (48.9% and 46 females (51.1%). There were 37 Asians in the sample

(41.1%), 42 were white or Caucasian (46.7%), 7 were black (7.8%), 1 identified as American Indian (1.1%), and 3 were multi-racial or some other race (3.3%).

The low attrition rate between time 1 and time 2 was accomplished by using principles of persuasion based around social proof (Cialdini, 2001). Each participant was asked during the experiment, “Will your return for the second phase of the experiment?” and asked to respond with a “yes” or “no” answer. Only 3 participants responded that they would not return for the second phase of the experiment. Additionally, when participants left the first phase of the study, they were asked to indicate the time and date of the session they would attend to complete the second phase of the study on a sign-out sheet.

Materials

Map-Learning Task. The map-reading task was adapted from Carpenter and Pashler’s (2007) stimuli. It is a software-based task that allows participants to learn a map by either testing or studying a map. The learning task was built in Adobe Captivate. The learning task uses a testing function as a manipulation for the testing effect, or in the absence of the testing effect, simply gives the participant a timed period to study the map. The map that participants viewed had 12 iconic feature representations such as a school and an airport. The testing function presented participants with a version of the map with only 11 features, so they could identify the missing feature from memory. Additionally, the colored maps were tested using Adobe Photoshop to ensure that participants with Protanopia and Deuteranopia (two forms of color-blindness) would be able to accurately see all map features.

Testing Effect Information. Approximately half of the participants received information about the testing effect (see Appendix C). This information was written so as to be understood by individuals who do not possess a background in psychology. Using the Flesch Reading Test scale (40.3) and the Flesch-Kincaid Grade Level scale (11.6), the reading level of the passage should be appropriate for college freshmen and above (Flesch, 1948; Kincaid, Fishburne, Rogers, & Chissom, 1975). The passage was pilot-tested. All participants in the pilot test (N = 14) were able to understand and correctly restate what the testing effect is.

Scoring Participant Map Drawings. The scoring method for the map drawings was adapted from the original Carpenter and Pashler (2007) study. For each map feature that was placed in the correct location, one point was awarded. However, one point was deducted for each extra feature placed by participants. For example, if a participant drew a factory (a feature not on the original map) on his or her map, this resulted in the loss of one point. Additionally, two types of accuracy (absolute accuracy and relative accuracy) were determined. *Absolute accuracy* relates to the correctness of feature placement without regard to other map features. This was operationalized as the feature being placed in the correct quadrant (NE, NW, SE, SW) of the map. As a part of absolute accuracy, liberal and stringent criteria were considered in scoring. A feature was placed with *liberal absolute* (LA) accuracy if it was placed in one of the correct quadrants that it appears in. For instance, if a river that runs through multiple quadrants was placed in at least one of the correct quadrants then that feature had LA accuracy. Additionally, *stringent absolute* (SA) accuracy was achieved if the feature was correctly placed in all of the quadrants it should appear in. For example, consider the same river as above. This feature would only achieve SA accuracy if it was drawn to pass through every quadrant that it originally appeared in.

Another form of accuracy that was considered in scoring was relative accuracy. *Relative accuracy* was defined as feature placement which was correct in relation to the other features around it. As above with absolute accuracy, there was also liberal relative accuracy and stringent relative accuracy. *Liberal relative* (LR) accuracy was achieved when a feature was placed in such a way that it appeared in correct relation to at least one of the other features around it. For example, if a house appeared on an original map with an airport to the north, a phone booth to the west, and a church to the east, as long as the house appeared in correct orientation with one of the other features, then LR accuracy was achieved. *Stringent relative* (SR) accuracy was achieved when a feature was placed in such a way that it appeared in correct relation to all of the other features around it. For example, consider the same house as above. This house must appear in correct orientation to all of the other features (airport, phone booth, and church), not just one of the other features, to achieve SR accuracy.

In addition to previously mentioned scoring of maps, one point was awarded for the LA accuracy, SA accuracy, LR accuracy, and then SR accuracy. Each type of accuracy had a score range of 0-18 points. A complete list of scoring decisions is listed in Table 1 and descriptive statistics by accuracy type can be found in Tables 2 and 3.

Rater Training & Interrater Agreement. The two raters were trained and given practice with sample maps upon completion of training. Raters were trained on the four types of accuracy (LA, SA, LR, and SR), and then scored 10 test maps to examine if they could score each map correctly. A "correctly" scored map was defined as adhering closely to the criteria for map reproduction accuracy. Finally, raters were asked to score maps created by participants from the pilot test, to check further for understanding of the four types of accuracy.

Each participant's map drawing was scored by the two trained raters who remained blind to the conditions. Interrater agreement was assessed by calculating r_{wg} (James, Demaree, & Wolf, 1984; Kozlowski & Hattrup, 1992). As can be seen in Table 4, most of the r_{wg} scores for each type of scored accuracy were all high (r_{wg} > .70). The score for time 1 SR accuracy is slightly low (r_{wg} = .68). However, time 1 SR accuracy approaches the suggested cutoff presented by James, et al. of .70 or higher (1984). Thus, these scores are considered to be in agreement and were averaged.

Measures

Judgment of Learning Measure. Participant JOL's were measured with four items that asked participants to rate what percentage (1-100%) of the presented information they felt they would be able to draw (see Appendix B). Thus, higher numbers indicated a higher JOL. Reliability for immediate JOL was acceptable (α = .82), and delayed JOL reliability was good (α = .90). The items measured immediate, as well as delayed JOLs during the first phase of the study, since the delayed information retention was measured during the second phase. During the second phase of the study, JOLs were assessed again for the map information that was about to be reproduced. Research has provided evidence that more accurate (or slightly inflated) judgments are provided by item-by-item JOLs, whereas aggregate JOLs tend to result in underestimations of actual learning (Koriat, 2007). Thus, the present study measured item-by-item JOLs and aggregate JOLs to account for both possibilities.

Self-Efficacy Measure. The present study measured immediate and delayed post-training self-efficacy (PTSE) using a 4-item measure asking participants to rate their confidence about ability to recall learned information (see Appendix B). Reliability for immediate PTSE needs improvement (α = .71), but delayed PTSE reliability was acceptable (α = .81). Participants rated PTSE on a 5-point Likert scale (1 = *not confident at all* to 5 = *completely confident*). To ensure equivalence of PTSE items and JOL items, PTSE items also measured item-by-item and aggregate PTSE. In addition, PTSE items assessed immediate and delayed PTSE during the first phase of the study, and once again during the second phase of the study for the same reasons listed above.

Motivation Measure. The link between self-efficacy and motivation is generally accepted in psychology research (Bandura, 1985), but motivation was not a focal variable of the current study. Nonetheless**,** participant motivation was assessed as an exploratory measure at different points throughout the study. This 2-item measure gauged the effort and perseverance that participants attempted to expend in the course of the study (see Appendix B). Reliability for this measure was good ($\alpha = .90$). Participants made ratings on a 5-point Likert scale ranging from 1 = *strongly disagree* to 5 = *strongly agree*.

Demographic Measures. Previous research has found that some individual difference variables, such as gender (Taylor, 2005), age, geography knowledge, and artistic ability (Halpern & Collaer, 2005) can affect how people perform visuospatial tasks. Therefore, items were used to assess these and other individual difference variables that could have had an impact on the results (see Appendix A).

Study Variables

The design of the current study incorporated 2 independent variables (IV) and 2 repeated measures dependent variables (DV). The two IVs were information condition (information versus no information) and learning condition (study versus test). The two repeated measures DVs were information retention (measured as map performance) and self-efficacy (measured as JOL and PTSE). Dependent variables were measured immediately (30 minutes post-learning) and delayed (7-10 days post-learning).

Procedure

The procedure was adapted from the original Carpenter and Pashler (2007) study to provide a replication of their work with an extension of their basic study. Carpenter and Pashler's focus was on the confirmation of increased participant retention of visuospatial information using the testing effect, whereas the current study extends this study by examining the role of self-efficacy. Additionally, Carpenter and Pashler explored test and study processing within subjects by utilizing two maps to create counterbalancing conditions. However, because the exploration of the testing effect in the current study used a between-subjects design, counterbalancing was not relevant.

Participants were randomly assigned to one of the two learning conditions (study or test) and one of the two information conditions (information about the testing effect or no information). Upon arrival, participants completed informed consent forms, which contained brief information about the study. This information explained to participants that they were part of a study involving map-reading, and that they would be tested over both maps without clues or prompting to help them, and that the study involved two phases (7-10 days apart). Participants in the information condition also received and read the information about the testing effect (see Appendix C).

All participants completed their tasks individually in a semi-private setting, so they would not be disturbed while performing the tasks. Prior to beginning the experiment, all participants spent approximately 60 seconds performing a practice task using colored shapes instead of map features. This practice task allowed participants an opportunity to see an example of what they would experience and to familiarize them with the operation of the software.

Participants in the study condition were given 20 seconds to study the map and all of the 12 features on the map (see Figure 2). Automatically following this, the screen went blank, and displayed instructions that upon pushing the space bar to begin, they would be able to study the map for an additional 100 seconds. Upon striking the space bar, the extra 100 seconds of study time began. Thus, for the study condition, participants were given a total of 120 seconds to study the map.

In the test condition, participants were given an initial 20 seconds to study the entire map with all 12 features. Next, they were given instructions telling them that the following screens (after pressing the indicated key) would test them over the information they had just seen. Upon pressing the key, they were shown a version of the map where one of the 12 features was missing. Participants were told to mentally visualize where the feature should be. When ready, they were prompted to press the space bar to make the missing feature appear. Once the feature appeared, a message appeared asking the participant to note whether or not their choice was correct or incorrect by pressing "Y" or "N." This was not scored, but served to help participants mentally note their performance. A one-second blank screen appeared, and then the same process was repeated with each map feature. Once the participant had completed all 12 map features, the process completed, to match the amount of time in the study experimental condition. To verify, a 2 (study:test) x 2 (information:no information) factorial ANOVA was conducted on the number of seconds each participant took to complete the procedure. There were no significant main effects by condition, F (1,90) = .01, p = .92, or information, F (1,90) = .11, p = .74, or interaction, F (1,90) = .94, p = .33, in time spent on task between any of the conditions: Study- no information (M = 628.68, SD = 294.21) , study- information (M = 565.60, *SD* = 222.62), test-no information (*M* = 586.60, *SD* = 161.80), test-information (*M* = 617.59, *SD* = 244.31).

Upon completion of either the study or the test conditions, all participants completed the JOL and PTSE measures (see Appendix B). To provide cognitive interference and conform to Carpenter and Pashler's (2007) original study design, participants then spent 20 minutes performing a visual attention task. After this, each participant was given a blank sheet of paper and told to reproduce the map from memory. Following map reproduction, participants filled out the motivation measure (see Appendix B) and the demographics measure (see Appendix A). Participants were also given an appointment to complete the second phase of the experiment 7-10 days later.

