Transitive Inference and The Testing Effect:
The Effects of Testing on Knowledge Structure Formation

Milton Picklesimer

A thesis submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Arts in the Department of Psychology

Chapel Hill
2012

Approved by:

Neil W. Mulligan

Kelly S. Giovanello

Peter C. Gordon

ABSTRACT

MILTON PICKLESIMER: Transitive Inference and The Testing Effect: The Effects of Testing on Knowledge Structure Formation
(Under the direction of Neil W. Mulligan)

Compared to restudying, testing has often been found to enhance memory. This is called the *testing effect*. However, the causes of this effect are not entirely understood. Testing could merely enhance isolated stimulus-response associations (i.e., item memory) or also enhance the unifying structure of the memoranda (i.e., relational memory). Recent studies have examined these issues with mixed results. The current study employed a transitive inference paradigm to teach participants a novel, highly inter-related knowledge structure comprised of several basic premises. Encoding strategy was manipulated between subjects. Both groups took a final test that assessed memory for the basic premises and their ability to make transitive inferences about them. Experiment 1 found no differences between the groups. After using a stronger manipulation in Experiment 2, it was found that, for participants who indicated awareness of the hierarchical structure of the materials, there were no differences between encoding conditions. For unaware participants, however, the restudying group showed superior performance on transitive inference problems. Thus, the current study identified conditions under which testing does not enhance inferential ability or memory for the unifying structure of the memoranda.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

The Testing Effect

Educators and memory researchers frequently employ tests as a form of learning assessment to see how much of the imparted knowledge has been retained by their pupils or subjects. Much to their surprise, researchers have found that tests can also be a form of learning *enhancement*. That is, compared to a baseline condition where one is given only one exposure to the material before the final test, taking an intervening test can improve one's memory for that material. In addition, this enhancement is often even greater than the enhancement produced by re-studying the material after the initial learning episode. This is called the *testing effect* (for reviews, see Dempster, 1996; Roediger & Karpicke, 2006a).

In the prototypical testing effect paradigm, there are three phases: the initial learning episode, the intermediate phase, and the final test. The phases of most interest are typically the latter two. In the intermediate phase, the subjects are asked to restudy all or part of the material from the initial learning episode, or take a test over some or all of the material. The final test is usually comprehensive—covering all material from the initial learning episode—and is used to reveal the later mnemonic effects of repeated studying or repeated testing. These two conditions are often compared to a baseline condition of items that were only presented during the initial learning phase.

Basic Findings

The testing effect has often been examined with paired-associate

learning tasks (Carrier & Pashler, 1992; Carpenter, Pashler, & Vul, 2006; Carpenter, Pashler, Wixted, & Vul, 2008, Experiment 3; Karpicke & Zaromb, 2010). For example Carrier and Pashler (1992, Experiment 2) presented subjects with Eskimo words and their English equivalent. Half of these pairs were restudied and the remaining half were subjected to retrieval practice. Memory for Eskimo words and their English equivalents was better for those pairs that received retrieval practice.

Testing can also enhance memory for paired associates that are weakly related (e.g. *factory - plant*) (Kornell, Hays, & Bjork, 2009, Experiments 3-6; Carpenter, 2009). At one level, this is counterintuitive because one might think that weakly related items are difficult to encode and are not processed with as much fluency as strongly related items—possibly making retrieval even more difficult. Under these situations, it would seem that restudying would be more fruitful than testing. Surprisingly, testing can potentiate learning of weakly related items (especially when coupled with feedback). In addition, testing has been shown to enhance retention of obscure, unfamiliar facts (e.g., "fake pearls were once made out of fish scales") (Carpenter et al., 2008, Experiments 1 & 2). It is also of interest to note that Carpenter et al. (2008) examined retention at unconventionally long intervals (i.e., up to 42 days) and found that retention of the material at longer intervals (i.e., 1 day or more) was better when the items were tested in the intermediate phase of the experiment.

Although the notion of test-enhanced retention of paired associates and obscure facts is informative in its own right, it would be practically and theoretically informative to know if testing can enhance memory for "natural" or at least richer materials. Roediger and Karpicke (2006b) showed test-enhanced memory for short passages taken

from the Test of English as a Foreign Language (TOEFL). Here subjects read one (Experiment 2) or several (Experiment 1) passages. In Experiment 1, restudying and retrieval practice were manipulated within-subjects—whereas a between-subjects design was used in Experiment 2. In both experiments, the final test was free recall. On the test, subjects were instructed to write down as much of the material as they could remember (exact wording and order were not required). Their responses were then compared against 30 pre-specified idea units for scoring. The results showed that testing enhanced memory compared to restudying but this enhancement only emerged after a delay (i.e. 2 days or 1 week). When memory was assessed merely 5 minutes after the initial session, performance was actually better in the repeated studying condition. Although performance on an immediate test favored repeated studying, repeated testing was more effective at retarding forgetting over longer retention intervals. For passages studied repeatedly, memory fell off precipitously while passages subjected to retrieval practice were forgotten at a much slower rate. In short, this implies that although repeated studying may yield short-term gains in retention, repeated testing can enhance retention over the long term. Despite this result, there is some variability across studies regarding when the testing effect emerges. Some studies report a testing effect on an immediate test (e.g., Carrier & Pashler, 1992; Kuo & Hirshman, 1996; Carpenter et al., 2008, Experiment 3; Carpenter, 2009) and other studies indicate that the effect does not arise until after a delay of approximately 24 hours or more (e.g., Roediger & Karpicke, 2006b; Carpenter et al., 2008, Experiments 1 & 2).

At this time the relationship between retention interval and the testing effect is unclear. One would think that providing feedback during retrieval practice should

promote an immediate benefit of testing because of the metacognitive benefits of feedback. However, some studies with feedback show an immediate testing effect (e.g., Carrier & Pashler, 1992; Carpenter et al., 2008, Experiment 3) while others do not (e.g., Carpenter et al., 2008, Experiments 1 & 2). An immediate testing effect has also been found in studies without feedback (e.g., Kuo & Hirshman, 1996; Carpenter, 2009). Other factors such as the number of retrieval practice trials, the type or inherent difficulty of the materials, and/or the sheer number of memoranda might influence the time course of the testing effect but these aspects will need to be explored more directly by other studies.

By now, many of the salient effects of testing should be evident but it should also be evident that all of the research described thus far has utilized verbal materials exclusively. A few studies have examined the effects of testing on non-verbal materials (e.g., Carpenter & Pashler, 2007; Rohrer, Taylor, & Sholar, 2010). Carpenter and Pashler (2007) used a map-learning paradigm where they manipulated restudying and retrieval practice within subjects—but across maps (e.g., restudy Map A then take a test over Map B). The maps were fictitious and comprised of 12 target features. All subjects were given 20 sec. to study the target map before moving on to the respective intermediate phase. For a map in the restudying condition, subjects simply studied the target map for 100 additional seconds. If the map was in the retrieval practice condition, the subject then saw a version of the map with a feature removed. They were asked to identify the missing feature and covertly form a mental image of it. To assess their own accuracy on the trial, they were shown the map in its entirety and then reported whether or not they were able to retrieve the correct missing feature (this self-report served as the dependent measure for the retrieval practice phase). This sequence was repeated until all 12 features

were assessed 1 to 2 times. After a 30-minute filler task, they were given a blank sheet of paper and asked to draw both maps as well as they could. Given the inherent difficulty in setting criteria for a map reconstruction test, all drawings were scored against two different sets of liberal and stringent criteria. By all scoring measures, map reconstructions were more accurate in the retrieval practice condition.

Rohrer et al. (2010) also tested 4th- and 5th-grade students on a map-learning paradigm but under slightly different conditions. The largest difference being that the retention interval was much longer (1 day instead of 30 min.). Also more objective measures were used during the intermediate phase and on the final test. In addition the intermediate phase was longer because there were more retrieval practice opportunities per map and, to equate processing time, the restudying condition was designed to be of equal length. Despite the different population, retention interval, and method, Rohrer et al. (2010) also showed that testing enhanced retention of maps and their feature arrangements.

Understanding the Testing Effect

Much of the research on the testing effect explores its generalizability and applied utility in educational settings. However, some recent research has begun to explore theoretical accounts of the phenomenon. Initial research on the testing effect helps rule out one simple account, the possibility that testing merely provides another presentation of the items (i.e., those which can be retrieved) and that this additional processing time drives the effect. Consider a case in which a baseline condition is compared to a condition in which subjects received an intervenient test. Those in the baseline condition will encounter the memoranda only once. However, those who received the intervenient

test will be incidentally afforded another presentation of any item that they can retrieve. To determine if the testing effect reduces to an effect of re-presentation, one needs to compare re-studying and retrieval practice when equated on processing time. Carrier and Pashler (1992) did so by providing feedback in the retrieval practice condition of a paired-associate learning paradigm. In the re-study condition, paired associates were presented in their entirety for the duration of the trial, 10 seconds. In the retrieval practice condition, the first word in the pair was presented for only the first half of the trial (5 sec.). In this half, subjects were instructed to try to recall the second word in the pair. In the second half (5 sec.), both words were presented regardless of the subject's ability to correctly retrieve the second word. In this design, *overall* processing time was equated because both re-study and retrieval practice trials lasted 10 seconds. Under these conditions, the testing effect still emerged, indicating that the testing effect is not simply due to the effect of an additional presentation of the items (see Dempster, 1996; Roediger & Karpicke, 2006b; Butler, 2010 for additional evidence on this point).

Given that the testing effect is not merely an artifact of additional processing time, it must be attributable to more complex mnemonic processes. Several studies have proposed that retrieval difficulty plays a role (Glover, 1989; Carpenter & DeLosh, 2006; Kornell et al., 2009). This makes some sense when comparing typical restudying and retrieval practice conditions. The former has virtually no retrieval difficulty because the materials are presented intact whereas the latter will always be more difficult because the materials are not presented intact throughout. This simple explanation could be part of the foundation of the testing effect but Kornell et al. (2009) offer a supplement that is a little more precise. Their idea is that, the more difficult the retrieval task, the more

exploration and elaborative processing one must undergo before reaching the correct answer. Even if the correct answer is not reached, memory will often be better for items for which a retrieval *attempt* was made (Experiments 3-6). However, one should note that this theory's criteria for establishing the optimal level of difficulty are not comprehensive or explicit. Kornell et al. (2009) manipulated difficulty by varying the associative strength of paired associates (e.g., *train - track* vs. *train - caboose*) but prior studies employed alternative difficulty manipulations.

Glover (1989, Experiment 4) manipulated difficulty by varying the level of self-initiated processing demanded by the type of initial test. Before the final test, subjects were randomly assigned to take a recognition, cued recall, or free recall test over previously read passages (there was also a control condition that had only one exposure to the material). The type of final test (i.e., recognition, cued recall, or free recall) was also crossed factorially with the type of initial test to see if the effect of the difficulty manipulation persisted regardless of final test type. Glover predicted that the size of the testing effect should be positively related to the level of self-initiated processing required during retrieval practice (regardless of final test type). This means that an initial free recall test should create a stronger testing effect than an initial cued-recall or recognition test. Glover observed this exact pattern—regardless of final test type.

Carpenter and DeLosh (2006) supplemented Glover's (1989) account with an alternative difficulty manipulation. They varied the number of cues presented with an item during initial testing. In their third experiment, Carpenter and DeLosh had participants study a series of words that were found to be easily retrievable in one of their prior experiments. These items were selected to equate inherent item difficulty because

Carpenter and DeLosh's focus was on an independent manipulation of difficulty—so they had to ensure that item selection effects did not drive their results. To independently manipulate difficulty, during retrieval practice they varied the number of letters presented for a previously studied word (i.e., 1, 2, 3, or 4 letters). For example, if a subject had previously studied the word *cabin*, then in the most difficult retrieval condition they would be presented with the stem *c* _ _ _ _ and be asked to recall the correct word. Using this manipulation, Carpenter and DeLosh found that retention on a final test was inversely related to the number of letters provided during initial testing. This is presumably because cue scarcity requires more elaborative processing. In sum, it seems that retrieval difficulty moderates the testing effect and that the benefits of retrieval difficulty can be conferred through manipulations of cue availability, associative strength, or test type.

In additional to the role of difficulty, other studies have examined how testing affects certain kinds of processing. One approach to this level of analysis would be to determine the similarities and differences between the testing effect and other encoding manipulations whose effects on processing are more thoroughly understood. For instance, in a between-subjects design, generation—relative to reading—can enhance memory for the occurrence of an item but can simultaneously disrupt memory for the order in which items were presented (Nairne, Reigler, & Serra, 1991). This pattern of findings indicates that generation typically enhances item specific but disrupts relational processing (for a review, see Mulligan & Lozito, 2004). Because testing also yields clear gains in item memory, Karpicke and Zaromb (2010) examined the degree to which testing might disrupt order memory—and whether it qualitatively differs from generation.

They found that testing enhanced item memory (as measured by a recognition memory test) more than generation (Experiment 4), but that it disrupted order memory to the same extent (Experiment 3). So while retrieval practice and generation may differ in the degree to which they affect item memory, it can be argued that they both exhibit opposite effects on item and order memory.

By the same token, the testing effect seems to behave like other "unusual" encoding manipulations. In a recent review, McDaniel and Bugg (2008) noted that unusual encoding conditions (e.g. generation, enactment, and perceptual interference) often enhance item-specific processing relative to their more common control condition (e.g., read items, observed actions, and intact items, respectively), but they do so at the expense of disrupting order memory (and possibly other forms of relational memory). In a mixed-list design where the unusual encoding condition is intermixed with items from the common encoding condition, disruptions in order processing are mitigated by the presence of the common items—resulting in the unusual encoding manipulation producing better item memory than reading. However, in a pure list design where all of the items are in the unusual encoding condition, the disruption in order processing is so severe that the typical enhancement in item memory is washed out and performance is no different or worse than in a pure list of the common encoding condition. One should note that this property holds most reliably when the final test is free recall. This is because both item and order information contribute to successful free recall. If the final test is item recognition, the deleterious effects of the unusual encoding condition are often not found because successful item recognition is virtually independent of order memory. All of that aside, the testing effect may seem more robust than other unusual encoding

manipulations because it occurs in free recall in studies employing pure list designs (e.g., Roediger & Karpicke, 2006b, Experiment 2; Carpenter, 2009, Experiment 2; Karpicke & Zaromb, 2010, Experiment 1) as well as mixed list designs (e.g., Roediger & Karpicke, 2006b, Experiment 1; Carpenter, 2009, Experiment 1; Karpicke & Zaromb, 2010, Experiment 2). However, the item-order account provided by McDaniel and Bugg (2008) posits only that the unusual encoding condition produces worse memory in a between-subjects design. The benefits of unusual encoding manipulations often disappear under such conditions, but this is not a requisite of the item-order tradeoff account. It has been shown that the testing effect is found even in a between-subjects design, but its magnitude is still larger in a within-subjects design (see Karpicke & Zaromb, 2010, Experiments 1 & 2). The item-order account predicts this pattern of decline.

