When Practice Does Not Make Perfect:

Investigations into the Underconfidence-With-Practice Effect on Judgments of Learning

by

Heather Tiede

University of Manitoba

Thesis Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Arts

in the

Department of Psychology

UNIVERSITY OF MANITOBA

THE UNIVERSITY OF MANITOBA

FACULTY OF GRADUATE STUDIES
*****
COPYRIGHT PERMISSION

When Practice Does Not Make Perfect: Investigations into the Underconfidence-With-Practice Effect on Judgments of Learning

BY

Heather Tiede

A Thesis/Practicum submitted to the Faculty of Graduate Studies of The University of

Manitoba in partial fulfillment of the requirement of the degree

Of

MASTER OF ARTS

Abstract

Success in learning new material depends in part on accurate judgments about how well the information has been learned. A common method for measuring accuracy in monitoring progress in learning involves asking participants to make Judgments of Learning (JOLs) or estimates of future recall for recently studied material. In turn, these estimates are compared to actual success in future recall. This research reveals that people are fairly accurate, if somewhat overconfident, in judging future memory performance for material studied once (Koriat & Levy-Sadot, 1999). Recently, however, Koriat, Sheffer, and Ma'ayan (2002), have presented evidence that repeated presentation and recall of a list of words reduces JOL accuracy, producing a shift toward underconfidence. One possible explanation for this Underconfidence-With-Practice (UWP) effect is that people discount the benefit of repeated study when each exposure is highly similar. If participants use the extrinsic cues provided by encoding a word differently (i.e., distinctive learning), or increasing the effortfulness of the encoding process, then the correspondence between JOLs and actual recall should be closer, thereby eliminating the UWP effect. Using a list-learning paradigm, Experiments 1 and 2 revealed that encoding words differently, or engaging in effortful encoding, did not eliminate the UWP effect. In Experiment 3, participants were explicitly informed either that (a) repetition is beneficial or (b) repetition is not beneficial. Although participants in the Benefit condition did not recall more words and were not better calibrated than participants in the No-Benefit condition, they reported significantly higher ratings for the benefit of repetition. The results are discussed within the framework of Koriat's (1997) cue-utilization theory.

Table of Contents

List of Tables

List of Figures

Introduction

In recent years, there has been a growing interest in metacognition and metacognitive processes across the lifespan (e.g., Hertzog & Hultsch, 2000; Plude, Nelson, & Scholnick, 1998). Metacognition is an all encompassing term that essentially refers to thinking about thinking. Although no single agreed-upon definition of metacognition has emerged, within the domain of cognitive psychology metacognition usually refers to the processes of monitoring mental states, control over cognitive processes, and strategy selection in guiding problem-solving, learning, and memory (Paris, 2002). An important component of metacognition is metamemory. Specific aspects of metamemory include memory-related beliefs and strategies, memory self-efficacy, and memory monitoring. Memory beliefs and strategies refer to one's general knowledge of memory functioning and the processes that should be engaged in to maximize acquisition of knowledge for later use. Memory beliefs and strategies may have a significant impact on behaviour. For example, consider an elderly individual who believes that it is natural for 'everyday memory' to worsen with age. The belief that memory deterioration is a 'normal' part of the aging process may affect behaviour in a number of ways. First, this person might become sensitive to memory failures and yet ignore the perceived memory 'problem' because of the belief that nothing can be done about it. On the other hand, this person might instead engage in memory strategies such as the use of mnemonics (e.g., imagery) and memory aids (e.g., note-taking) in order to maximize memory performance. In either case, memory beliefs and the consequential use (or nonuse) of strategies are important determinants of eventual memory performance.

Memory self-efficacy relates to how one 'feels' about their memory ability and performance. Consider another older adult who, like the person described above, also holds the belief that memory gets worse with age. This individual may be less concerned about minor instances of forgetting because they feel their memory is generally quite good.

Memory monitoring is an important component in acquiring new skills, learning new information, and effectively utilizing strategies for learning. Memory monitoring involves the ability to study new information and make a judgment concerning how well the new information has been processed and acquired for future use. For example, when attempting to memorize items to pick up at a grocery store, the task is to study that information enough so that the items can be recalled later. Success in this task depends on accurate judgments about how well the information has been learned because such judgments are critical in guiding the learning process. If a learner judges the grocery list as not yet committed to memory, they will devote more time to reviewing the list or will develop alternative encoding strategies. The consequence is that error in monitoring learning progress will lead either to incomplete acquisition of the list or to devoting more time than necessary to the task.

Factors that lead to underestimations of learning success have potentially dramatic implications for adults of all ages in their efforts to acquire new skills and to commit novel information to memory. The particular focus of this project is to explore underestimations of learning due to inaccurate memory monitoring by young adults.