When participants returned for the second phase of the experiment they filled out the motivation, JOL, and PTSE measures. Finally, participants were given a blank sheet of paper and told to reproduce the map.

Finally, participants were given three items to answer that served as a manipulation check (see Appendix D). It was proposed that participants who have received information about the testing effect would be able to answer these questions more accurately. If they could, this indicated that the testing effect information manipulation was successful between conditions. After answering these items, participants were debriefed about the study and thanked for their time.

Results

Descriptive statistics including means, standard deviations, correlations, and skewness/kurtosis are reported for each key variable in Table 5. While the main study variables are normally distributed, some of the individual difference variables show some degree of skewness (i.e. age, high school and college art classes, and high school and college geography classes). The non-normality in age is likely due to the undergraduate sample. Additionally, the non-normality seen in the other variables is likely due to a smaller number of people expected to have experience in those areas. For example, because the sample population was primarily from the business and psychology departments, the number of participants who would have taken a number of art or geography classes in high school and college would be minimal, which would exhibit skewness.

Univariate and multivariate outliers were examined before hypothesis testing. Four participants' data were identified as univariate outliers. Upon further examination, these participants' data varied systematically and were removed. For example, one participant assigned sequential JOL scores (e.g.: 1, 2, 3, 4, 4, 3, 2, 1), indicating a lack of attention or motivation to perform the task. Multivariate outliers were also investigated using mahalanobis' distances. One participant was identified as a multivariate outlier and removed from analyses because she recorded 600 months working in a map-reading job, although this was not possible given her age. Levene's tests for equality of variances indicated homogeneity of variance. Linearity of the dependent and the independent variables was tested. Using Cook's distances no cases were found to be influential (cutoff of 1.0; Cohen, Cohen, West, & Aiken, 2003). The data appear to be linear and homoscedastic. Thus, multivariate normality can be assumed.

Manipulation Checks. To examine how well participants had recognized the testing condition manipulation they were asked, "During the map-learning task, was the map you looked at shown with features removed?" Four participants in the study condition indicated that they had seen the map with some features removed, even though they should not have. Four other participants from the testing condition stated that they had not seen a map with features removed.

Only one of these four participants from the testing condition mentioned “the map features changing” in another debriefing question. Thus, this one participant’s data was retained, but the remaining seven participants’ data were removed from subsequent analyses. Results of data analyses were not significantly different with or without these participants’ data.

Another manipulation check was performed to assess if participants had recognized the presence of the testing effect information. Participants were asked: “Did you read information about how to remember more information by quizzing yourself?” Seven participants (three from the information condition, and four from the no information condition) failed this manipulation check. However, a follow-up question asked, “How is quizzing yourself over information like learning to play basketball?” This question referenced the specific metaphor used in the testing-effect information provided to participants (see Appendix C). Only two of the seven participants (one from the information condition and one from the no information condition) failed this follow-up question. These two participants’ data were excluded from subsequent analyses. In general though, the manipulations worked as intended

Individual Difference Variables. The individual difference variables of age, gender, artistic ability, and map-reading ability/familiarity were examined for their relationship with key study variables and for possible ancillary analyses (see Table 5). Consistent with previous research into visuospatial learning, participant knowledge of art and geography had significant relationships with variables relating to self-efficacy. Thus, variables relating to artistic ability and map-reading knowledge were considered in ancillary analyses of the data, following hypothesis testing. Also consistent with previous research into visuospatial learning, gender had significant relationships with delayed information retention variables. Age did not have significant relationships with any key study variables.

In addition, due to the concentration of participants who had majored in business and psychology, a factorial ANOVA was performed with the key variables in the study. The only variable that business students and psychology students exhibited significant differences on was delayed retention, F (1,68) = 4.08, $p < .05$, such that psychology students exhibited significantly better delayed retention. This relationship had a small effect (*partial* $\eta^2 = .06$). Other majors of study had no significant effects on the key variables in the study.

Hypothesis Testing. The current study attempted to replicate the testing effect. Thus, two independent samples t-tests using learning condition (study:test) as the IV were conducted on information retention. To test hypothesis 1a, immediate information retention was used as the DV. Similarly, delayed information retention was used as the DV to address hypotheses 1b. These t-tests were conducted using only the no information conditions to provide a replication of previous research.

The mean immediate retention for participants in the study condition (M = 35.49, SD = 11.74) was not significantly different from that of participants in the test condition (M = 37.29, SD = 13.97). Thus, hypothesis 1a was not supported, such that information retention for those who learned through studying did not exhibit more immediate retention than those who learned through testing, $t(88) = -.66$, $p = .51$. Additionally, the mean differences between delayed retention for participants in the study condition (M = 27.60, SD = 11.87) and test condition (M = 29.09, SD = 12.65) were not significantly different, $t(88) = -.58$, $p = .57$. Thus, hypothesis 1b was not supported.

Researchers have found in previous studies that participants' ratings of JOL and PTSE are highly related (Cuevas, 2004; Ford et al., 1998). Additionally, there is a distinction between these two concepts that provides logic for further exploration. A JOL is a metacognitive evaluation or internal test by a person about information that has been learned. This is different from PTSE which is an attitudinal feeling of confidence about future recall of learned information. This provided rationale for the current study to examine these two constructs for similarity or difference to direct subsequent hypothesis testing. Thus, the current study measured both JOL and PTSE to explore the dimensionality of the two constructs.

All of the aggregate JOL and PTSE measures are highly correlated with one another (see Table 5). However, the variables that are suspected to be measuring the same construct are of specific interest to the current study (e.g.: immediate aggregate PTSE and immediate aggregate JOL). These relationships are some of the highest that are represented in Table 5. Immediate aggregate PTSE is significantly related to immediate aggregate JOL (r = .57, p < .01). Likewise, delayed aggregate PTSE and delayed aggregate JOL are also highly correlated (r = .70, p < .01). Additionally, participant ratings of PTSE and JOL at the time 2 of the study are also significantly related (r = .74, p < .01).

Due to the high correlations of each variable and theoretical background (Cuevas, 2004; Ford et al., 1998), a confirmatory factor analysis (CFA) was conducted using LISREL 8.6.2, to test possible models suggested by the data (see Table 6). The first model tested was a one-factor model (see Figure 3) that would suggest that immediate and delayed measurements of self-efficacy (i.e.: PTSE and JOL). This model exhibited poor fit to the data (RMSEA = .24, SRMR = .17). The second model tested was a two-factor model that would suggest a factor structure based on PTSE and JOL being associated, such that the two common factors would be measures of immediate and delayed self-efficacy (see Figure 4). This model exhibited poor fit to the data (RMSEA = .24, SRMR = .17). The third model tested was a two-factor model that suggested a factor structure based on PTSE and JOL being separate factors (see Figure 5). This model also did not exhibit a good fit to the data (RMSEA = .16, SRMR = .11). The fourth model was a four-factor model focused on JOL and PTSE being separate factors and delineated by immediate and delayed self-efficacy (see Figure 6). Similar to previous models, the fit indices were not acceptable (RMSEA = .16, SRMR = .10). The fifth and final model was a four-factor model focused on PTSE and JOL being associated, such that each the factor structure would be associated with map features (see Figure 7). This model also exhibited poor fit with the data (RMSEA = .24, SRMR = .17). Examination of χ^2 difference tests for model fit revealed that model fit improved significantly from Model 2 to Model 3, $\chi^2(1, N = 90) = 237.28$, $p < .05$, and from Model 3 to Model 4, $\chi^2(5, N = 90) = 20.50$, $p < .05$, with Model 4 exhibiting the best fit of all models. However, it should be noted that all models exhibited overall poor fit to the data based on examination of the fit indices.

To understand more thoroughly the results of the CFAs, an exploratory factor analysis (EFA) was conducted using SPSS 15. The EFA was conducted on all PTSE and JOL variables using maximum likelihood estimation with direct oblimin factor rotation. Kaiser normalization, scree plots, and parallel analysis suggested 4 factors. Factor loadings exhibited a relatively clear grouping of delayed JOL and delayed PTSE items under a single factor. Additionally, immediate JOL and PTSE for map icons loaded under another factor. However, all other items exhibited numerous cross-loadings.

After examination of the high correlations between PTSE and JOL and the EFA and CFAs, the true relationship between PTSE and JOL is still unclear. Thus, all hypotheses testing involving self-efficacy are examined using PTSE and JOL data separately.

To analyze hypotheses 2a and 2b, two separate 2 (study:test) x 2 (information:no information) factorial ANOVAs were conducted on the two measures of map-task self-efficacy (immediate and delayed). The General Linear Model (GLM) function for ANOVA in SPSS 15.0 was used to address the differences in sample size between IV and moderator conditions (Frazier, Tix, & Barron, 2004; Tabachnick & Fidell, 2006).

Hypothesis 2a suggested that testing effect information moderates the relationship between learning condition and immediate map-task self-efficacy, such that the inclusion of testing effect information will produce a significant simple main effect on immediate map-task self-efficacy. Using JOL as the outcome variable, there was no significant main effect of study/test condition, $F(3,86) = 3.02$, $p = .09$, *partial* $\eta^2 = .03$, or of information, $F(3,86) = .87$, $p = .35$, *partial* $\eta^2 = .01$. No interaction of learning condition and information condition was found, $F(3,86) = .08$, $p = .77$, *partial* $\eta^2 = .00$. However, using PTSE as the outcome variable, there was a significant main effect of study/test condition, $F(3,86) = 5.51$, $p < .05$, *partial* $\eta^2 = .06$, such that those who learned from testing ($M = 2.24$, $SD = .80$) exhibited higher self-efficacy than those who learned through studying ($M = 2.67$, $SD = .83$). There was no significant effect of information, $F(3,86) = .45$, $p = .51$, *partial* $\eta^2 = .01$, or interaction, $F(3,86) = .12$, $p = .73$, *partial* $\eta^2 = .00$. Thus, hypothesis 2a was partially supported when using PTSE.

Examination of hypothesis 2b using ANOVA analysis of JOL revealed similar results to hypothesis 2a. There was no significant main effect of study/test condition, $F(3,86) = .26$, $p = .61$, *partial* $\eta^2 = .00$, or of information, $F(3,86) = 1.99$, $p = .23$, *partial* $\eta^2 = .02$, and no significant interaction, $F(3,86) = .00$, $p = .96$, *partial* $\eta^2 = .00$. Using PTSE as the outcome variable, however, there was a significant main effect of study/test condition, $F(3,86) = 4.33$, $p < .05$, *partial* $\eta^2 = .05$, such that those who learned from testing ($M = 2.67$, $SD = .80$) exhibited higher self-efficacy about delayed information retention than those who learned through studying ($M = 2.24$, $SD = .83$). Additionally, there was a significant main effect of information, $F(3,86) =$

5.03, $p < .05$, *partial* $\eta^2 = .08$, such that those who were not provided with information about the testing effect ($M = 3.45$, $SD = .89$) exhibited higher self-efficacy about delayed information retention than those who were provided with information ($M = 2.96$, $SD = .97$). No significant interaction of learning condition and information condition was found, $F(3,86) = .00$, $p = .98$, *partial* $\eta^2 = .00$. Given that analysis of JOL produced non-significant results, but PTSE produced a significant main effect of condition and information provided to participants, hypothesis 2b was partially supported.