The interim conclusion is that Karpicke and Zaromb's (2010) results can be explained by the item-order account but additional evaluation of their study is required in regards to the kind of relational information they examined. Their results show how testing affects relational memory but, as they mentioned, their design examined only one kind of relational memory. Order memory differs from other kinds of relational information like semantic associative strength or category membership, and may be more sensitive to disruption. Zaromb and Roediger (2010) proposed that testing enhances semantic organizational processing—which facilitates recall. They presented subjects with a list of categorized words (i.e., the lists contained equal numbers of words from several categories). In their first experiment, the restudying and retrieval practice conditions were manipulated within subjects. The restudying condition consisted of 8

study trials (SS SS SS SS). There were 2 retrieval practice conditions—one with 2 intervenient tests (SS ST SS ST) and one with 4 (ST ST ST ST). Subjects encountered the same set of words for the duration of the experiment. Lastly, the final free recall test took place two days later.

To assess organizational processing, two measures were used: the adjusted ratio of clustering (ARC) and pair frequency (PF). ARC measures the extent to which the category structure of the list is reflected in the order of recall; by measuring the number of times members from the same category are recalled in succession (relative to the amount of category repetitions expected by chance). The PF measure assesses subjective rather than objective organization by examining how often two words are recalled in adjacent positions across recall tests. This measure of organization does not require that adjacent items be from the same category, allowing the assessment of idiosyncratic organization of the material. In passing, it should be noted that at least two recall trials are required for the PF measure, so this measure could not be assessed in the restudy condition.

The results showed that, compared to the restudy condition, total recall on the final test was higher in both retrieval practice conditions—so a testing effect *was* found. Surprisingly one measure of organization showed no significant differences among the conditions while another measure did. The measure showing no differences was the ARC, average ratio of clustering, and the one showing significant differences was the pair frequency (PF) measure. Recall that PF could not be computed for the restudy condition so it was only computed for the 2- and 4-test conditions. It was found that, on the final test, the PF score of the 4-test condition was higher than that of the 2-test

condition. In short, only subjective organization (as measured by PF) increased with the number of test trials.

Zaromb and Roediger speculated that ARC scores did not differ because there were so many encoding trials and encoding was manipulated within subjects. To address these matters, a second experiment was conducted in which encoding was manipulated between subjects. Those in the repeated studying condition studied a list twice and those in the retrieval practice condition studied a list once and then took a free recall test over it. Both groups studied three different lists of similar semantic composition to that of the first experiment. The final test occurred one day later. Overall recall was higher in the retrieval practice group so a testing effect was found. ARC scores were also significantly higher in the retrieval practice group. PF scores were not obtained. These results confirm Zaromb and Roediger's (2010) original hypothesis that testing can enhance organizational processing, which facilitates free recall.

The Effects of Testing Beyond Material That Is Retrieved

Most studies on the testing effect have focused on the fate of information that is successfully retrieved during initial learning or at least subjected to retrieval practice. Though recently some researchers have ventured into examinations of retrieval practice's more systemic effects. For example Chan (2009) examined the determinants of retrieval-induced facilitation and retrieval-induced forgetting, but his analysis focused on the information that was *overtly* subjected to retrieval practice (i.e. the presented items). However, one must be aware that encoding manipulations can affect information that is not overtly retrieved but still central to the task. Karpicke and Zaromb (2010) and Zaromb and Roediger (2010) made it clear that testing can affect relational information

but the specifics of their results were not uniform. Karpicke and Zaromb (2010) found that testing impaired one kind of relational memory (serial order), while Zaromb and Roediger (2010) found evidence that testing enhanced another kind of relational memory (category clustering)—but only under some conditions.

The discrepancy between these two studies could be a result of the inherent differences in the difficulty of encoding one kind of information over the other; serial order is often arbitrary but category membership is often meaningful. The discrepancy could also be due to the manner in which the relational information was processed during retrieval practice. Karpicke and Zaromb (2010) did not test order memory during retrieval practice so it could be that order memory was impaired because it initially received so little emphasis from the experimenter or the subjects, for that matter. Zaromb and Roediger (2010) did not overtly emphasize category membership in their recall task but it is very plausible that the subjects did. Again, category membership is meaningful—which makes it a very useful cue during retrieval. Given this notion—and considering that ARC scores increased at a negatively accelerating rate with each successive test—it is reasonable to believe that subjects became more reliant on semantic organizational processing over the course of the study. Of course a direct reconciliation of the discrepancies between these studies remains untested.

Another manner by which one can examine the fate of non-tested information is to investigate the effects of testing on transfer of learning. Transfer requires that one apply previously learned knowledge to novel contexts. These novel contexts must be, by definition, previously untested so these settings provide the proper grounds for examining the effects of testing on information that did not overtly receive retrieval practice. Butler

(2010) used this approach with verbal materials and he had subjects study several passages that contained facts and concepts[1] over various topics. In his second experiment, the restudying and retrieval practice conditions were manipulated within-subjects. The final test consisted of inferential questions over previously learned facts or concepts. These questions were not verbatim re-presentations of previous questions; they required that a fact or concept be applied in a novel way. He found that repeated testing enhanced transfer of previously learned facts and concepts. For present purposes, this finding will be termed *test-enhanced transfer*. A related study by Rohrer et al. (2010) found that testing enhanced transfer on a map-learning task. On the transfer test of their second experiment, Rohrer et al. presented children with the names of two previously-studied cities and then asked the children, when traveling along a specified route, though which cities would they pass when going from the origin to the destination? Rohrer et al. found that testing improved performance on this type of question. Unfortunately, it was not entirely clear how much of the test-enhanced transfer was due to *initial* testing. On the day of the final test, subjects performed a test over the individual map locations first and the transfer test followed. Performance on the transfer test still favored repeated testing but the introduction of an intervening test most certainly affected transfer test performance and therefore made it difficult to isolate the unique influence of initial testing. What these studies (i.e., Rohrer et al., 2010; Butler, 2010; Zaromb & Roediger, 2010; Karpicke & Zaromb, 2010) implicate—in one way or another—is the influence of testing on memory for relationships among items. Butler (2010) showed that testing could enhance memory enough to permit the organized application of multiple pieces of

[1] In his study, Butler (2010) defined facts as information gathered from one sentence and concepts as information abstracted from multiple sentences.

information downstream. It can also be inferred that Rohrer et al. (2010) provided at least some evidence that testing can create a representation coherent enough to permit the derivation of relationships from previously learned information. Zaromb and Roediger (2010) suggested that testing promotes organizational processing but they found that the level of observed organization depended on the experimental design and the dependent measure. Finally, Karpicke and Zaromb (2010) demonstrated that testing can enhance memory for individual items a great deal, but such enhancement comes at the cost of significantly disrupting memory for serial order. However, it is possible that serial order is a kind of relational information that is especially sensitive to disruption. Therefore a test of serial order might be too conservative a metric of the amount of preserved relational processing. Furthermore, it has been found that another strong encoding manipulation, generation, depletes serial order memory to a much smaller degree when the items have a high degree of inter-item relatedness (for a review, see Mulligan & Lozito, 2004). Perhaps this same property applies to the testing effect. Karpicke and Zaromb (2010) did acknowledge that their paradigm was designed to examine only one kind of relational information, so generalizations to other kinds of inter-item relationships could not be made.

In sum, the degree to which testing enhances a knowledge structure (i.e. the coherence and organization of memoranda) has not been made clear. To assess such matters (and some other basic empirical questions) a different paradigm would be useful—possibly one that examines more tractable properties like *abstractions* of representations. By abstraction, I mean the process by which one learns a structure that relates a set of items to one another (as opposed to simply representing them in isolation).

There are of course various degrees and classes of abstraction. There are instances in which one needs to abstract more conceptual properties through induction (such as in category/exemplar learning). There are also narrower instances of abstraction in which one need only learn an ordinal hierarchy that ranks items along a single dimension (e.g., A > B, B > C, etc.). Our study takes the first steps in assessing the impact of retrieval practice on abstraction so, for reasons to be made self-evident in forthcoming sections, our research will focus on the latter kind of abstraction. As will also be shown later on, previous studies on the testing effect have not truly examined abstraction. Some studies (i.e., Rohrer et al., 2010; Butler, 2010) had conditions that approximate the necessary examinations for this topic but no unambiguous conclusions can be drawn from them. The transitive inference paradigm might prove a viable approach to this problem. A review of transitive inference research immediately follows this section. After the review, the rationale for our current study is explained in greater detail and the design of our two novel experiments is provided.

Transitive Inference

Transitive inference is, in many respects, reasoning about concepts or elements within a hierarchy. For example, if one is told that Bill is taller than Steve, who is taller than Al, one can easily infer that Bill is taller than Al. This is transitive inference (TI). For pedagogical purposes, TI is normally introduced with an example that employs superlatives regarding a nameable dimension (such as height), but laboratory settings often examine TI with more demanding premises. Most importantly for our purposes, the TI paradigm provides a controlled venue for examining the influence of repeated testing on relational processing and representation abstraction because 1) the relationships

among the stimuli can be fixed along a single dimension and 2) novel problems can be easily constructed from non-adjacent elements of the series.

When researchers use TI to examine the abstraction of relationships, the paradigm is typically more complex than the example described above. First of all, TI paradigms often involve learning about the relationships among five or more elements. The elements of the stimulus set are usually novel and arbitrarily ranked (i.e., A > B > C > D > E). The adjacent elements make up the premise pairs—whose relationships are taught overtly. That is, participants learn over several trials that A > B, B > C, and so forth. The structure of the hierarchy creates inherent differences in the reinforcement histories of the elements. A is always reinforced because it sits at the top of the hierarchy and E is always avoided because it rests at the bottom. The internal elements (B, C, and D) have different reinforcement histories; half of the time they are rewarded (e.g. C > D) and, for the other half, they are not rewarded (e.g. B > C). TI paradigms have been used with humans and animals (typically rats). Similar reinforcement histories do arise in both groups albeit through different circumstances. With rats, selection of the correct stimulus from a pair yields a tangible reward like food, whereas with humans selection of the correct stimulus is typically not accompanied with a reward (though feedback is usually provided).

In either case, subjects learn the premise pairs and then advance to a final test. On the final test, their memory for the basic premise pairs (e.g. B > C) is assessed and they are also presented with novel combinations of the elements to determine the extent to which they demonstrate TI (i.e., to measure the extent to which they have abstracted the relationships among items never presented together). Finally, it should be noted that for a

five-element stimulus set, there is only one combination that provides a true test of transitive inference. The AE pair does not provide this test because both elements have consistent reinforcement histories. In fact, any transitive test pair involving A can be solved by merely remembering that A was always reinforced. Conversely, any transitive test pair involving E can be solved by merely remembering that E was always unrewarded. Therefore the only appropriate transitive test pair is BD. In this context, transitive inference is operationalized by the choice of B in the BD pairing.

Basic Findings & Theoretical Accounts of TI

Using this basic paradigm, transitive inference has been observed with rats using odors as the elements (Dusek & Eichenbaum, 1997; Van Elzakker, O'Reilly, & Rudy, 2003). With humans it has been found when using non-verbal materials like Japanese Hiragana characters (Greene, Spellman, Dusek, Eichenbaum, & Levy, 2001; Frank, Rudy, Levy, & O'Reilly, 2005), abstract images (Libben & Titone, 2008), and basic shapes (Lazareva & Wasserman, 2010). Dusek and Eichenbaum (1997) proposed that transitive inference is the result of explicit, declarative representational flexibility because lesions to structures connecting the hippocampus to cortical or sub-cortical areas in rats impairs performance on transitive pairs while leaving performance on the overtly learned adjacent pairs intact. Of course their conclusion largely hinges on the notion that the hippocampus is involved exclusively in explicit relational processing (for an alternative account, see Frank, Rudy, & O'Reilly, 2003). Dusek and Eichenbaum's (1997) explicit processing-based account of TI assumes that all elements of the premise pairs have equal associative strengths. This means that, in a paradigm with five elements, A through E, the strength of the association between the first premise pair, AB, should be

equal to the associative strength between all other adjacent pairs. Furthermore—in a paradigm with *six* elements, A through F, the strength of the association between elements of the internal pairs (BC, CD, and DE) should all be equal. So, according to the explicit processing account, TI performance should be equal on the novel BD, CE, and BE pairs because the weights connecting each element in the chain have equal strength. To the contrary, several studies have found this to be false; performance on the BE pair tends to be much better than the other two novel TI pairs, CD and BD (Van Elzakker et al., 2003; Frank et al., 2005; Libben & Titone, 2008). All of this aside, one could still argue that the TI paradigm measures abstraction—though this abstraction need not form a declarative memory.

Though, to be clear, this alternative theoretical account of transitive inference is not proposing that transitive inference is devoid of relational processing; it only means that TI is not always achieved through an entirely explicit and logical process. What this means for our yet-to-be-proposed study is that any enhancement in performance cannot be attributed to an augmentation of explicit logical processes. If, on the other hand, our experimental manipulation promotes awareness of the hierarchy among the stimuli, then the use of explicit logical processes can ensue.

On a related note, some studies have shown that declarative awareness of the hierarchy is not necessary for successful performance but it is associated with *better* performance (Green et al., 2001; Frank et al., 2005; Libben & Titone, 2008; Lazareva & Wasserman, 2010). This means that, when TI performance is at ceiling, a declarative strategy is at play. If there is a gradient in TI, where performance is better when the elements are farther apart, then implicit/non-declarative reasoning probably takes place.

This spectrum of performance is a potential boon for our study because, under the right conditions, we can examine different sorts of transitive inference.

The Double-Function List: Relative of the TI Paradigm

As stated earlier, the TI paradigm has not been heavily used in conventional episodic memory research. However, there are a few studies employing the double-function list paradigm (which is a close relative of the TI paradigm). Consideration of these studies might inform our expectations about the effect of testing on transitive inference. In the double-function list paradigm, subjects learn a series of cue-target pairs in which the cue for each pair serves as the target for another pair. For example, one might have to learn pairs such as *eye-hat, hat-jug, jug-cat,* etc. (Primoff, 1938). These pairs are presented in a random order to cloud the relationships among them. One critical difference between this paradigm and the TI paradigm is that in a double-function list, the target of the last pair in the list serves as the cue in the first pair. This effectively links the stimuli into a loop of paired associates; there is no hierarchy.

Primoff's (1938) classic study was originally designed to investigate why learning overlapping associations (e.g. A-B, B-C, C-D, & D-A) is more difficult than learning only isolated associations (e.g. A-B and C-D). In some cases, learning double-function lists can take twice as long as single-function lists (Primoff, 1938; Slamecka, 1976). What is most important for present purposes is that the overlap among the stimuli and the dilemma of encoding such overlap are integral properties shared by the double-function list and transitive inference paradigms. This commonality between the two paradigms can guide our predictions about an amalgam of the TI and testing effect paradigms.

Slamecka's (1976) re-evaluation of double-function lists incidentally provided such guidance.

Before delving into Slamecka's (1976) work, it should be noted that Primoff (1938) concluded that learning a double-function list is so arduous because of the pervasive interference. Slamecka (1976) acknowledged that this conclusion was reasonable but that Primoff (1938) did not identify the locus of interference. Did it originate in encoding, retrieval or both? Over the course of three experiments, Slamecka (1976) concluded that the interference arose from processes at encoding and retrieval, but that retrieval was more culpable. At encoding, the immediate backward associate was significantly problematic; no other associate had a significant effect. For example, if one had already learned the pair *eye-hat*, then learning *hat-jug* would be impeded by *eye* in the *eye-hat* pair. Retrieval proved particularly problematic because, upon retrieving a given paired associate, interference arose from immediate *and* distant associates. The immediate backward associate was the primary source of impediment. More distant associates in both the backward and forward directions contributed to interference as well, but to a lesser degree. So both encoding and retrieval were both deemed sources of difficulty, but difficulty at retrieval (i.e., production) was the most observable (and most comprehensively evaluated in the three experiments). Insofar as double-function list learning is similar to transitive learning, Slamecka's (1976) results suggest that repeated testing would impair transitive learning more than restudying.

Rationale for the Current Study

Indirect support for our current study can be found in Butler's (2010) study on test-enhanced transfer. Although Butler's (2010) study is not directly comparable to a

standard transitive inference paradigm, Butler's use of inferential questions on the final test is relevant[2]. The inferential questions required the use of only one previously learned fact to resolve a question that was related to the studied domain—but whose answer was not overtly taught. Butler provided the following example: participants might read in a passage "there are over 1,000 bat species." On a later inferential question, they would be asked, "If there are about 5,500 species of mammals in the world, approximately what percent of all mammal species are species of bat?" To answer successfully, a participant would first recall that there are at least 1,000 bat species and then divide that number by 5,500 to then conclude that bats account for approximately 20% of all mammal species. This inferential aspect is akin to a transitive inference problem but recall that a TI problem incorporates two previously learned elements—whereas Butler's (2010) inferential questions incorporated one—and TI problems do not require one to respond to questions providing *completely* novel information (i.e., that there are 5,500 species of mammal in the world). Even though Butler's (2010) design is not directly comparable to a transitive inference paradigm, it implies that repeated testing can facilitate applying previously learned knowledge to situations that require one to incorporate that knowledge in the production of a novel response to a novel question. If this is the case, then repeated testing should enhance TI performance compared to restudying.

However, a central demand of transitive learning is overcoming the deleterious effects of interference accumulation. One must at least understand that there are overlapping associations among the stimuli (e.g., B > C and C > D), but one must also make sure not to confuse them and/or produce the incorrect response. While the impact

[2] One might think that the conceptual questions would be most comparable to TI questions. However, the concepts were taught overtly and, in TI paradigms, the transitive test pairs are not. Therefore the conceptual questions would not be the best analogue of TI test pairs.

of retrieval practice on this exact kind of interference has not been studied, there is some evidence that testing can at least reduce *proactive* interference. Szpunar, McDermott, and Roediger (2008, Experiment 1A) had participants learn 5 lists of interrelated words. The lists contained words from the same semantic categories but no word appeared in more than one list. In the retrieval practice condition, participants took a free recall test after studying each list. Those in the restudying condition solved math problems after each list. The fifth and final list was the only one for which both groups received a free recall test. Here it was found that those in the retrieval practice condition recalled more words overall and committed fewer prior list intrusions. While these results bode well for instances akin to list-learning, the conclusions drawn from them cannot be directly applied to transitive learning because the interference observed in Szpunar et al. (2008) was across multiple lists rather than within a single list. Therefore it is not entirely clear if in their case testing actually helped resolve interference among overlapping associations, if it promoted temporal event segmentation, or both. On the other hand, Slamecka's (1976) double-function list-learning study, by its very nature, did focus on within-list associations and found that retrieval impaired list acquisition more than studying. However, Slamecka's (1976) study was designed in the verbal learning tradition of the time so it used alternating study and test trials; participants were not randomly assigned in a between-subjects fashion to a middle phase of mass restudying or testing. As will be shown later, our study manipulates restudying and retrieval practice between subjects to more directly assess the impact of intervenient encoding manipulations like these. Though, insofar as our current study is comparable to a double-function list paradigm, we should find that repeated testing enhances memory for the

premise pairs but does not facilitate the formation of hierarchical representations—and may even hinder it.

The Importance of Abstraction

Recall that Butler (2010) demonstrated that testing could enhance transfer of learning. While his findings are meritorious in their own right, they do not demonstrate abstraction. Butler (2010) did not aim to examine that but the reader should be made aware of the differences so that the unique contribution of our study will be evident. One might think that performance on the conceptual questions demonstrates abstraction because successful completion of them required combining information learned from multiple sentences of a passage. However, answers to the conceptual (and factual) questions were taught overtly during retrieval practice. Although a retrieval attempt was required during these intervenient tests, feedback was given regardless of accuracy. This likely promoted a form of stimulus-response learning for both the conceptual and factual questions. Also the inferential questions did not require the flexible recombination of previously learned elements. In essence, the subjects were not left entirely up to their own devices to derive the conceptual information. So it would appear that testing only increased the likelihood of transfer of overtly taught information. The only requirement on the subject's part was that they remember and recognize the relevance of previously learned units information. The effect of repeated testing on one's ability to form new knowledge structures and derive relationships from them is still uncertain.

Rohrer et al.'s (2010) map learning study might also appear to demonstrate abstraction but aspects of their design leave the matter uncertain. One critical issue was that an unlabeled map was present for the entirety of the transfer test. This was probably

done because their subjects were children and it makes the test much more manageable. Regardless, it does not allow one to see if their subjects truly abstracted the topography. On a transitive inference test, the analogue of that approach would be presenting a hierarchy with empty slots and then asking the subjects to make transitive inferences; they would then be made aware of the knowledge structure by an external agent. Furthermore Rohrer et al.'s (2010) design does not allow one to easily examine the systematic effects retrieval practice bears on a knowledge structure. Since TI paradigms with 6 or more elements contain TI pairs of varying difficulty, one could determine if testing affects the finer and/or coarser aspects of a knowledge structure.

Additional Advantages of the TI Paradigm

An amalgam of the TI and testing effect paradigms provides a different *angle* for examining the effects of testing on relational processing but this alone does not necessitate such an undertaking. However, there are several additional advantages of such an amalgam.

First of all, this approach allows one to examine the effect of testing on a specific kind of reasoning: hierarchy abstraction. Also one can specify and hold constant the relationship(s) among the stimuli. The use of richer or more naturalistic materials often does not permit this kind of specification so it then becomes difficult to identify which kinds of relationships are influenced by testing. One can bluntly assess the impact of an encoding manipulation on associative memory formation, but one must acknowledge that associations are not uniform in their classification. The kinds of associations one must learn are usually defined by specific relationships (e.g., *x* is greater than/less than *y*; *x* is an exemplar of *y*; *y* is a function of *x*; *x* is necessary for *y* to occur; *x* came before *y*).

Richer materials often contain an assortment of several classes of relationships. Given the variety of potential associations, it is important to specify (and control for) the kind of relationship being examined.

An amalgam of the two paradigms also provides a practical advantage. The testing format of a typical TI paradigm allows one to hold the response format constant while varying the question difficulty (i.e., adjacent pair vs. TI pair). This ensures that differences in the difficulty among question types are not confounded by response output demands. For example if one wanted to assess the effect of item difficulty, one would not use a recognition test for the easy items and a recall test for difficult ones. Such a design conflates item difficulty with response demands. This is not a problem in the TI paradigm because the response demands are the same for every question: press a key. However, some questions are over adjacent pairs and others are over transitive pairs. Thus the two types of questions differ only on two dimensions: 1) whether or not the relationship between the elements is direct and 2) whether or not the relationship being assessed was overtly taught.

Experiment 1

Methods

Participants. Fifty-six undergraduates were recruited in exchange for course credit. All were young adults who reported having normal or corrected-to-normal 20/20 vision. Also none of the participants reported having color blindness.

Materials. In our adaptation of the TI paradigm, the stimuli were paintings by largely unknown artists. All paintings were of ordinary landscapes and skyscapes[3].

[3] These paintings were taken from stimuli compiled by Kornell and Bjork (2008) in a study on induction. They were found to not be overly distinctive but subjects could remember them at above chance levels.

Paintings were chosen as the stimuli because the standard TI paradigm uses nonsense figures with many, many trials. In these settings, the goal is to use stimuli that are not easily labeled, and whose properties are not easily placed on any sort of discrete continuum. For this reason we wanted to use non-verbal materials. However, we also needed to have intermediate levels of performance and room for increase due to re-study or retrieval practice. Consequently, we sought unfamiliar paintings because they are easier to learn than non-sense shapes (and so can be learned in a few trials), but are not easily labeled or given verbal codes. We randomly chose seven paintings from the bank provided by Kornell and Bjork (2008). Each painting came from a different artist. The paintings were counterbalanced across subjects so that they appeared in all positions of the hierarchy and both sides of the screen equally often.

Design & Procedures. During initial learning, the paintings were presented in pairs on a computer screen to convey which was the more valuable of the two. Three green dollar signs appeared below the more valuable one at the onset of the trial and remained for its duration. The subjects' task was to press a corresponding key to confirm that they understood which painting was the more valuable (the *z* key for paintings appearing on the left and the *m* key for those on the right). For every trial, a pair appeared for 7.5 sec and then a cross hair appeared for 500 ms; the next pair followed in the same manner. All pairs were presented 3 times in a block-randomized manner. This means that all pairs were presented in a random order and then this process repeated two more times (each time in a new random order). During initial learning, the participants' task was to memorize the relationships in these premise pairs because their memory for them would be tested later. The nature of the final test was not conveyed. After the

learning phase, participants restudied the premise pairs or were tested over all of the pairs they had just learned. These phases were the restudying and retrieval practice conditions, respectively, and they were manipulated between subjects. This was necessary to examine the pure effects of one strategy versus another. Not to mention that, given the inter-related nature of the knowledge structure, manipulating encoding within subjects would have likely created diffusion across the two conditions.

The restudying phase was identical to the initial learning phase with only a few changes. The subjects were reminded of the initial instructions but were also told that they were being given additional time to study the premise pairs before the memory test. The retrieval practice condition proceeded differently. Subjects were told that this phase was practice for the later memory test. The overall time for each trial was the same as the other phases (7.5 sec), but 5 sec of that time were devoted to retrieval practice and the remaining 2.5 sec were devoted to computer-provided feedback. During the first 5 seconds, they were presented with one of the premise pairs and asked to indicate which was the more valuable of the two by pressing the appropriate key (*z* or *m*). After 5 seconds, three green dollar signs appeared below the correct painting for 2.5 seconds—regardless of whether or not they could retrieve the correct answer. A crosshair then appeared for 500 ms and the next pair followed. All premise pairs were tested (or restudied) 3 times in a block-randomized manner.

A mental arithmetic distractor task followed the second phase. Here subjects had to complete true/false arithmetic problems for 6 minutes. For each problem, they were shown an answer that could be true or false. They had up to 15 sec to decide. In addition the computer provided feedback after every trial. The feedback indicated whether or not

they solved the problem correctly, and a running percent correct was displayed as well. This was provided to help them gauge their accuracy and to ensure that they were engaging in the task. However, no fixed level of accuracy was desired. They were simply told to do their personal best.

Immediately after this distractor phase came the final memory test. Subjects were tested over all of the relationships they were taught explicitly (e.g. "Which is worth more: A or B?"), and over the relationships they were not taught explicitly (e.g. "Which is worth more: B or E?"). The latter kind of question required transitive inference and such questions appeared only on the final test. All possible combinations of the elements were tested 4 times in a block-randomized fashion. This is typically done in TI studies to provide reliable accuracy estimates of the pairs. Of course no feedback was given on the final test and the questions were self-paced. The subjects were told that they would see pairs of paintings multiple times and that they should try to recall the correct answer. If they were not sure, they were instructed to give their best guess. Once the final test was complete, they were asked to fill out an awareness questionnaire described in Appendix A. Awareness of the hierarchy can facilitate performance and it can be assessed with a post-experiment questionnaire. The questionnaire asked participants if they were generally aware of a hierarchy and if they could explicate it. We used questionnaires adapted from another transitive inference study (Frank et al., 2005). In this experiment, the questionnaire was only used for potential post-hoc comparisons. After they completed the questionnaire, subjects were debriefed and dismissed.

Results

Two participants were excluded from the full analysis because their premise pair memory on the final test was more than 2 standard deviations below their respective group means. One participant was dropped from the restudy group and one from the retrieval practice group. All subsequent analyses are based on the remaining 54 participants.