Hypothesis 3a suggested an interaction between learning condition and information condition, such that test condition will produce better delayed information retention than study condition when testing effect information is provided. Thus, the effect between learning condition and information condition as it related to immediate retention was examined using a 2x2 between groups factorial ANOVA. There was no main effect of study/test condition, $F(3,86) = .19$, $p = .66$, *partial* $\eta^2 = .00$, or of information, $F(3,86) = 1.28$, $p = .26$, *partial* $\eta^2 = .02$. No interaction of learning condition and information condition was found, $F(3,86) = 1.07$, $p = .31$, *partial* $\eta^2 = .02$. Thus, hypothesis 3a was not supported.

Hypothesis 3b stated that the relationship between learning condition and delayed retention is mediated by self-efficacy. Because of the previous discussion concerning PTSE and JOL, these analyses included both variables. Conforming to path analytic theory, a linear regression was performed on the path from learning condition to JOL and was non-significant ($t = -.26$, $p = .80$). The next step in path analytic theory was to check the relationship between JOL and delayed retention, and this relationship was non-significant ($t = -.39$, $p = .70$). Because both of these relationships are non-significant, there is no need to proceed further with examination of self-efficacy in the form of JOL as a potential mediator (see Figure 8).

To examine PTSE as a potential mediator for the relationship between learning condition and delayed retention, path analytic theory was utilized again. Thus, a linear regression was performed on the path from learning condition to PTSE and was non-significant (t = 1.67, p = .10). The next step in path analytic theory was to check the relationship between JOL and delayed retention, and this relationship was non-significant (t = -.22, p = .83). Once again, because both relationships are non-significant, there is no need to proceed further. Hypothesis 3b was not supported (see Figure 9).

Ancillary Analyses. Applied psychology has examined some important elements of individuating information that may explain the results of the current study. A key question to be answered is, "why was the robust testing effect not displayed in this study?" This question is especially critical, given that the current study used a design adapted from Carpenter and Pashler (2007), wherein the testing effect was clearly displayed (partial η^2 = .80-.84). One possible explanation may stem from the motivation of the participant learners in this study. Research has found that motivation to learn has the greatest impact on training outcomes (Tziner et al., 2007), with low motivation degrading learning comprehension and memory for the learned material(Ackerman, Kanfer, & Goff, 1995). Thus, one possible reason for why the current study failed to find the testing effect may be the level of learner motivation.

Unlike the current study, Carpenter and Pashler (2007) used a motivation manipulation in which they told participants that the top one-third of scorers would be paid $10. It is likely that this manipulation led to high motivation among participants to learn (although they did not explicitly measure participant motivation). If the testing effect is a learning method, it is possible that participants who are unmotivated to learn may not benefit from it. Cognitive psychology, however, is more focused on the mechanical elements of the mind and less on the individuating differences a person may express. Thus, it is not surprising that throughout the testing effect literature inferences are made about the impact that individual motivation may have, yet no study explicitly measured motivation in regard to the testing effect (Roediger & Karpicke, 2006a) including the study conducted by Carpenter and Pashler (2007).

The current study asked participants to self-report their motivation to work through the exercises of the study with two items of effort and persistence (5-point scale; 1 = “strongly disagree” to 5 = “strongly agree”). Regarding effort-based motivation, participants on average indicated moderately strong disagreement (M = 1.87, SD = .80) with the statement, “I have tried my hardest to learn what was presented to me”. Likewise, participants on average indicated moderately strong disagreement (M = 1.70, SD = .66) with the persistence-based motivation statement, “Even when the task was difficult, I did my best to perform the task”. Statistics to examine how motivated participants differed from unmotivated participants were attempted, however only three participants gave ratings greater than three on each of these variables rendering an analytic comparison impossible. Thus, since motivation may be a mediating or moderating variable between learning condition and information retention, future research should focus on comparing motivated versus unmotivated learners using the testing effect to explore this relationship.

A further piece of evidence about this possible relationship comes from findings during ancillary analyses of participants’ individual differences. While people need to be motivated to engage in a learning condition, the use of previously acquired knowledge is much more innate and automatic (Howell, 2003). In the case of the current study, the number of college geography classes taken by a participant was found to be a significant predictor of delayed LA accuracy, β = .83, $t(41) = 1.72$, $p < .10$, SA accuracy, β = 1.38, $t(41) = 2.17$, $p < .05$, LR accuracy, β =

1.17, $t(41) = 2.81$, $p < .10$, and SR accuracy, $\beta = .59$, $t(41) = 2.50$, $p < .05$, regardless of whether one was in the testing or study condition. Additionally, geography classes explained a significant amount of variance in delayed LA accuracy, $R^2 = .07$, $F(1, 40) = 2.96$, $p < .10$, SA accuracy, $R^2 = .11$, $F(1, 40) = 4.72$, $p < .05$, LR accuracy, $R^2 = .07$, $F(1, 40) = 2.81$, $p < .10$, and SR accuracy, $R^2 = .14$, $F(1, 40) = 6.24$, $p < .05$. Thus, even though participants were unmotivated to engage in the learning process, they still naturally performed the task using their previously learned skills.

Discussion

The testing effect has shown strong effects in a number of research contexts and classroom applications, and shows promise in other applied settings. The present study failed to find support for any of the hypotheses regarding this phenomenon. It also only found partial support for the implications of providing information to participants to increase participant self-efficacy. Nonetheless, the present experiment found information that could have implications for the field of research on the testing effect. Specifically there are important findings that point to learner motivation being a critical component of any learning strategy.

A key difference between the methodology used in the current study and previous testing effect literature is that the current study used a method of testing that provided immediate feedback. That is, participants in the testing conditions self-scored their responses while testing. Carpenter and Pashler (2007) used this same method, but did not measure learner self-efficacy. Learners who are provided immediate feedback about their performance are more likely to feel more confidence about their ability to recall correct responses in the future (Butler et al., 2007). So, participants in the testing condition were able to feel more confident about their ability to respond correctly if tested immediately, because they received self-scored feedback about their correct and incorrect responses.

Another difference between Carpenter and Pashler's study (2007) and the current study is that the current study used a between-subjects design opposed to Carpenter and Pashler's within-subjects design. A between-subjects design is more vulnerable to error variance introduced in the form of unaccounted-for individual differences. Thus, the current study's replication of the testing effect may have been affected by this extra error variance. However, due to the current study's focus on measuring information retention at two different time points, a between subjects design was necessary to avoid carryover practice effects.

A final difference between Carpenter and Pashler's study (2007) and the current study is that analyses and means were based on raw scores from participant map reproduction in the current study (see Table 2). However, Carpenter and Pashler presented their map scores as proportions. Thus, the current study's raw scores were transformed into proportion scores for comparison (see Table 6), and analyses were performed again, with no significant main effects.

Similarly, an explanation for the pattern of results in the current study for participant self-efficacy for delayed retention can be found in the previous testing effect literature. Roediger and Karpicke (2006b) noted that participants who learned by studying reported feeling more confident about their ability to remember information if tested later. Thus, the information reported anecdotally by Roediger and Karpicke was supported with data in the current study.

The results of delayed self-efficacy for participants who were provided with information about the testing effect were opposite of the proposed direction. That is, those who were given information about the testing effect were not as confident about their ability to recall and reproduce the maps in the future. One of the debriefing questions that participants were asked provides some insight into a possible reason for these results. Participants were asked if they had a test in the future, would they study by a traditional study method or use testing to prepare. It was assumed that if the information provided about the testing effect was effective that participants would advocate the testing method more than those who had not been provided the testing effect information. Even though more participants who were given information advocated testing as a personal learning method (64.6%) than participants who did not receive this information (45.2%), this difference was not significantly different $\chi^2(1, N = 90) = 0.90$, $p = .39$. Thus, a likely explanation for this result is that the testing effect information manipulation was not strong enough to circumvent the typical response to testing (i.e. opposite to the hypothesized direction of relationship), which is to believe that studying is a superior method.

Research Implications

From previous research, there appears to be little question that the testing effect produces the effects that it is purported to (Gates, 1917; Roediger & Karpicke, 2006a; Spitzer, 1939; Tulving, 1967). However, the current study found that two individual difference variables (motivation and previous ability) may change how the testing effect manifests.

A key variable of interest is trainee motivation. Motivation predicts how much effort and perseverance trainees will have in a given training program (Quiñones, 1997). A trainee’s level of motivation shows clear, positive connections with the inherent success of a training program (Goldstein & Ford, 2002). Quite simply, trainees with higher pre-training motivation are known to significantly outperform less motivated trainees (Tannenbaum & Yukl, 1992). The link between motivation and self-efficacy is also well-known and well-studied (Bandura, 1985; Bandura & Locke, 2003). Previous testing effect studies have provided motivation manipulations such as monetary incentives for high performers (Carpenter & Pashler, 2007), or presented the experiments in educational contexts where participants' performances could have an impact on academic achievement (Leeming, 2002; Lyle & Crawford, in press). Thus, given that the current study's attempt at replication of the testing effect may have been impacted by the low motivation of the participants, future research should explore the testing effect as a learning method and examine how people learn differentially with and without motivation to perform. Additional support for this perspective can be seen in Table 6, by comparing the mean proportion scores for immediate information retention from the current study and those from Carpenter and Pashler's (2007). Carpenter and Pashler's motivated participants outperformed the current study's unmotivated participants in mean proportion of information recalled. Specifically, in the future more data should be collected from a highly motivated sample to compare with these data to examine the impact of motivation on the testing effect.

Carpenter and Pashler (2007) found the robust testing effect in a study using visuospatial learning. Research has found that those with geography knowledge perform better on visuospatial tasks and visuospatial learning (Taylor, 2005). Geography knowledge was examined in the current study, but not in Carpenter and Pashler's study (2007). Results of the ancillary analyses of geography knowledge show that those with high geography knowledge are at an advantage when presented with visuospatial information in a learning context. That is, the number of college geography classes that a person has taken is a significant predictor of delayed retention of visuospatial information, although this knowledge only accounted for 7% to 14% of the variance in recalled information leaving more variance to be explained by other factors. Most testing effect studies have relied upon prose and text-based information. It may be that participants who have experience with language or other associated abilities may exhibit similar differences in ability to retain this kind of information as found with how geography operates for visuospatial information. Future studies should include abilities or skills that may impact performance depending on the type of information to be retained.