Phase 2 Retrieval Practice Performance. We first examined performance during the retrieval practice phase in order to check our manipulation. Performance on the first block (M = .83, SD = .17) was significantly above chance, $t(26) = 10.365$, $p < .001$. Therefore the 3 blocks dedicated to initial learning were sufficient for inducing above chance premise pair memory. This ensured that participants had acquired a representation sufficiently stable for engaging in retrieval practice. Average performance across all three blocks was also good but not at ceiling (M = .82, SD = .14).

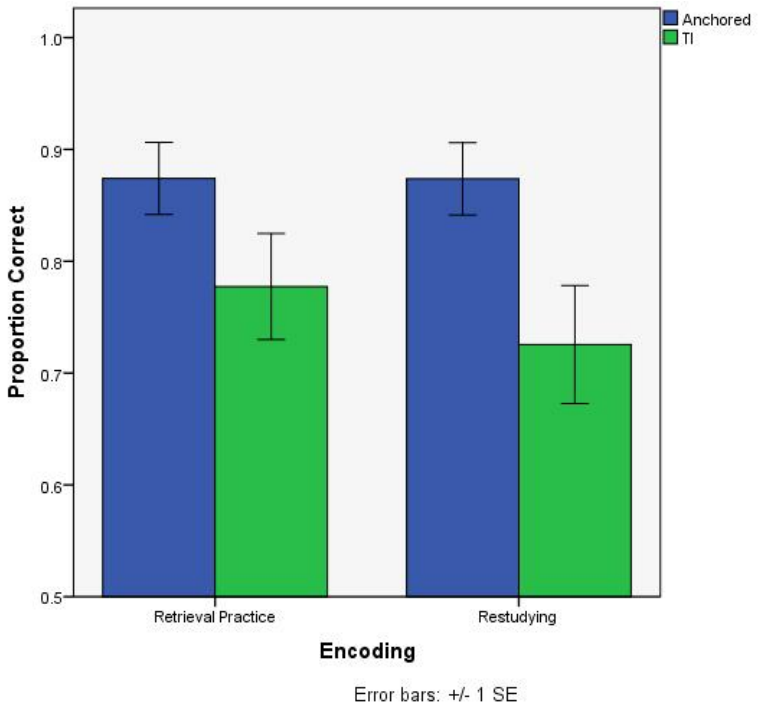
Overall Final Test Performance. For premise pair memory, we found no significant difference between restudying and retrieval practice, $t(52) < 1$. This result facilitates interpreting potential differences in representational flexibility because they cannot be attributed to differences in overall memory strength of the premises.

We then analyzed performance on the non-adjacent pairs. For the first such analysis, we broke these pairs down into two general types of problems: those involving a combination of an anchoring element (i.e. painting A or painting G) and a non-adjacent internal element, and those involving a combination of only non-adjacent internal elements (e.g., B v. D; C v. E, etc.). The former type will henceforth be referred to as anchored pairs and the latter, TI pairs. This analysis served two purposes: 1) to act as a

manipulation check to ensure that we were examining the typical pattern of results in TI studies and 2) to examine group differences in overall representational flexibility.

A 2 x 2 (Encoding by Problem Type) repeated measures ANOVA showed that there was no main effect of encoding, $F < 1$. There also was no encoding by problem type interaction, F (1, 52) = 1.329, MSE = .018, p =.248. However, we did find a significant main effect of problem type, F (1, 52) = 30.873, MSE = .405, $p < .001$. Overall percent correct was higher for anchored pairs than TI pairs. This is typical of TI studies (Van Elzakker et al., 2003; Frank et al., 2005). The critical result, however, was that there was no difference in TI between encoding conditions.

Figure 1. Experiment 1 Performance on Anchored & TI Pairs

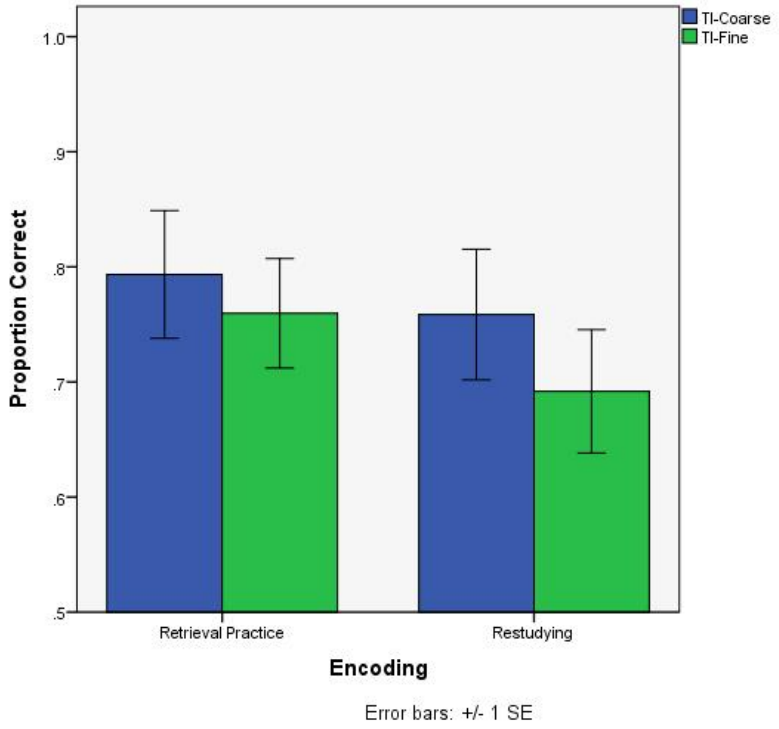


A non-significant difference in performance on the anchored pairs is to be expected because they can be solved via simple approach/avoid strategies. Therefore, any encoding manipulation is unlikely to confer any additional benefits to performance.

Recall, however, that problems involving the non-adjacent internal elements of the hierarchy are the best tests of TI because these elements have variable reinforcement histories, and so cannot be solved via simple approach/avoid strategies. Since we found no significant differences in overall TI, we concluded that there were no significant differences in overall representational flexibility. However, it is possible that one encoding condition may have exhibited superior performance on a certain kind of TI problem. To examine this, we assessed TI for coarse- and fine-grained problems. Recall that TI problems BD, CE, and DF require evaluating the minimum number of premises for a true TI problem: two. They also typically require a more discerning evaluation because there is less symbolic distance between the elements. As such, these problems were collapsed into the fine-grained condition. The remaining TI problems (BF, BE, and CF) encompass 3 or more premises so they were collapsed into the coarse-grained condition. The following analyses are based on these categorizations.

A 2 x 2 (Encoding by Grain Size) repeated measures ANOVA revealed that there was no main effect of encoding and no encoding by grain size interaction (both F's < 1). However, there was a marginally significant main effect of grain size, F (1, 52) = 3.736, MSE = .0068, p = .059. Of the TI problems, coarse-grained problems were somewhat easier to solve than fine-grained ones. Other TI studies involving more than 5 elements have found a similar pattern (Van Elzakker et al., 2003; Frank et al., 2005).

Figure 2. Experiment 1 Performance on Coarse- & Fine-Grained Pairs



These results show that the previous non-significant difference in overall TI extended to both grain sizes. Therefore, we can conclude that, under these conditions, both encoding manipulations yield comparable profiles in representational flexibility—regardless of grain size. For a summary of performance on all measures analyzed thus far broken down by encoding condition, see Table 1.

Table 1. *Experiment 1 Final Test Performance*

	Retrieval Practice	**Restudying**
Premise Pairs	.81 (.13)	.81 (.15)
Problem Type:		
Anchored	.87 (.17)	.87 (.17)
Transitive Inference	.78 (.25)	.72 (.27)
TI by Grain Size:		
Coarse-Grained Problems	.79 (.29)	.76 (.29)
Fine-Grained Problems	.76 (.25)	.69 (.28)

Note. Means are displayed with standard deviations in parentheses.

Final Test Performance Broken Down by Test Half. A potential concern for the typical TI paradigm is that any differences between conditions could wash out over the course of the final test. Transitive inference trials are conventionally assessed multiple times over multiple test blocks. However, as the test progresses, the learning histories of the encoding conditions begin to develop more commonality and thus their initial differences might become obscured. To investigate this possibility, we broke down performance by test half to see if the pattern of differences changed qualitatively over the course of the test.

In short, we found that the pattern of differences we observed in our initial analysis did not differ qualitatively across the test halves. We only found a small (~4%) drop in premise pair memory across the test halves. Most importantly, this effect did not interact with encoding; thus the rate of decline was the same for both conditions. For a more detailed discussion of the analysis, the interested reader is directed to Appendix B.

Interim Conclusions. Our initial analyses showed that there were no differences in representational flexibility—regardless of grain size. Our exploratory analyses of performance across test halves also found that this pattern did not qualitatively change as the test progressed. Although premise pair memory did show a small but significant drop in the second half, the rate of decline did not differ between groups. Therefore, for every epoch of the test, it appears that there was no qualitative change in the pattern of non-significance.

Awareness Questionnaires

Recall that Frank et al. (2005) and Libben and Titone (2008) have shown that the degree of awareness of the hierarchy is positively associated with TI. Those who

demonstrate awareness of the hierarchy perform better than those who show no awareness. As such, we wanted to examine TI differences between the conditions when controlling for awareness.

Awareness Criteria. The questionnaire we adapted from Frank et al. (2005) has no criteria for weighting responses to compute an "awareness score"; rather it simply asks increasingly specific questions to determine if someone is aware of the hierarchical nature of the memoranda. Participants were deemed aware if they could clearly convey that they noticed the hierarchy among the paintings. However, they need not use the word hierarchy, they could simply state that they noticed that the paintings were ranked from least valuable to most. Also—a participant did not need to be able to explicate the hierarchy in its entirety. It is possible that a participant could be aware of the fact that there is a hierarchy, but have imperfect memory of the entire structure. This is not an unreasonable distinction given that TI performance overall was well above chance but below ceiling.

Participants who were deemed unaware demonstrated no clear knowledge of the hierarchy. Those who said that they went with instinct or guessed were deemed unaware. Those who said they based their judgments on the perceptual characteristics were also deemed unaware because the perceptual characteristics of the paintings (e.g., color, vividness, style) were orthogonal to their rank. Some participants reported that they thought the paintings were organized around a structure, but reported the wrong structure. For this, they were deemed unaware because endorsing the wrong structure leads to incorrect inferences. Finally, some participants just reported that they did not notice anything and for this they were also deemed unaware.

Reliability Analyses. Of the 54 subjects, the responses for 46 of them were evaluated by two raters (the author, M. P., and a research assistant). The raters were always blind to the encoding condition. Eight responses were not included in the reliability analysis because they were discussed as examples by both raters in order to establish and evaluate criteria. Of the remaining 46 responses, we observed good reliability (proportion agreement = .90). For the eight example responses, their agreed upon evaluation (i.e., aware or unaware) was entered into the subsequent analyses of awareness. For the remaining 46 responses, those that received unanimous agreement from both raters were automatically entered into the subsequent analyses. Those for which the raters disagreed were discussed until an agreement was reached. This final evaluation was then entered into the subsequent analyses.

Awareness Rates. The proportions of aware and unaware participants were not significantly different between groups, $\chi^2 < 1$. Therefore, the majority of participants in both groups became aware of the hierarchical structure and the rate of awareness did not differ significantly between groups. This is not too surprising given that the nature of the hierarchy was likely more obvious than in many TI studies. The paintings were presented in terms of worth. Even though no dollar amounts were appended to any painting, worth is a familiar, meaningful dimension. Most TI studies provide no obvious dimension on which one can rank the stimuli—only their relative preferences are reinforced through massed trial-by-trial feedback.

Effects of Serendipitous Awareness. Before evaluating TI differences as a function of awareness, we needed to assess differences in premise pair memory. If present, they could mediate differences in TI. A 2 x 2 (Encoding by Awareness)

ANOVA showed that there was no main effect of encoding ($F < 1$) and no encoding by awareness interaction ($F < 1$). We did, however, find a significant main effect of awareness, F (1, 50) = 8.075, MSE = .148, $p < .01$. Participants deemed aware had better premise pair memory than those deemed unaware. This is somewhat consistent with prior studies. When awareness is experimentally induced, it can lead to greater premise pair memory (Greene et al., 2001) or expedite acquisition of the premise pairs (Libben & Titone, 2008). The rates of awareness and a breakdown of premise pair memory are all shown in Table 2.

Table 2. *Awareness Rates & Premise Pair Memory*

	Retrieval Practice	**Restudying**
Aware	20	19
Unaware	7	8
Aware	.84 (.12)	.86 (.14)
Unaware	.75 (.15)	.71 (.13)

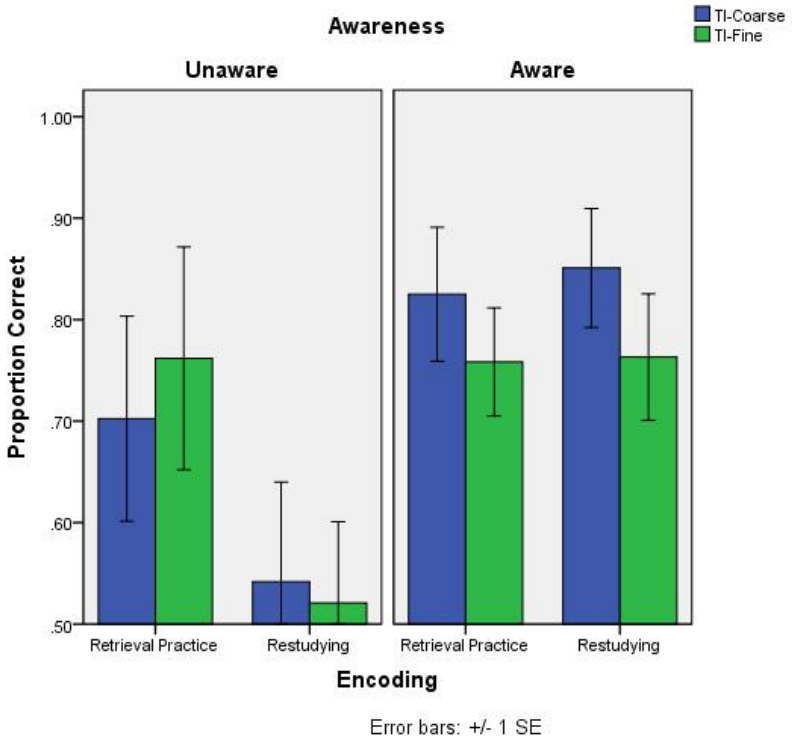
Note. The upper half of the table displays the raw numbers of participants in each encoding condition who were classified as aware and unaware. The lower half of the panel displays the premise pair memory means and standard deviations (in parentheses) for each encoding condition broken down by awareness classification.

We then examined differences in overall TI (regardless of grain size). A 2 x 2 (Encoding by Awareness) ANOVA revealed no main effect of encoding, F (1, 50) = 1.505, MSE = .093, p = .226. There was also no encoding by awareness interaction, F (1, 50) = 2.045, MSE = .126, p = .159. However, there was a significant main effect of awareness, F (1, 50) = 4.916, MSE = .303, $p < .05$. By and large, aware participants outperformed unaware participants on TI pairs. This did not differ significantly between encoding conditions.

Finally, since awareness has been found to modulate grain size effects in TI (Frank et al., 2005; Libben & Titone, 2008), we structured our last analysis of awareness

around coarse- and fine-grained TI problems. Using a 2 x 2 x 2 (Encoding by Awareness by Grain Size) repeated measures ANOVA, we found no main effect of encoding, F (1, 50) = 1.505, MSE = .186, p = .226, and no main effect of grain size, F (1, 50) = 1.012, MSE = .018, p = .319. Like before, there was a significant main effect of awareness, F (1, 50) = 4.916, MSE = .607, $p < .05$, but no interaction between encoding and awareness, F (1, 50) = 2.045, MSE = .252, p = .159. There was, however, a marginally significant interaction between grain size and awareness, F (1, 50) = 2.818, MSE = .05, p = .099. No other interaction approached significance (all F's < 1, all p's > .38).

Figure 3. Experiment 1 – TI by Awareness



As in the preceding analysis, aware participants overall outperformed unaware participants. The current analysis extended this finding to both grain sizes. Although there were hints of interactions between awareness and encoding, as well as awareness

and grain size, neither of these reached significance. This is likely due to the same reasons encoding and awareness did not interact in the preceding analysis when TI was collapsed across grain sizes.

Conclusions

Under these conditions, retrieval practice and restudying yielded no differences in transitive inferences; nor did they differ across grain sizes, awareness, or any combination thereof. Also note that premise pair memory was equal between the two conditions so the interpretations of TI ability were rather straightforward. Furthermore, under the current experimental design, both encoding conditions made most participants aware of the hierarchy, thereby facilitating transitive inferences. This facilitation seemed to extend to both coarse- and fine-grained problems. As one might intuit, aware participants had better premise pair memory and outperformed unaware participants on all forms of TI. Critically, this did not differ between the two encoding conditions. By all measures, this design yielded no significant or meaningful differences in representational flexibility between the groups.

The non-significant differences observed in this experiment could be the true absence of a difference or a result of our experimental design. Given that there were relatively few blocks in the intermediate phase (i.e., retrieval practice or restudying), we felt that our null results could be attributable to an insufficient manipulation of retrieval practice. It is possible that using only 3 blocks in the intermediate phase rendered the retrieval practice and restudy groups too similar. Therefore, our next experiment strengthened the manipulation by increasing the number of blocks in the intermediate phase to 6. In addition, performance was relatively high following initial learning—

raising the possibility that there was not enough room to examine the effects of the encoding manipulation. Thus, we cut down the number of initial learning blocks from 3 to 2. Note that this still leaves a total of 8 learning blocks (a 33% increase in total learning time). An even number of blocks was required to ensure that the number of presentations on each side of the screen was equal for each painting in a premise pair. Finally, also note how this design structures the learning conditions so that participants now spend the majority of their learning time (75%) in their respective encoding condition. In the previous experiment, they spent only 50% of their learning time in their respective condition. As we shall see, this change was sufficient to induce differences between the two groups.

Experiment 2

Methods

Participants. Fifty-six undergraduates were recruited in exchange for course credit. None of them had participated in the previous experiment. All were young adults who reported having normal or corrected-to-normal 20/20 vision. None of the participants reported having color blindness.

Materials. The materials were the same as those used in Experiment 1.

Design & Procedures. The design was the same as Experiment 1 (i.e., between-subjects). We only modified the way in which study time was distributed. The number of initial learning blocks was reduced to 2 and the number of intervenient learning blocks was increased to 6; yielding 8 total blocks. All other aspects of the encoding, distractor, and final test phases were the same as Experiment 1.

Results

Two participants were excluded from the full analysis because their premise pair memory on the final test was more than 2 standard deviations below their respective group means. One participant was dropped from the restudying group and one from the retrieval practice group. All subsequent analyses are based on the remaining 54 participants. When necessary, the degrees of freedom were corrected for unequal variances.

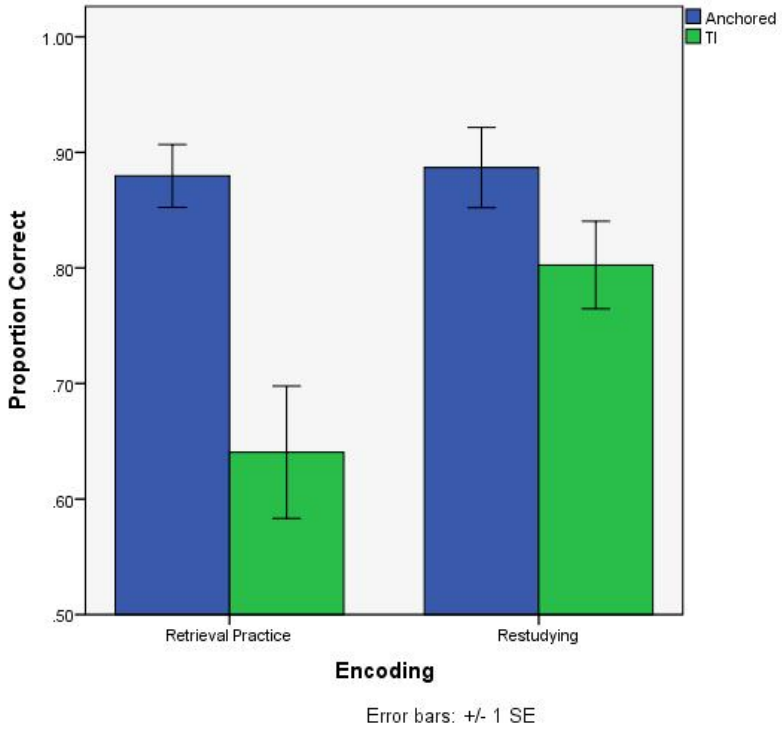
Phase 2 Retrieval Practice Performance. We first examined performance during the retrieval practice phase in order to check our manipulation. Performance on the first block (M = .72, SD = .23) was significantly above chance, $t(26) = 4.996$, $p < .001$. This shows that the 2 blocks dedicated to initial learning were still sufficient for inducing above chance premise pair memory before moving on to the intermediate phase. Mean performance across all blocks was also good but not at ceiling (M = .81, SD = .15).

Overall Final Test Performance. For premise pair memory, there was no significant difference between encoding conditions, $t(52) = 1.004$, $p = .32$. Because the groups did not differ in this regard, any subsequent differences in transitive inference can be interpreted solely in terms of representational flexibility and not as mediated by differences in overall memory strength of the premises.

As in the first experiment, we analyzed performance on the non-adjacent pairs. Recall that this analysis serves as a manipulation check on anchored and TI pairs, and as a test for any differences in overall TI. A 2 x 2 (Encoding by Problem Type) repeated measures ANOVA revealed a marginally significant main effect of encoding [$F(1, 52) = 3.268$, $MSE = .193$, $p = .076$], a significant main effect of problem type [$F(1, 52) = 22.975$, $MSE = .707$, $p < .001$], and a significant encoding by problem type interaction,

$F(1, 52) = 5.261$, $MSE = .162$, $p < .05$. Pair-wise comparisons revealed that performance on the anchored pairs was equivalent across groups, $t(52) < 1$. However, the restudying group outperformed the retrieval practice group on the TI pairs, $t(45.071) = 2.360$, $p < .025$.

Figure 4. Experiment 2 Performance on Anchored & TI Pairs



Although the previous analysis demonstrated an advantage in the restudying group for overall TI, it is possible that their advantage is exclusive to just one type of TI problem. Therefore we broke down the TI problems into coarse- and fine-grained problems (as in Experiment 1). A 2 x 2 (Encoding by Grain Size) repeated measures ANOVA revealed only a significant main effect of encoding, $F(1, 52) = 5.572$, $MSE = .709$, $p < .025$. There was no main effect of grain size [$F(1, 52) = 2.278$, $MSE = .047$, $p = .137$], nor was there an encoding by grain size interaction, $F < 1$. The restudying group outperformed the retrieval practice group on all TI pairs—regardless of grain size. For a

summary of performance on all measures analyzed thus far broken down by encoding condition, see Table 3.

Table 3. *Experiment 2 Final Test Performance*

	Retrieval Practice	**Restudying**
Premise Pairs	.81 (.13)	.85 (.13)
Problem Type:		
Anchored	.88 (.14)	.89 (.18)
Transitive Inference	.64 (.30)	.80 (.20)
TI by Grain Size:		
Coarse-Grained Problems	.64 (.35)	.84 (.20)
Fine-Grained Problems	.63 (.28)	.77 (.23)

Note. Means and standard deviations (in parentheses)

Final Test Performance Broken Down by Test Half. To mitigate any potential concern that the pattern we observed could be have actually been stronger (or qualitatively different) at the beginning of the test and then washed out later on, we broke down performance by test half. Similar to Experiment 1, the pattern we found in our initial analyses collapsing across all test trials did not differ qualitatively across the test halves. We only found a small (~4%) but significant drop in premise pair memory that was the same for both encoding conditions. The overall advantage in TI we observed in the restudying condition was the same across both halves. Therefore, the nature of the restudying advantage we found in our initial analyses was not occluded when aggregating across all trials. For a full report of the analysis, the interested reader is directed to Appendix B.

Interim Conclusions. The change in our experimental manipulation proved sufficient for inducing a difference between the groups in TI. Restudying yielded superior performance on all TI problems both coarse and fine-grained. This advantage was also in the presence of a non-significant difference in premise pair memory—

facilitating our interpretation of the difference in TI. Our additional analyses broken down by test half also showed that this advantage did not change significantly as the test progressed. Therefore we can conclude that restudying yielded better performance on TI problems because the representation it created was significantly more flexible.

Awareness Questionnaires

As in Experiment 1, we also examined TI as a function of awareness. The same awareness scale and criteria used in Experiment 1 were used here. The same 2 raters from Experiment 1 also graded all of the responses. Both raters were blind to a participant's condition and the scale had good reliability (proportion agreement = .93). As done previously, evaluation disagreements were settled through discussion to arrive upon a final label of aware or unaware. A chi-square test revealed that the proportion of aware participants was higher in the restudying than retrieval group, χ^2 (1) = 4.441, $p <$ .05. Therefore the current design rendered restudying a more effective agent at promoting awareness.

Before evaluating TI differences as a function of awareness, we needed to assess differences in premise pair memory. If present, they could mediate differences in TI. A 2 x 2 (Encoding by Awareness) ANOVA revealed no main effect of encoding ($F < 1$) and no encoding by awareness interaction, $F < 1$. We found only a significant main effect of awareness, $F(1, 50) = 17.603$, $MSE = .228$, $p < .001$. Participants deemed aware had better premise pair memory than those deemed unaware. This is consistent with Experiment 1 and aligns somewhat with studies that have induced awareness experimentally (e.g., Greene et al., 2001; Libben & Titone, 2008). Controlling for awareness, there is still no significant difference in premise pair memory between the

encoding conditions, although premise pair memory is better for the aware participants. The rates of awareness and a breakdown of premise pair memory are all shown in Table 4.

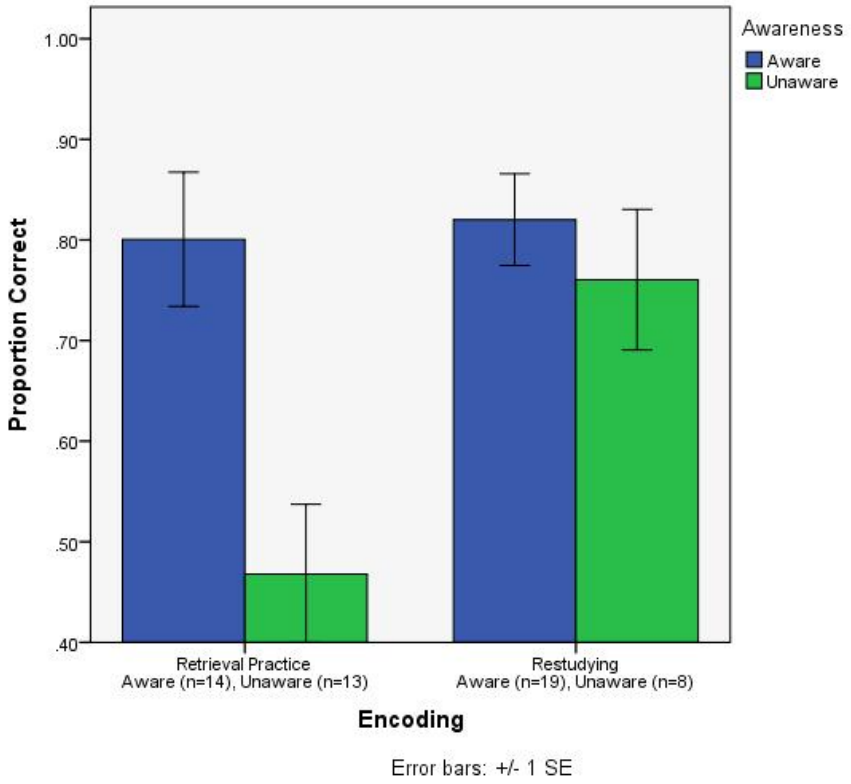
Table 4. *Awareness Rates & Premise Pair Memory*

	Retrieval Practice	**Restudying**
Aware	14	19
Unaware	13	8
Aware	.88 (.12)	.89 (.10)
Unaware	.74 (.10)	.75 (.14)

Note. The upper half of the table displays the raw numbers of participants in each encoding condition who were classified as aware and unaware. The lower half of the panel displays the premise pair memory means and standard deviations (in parentheses) for each encoding condition broken down by awareness classification.