An additional research implication of the current study regards the findings about the relationship of JOL and PTSE. Previous research studies have suggested that these two constructs may be very closely related (Cuevas, 2004; Ford et al., 1998). The current study found a pattern of correlations that suggested relationships similar to previous research (see Table 7). But, a CFA performed on these variables presented unclear results. However, the results of this analysis may be due of insufficient power (Hu & Bentler, 1999). Thus, future research should explore this question with a larger sample.

Practical Implications

A practical implication of this study surrounds the findings regarding knowledge of geography. Practitioners may want to ensure that individuals to be trained on visuospatial skills are given a pre-assessment to discover their geographic knowledge. In this way, practitioners could consider if people with fewer geographic skills require extra attention, and set appropriate expectations about the success of training programs.

Limitations

The present study is not without limitations. The first limitation is that the application of the learned knowledge is expressed by reproduction of the map. Ideally, exhibition of learned knowledge should be in the application of that knowledge to a task associated with the learned material, instead of reproduction or recall. However, the focus of this study was on exploring how trainee self-efficacy could affect information retention. Thus, attempting to replicate the testing effect using an established task was a key component. Creation of an applied task involving participants' transfer of visuospatial knowledge to a task such as route-mapping, or following directions could be an important direction for future research.

Another limitation of the study involved the visual attention task that was chosen for the study. While anecdotal, some participants reported being frustrated by it, potentially resulting in less motivation to complete the study or give accurate responses on later measures. Future research should pilot test any visual attention tasks and probe participants for feedback to ensure adequate engagement throughout the study.

Conclusion

The testing effect has shown promise in the exploration of cognitive processes, as well as in application to classroom settings. While this study did not find support for the proposed hypotheses, the results of the analyses provided other avenues for future research as well as other practical considerations. In the past, the testing effect has been regarded as a psychological process that can be tapped to improve and enhance learning. However, the current study has presented some evidence that highlights it more as a learning method, and as such comes with some of the same considerations of other learning methods.

References

Ackerman, P. L., Kanfer, R., & Goff, M. (1995). Cognitive and noncognitive determinants and consequences of complex skill acquisition. *Journal of Experimental Psychology: Applied, 1*(4), 270-304. doi: 10.1037/1076-898X.1.4.270

Arbuckle, T. Y., & Cuddy, L. L. (1969). Discrimination of item strength at time of presentation. *Journal of Experimental Psychology, 81*(1), 126-131. doi: 10.1037/h0027455

Bandura, A. (1985). *Social foundations of thought and action: A social cognitive theory*: Prentice-Hall.

Bandura, A., & Locke, E. A. (2003). Negative self-efficacy and goal effects revisited. *Journal of Applied Psychology*, *88*, 87-99. doi: 10.1037/0021-9010.88.1.87

Bangert-Drowns, R., Kulick, J., & Kulick, C.-L. (1991). Effects of frequent classroom testing. *Journal of Educational Research, 85*(2), 89-99.

Bouffard-Bouchard, T. (1990). Influence of self-efficacy on performance in a cognitive task. *The Journal of Social Psychology, 130*(3), 353-363.

Budworth, M.-H., & Sookhai, F. (2008). *Training supervisors to support transfer of training: Investigating the effects of supervisor support, self-efficacy, and transfer climate*. Paper presented at the 24th Annual SIOP Conference, New Orleans, LA.

Butler, A., Karpicke, J., & Roediger, H. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology, 19*(4/5), 514-527. doi: 10.1080/09541440701326097

Carpenter, S. K., & Pashler, H. (2007). Testing beyond words: Using tests to enhance visuospatial map learning. *Psychonomic Bulletin & Review, 14*(3), 474-478.

Cialdini, R. B. (2001). *Influence: Science and practice* (4th ed.). Needham Heights, MA: Allyn & Bacon.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). Outliers and Multicollinearity: Diagnosing and Solving Regression Problems II. In J. Cohen (Ed.), *Applied Multiple Regression: Correlation Analysis for the Behavioral Sciences* (Vol. 3, pp. 390-430). Taylor & Francis, Inc.

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

Cuevas, H. M. (2004). *Transforming learning into a constructive cognitive and metacognitive activity: Use of a guided learner-generated instructional strategy within computer-based training.* Doctor of Philosophy Dissertation, University of Central Florida, Orlando, Florida.

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition* (1st ed.). Los Angeles, California: Sage.

Dunlosky, J., & Nelson, T. O. (1992). Importance of the kind of cue for judgments of learning (JOL) and the delayed-JOL effect. *Memory & Cognition, 20*(4), 374-380.

Ebbinghaus, H. (2003). Memory: A contribution to experimental psychology. In M. P. Munger (Ed.), *The history of psychology: Fundamental questions* (pp. xi, 514). New York: Oxford University Press.

Finn, B., & Metcalfe, J. (2008). Judgments of learning are influenced by memory for past test. *Journal of Memory and Language, 58*(1), 19-34. doi: 10.1016/j.jml.2007.03.006

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*(3), 221-233. doi: 10.1037/h0057532

Ford, K., Kraiger, K., & Merritt, S. (2009). An updated review of the multidimensionality of training outcomes: New directions for training evaluation research. In S. Kozlowski & E. Salas (Eds.), *Learning, Training, and Development in Organizations*. Mahwah, NJ: SIOP Frontiers Book Series.

Ford, K., Smith, E. M., Weissbein, D. A., Gully, S. M., & Salas, E. (1998). Relationships of goal orientation, metacognitive activity, and practice strategies with learning outcomes and transfer. *Journal of Applied Psychology, 83*(2), 218-233. doi: 10.1037/0021-9010.83.2.218

Frayne, C. A., & Geringer, J. M. (2000). Self-management training for improving job performance: A field experiment involving salespeople. *Journal of Applied Psychology, 85*(3), 361-372.

Frazier, P. A., Tix, A. P., & Barron, K. E. (2004). Testing moderator and mediator effects in counseling psychology research. *Journal of Counseling Psychology, 51*(1), 115-134. doi: 10.1037/0022-0167.51.1.115

Gates, A. (1917). Recitation as a factor in memorizing. In R. Woodworth (Ed.), *Archives of Psychology* (Vol. 6, pp. 1-140). New York: Stanford University.

Geringer, J. M., & Frayne, C. A. (1993). Self-efficacy, outcome expectancy, and performance of international joint venture general managers. *Canadian Journal of Administrative Sciences, , 10*, 322-333. doi: 10.1111/j.1936-4490.1993.tb00037

Gist, M. E., & Mitchell, T. B. (1992). Self-efficacy: A theoretical analysis of its determinants and malleability. *Academy of Management Review, 17*, 183-211. doi: 10.2307/258770

Goldstein, I. L., & Ford, J. K. (2002). *Training in organizations: Needs assessment, development, and evaluation (4th ed.)*. Belmont, CA: Wadsworth/Thomson Learning.

Halpern, D. F., & Collaer, M. L. (2005). Sex differences in visuospatial abilities: More than meets the eye *The Cambridge Handbook of Visuospatial Thinking* (pp. 170-212). New York, NY: Cambridge University Press.

Howell, W. C. (2003). Human factors and ergonomics. In W. C. Borman, D. R. Ilgen & R. J. Klimoski (Eds.), *Handbook of psychology: Industrial and organizational psychology* (Vol. 12, pp. 541-564). Hoboken, NJ: John Wiley & Sons Inc.

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

James, L. R., Demaree, R. G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology, 69*(1), 85-98. doi: 10.1037/0021-9010.69.1.85

Kalakoski, V., & Saariluoma, P. (2001). Taxi drivers' exceptional memory of street names. *Memory & Cognition, 29*(4), 634-638.

Kincaid, J. P., Fishburne, R. P., Jr, Rogers, R. L., & Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel (pp. 49). Springfield, Virginia: U.S. Navy.

Koriat, A. (2007). Metacognition and consciousness. In P. D. Zelazo, M. Moscovitch & E. Thompson (Eds.), *The Cambridge Handbook of Consciousness* (pp. 289-325). Cambridge: Cambridge University Press.

Koriat, A., Sheffer, L., & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: Judgments of learning exhibit increased underconfidence with practice. *Journal of Experimental Psychology: General, 131*(2), 147-162. doi: 10.1037/0096-3445.131.2.147

Kozlowski, S. W., & Hattrup, K. (1992). A disagreement about within-group agreement: Disentangling issues of consistency versus consensus. *Journal of Applied Psychology, 77*(2), 161-167. doi: 10.1037/0021-9010.77.2.161

Kraiger, K. (2003). Perspectives on training and development. In W. C. Borman, D. R. Ilgen & R. J. Klimoski (Eds.), *Handbook of psychology: Industrial and organizational psychology* (Vol. 12, pp. 171-192). Hoboken, NJ: John Wiley & Sons Inc.

Latham, G. P., & Frayne, C. A. (1989). Self-management training for increasing job attendance: A follow-up and replication. *Journal of Applied Psychology, 72*, 411-416. doi: 10.1037/0021-9010.74.3.411

Lee, C. (1984). Accuracy of efficacy and outcome expectations in predicting performance in a simulated assertiveness task. *Cognitive Therapy and Research, 8*, 509-516. doi: 10.1007/BF01315096

Leeming, F. (2002). The exam-a-day procedure improves performance in psychology classes. *Teaching of Psychology, 29*(3), 210-212.

Lyle, K. B., & Crawford, N. A. (in press). *The power of PUREMEMs: Daily retrieval practice increases performance on statistics for psychological exams*. University of Louisville. Louisville.

Metcalfe, J. (2000). Metamemory theory and data. In E. Tulving & F. Craik (Eds.), *The Oxford Handbook of Memory* (1st ed., pp. 197-211). Oxford: Oxford University Press.

Metcalfe, J., & Finn, B. (2008). Familiarity and retrieval processes in delayed judgments of learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 34*(5), 1084-1097. doi: 10.1037/a0012580

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect.". *Psychological Science, 2*(4), 267-270. doi: 10.1111/j.1467-9280.1991.tb00147.x

Quiñones, M. (1995). Pretraining context effects: Training assignment as feedback. *Journal of Applied Psychology, 80*(2), 226-238. doi: 10.1037/0021-9010.80.2.226

Quiñones, M. (1997). Contextual influences on training effectiveness. In M. A. Quinones & A. Ehrenstein (Eds.), *Training for a rapidly changing workplace: Applications of psychological research* (pp. 177-199). Washington, DC: American Psychological Association.

Roediger, H., & Karpicke, J. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*(3), 181-210. doi: 10.1111/j.1745-6916.2006.00012.x

Roediger, H., & Karpicke, J. (2006b). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249-255.

Schwoerer, C. E., May, D. R., Hollensbe, E. C., & Mencl, J. (2005). General and specific self-efficacy in the context of a training intervention to enhance performance expectancy. *Human Resource Development Quarterly, 16*(1), 111-129. doi: 10.1002/hrdq.1126

Sitzmann, T., Brown, K. G., Casper, W. J., Ely, K., & Zimmerman, R. D. (2008). A review and meta-analysis of the nomological network of trainee reactions. *Journal of Applied Psychology*, *93*(2), 280-295. doi: 10.1037/0021-9010.93.2.280

Spitzer, H. (1939). Studies in retention. *The Journal of Educational Psychology, 30*(9), 641-656. doi: 10.1037/h0063404

Strack, F. (1992). The different routes to social judgments: Experiential versus informational strategies. In L. L. Martin & A. Tesser (Eds.), *The construction of social judgments* (pp. 249-275). Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

Szpunar, K. K., McDermott, K. B., & Roediger, H. L., III. (2008). Testing during study insulates against the buildup of proactive interference. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 34*(6), 1392-1399. doi: 10.1037/a0013082

Tabachnick, B. G., & Fidell, L. S. (2006). *Using Multivariate Statistics* (5th ed.). Boston: Allyn & Bacon.

Tannenbaum, S. I., & Yukl, G. (1992). Training and development in work organizations. *Annual Review of Psychology, 43*(1), 399. doi: 10.1146/annurev.ps.43.020192.002151

Taylor, H. A. (2005). Mapping the understanding of understanding maps. *The Cambridge Handbook of Visuospatial Thinking* (pp. 295-333). New York, NY: Cambridge University Press.

Toppino, T. C., & Cohen, M. S. (2009). The testing effect and the retention interval: Questions and answers. *Experimental Psychology, 56*(4), 252-257. doi: 10.1027/1618-3169.56.4.252

Tulving, E. (1967). The effects of presentation and recall of material in free-recall learning. *Journal of Verbal Learning and Verbal Behavior, 6*, 175-184. doi: 10.1016/S0022-5371%2867%2980092-6

Tziner, A., Fisher, M., Senior, T., & Weisberg, J. (2007). Effects of trainee characteristics on training effectiveness. *International Journal of Selection and Assessment, 15*(2), 167-174. doi: 10.1111/j.1468-2389.2007.00378.x

Wheeler, M., Ewers, M., & Buonanno, J. (2003). Different rates of forgetting following study versus test trials. *Memory, 11*(6), 571-580. doi: 10.1080/09658210244000414

Table 1

Scoring for Participant Maps

Scoring Criteria	Score
Correct Feature	1
Extra Feature	-1
LA Accuracy	1
SA Accuracy	1
LR Accuracy	1
SR Accuracy	1

Note: Adapted from Carpenter & Pashler (2007)

Table 2.

Scoring Criteria Descriptives (No Information)

Scoring	Immediate		Delayed	
Procedure	M	SD	M	SD
Test/Study				
LR	9.92	2.83	6.67	2.50
SR	6.67	2.82	8.58	3.37
LA	9.42	3.26	7.08	3.39
SA	3.42	2.36	1.83	1.37
Total	37.42	12.79	29.75	10.85
Study				
LR	9.83	2.04	6.71	3.12
SR	7.06	2.48	8.59	4.20
LA	10.06	2.01	6.88	4.78
SA	4.50	2.09	1.47	1.46
Total	39.06	10.25	29.94	13.38

Table 3.
Scoring Criteria Descriptives (Information)

Scoring Procedure	Immediate M	Immediate SD	Delayed M	Delayed SD
Test/Study				
LR	9.62	3.60	6.26	3.62
SR	6.57	2.79	8.21	4.87
LA	9.43	3.80	6.84	4.39
SA	3.57	3.12	1.26	0.99
Total	37.14	15.52	28.37	15.47
Study				
LR	8.41	2.72	5.52	2.59
SR	6.37	2.57	7.22	3.51
LA	8.22	3.27	5.67	3.46
SA	3.11	2.15	1.07	1.14
Total	33.11	12.24	26.11	11.06

Table 4.
Interrater Agreement Statistics

Scoring Criteria	r_{wg}
Time 1	
Correct Items	0.94
Liberal Absolute Accuracy	0.94
Stringent Absolute Accuracy	0.91
Liberal Relative Accuracy	0.86
Stringent Relative Accuracy	0.68
Overall Accuracy	0.95
Time 2	
Correct Items	0.97
Liberal Absolute Accuracy	0.98
Stringent Absolute Accuracy	0.97
Liberal Relative Accuracy	0.92
Stringent Relative Accuracy	0.88
Overall Accuracy	0.98

Note: r_{wg} calculated using procedures outlined by James, Demaree, & Wolf (1984)

Table 5

Descriptive Statistics and Correlations for Key Study Variables

Variable	N	Mean	SD	Skewness	Kurtosis	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Gender	90	1.51	.50	-.05	-2.04	--												
2. Race	90	5.01	2.92	.00	-1.99	-.14	--											
3. Age	90	24.48	6.12	2.21	6.18	-.21*	.20	--										
4. High School Art Classes	90	1.99	1.88	1.22	.93	.07	.08	-.18	--									
5. College Art Classes	90	.79	1.50	4.09	23.72	.08	.07	-.15	.43**	--								
6. Hrs per Week Spent on Art	90	1.07	1.83	2.48	7.21	.15	.05	-.13	.24*	.53**	--							
7. High School Geography Classes	90	1.62	1.50	2.67	11.08	.07	.11	-.20	.29**	.05	.23*	--						
8. College Geography Classes	90	.41	.72	2.38	7.55	-.18	.17	-.02	-.02	-.01	.22*	.38**	--					
9. Months in Map-Reading Jobs	90	6.81	23.98	4.77	23.64	-.24*	.01	.41**	-.02	-.10	.01	-.08	.07	--				
10. Motivation - Effort	90	1.87	.80	.66	-.01	.00	-.23*	.09	-.16	-.10	.04	.00	.06	.26**	--			
11. Motivation - Persistence	90	1.70	.66	.42	-.73	-.07	-.18	.03	-.12	-.11	.08	.04	.07	.28**	.80**	--		
12. Immediate Aggregate JOL	90	64.56	22.91	-.34	-.85	-.12	.06	-.03	.26**	.26**	.18	.16	.13	.05	-.18	-.09	--	
13. Delayed Aggregate JOL	90	2.46	.84	.11	-1.19	.21*	-.14	.04	-.43**	-.28**	-.16	-.18	-.20	-.12	.24*	.17	-.57**	--
14. Immediate Aggregate PTSE	90	46.38	28.25	.38	-.45	.00	-.12	-.12	.25*	.28**	.15	.31**	.24*	.04	-.17	-.13	.61**	-.50**
15. Delayed Aggregate PTSE	90	3.19	.96	.31	-.52	.03	.16	.13	-.32**	-.23*	-.03	-.20	-.23*	-.13	.15	.13	-.44**	.59**
16. Aggregate JOL - T2	90	37.09	26.37	.46	-.96	.04	-.33**	-.14	.22*	.21*	.15	.06	.19	.07	-.05	.04	.49**	-.38**
17. Aggregate PTSE - T2	90	3.37	1.01	-.06	-.61	.09	.25**	.12	-.28**	-.13	-.01	.06	-.10	-.07	.12	.03	-.31**	.38**
18. Liberal Absolute Accuracy - T1	90	9.38	2.89	.34	.52	-.10	.22*	-.02	.16	.01	.13	.06	-.01	.00	-.01	-.02	.03	-.12
19. Stringent Absolute Accuracy - T1	90	6.63	2.64	.09	-.73	-.14	.35**	.12	.14	.01	.04	.06	.16	.10	-.01	-.03	.02	-.27
20. Liberal Relative Accuracy - T1	90	9.19	3.22	.10	.22	-.07	.24*	.03	.18	.04	.12	.09	.01	-.04	-.06	-.08	.10	-.18
21. Stringent Relative Accuracy - T1	90	3.58	2.46	1.18	1.63	-.16	.26**	.11	.17	.02	.00	.13	.14	.10	.09	.11	.03	-.13
22. Immediate Retention - Total Score	90	36.39	12.86	.30	-.33	-.13	.35**	.09	.16	.00	.08	.09	.09	.06	-.01	-.02	.05	-.20*
23. Liberal Absolute Accuracy - T2	90	6.22	2.87	.39	-.27	-.27**	.15	-.06	.13	-.02	.04	.04	.17	.10	-.03	-.04	.07	-.16
24. Stringent Absolute Accuracy - T2	90	8.08	3.85	.33	-.41	-.33**	.19	-.03	.09	-.02	.02	.02	.21*	.13	-.01	-.02	.06	-.21*
25. Liberal Relative Accuracy - T2	90	6.53	3.85	.30	-.27	-.21*	.20	.00	.06	.01	.03	-.01	.18	.09	.04	.00	-.02	-.13
26. Stringent Relative Accuracy - T2	90	1.39	1.24	.73	.18	-.20	.10	.06	-.15	-.08	-.04	.01	.29**	.21*	.01	-.01	-.11	-.08
27. Delayed Retention - Total Score	90	28.34	12.22	.31	-.37	-.31**	.20	-.04	.12	-.03	.02	.04	.19	.09	-.03	-.03	.06	-.20

Note: ** $p < .01$; * $p < .05$

	14	15	16	17	18	19	20	21	22	23	24	25	26	27
	--													
	-.70**	--												
	.57**	-.49**	--											
	-.46**	.51**	-.74**	--										
	-.13	.04	.05	.04	--									
	-.09	-.04	.11	-.04	.78**	--								
	-.06	.01	.04	.08	.92**	.73**	--							
	-.09	.00	-.01	.05	.46**	.52**	.59**	--						
	-.11	.01	.03	.04	.92**	.88**	.93**	.69**	--					
	-.01	-.03	.17	-.11	.62**	.65**	.60**	.41**	.66**	--				
	.00	-.06	.17	-.11	.59**	.68**	.56**	.40**	.65**	.97**	--			
	-.05	-.02	.13	-.07	.54**	.66**	.51**	.36**	.62**	.90**	.95**	--		
	.05	-.13	.08	-.03	.22*	.38**	.26**	.25*	.34**	.45**	.58**	.55**	--	
	.00	-.07	.16	-.11	.63**	.69**	.61**	.41**	.68**	.98**	.99**	.93**	.52**	--

Table 6.
Mean Proportion Correct

Scoring Procedure	Immediate M	Immediate SD	Delayed M	Delayed SD	Carpenter & Pashler (2007) M	Carpenter & Pashler (2007) SE
Test/Study						
LA	0.55	0.16	0.37	0.14	0.74	0.03
SA	0.37	0.16	0.48	0.19	0.61	0.02
LR	0.52	0.18	0.39	0.19	0.68	0.03
SR	0.19	0.13	0.10	0.08	0.61	0.03
Study						
LA	0.55	0.11	0.37	0.14	0.63	0.03
SA	0.39	0.14	0.47	0.23	0.53	0.03
LR	0.56	0.11	0.38	0.26	0.62	0.03
SR	0.25	0.12	0.08	0.08	0.54	0.03

Note: Information is adapted from Carpenter and Pashler (2007; Table 1)

Table 7.
Global Fit Indices

Model	χ^2	df	CFI	NNFI	SRMR	RMSEA
One-Factor SE Model	529.48	84	.75	.64	.17	.24
Two-Factor Immediate/Delayed Model	504.56	83	.77	.67	.19	.24
Two-Factor PTSE/JOL Model	267.28	83	.85	.78	.11	.16
Four-Factor PTSE/JOL Model	246.78	78	.87	.80	.10	.16
Four-Factor Map Features Model	471.86	78	.81	.70	.17	.24

Note: All χ^2 values are significant at $p < 0.01$. CFI = comparative fit index; NNFI = non-normed fit index; SRMR = standardized root-mean-square residual; RMSEA = root-mean-square error of approximation.