We next examined differences in overall TI (regardless of grain size). A 2 x 2 (Encoding by Awareness) ANOVA revealed a significant main effect of encoding [$F(1, 50) = 5.875$, $MSE = .299$, $p < .025$] and a significant main effect of awareness [$F(1, 50) = 9.291$, $MSE = .472$, $p < .005$.]. These main effects were qualified by a significant interaction [$F(1, 50) = 4.493$, $MSE = .228$, $p < .05$]. For aware participants, TI did not significantly differ as a function of encoding, $t < 1$. For unaware participants, however, TI was significantly better in the restudying group, $t(19) = 2.808$, $p < .025$. At this juncture, it would seem that the restudying group's advantage is exclusive to the subset of participants who were unaware.

Figure 5. Experiment 2 – TI by Awareness



Recall that awareness has been found to modulate grain size effects in TI (Frank et al., 2005; Libben & Titone, 2008). As such we structured our last TI analysis around coarse- and fine-grained problems. A 2 x 2 x 2 (Encoding by Awareness by Grain Size) repeated measures ANOVA revealed a significant main effect of encoding, $F(1, 50) = 5.875$, $MSE = .597$, $p < .025$. There was also a significant main effects of awareness [$F(1, 50) = 9.291$, $MSE = .945$, $p < .005$] and a marginally significant main effect of grain size [$F(1, 50) = 3.349$, $MSE = .069$, $p = .073$]. Grain size did not interact with any variables (F's < 1.66, p's > .20). However, as in the previous analysis, we found a significant interaction between encoding and awareness, $F(1, 50) = 4.493$, $MSE = .457$, $p < .05$.

For aware participants, TI performance for both grain sizes did not significantly differ as a function of encoding (both *t*'s < 1). For unaware participants in the restudying group, there appeared to be an advantage for coarse-grained pairs but this was non-significant (recall that there was no significant 3-way interaction between encoding, awareness, and grain size). The crucial result was that, for the unaware participants, TI was significantly better in the restudying group regardless of grain size—both *t*(19)'s > 2.24, *p*'s < .04. Performance on TI problems at all levels of analysis is displayed in Table 5.

Table 5. *Experiment 2 TI Performance*

	Retrieval Practice		**Restudying**	
	Aware (n = 14)	Unaware (n = 13)	Aware (n = 19)	Unaware (n = 8)
TI (overall)	.80 (.25)	.47 (.25)	.82 (.20)	.76 (.20)
Coarse-Grained	.80 (.32)	.48 (.32)	.84 (.20)	.83 (.19)
Fine-Grained	.80 (.23)	.46 (.23)	.80 (.22)	.69 (.23)

Note. The rates of awareness for both encoding conditions are displayed below their respective column headings. In addition, performance for overall TI and TI broken down by grain size are displayed in the bottom three rows. Means and standard deviations are shown (in parentheses).

Conclusions

As was the case in Experiment 1, it is clear that greater awareness is associated with greater premise pair memory and better TI. There also seems to be no uniform effect of grain size in and of itself—but a hint of interaction with encoding and awareness. Most importantly, when one can control for awareness post-hoc, the TI advantage associated with restudying changes. We found that such an advantage is evident only when examining unaware participants. Also—recall that premise pair memory was still equal between the encoding conditions when controlling for awareness. So the TI advantage for unaware participants in the restudying condition must be due to differences in representational flexibility and not overall premise pair memory.

Therefore, even when controlling for awareness, there is evidence of unambiguous differences in representational flexibility.

One potentially lingering concern is over our assessment of awareness. At first glance, it is somewhat puzzling to find that TI for aware participants in the retrieval practice group was far superior to the unaware sub-group; meanwhile TI did not significantly differ as a function of awareness for those in the restudying group. Based on these results, one might argue that our ability to discern between aware and unaware participants was somewhat compromised. However, we did find that unaware participants in the restudying group showed a hint of the grain size effect typically found with unaware participants in other studies (e.g., Frank et al., 2005; Libben & Titone, 2008). Note that we also found that unaware participants in the retrieval practice group exhibited at-chance TI. They were essentially guessing. Also note that this pattern differs slightly from Experiment 1. There, we found only a simple main effect of awareness (and a marginally significant awareness by grain size interaction). The aware group showed a trend for the typical grain size effects while the unaware group did not; they merely showed overall impoverished performance. Taken together, these results convey that we were able to discern between aware and unaware participants. In fact the different profiles across both experiments for the aware and unaware participants of both encoding conditions would suggest that we were able to successfully detect *degrees* of awareness.

General Discussion

The current study provides a novel analysis of the effects of testing on abstraction and knowledge structure formation. This was due to an amalgam of paradigms from the

testing effect and transitive inference literatures. Unique to this paradigm was the fact that it allowed for the creation of a novel knowledge structure whose relationships were intrinsic to the materials (i.e., A > B > C, etc.). Zaromb and Roediger's (2010) materials had intrinsic relationships but they capitalized on pre-existing associations in long-term memory. In regards to Karpicke and Zaromb (2010, Experiment 3): they tested for serial order memory so the associations among the materials were novel, but their subsequent importance was not emphasized during encoding. These properties made serial order information of little value during learning. Conversely, the relational properties of the stimuli in the current paradigm were of immense value because realizing them would help consolidate one's representation of the materials.

Most importantly, the current paradigm also allowed for the creation of novel inference problems. The paradigms used by Zaromb and Roediger (2010) and by Karpicke and Zaromb (2010) did not allow for this. Other studies of the testing effect have examined performance on transfer or inference problems (Butler, 2010; Rohrer et al., 2010; Karpicke & Blunt, 2011)—all finding positive effects. The present experiments are the first to find no such effect. The current paradigm, however, permitted control of the grain size or "difficulty" of a problem (where fine-grained and coarse-grained problems corresponded to easy and hard problems, respectively) and still found no positive effect. To my knowledge, no other study of the testing effect has exhibited such control.

Recap of Results

Across both experiments, there was no test-related enhancement of TI, regardless of grain size. Recall that premise pair memory was also equal between the groups in both

experiments. This was fortuitous in that it permitted interpreting the TI data solely in terms of differences in representational flexibility. In Experiment 1, there was no difference in TI. This was attributed to the strength of the manipulation. Thus, in Experiment 2 one initial learning trial was removed and 3 intervenient ones were added, bringing the grand total to 8. After increasing the number of encoding trials, testing was actually found to impede TI. In addition, controlling for awareness in Experiment 2 revealed that the negative effect of testing was exclusive to the unaware participants. Furthermore, TI was at chance for unaware participants who had engaged in retrieval practice. Their counterparts who had engaged in restudying exhibited TI that was well above chance. The fact that the effects of the encoding manipulations only differed for unaware participants is a surprising revelation and it also supports our decision to do a post-experiment assessment of awareness. Most people in fact became aware of the hierarchical relational structure of the materials. TI for aware participants was also well above chance and equal for both encoding conditions. Keep in mind that this still means that by all measures, there were no positive effects of testing. Whether or not the effects were negative or null depended on awareness.

Interestingly, these results can also be taken to imply that, if the effects of awareness are robust, then conditions with low levels of awareness should show strong negative testing effects on TI. More common TI paradigms use hierarchies that are less apparent because no dimension (such as value) is overtly provided on which participants can rank the stimuli (e.g., Greene et al., 2001; Frank et al., 2005; Libben & Titone, 2008; Lazareva & Wasserman, 2010). Under these conditions (ones that require a great deal of self-initiated relational processing), awareness rates are quite low. Based on the results

of the current experiments, one would predict that negative testing effects on TI would abound under such conditions.

Theoretical Implications for the Testing Effect

Item-Order (Item-Relational) Account

In the introduction it was mentioned that "unusual" encoding manipulations that promote item-specific processing have varying effects in free recall and order reconstruction. In mixed lists where items are subjected to common and unusual encoding manipulations (e.g., reading and generation, respectively), overall free recall is highest and overall order memory is intermediate: better than a pure list in the unusual condition, and worse than a pure list in the common condition. In pure lists, where items are subjected entirely to a common or unusual manipulation, free recall is usually equal between the two encoding conditions but order memory favors the common condition. In fact compared to the mixed list, order memory is usually impaired significantly in the pure unusual condition and significantly improved in the pure common condition. This ubiquitous pattern of results has been captured by the item-order account (McDaniel & Bugg, 2008). In the strictest sense, this account is designed to predict performance on free recall and order reconstruction tests. One can, however, extend it to predict the effects of an encoding manipulation on item and relational memory, respectively (e.g., Hunt & McDaniel, 1993; Mulligan & Lozito, 2004).

Recent studies have demonstrated that the effects of testing on relational memory are not uniform. Zaromb and Roediger (2010, Experiment 2) found that testing enhanced relational memory (as measured by category clustering in free recall). Karpicke and Zaromb (2010, Experiment 3), on the other hand, found that testing disrupted relational

memory (as measured by an order reconstruction test). The aforementioned experiments from both studies found an item memory enhancement and used a between-subjects design. Although seemingly paradoxical, both sets of results can be explained by the item-order account. One of the more nuanced tenets of this account is that, in the presence of more salient or meaningful relational information (like semantic category membership), pure lists of items in the unusual condition will not have a significant disruption of relational memory (McDaniel & Bugg, 2008). Under such conditions, relational memory in the unusual condition is spared or even enhanced. This is because salient or more meaningful relational information is easier to notice during encoding and, at retrieval, helps one generate candidate targets and/or narrow the search set (McDaniel & Bugg, 2008). Such dynamics might explain the disparity between Zaromb and Roediger (2010, Experiment 2) and Karpicke and Zaromb (2010, Experiment 3). The former study used lists of medium frequency nouns from familiar taxonomic categories—resulting in a familiar, meaningful relational structure to the materials. Although the latter study did use paired associates with familiar relationships (e.g., *heart – love*), they were arranged arbitrarily in unrelated lists—creating an unfamiliar, meaningless relational structure. Thus, one contributor to the disparity between these two studies may lie in the familiarity and meaningfulness of their relational structures. Another contributor might also be the way in which the relational information was used to guide retrieval during retrieval practice. Zaromb and Roediger (2010) used free recall during retrieval practice. In free recall, one can use category information as a cue for recall. Practice of this strategy also translates well to the final recall test. In Karpicke and Zaromb (2010), retrieval practice was implemented with cued recall of individual items.

Therefore, any order information one might have gleaned during encoding could not be recapitulated during retrieval practice. This would also impair one's ability to reinstate item order on the order reconstruction test. In sum, although the relational structures differed between the two studies, this was also conflated with the degree to which one could reinstate it during retrieval practice. Future studies should take this into consideration. Nonetheless, as will be shown later in this section, the nature of the relational structure is still a key culprit.

The results of the current study are similar to those of Karpicke and Zaromb (2010, Experiment 3) in that they show how testing can disrupt relational processing. However, because we did not observe an enhancement in premise pair memory, we technically cannot conclude that testing enhances item processing at the expense of disrupting relational processing. Instead, the current study revealed conditions under which testing will still disrupt relational memory but not enhance item memory. Similar results were found by Mulligan (2002, Experiment 3) for the generation effect when using unfamiliar paired associates (e.g., *tomato - cake*). Under these conditions, he found that generation yielded a negative effect on order memory and no effect on item memory.

Mulligan (2002), however, found null item and negative order memory effects when encoding was manipulated within subjects—we manipulated encoding between subjects. Because the testing effect is reduced in a between-subjects design (Roediger & Karpicke, 2006b, Experiment 2; Karpicke & Zaromb, 2010, Experiment 3), it is possible that using stimuli with unfamiliar associations within a between-subjects design induces a null effect on item memory while maintaining the relational memory disruption. Recall, however, that Roediger and Karpicke (2006b) and Karpicke and Zaromb (2010) used

materials covering several familiar concepts and associations in long-term memory. Therefore, it appears that, under a between subjects design, it is possible to get a null testing effect on item memory, but only when using materials with unfamiliar intra-item associations. Future studies of the testing effect could experimentally manipulate the familiarity of the intra-item associations in a manner similar to Mulligan (2002, Experiment 3) to see if this is indeed the case.

Effects of Interference

It has been shown that testing can mitigate the effects of proactive inference *across* lists of semantically categorized words (Szpunar et al., 2008; Zaromb & Roediger, 2010). However, recall that in a double-function list-learning paradigm, Slamecka (1976) found evidence that testing might exacerbate interference *within* a list. Given that our paradigm is similar to double-function list learning, one might expect interference to be abound—therefore making testing a less optimal learning strategy. Specifically, Slamecka (1976) found that retrieval-based learning promoted interference from near *and* remote associates in both forward and backward directions. If similar mechanisms were operating in our paradigm, one would expect performance to be relatively worse in the retrieval practice group for the premise pair with the most remote associates in both directions. This would be the DE pair. Element D rests in the center of the hierarchy and thus has the most potential for interference. In a post-hoc analysis for Experiment 2, we in fact found significantly worse performance on this pair, compared to the restudying group, $t(52) = 2.615$, $p < .025$.

It would appear that testing does not always mitigate the effects of interference. The deleterious effects that testing exerts on hierarchy abstraction seem similar to those it

exerts on double-function list learning (e.g., Slamecka, 1976). However, to test this conclusively, one would have to examine response production errors in an open-ended manner. Slamecka (1976) concluded that testing generates more interference only because the testing subjects were more likely to produce within-list intrusion errors from near *and* remote associates to a given cue. Our paradigm only involved a 2AFC test so we could not examine intrusion errors in a more open-ended manner. However, the accuracy data on the DE pair strongly suggest that interference was higher in the retrieval practice group.