Figure 1. Percentage of information recalled by participants in study conditions and test conditions. Adapted from data reported by Roediger, et al. (2006b).

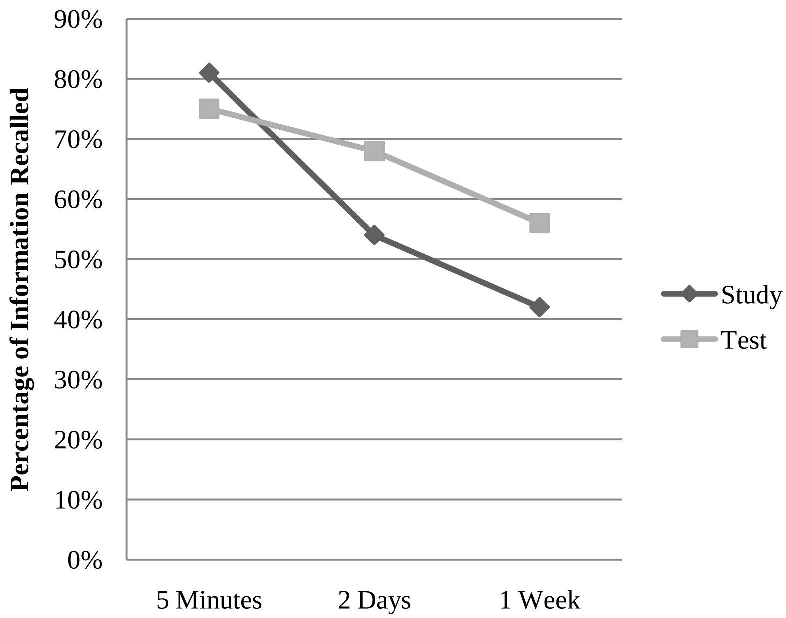


Figure 2. Map with all 12 iconic map features. Adapted from Carpenter and Pashler (2007).

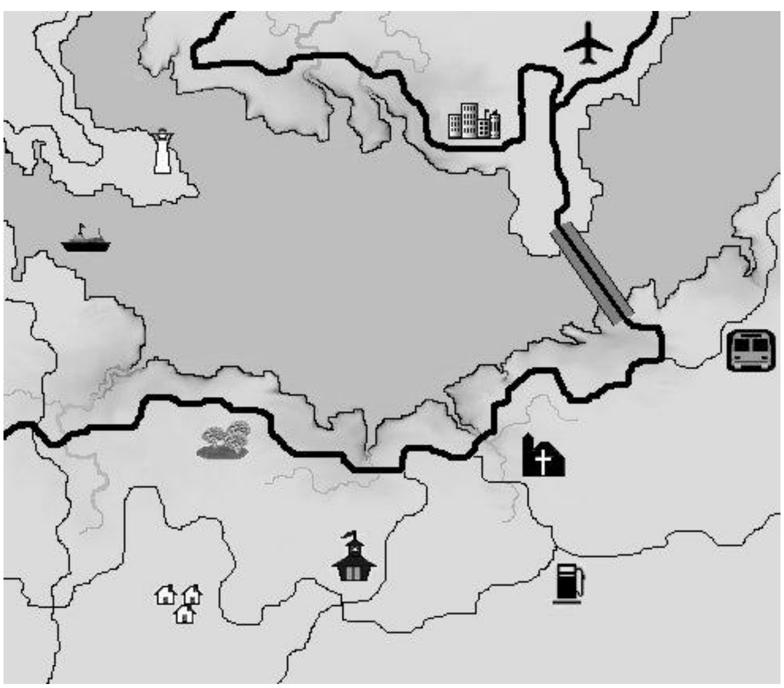


Figure 3. One-factor model of Self-Efficacy

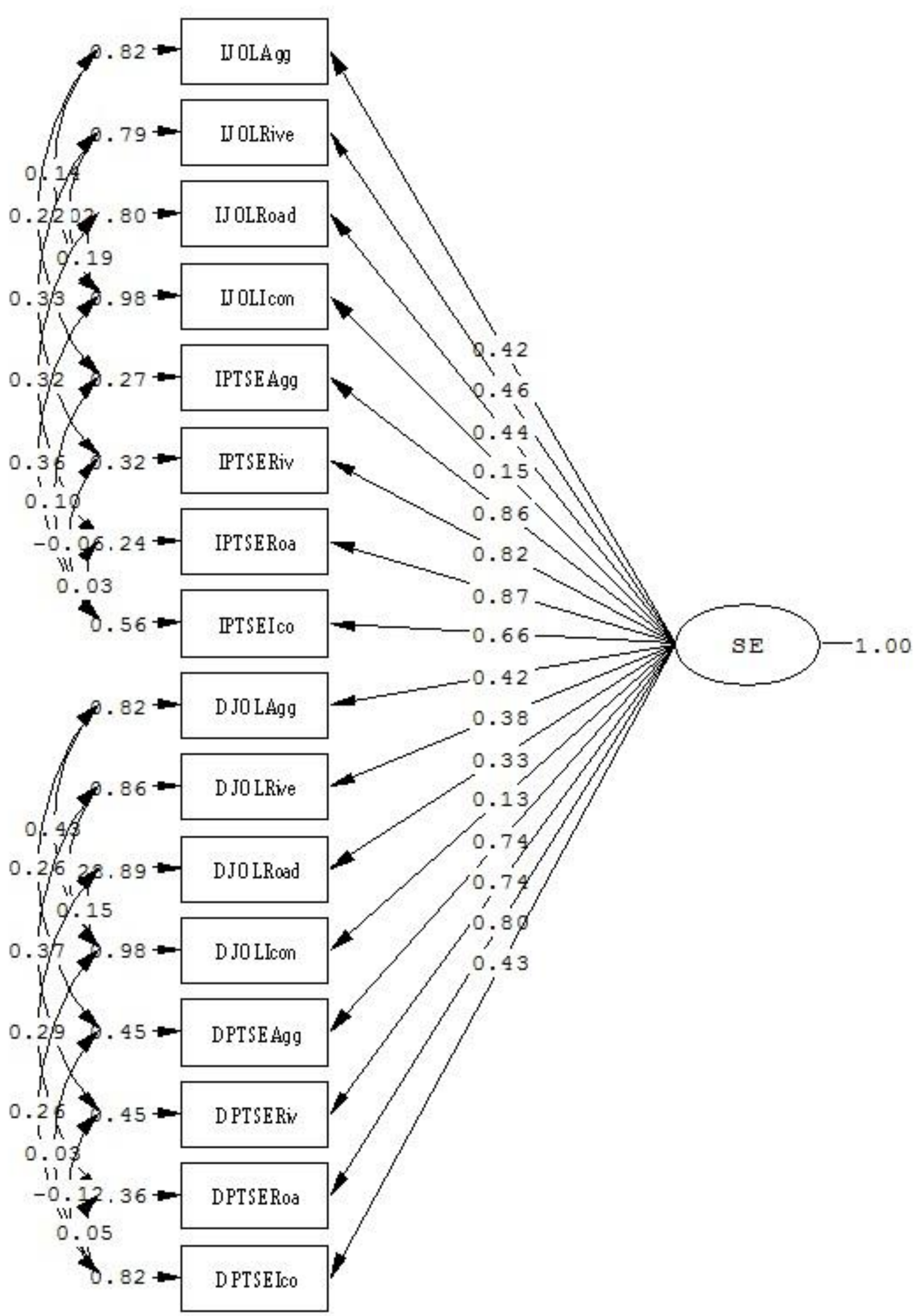


Figure 4. Two-factor model of Immediate and Delayed self-efficacy

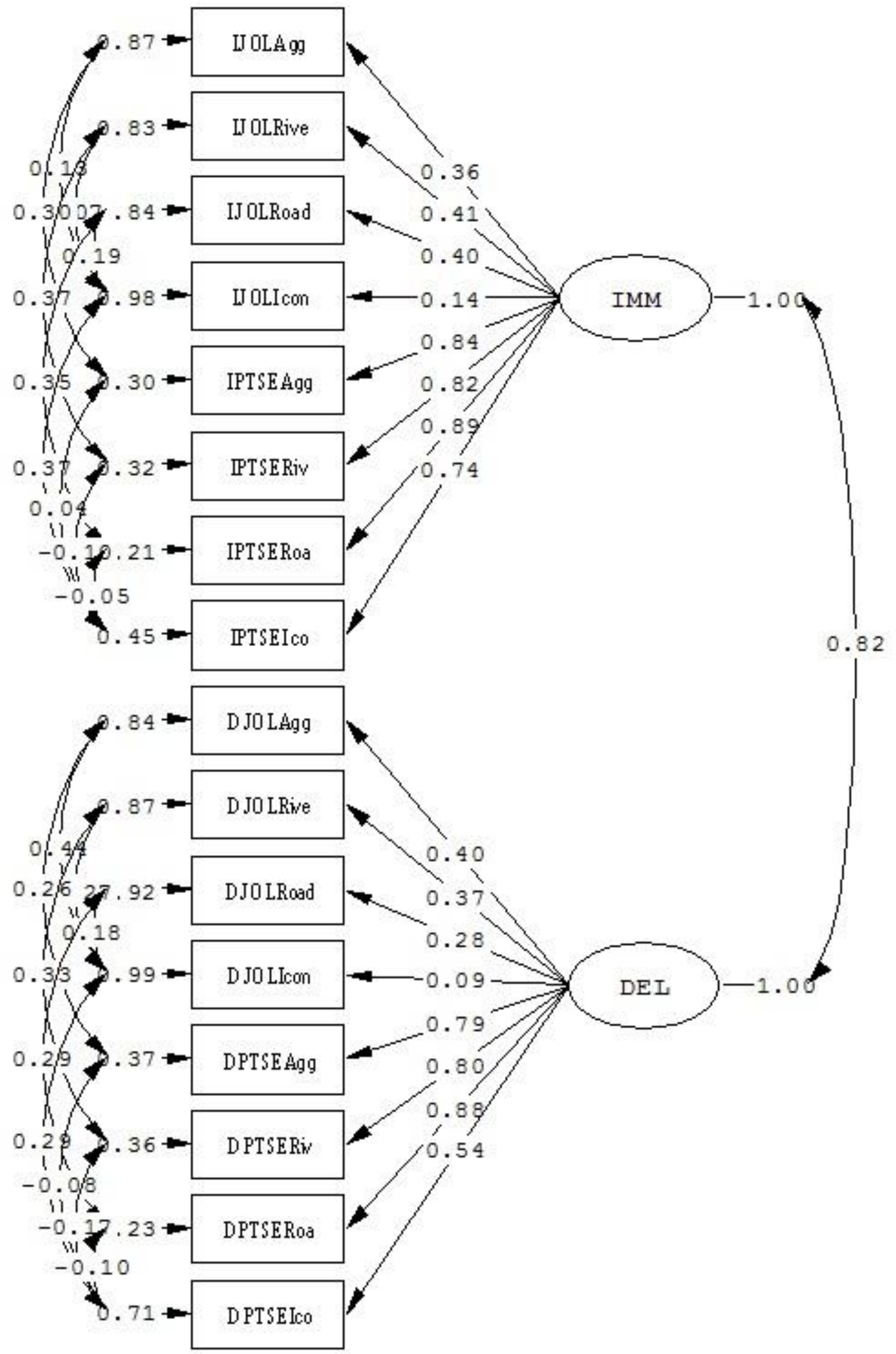


Figure 5. Two-factor model of PTSE and JOL

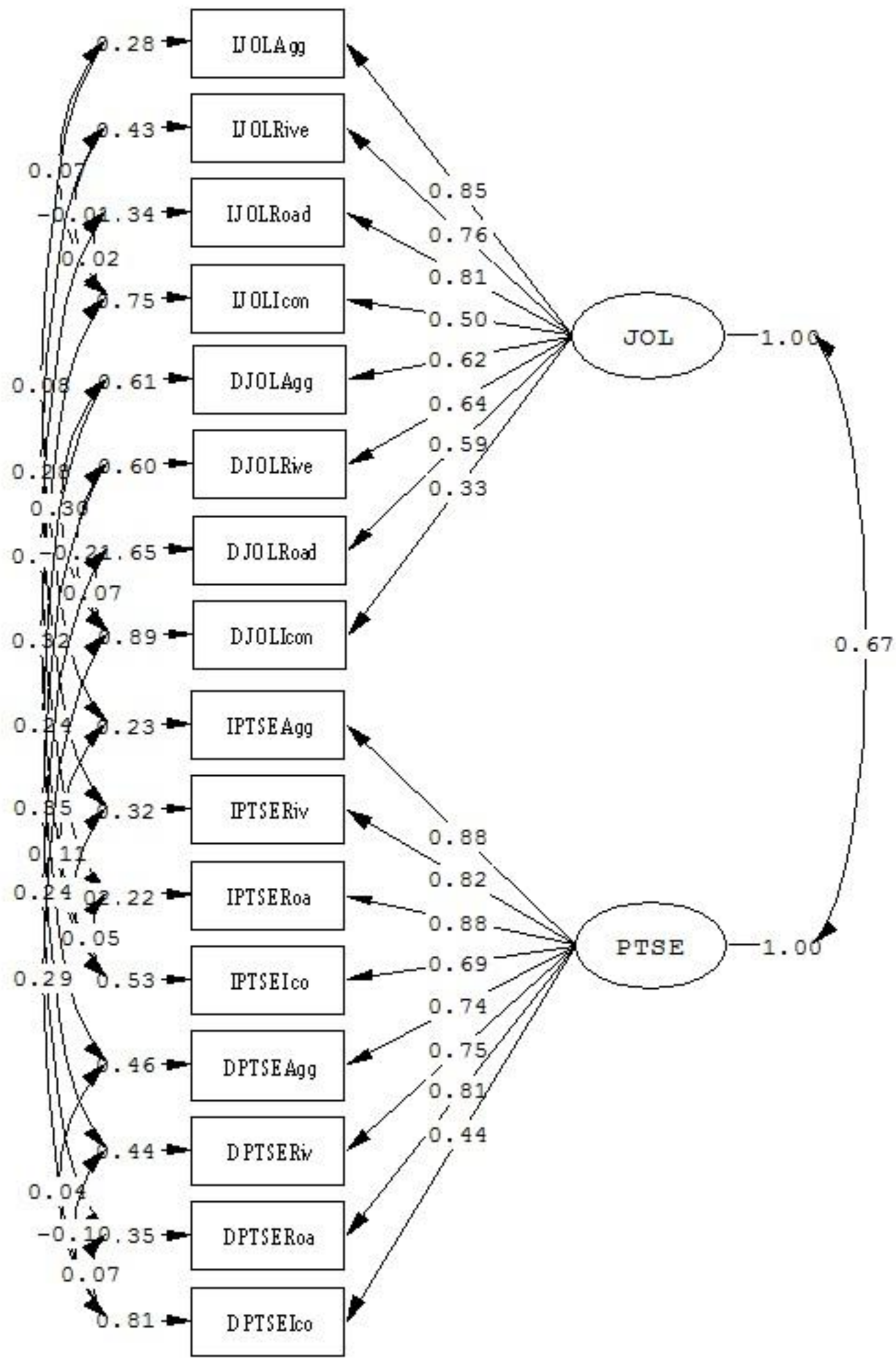


Figure 6. Four-factor model of PTSE and JOL (Immediate and Delayed)