Role of Elaborative Rehearsal

The primary question posed by our results is why, for unaware participants, did the restudying group perform better on the TI problems even when overall premise pair memory was equal? I have provided suggestive evidence that restudying yields less within-list interference than retrieval practice. I have also asserted that the current results support the notion that restudying in general allows more room for relational processing than "unusual" encoding strategies (i.e., the item-order account, McDaniel & Bugg, 2008). That is, because restudying does not place as much emphasis on item-level information as other encoding strategies, it provides more opportunities to take notice of the relationships among the items. This is a form of elaborative rehearsal. It is possible that the luxury of elaborative rehearsal was less accessible in the retrieval practice condition because more time was spent on local, premise-based learning. Recall that, during the intervenient phase, participants had 5 sec to devote to recall and were given 2.5 sec for feedback. Because the majority of the trial time was dedicated to retrieving an isolated cue-target association (i.e., item-level information), participants had little time to

engage in elaborative rehearsal. If this is indeed the case, then retrieval practice should yield worse performance than any strategy that promotes elaborative rehearsal.

Surprisingly, a recent study by Karpicke and Blunt (2011, Experiment 1) found that retrieval practice yielded better memory than restudying or concept mapping—all the while using a between-subjects design. Concept mapping is when one takes the information contained in a text and then rewrites it as a series of nodes and links that convey the overall, abridged idea structure of the passage. This is obviously an active, elaborative encoding strategy. It is therefore all the more surprising that retrieval practice yielded better performance. There are several differences between their study and the current one (e.g., their memoranda were scientific text passages and their retention interval was 1 week). However, one crucial difference was how they operationalized retrieval practice. In this condition, participants read a passage for 5 minutes and then took a free recall test lasting 10 minutes. Afterwards, they actually *re-read* the passage for another 5 minutes and then engaged in another 10 minutes of free recall. This additional re-reading period is probably a key factor. It likely yielded at least two benefits: 1) it provided corrective feedback and 2) it gave more room for elaborative rehearsal. Although the present study provided feedback, it was intermixed with retrieval practice on a trial-by-trial basis—which is a very different experience than a protracted period of restudying. Perhaps retrieval practice intermixed with protracted restudying periods is a more optimal learning schedule—possibly because of the supplementary elaborative rehearsal. If such rehearsal was a driving factor in our study, future studies could experimentally manipulate the level of elaborative rehearsal to see if this makes one group's performance mimic the other. One could supplement retrieval practice with

the appropriate amount of restudying time to see if this yields TI behavior similar to that produced by pure restudying.

Transfer, Abstraction, Flexibility, & Inference

One of the most important contributions of the current study was that it disambiguated the effects of testing on abstraction and representational flexibility. Recall that this is different than the effects of testing on the transfer of knowledge. Abstraction involves deriving the relational structure of the memoranda from its basic premises. Representational flexibility describes the ease with which one can recombine parts of a knowledge structure in a way that was not overtly taught to them. Transfer, on the other hand, involves realizing that information learned in one domain is applicable to a different domain, and then applying that information successfully.

As mentioned in the introduction, Butler (2010) demonstrated that, compared to restudying, testing can facilitate transfer. However, recall that the retrieval practice condition used factual and conceptual questions to test the participants' knowledge of the passages, and that the answers to these questions were reinforced via feedback. This likely promoted stimulus-response learning—obscuring the effects of testing on abstraction of the concepts. Also the transfer tests did not require participants to recombine previously learned facts and concepts. Therefore, although Butler's (2010) study was meritorious in many respects, it did not examine the effects of testing on abstraction and representational flexibility. Rohrer et al.'s (2010) study of test-enhanced transfer likewise did not clarify the effects of testing on abstraction and representational flexibility. Recall that Rohrer et al. (2010) required participants to learn locations on a map. On the transfer test, participants were asked "When traveling from [X town] to [Z

town], which town would you pass through?" This paradigm provided a better examination of abstraction and flexibility than Butler (2010). However, the effects of testing on those two properties were still unclear in Rohrer et al. (2010) because the participants were provided with a map on the transfer test, minus the town names. Therefore, Rohrer et al. could not examine the effects of testing when one must act entirely on his or her own abstracted, *internal* representation. The current study, however, did not provide any such aide on the final test and therefore could unambiguously examine the effects of testing on an abstracted, internal representation. Under our conditions, these effects turned out to be null for aware participants and negative for unaware participants.

The results of Karpicke and Blunt (2011) present a more puzzling case. Recall that, in their first experiment, they too manipulated encoding between subjects. They also found that testing enhanced performance on inference problems. While this is a demonstration of test-enhanced inference, it is unclear as to whether or not this reflects the same kind of flexibility and abstraction required in the current study. The following example provided from their study suggests that the flexibility they observed may be more local.

In their first experiment, participants studied a passage about sea otters taken from a preparatory guide for the Test of English as a Foreign Language (TOEFL; cf. Rogers, 2001). The final test contained 14 verbatim questions and 2 inference questions. An example of an inference question and an acceptable answer is as follows:

> Q: "What would be the consequences of removing sea otters from their environment?" A: "There would be a lack of protection of kelp and seaweed, because fewer otters would eat the invertebrates that destroy kelp and seaweed. The presence of more invertebrates would change the ecosystem."

It should also be known that most of the necessary information for the answer could be gleaned from part of the original passage:

> "Sea otters play an important environmental role by protecting forests of seaweed called kelp, which provide shelter and nutrients to many species. Certain sea otters feast on invertebrates, like sea urchins and abalones, that destroy kelp." (Rogers, 2001).

It would appear that answering the above inference question successfully requires some degree of flexibility. However, it is not clear if this represents the same scale of flexibility and abstraction required in our paradigm. Because the inference in question was based on a few facts gleaned directly from one part of the passage, one could argue that the requisite flexibility is more local than global. In addition there are other critical differences between their study and ours: Karpicke and Blunt (2011) had supplemental elaborative rehearsal, used materials with more familiar associations, had fewer inference problems, and used a longer retention interval. As stated before, I believe that the supplemental elaborative rehearsal played a unique role in their study and that the familiarity of the materials played a role in their study and similar ones.

Potential Criticisms of Our Design

Like other encoding manipulations that emphasize item-specific processing, the benefits of testing are largest in a within-subjects design (Roediger & Karpicke, 2006b; Karpicke & Zaromb, 2010). One might therefore argue that our between-subjects design put the retrieval practice group at a systematic disadvantage. While such a design will compress the size of the testing effect, recall that other studies have found the effect in a between-subjects design (Roediger & Karpicke, 2006b, Experiment 2; Karpicke & Zaromb, 2010, Experiment 3) and have even asserted that this is a quality on which the testing effect differs from the generation effect (Karpicke & Zaromb, 2010). In sum, the use of a between-subjects design alone could not have eliminated the positive effects of

testing. Implementing a within-subjects manipulation in a TI paradigm would also be problematic. It would have been difficult to subject one half of the premises to restudying and the other to retrieval practice without raising the potential of cross-condition contamination.

Another concern might be the demands of our retrieval manipulation. It has often been found that the testing effect is most pronounced when the intervenient test requires more self-initiated processing. Recall that Glover (1989) found that, compared to recognition or cued recall, an intervenient free recall test produced the largest testing effect—regardless of whether the final test was recognition, cued recall, or free recall. Based on Glover's (1989) results, one might argue that our design imposed another systematic disadvantage because the intervenient test was only 2AFC. Under normal circumstances, 2AFC *recognition* would require less self-initiated processing than free recall. However, the current paradigm was not 2AFC recognition because it did not require forced judgments over one old and one new item. The present 2AFC test requires more self-initiated processing than 2AFC recognition. The majority of the elements in the current paradigm had variable reinforcement histories. For the internal pairs, half the time a given painting was the right answer and half the time it was the wrong one. Successfully recalling the correct answer requires a substantial amount of interference control—which presumably requires a great deal of self-initiated processing. Therefore, it seems unlikely that our retrieval practice manipulation placed insufficient demands on self-initiated processing. Rather it would seem that the nature of the interference at hand, the between-subjects design, the relational structure of the materials, and the lacking familiarity for their associations contributed to the present results.

Finally, one might wonder if our design had too many encoding trials, resulting in over-learning in both groups and thus perverting the potential differences between them. However, Zaromb and Roediger (2010) had many encoding trials in their first experiment and found that testing did not increase objective organization (i.e., category clustering) but did increase subjective organization (i.e., pair frequency). If the number of trials in itself were a key factor, then there should have been no differences between the groups because our materials required more objective organizational memory[4]. To the contrary, it was abundantly clear in Experiment 2 that the testing group was impoverished on the TI problems (at least in the unaware group). Furthermore, the testing group received feedback on a trial-by-trial basis during the intervenient phase. This should have given them the opportunity to engage in a great deal of corrective learning. In spite of this, there was still no advantage in relational memory as indexed by TI. If anything, Experiment 2 showed that the number of encoding trials in the first one was insufficient for revealing potential group differences.

Necessary Qualifiers

One must also keep in mind that the negative effects observed in the current study are thus far restricted to an immediate test. The null effect or disadvantage of testing often fades within 1 or 2 days after initial learning (Roediger & Karpicke, 2006b; Carpenter et al., 2008). Therefore we should expect no benefit of restudying at longer retention intervals. The longer retention interval in Karpicke and Blunt (2011) might also explain why they observed a moderate, but significant advantage on their inference

[4] Zaromb and Roediger's (2010) first experiment also manipulated encoding within subjects. Nonetheless, the present results show that the number of encoding trials is not a key factor in and of itself.

problems[5]. A natural follow-up to the current study would be to examine performance at a longer retention interval.

Generalizing To Broader Aspects of Learning

In sum, the current results suggest that when one is learning a novel, highly inter-related knowledge structure, testing is less effective than restudying. The present design mimics conditions under which a knowledge structure is in its infancy and there is little support from pre-existing knowledge. There may be a point in the course of learning at which testing may enhance long-term retention, but the present results suggest this does not occur at the initial stages.

Educational Implications

The results of the retrieval practice condition demonstrated the potential deleterious effects of excessive premise-based learning. More specifically, an over emphasis on stimulus-response based learning actually lead to representational *in*flexibility. Problems such as this probably explain why effective second language learning is often supplemented with immersion exercises that promote flexibility (not just stimulus-response learning).

For example, when learning a language, it helps to learn the basic vocabulary and rules of syntax but, if one is to effectively make use of that knowledge and engage in active conversation, one must also practice recombining the basic elements of vocabulary and syntax—possibly in the form of practice compositions or elementary conversational exchanges that tax one's knowledge in a novel way. This approach provides the

[5] No exact mean difference was reported because, when verbatim and inference questions were analyzed, there were only main effects of encoding and question type. Retrieval practice was overall better than restudying and verbatim questions were easier than inference questions. Visual inspection reveals that the difference on inferential questions was on the order of 11% (restudying: $M \sim .58$; retrieval practice: $M \sim .69$).

opportunity to find new navigational routes of a knowledge structure that one would not otherwise discover during stimulus-response learning. This in effect might render the representation more flexible and thus facilitate solutions to future (linguistic) problems.

Conclusions

Testing has often been shown to produce remarkable effects on memory retention. The net result of these effects is usually positive—lending further support to its educational utility. However, several recent studies have shown that its effects on various aspects of memory are not uniformly positive. While testing often enhances item information, its effects on relational information are more complex. The current study revealed conditions under which testing will impair relational memory and have no effect on item memory. The results also suggest that the nature in which knowledge must be recombined is another key determinant of the effects of testing on relational memory. The present study, in combination with other recent ones, also suggests that the familiarity of the to-be-learned associations might be a key determinant of the positive effects of testing on item memory. Advocates of the educational utility of testing should carefully consider the potential systemic effects of testing (and possibly the nature of the materials) before deciding to implement additional memory tests. This can have a tremendous impact on the marginal returns of testing.

Appendices

Appendix A

Awareness Questionnaire (adapted from Frank et al., 2005)

1. Do you have any prior knowledge of the paintings used in the experiment?

2. If you answered "Yes" to question 1, please indicate to what extent you are familiar with these paintings.

3. Did you have the impression that some of the pairs were easier to learn than others?

4. Did you think any of the paintings were ALWAYS the *most* valuable (no matter what the other painting was)?

5. Did you think any of the paintings were ALWAYS the *least* valuable (no matter what the other painting was)?

6. Did you have the impression that there was some kind of logical rule, order, or hierarchy of the paintings in the experiment? If so, please explain briefly.

7. In the test phase, were there any new paintings?

8. Where there any new combinations of paintings?

9. If you answered "Yes" to question 8, how did you make your choice in these cases? (e.g., guessed, went with instinct, used some sort of rule—explain)

Appendix B

Experiment 1 Test-Half Analyses

For premise pair memory, a 2 x 2 (Encoding by Test Half) repeated measures ANOVA revealed no significant main effect of encoding ($F < 1$), nor a significant encoding by test half interaction, F (1, 52) = 1.009, MSE = .006, p = .32. There was, however, a significant main effect of test half, F (1, 52) = 5.809, MSE = .037, $p < .025$:

premise pair memory was better in the first half of the test (M = .83, SD = .14) than in the second half (M = .79, SD = .16). This could be due increasing interference or "unlearning" from probing the memory trace multiple times and from multiple directions of association. Additionally, simple forgetting might occur over time because the test trials do not provide feedback. However, most importantly, the rate of decline did not differ between the encoding conditions and the non-significant difference in premise pair memory we observed did not change over test blocks.

We then analyzed anchored and TI problems across the test. A 2 x 2 x 2 (Encoding by Test Half by Problem Type) repeated measures ANOVA revealed no main effect of encoding or test half (both F's < 1). However, like before, there was a significant main effect of problem type, F (1, 52) = 31.375, MSE = .826, p < .001. For both halves of the test, performance on anchored pairs was better than that for TI pairs [both t(53)'s > 4.8, both p's < .001]. There were no significant 2-way interactions (all F's < 1.33, all p's > .25). We did, however, observe a marginally significant 3-way interaction between encoding, problem type, and test half, F (1, 52) = 3.837, MSE = .026, p = .056. Pair-wise tests revealed that the change in the difference between anchored and TI pairs across the test halves differed between the two encoding conditions. In the retrieval practice condition, the difference appeared to shrink because it was significant in the first half [t(26) = 3.287, p < .005] but only marginally significant in the second half [t(26) = 2.359, p = .026]. For the restudy condition, however, the difference was significant in both the first [t(26) = 3.947, p < .005] and second halves [t(26) = 4.451, p < .001]. Critically, however, there were no differences in TI between the groups across the

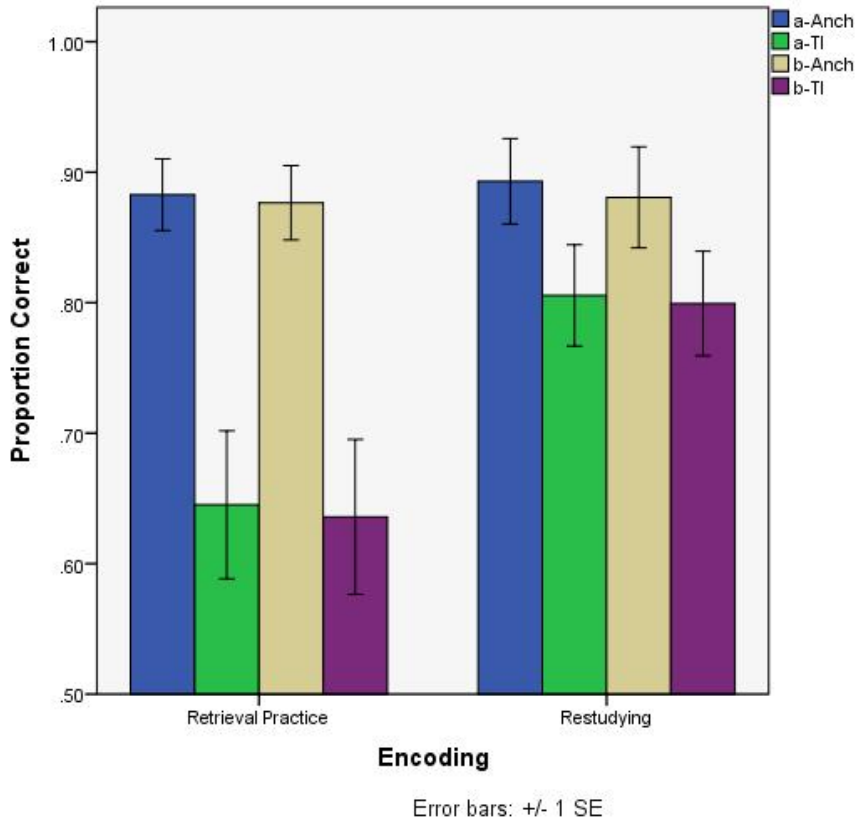
test. So, while there were minute changes across the test within groups, this did not qualitatively affect the general pattern of non-significant differences between groups.

Experiment 2 Test-Half Analyses

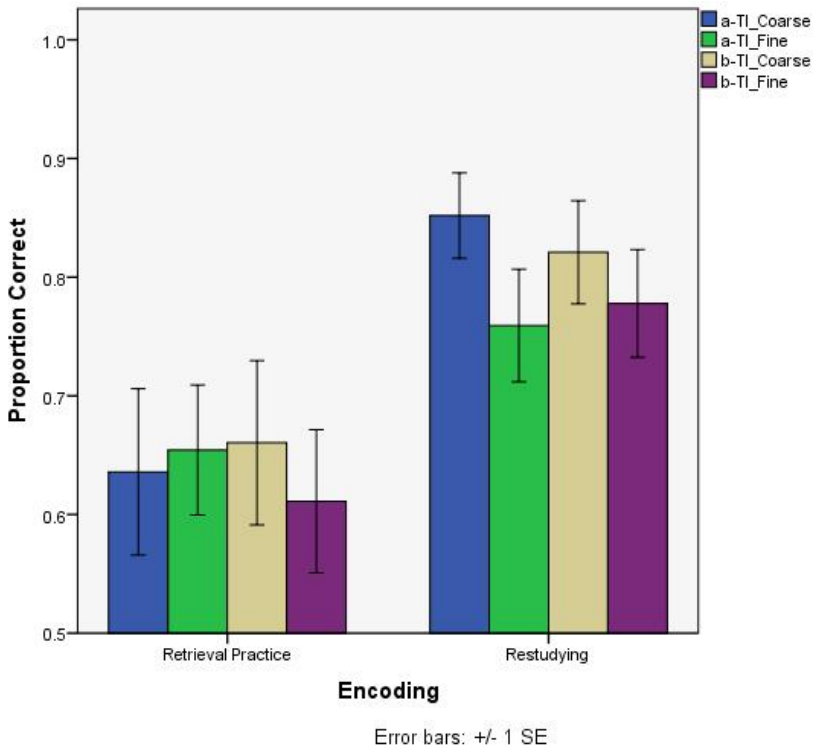
We first examined premise pair memory. A 2 x 2 (Encoding by Test Half) repeated measures ANOVA revealed no significant main effect of encoding [$F(1, 52) = 1.008$, $MSE = .034$, $p = .32$.] and no significant encoding by test half interaction, $F < 1$. We only observed a significant main effect of test half, $F(1, 52) = 17.695$, $MSE = .062$, $p < .001$. As in Experiment 1, for both groups, premise pair memory was better in the first half.

Next we examined performance on the non-adjacent pairs (the anchored and TI problems). A 2 x 2 x 2 (Encoding by Problem Type by Test Half) repeated measures ANOVA revealed the same pattern of results we found in our initial analyses. Critically, however, there was no main effect of test half, $F < 1$. Although we did find a main effect of problem type, $F(1, 52) = 22.975$, $MSE = 1.413$, $p < .001$. Anchored pairs were easier to solve than TI pairs. There was also a marginally significant main effect of encoding, $F(1, 52) = 3.268$, $MSE = .387$, $p = .076$. We also observed the critical interaction between encoding and problem type, $F(1, 52) = 5.261$, $MSE = .324$, $p < .05$. As in the first wave of analyses, performance was equal on the anchored pairs (first and second blocks: t's < 1) but TI was superior in the restudying group (first and second blocks: both $t(52)$'s > 2.28, p's < .03). There were no other significant interactions (all F's < 1).

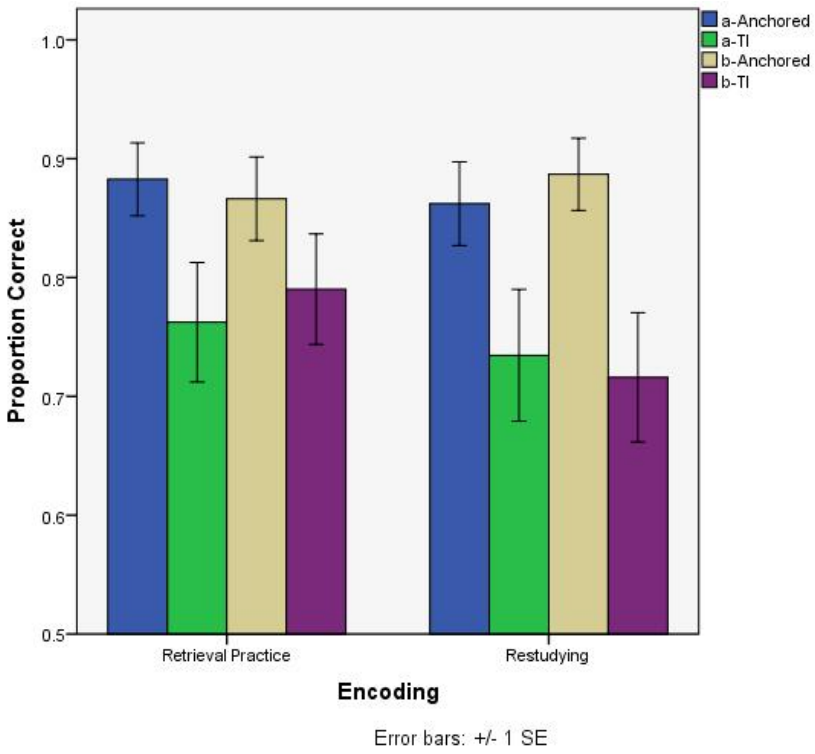
Since the effect of test half was non-significant and we observed the same interaction found in the prior analyses, we can conclude that the change in the restudy condition's advantage on the TI pairs was not significantly different in the second half.



To see if the restudying group's TI advantage on the test might have changed by grain size, we also analyzed performance on the coarse- and fine-grained pairs across test halves. A 2 x 2 x 2 (Encoding by Test Half by Grain Size) repeated measures ANOVA revealed a pattern of results very similar to that of our initial analyses. Like before there was no main effect of grain size, $F(1, 52) = 2.278$, $MSE = .094$, $p = .137$. We also observed a significant main effect of encoding, $F(1, 52) = 5.572$, $MSE = 1.418$, $p < .025$. Critically, there was no main effect of test half, $F < 1$. There was a marginally significant 3-way interaction, $F(1, 52) = 3.214$, $MSE = .046$, $p = .079$. No other interactions approached significance (all F's < 1).



Although these results suggest a bit more nuance than previously thought, the significant main effect of encoding is consistent with our prior analysis in which we collapsed across test halves. Most importantly, *encoding did not interact with test half*—showing that the restudying condition's advantage for both grain-sizes did not significantly change over test halves. The marginally significant 3-way interaction might be indicative of minute changes in the restudying condition's advantage. In any case, the interaction is non-significant and therefore only suggestive. As such the most parsimonious conclusion is that restudying yields TI advantages for both grain sizes—regardless of test half.



As in our primary analyses, we also broke down the test half by TI grain size. A 2 x 2 x 2 (Encoding by Grain Size by Test Half) repeated measures ANOVA revealed no main effect of encoding or test half (both *F*'s < 1). There was a marginally significant main effect of grain size, F (1, 52) = 3.847, MSE = .14, p = .055. As in the prior analysis, coarse-grained problems were somewhat easier to solve than fine-grained ones. There was no significant interaction between encoding and test half, F (1, 52) = 1.658, MSE = .029, p = .204. No other interactions were significant either (all *F*'s < 1). These results show that, across the test, TI did not differ between groups. This pattern extended to both grain sizes. In sum, the pattern of non-significance we observed in our initial analyses was not due to an initial effect washing out across the test halves.

References

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying, *Journal of Experimental Psychology: Learning, Memory, & Cognition, 36(5),* 1118-1133.

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval, *Journal of Experimental Psychology: Learning, Memory, & Cognition, 35(6),* 1563-1569.

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect, *Memory & Cognition, 34(2),* 268-276.

Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting, *Memory & Cognition, 36(2),* 438-448.

Carpenter, S. K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin & Review, 13(5),* 826-830.

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

Carrier, M. & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition, 20(6),* 633-642.

Chan, J. C. K. (2009). When does retrieval induce forgetting and when does it induce facilitation? Implications for retrieval inhibition, the testing effect, and text processing, *Journal of Memory and Language, 61,* 153-170.

Dempster, F. N. (1996). Distributing and managing the conditions of encoding and practice. In E. L. Bjork & R. A. Bjork (Eds.), *Human Memory* (pp.317-344). San Diego, CA: Academic Press.

Dusek, J. A. & Eichenbaum, H. (1997). The hippocampus and memory for orderly stimulus relations, *Proceedings of the National Academy of Sciences of the United States of America, 94(13),* 7109-7114.

Frank, M. J., Rudy, J. W., Levy, W. B., & O'Reilly, R. C. (2005). When logic fails: Implicit transitive inference in humans, *Memory & Cognition, 33(4),* 742-750.

Frank, M. J., Rudy, J. W., & O'Reilly, R. C. (2003). Transivity, flexibility, conjunctive representations, and the hippocampus: II. A computational analysis. *Hippocampus, 13,* 341-354.

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology, 81(3),* 392-399.

Greene, A. J., Spellman, B. A., Dusek, J. A., Eichenbaum, H., & Levy, W. B. (2001). Relational learning with and without awareness: Transitive inference using non-verbal stimuli in humans, *Memory & Cognition, 29(6),* 893-902.

Hunt, R. R., & McDaniel, M. A. (1993). The enigma of organization and distinctiveness. *Journal of Memory and Language, 32(4),* 421-445.

Karpicke, J. D. & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science,* 331, 772-775.

Karpicke, J. D. & Zaromb, F. M. (2010). Retrieval mode distinguishes the testing effect from the generation effect, *Journal of Memory & Language, 62(3),* 227-239.

Kornell, N. & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"?, *Psychological Science, 19(6),* 585-592.

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning, *Journal of Experimental Psychology: Learning, Memory, & Cognition, 35(4),* 989-998.

Kuo, T. M. & Hirshman, E. (1996). Investigations of the testing effect, *American Journal of Psychology, 109(3),* 451-464.

Lazareva, O. F. & Wasserman, E. A. (2010). Nonverbal transitive inference: Effects of task awareness and human performance, *Behavioral Processes, 83,* 99-112.

Libben, M. & Titone, D. (2008). The role of awareness and working memory in human transitive inference, *Behavioral Processes, 77,* 43-54.

McDaniel, M. A. & Bugg, J. M. (2008). Instability in memory phenomena: A common puzzle and a unifying explanation, *Psychonomic Bulletin & Review, 15(2),* 237-255.

Mulligan, N. W., & Lozito, J. P. (2004). Self-generation and memory. In B. H. Ross (Ed.), *Psychology of Learning and Motivation* (pp. 175-214). San Diego: Elsevier Academic Press.

Nairne, J. S., Reigler, G. L., & Serra, M. (1991). Dissociative effects of generation on item and order retention, *Journal of Experiment Psychology: Learning, Memory, & Cognition, 17(4),* 702-709.

Primoff, E. (1938). Backward and forward associations as an organizing act in serial and paired associate learning. *Journal of Psychology, 5,* 375-395.

Roediger, H. L. III & Karpicke, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice, *Perspectives on Psychological Science, 1(3),* 181-210.

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

Rogers, B. (2001). TOEFL CBT Success. Princeton, NJ: Peterson's.

Rohrer, D., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning, *Journal of Experimental Psychology: Learning, Memory, & Cognition, 36(1),* 233-239.

Slamecka, N. J. (1976). An analysis of double-function lists, *Memory & Cognition, 4(5),* 581-585.

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.

Van Elzakker, M., O'Reilly, R. C., & Rudy, J. W. (2003). Transivity, flexibility, conjunctive representations, and the hippocampus: I. An empirical analysis. *Hippocampus, 13,* 334-340.

Zaromb, F. M. & Roediger, H. L. III (in press). The testing effect in free recall is associated with enhanced organizational processes, *Memory & Cognition*