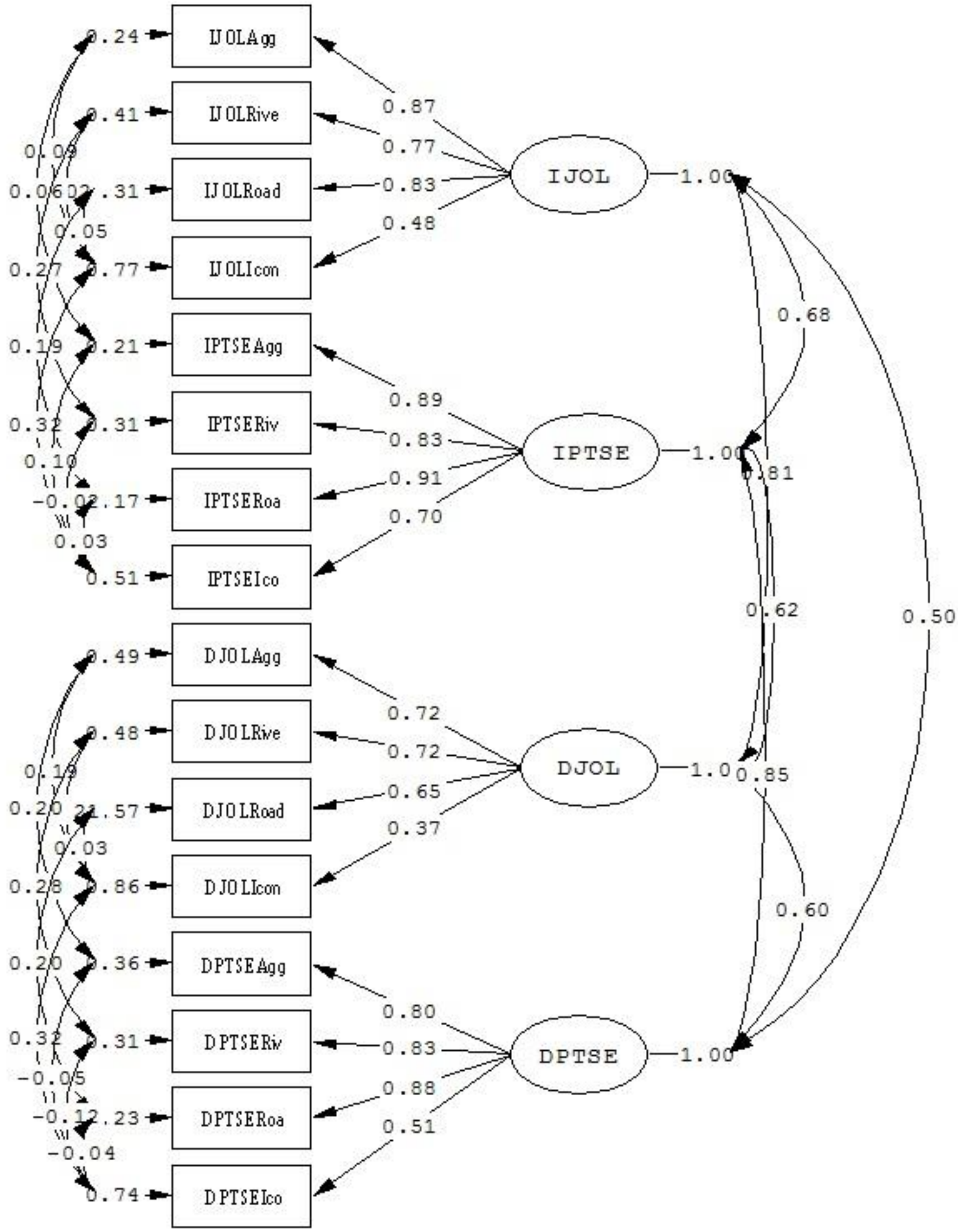


Figure 7. Four-factor model of map features

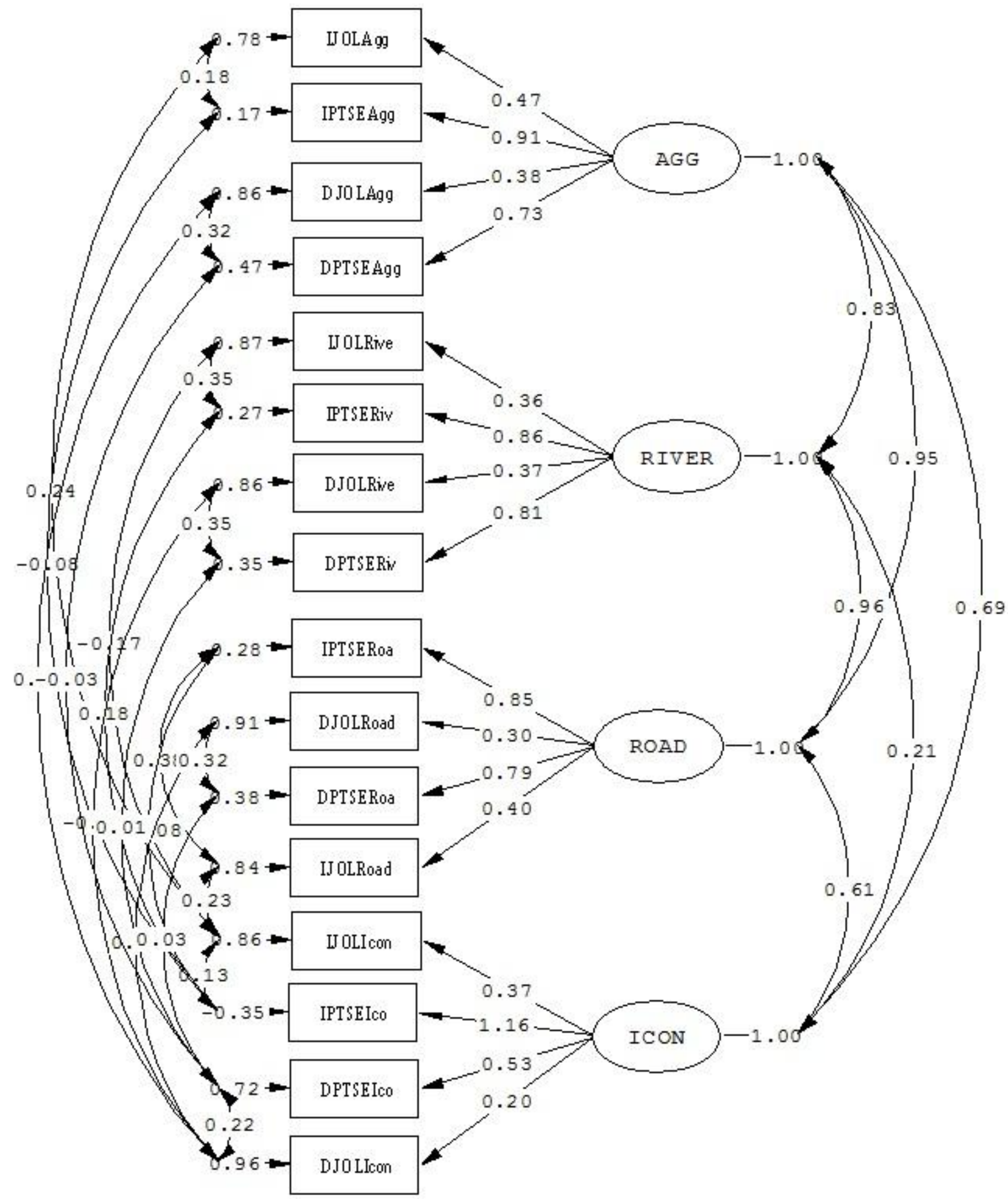


Figure 8. Decomposition of effects – mediation relationship described in hypothesis 3b (JOL)

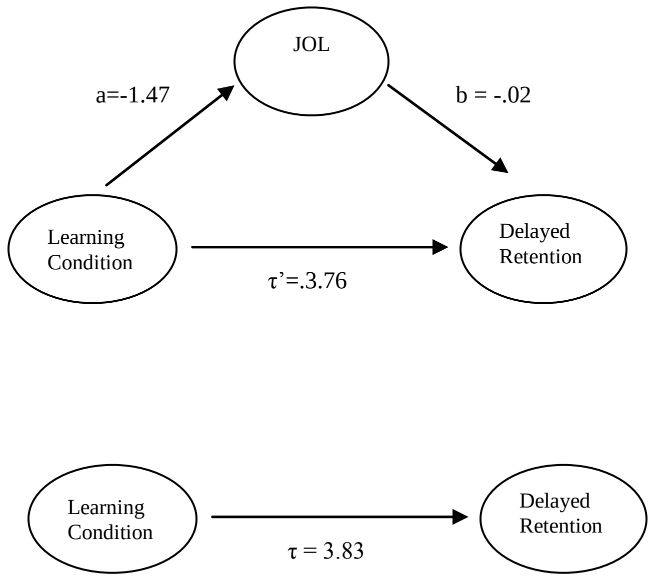
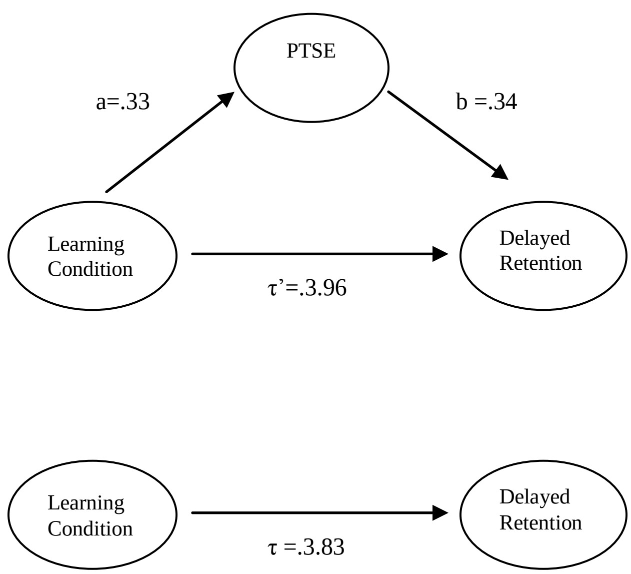


Figure 9. Decomposition of effects - mediation relationship described in hypothesis 3b (PTSE)



Appendix A

Demographics measure

1) What is your Gender?

- ❑ Male
- ❑ Female

2) What is your Race?

- ❑ American Indian or Alaska Native
- ❑ Asian
- ❑ Black or African American
- ❑ Hispanic
- ❑ Native Hawaiian or Other Pacific Islander
- ❑ Multi-racial
- ❑ Some other race
- ❑ White

3) What is your Age (in years)? ______________________

4) Please indicate your level of education.

- ❑ Freshman
- ❑ Junior
- ❑ Sophomore
- ❑ Senior

5) Please indicate your current major (if you have more than one, check all that apply).

- ❑ Psychology
- ❑ Education
- ❑ Sociology
- ❑ Communication
- ❑ Nursing
- ❑ Business
- ❑ Social Work
- ❑ Criminology
- ❑ Biology
- ❑ Other:____________________________

6) How many art classes did you take in high school? ______________

7) How many art classes have you taken in college? _______________

8) On average, how many hours per week do you spend doing creative projects, such as drawing or painting? _____________

9) How many geography classes did you take in high school? _____________

10) How many geography classes have you taken in college? _____________

11) How many months have to spent in jobs that require map-knowledge, such as delivery driving, truck driving, or taxi-driving? __________________

Appendix B

Self-Efficacy, Judgments of Learning (JOL), and Motivation Measures*

<u>JOL</u>

What percentage of the map items that you've learned do you think you'll be able to correctly draw if tested before you leave today? _____%

What percentage of the map items that you've learned do you think you'll be able to correctly draw in two (2) weeks? _____%

What percentage of the [insert map item] do you think you'll be able to draw correctly if tested before you leave today? _____%

What percentage of the [insert map item] do you think you'll be able to draw correctly if tested in two (2) weeks? _____%

<u>Self-Efficacy (1=strongly disagree, 5=strongly agree)</u>

I am confident in my ability to correctly draw the map items I’ve learned if tested today.

(1) (2) (3) (4) (5)

I am confident in my ability to correctly draw the map items I've learned if tested in two (2) weeks.

(1) (2) (3) (4) (5)

I am confident in my ability to correctly draw the [insert map item] I’ve learned if tested today.

(1) (2) (3) (4) (5)

I am confident in my ability to correctly draw [insert map item] I've learned if tested in two (2) weeks.

(1) (2) (3) (4) (5)

<u>Motivation (1=strongly disagree, 5=strongly agree)</u>

I have tried my hardest to learn what was presented to me.

(1) (2) (3) (4) (5)

Even when the task was difficult, I did my best to perform the task.

(1) (2) (3) (4) (5)

Appendix C

Reading for Participants in the "Testing Effect Information" Condition

Studying or reading information over and over can help you remember things for short periods of time. But, quizzing yourself over information can be a good way to remember it for a long time. Basically, the brain needs to practice getting things out of it. For example, think about what you would do to learn to play a sport. If you need to learn how to play basketball, you learn better by actually playing the game, instead of reading about how to play it. Your brain learns information the same way that your body learns the skills. In order to remember information, you should practice the information by quizzing yourself. When you try to remember the information, you are helping your brain learn the material. This method of quizzing will help you learn better than just reading or studying the information. When you test yourself, you're exercising your brain in the same way that it remembers.

Appendix D

Debriefing Survey

Please answer the following questions to the best of your ability

1. How is quizzing yourself over information like learning to play basketball? (manipulation check – information condition – looks for understanding of the testing effect)

2. Did you read information about the how to remember more information by quizzing yourself? (manipulation check – information condition – looks for simply receiving the information)

   - ❑ Yes
   - ❑ No

3. During the map-learning task, was the map you looked at shown with features removed? (manipulation check – learning condition)

   - ❑ Yes
   - ❑ No

4. (Only for Time 2) Before coming to the experiment session today, I rehearsed the map information from the last session.

   - ❑ Yes
   - ❑ No

Appendix E

Pilot Study

A pilot study will be conducted to determine if the testing effect information provided to participants creates a strong enough manipulation for the information condition of the full study. This will be assessed through the examination of three manipulation check questions embedded in other debriefing questions (see below). These questions will be:

- How is quizzing yourself over information like learning to play basketball?
- Did you read information about the how to remember more information by quizzing yourself?
- During the map-learning task, was the map you looked at shown with features removed?

Answers from participants in the information condition should show an understanding of the testing effect and a belief that testing assists in learning. Likewise, participants in the no information condition should have little or no understanding of the testing effect.

Additionally, the pilot study will assess the influence of individual difference variables from the demographics measure on the information retention measure (map-drawing procedure). To assess the influence of these variables, I will enter each into a regression equation as predictor variables (IV) with the map-drawing score as the criterion variable (DV). With each of these variables entered into the equation, I will look for significant variance in map-drawing scores explained by any of the individual difference variables. If there is significant variance explained by these variables, this information will be used during the full study to determine covariates or explore individual differences for the analyses described in the results section.

To examine any other problems with the study, participants will be asked questions that assess their impressions of the study. The following questions will be presented upon completion of the pilot study, and participants will give responses in essay format:

- What do you believe the purpose of the study was?
- What are your reactions to this study?
- Was there any part of the study that you found confusing? If so, please explain.