Memory Monitoring and Control over the Learning Process

Memory monitoring is critically important in guiding the learning process (see Nelson & Narens, 1990). This crucial role of memory monitoring derives from its use in guiding control processes people apply in their efforts to acquire, maintain, and later retrieve knowledge, thereby influencing behaviour at each stage of the learning process. According to the Nelson and Narens' framework, control processes operate at the meta-level in order to prompt initiation, continuation, or termination of an action. At the point of acquisition, control processes direct study based on information provided via monitoring processes. As learning progresses, people use memory-monitoring indicators to assess how well information is learned, which will then determine the amount of study time and effort that is placed on learning certain items, relative to others. During the learning process, metacognitive decisions also direct the termination of study. For example, when an item is judged to be well-learned, study efforts will cease.

Outside the laboratory, it is clear that efficient memory monitoring is essential for performance. Consider a university student who has two exams the following day. Several factors will determine which course material is studied and for how long. Ignoring for the moment any grade incentives (e.g., relative worth of each exam), to gain full benefit from study the student must apportion her time in the most efficient manner possible. Typically, this involves making a subjective judgment as to what material can be learned easily and what material is already well-learned. Based on these judgments, the student will devote more attentional resources to yet-unlearned material according to its level of difficulty. The student's performance on these two exams will naturally depend, in part, on her actual memory ability. In addition, however, performance on the exams will also depend on the student's efficiency in monitoring her own learning progress.

The contribution of memory monitoring to performance on remembering tasks is well documented, influencing strategy use during both study and retrieval (Koriat & Goldsmith, 1996, 1998; Plude et al., 1998). In a typical experiment of this kind, participants are shown a list of items (e.g., paired associates, sentences, answers to general knowledge questions, etc.) to study for a later memory test. During study, participants estimate the ease of learning the item or the likelihood of recalling the item later. For example, Mazzoni, Cornoldi, and Marchitelli (1990) found that participants will modify their learning strategy depending upon how well they believe they know the target item. In their research (Experiment 1), participants who were allowed to re-study the target items at their own pace after the first exposure to the list devoted more study time to items they were initially uncertain about being able to recall, thereby diverting resources away from items judged to be very-well learned or very-poorly learned. Participants in this study were fairly accurate in discriminating between items that they already knew, and therefore did not need to study further, and items that were still learnable given the time constraints of the study. The results of this experiment illustrate how memory-monitoring processes, such as assessing how well an item is learned and judging its potential to be learned, can influence learning strategies. In this example, the participants shifted their focus and attention toward items that had the most potential to improve their performance.

The significance of the role of memory monitoring in directing learning was further demonstrated by Nelson, Dunlosky, Graf, and Narens (1994). In a series of experiments, they found that how well items were judged to be learned directed strategy use during study, thereby influencing actual performance. Using 36 Swahili-English translation equivalents as stimulus-response pairs, participants studied the items, and then decided how likely they thought it was that they would be able to remember the target word later when cued with the stimulus. Half of the items were then restudied before final recall. Participants were randomly assigned to one of four conditions. In the Worst-Learned Items (WLI) condition, the 18 pairs chosen for restudy had received the lowest likelihood judgments from that individual. In the Best-Learned Items (BLI) condition, the 18 pairs with the highest judgments were restudied. In the Normative-Most-Difficult Items (NMDI) condition, the items chosen for restudy were the 18 most difficult word pairs based on group base-rate data concerning level of difficulty and recall performance. The fourth group consisted of the Self-Chosen Items (SCI) group. After each likelihood judgment, participants in the SCI condition could decide by pressing a button if they wanted to study that item later, up to maximum 18 items. Overall, recall performance was highest for participants in the WLI and SCI conditions, followed by the NMDI condition, and poorest for the BLI condition. Thus, additional study time was most effective when the items chosen for restudy were based on one's own judgments of learning, even relative to items considered to be most difficult for people in general.

To summarize, people use memory-monitoring processes both to infer the ease of learning new information and how well that material has been learned. In turn, these processes guide control over study strategies, allocation of study time, and study termination (e.g., Nelson & Leonesio, 1988). In the context of memory-monitoring processes, these Experiments are mainly concerned with people's accuracy in judging their own success in learning because of the importance of these judgments in directing allocation of learning resources.

Are individuals fairly adept at monitoring their own success in committing information to memory? Correspondence-oriented research examines both accuracy and error in remembering information about the past and in memory monitoring (Koriat, Goldsmith, & Pansky, 2000). With respect to memory monitoring, correspondence refers to the association between expectations of future success in remembering and actual success in future remembering. One way of assessing memory monitoring accuracy is by having participants explicitly make Judgments of Learning (JOLs) at the time of study and then comparing those judgments to actual success in remembering that information in the future (Dunlosky & Nelson, 1994; Koriat & Shitzer-Reichert, 2002).

Judgments of Learning

Judgments of Learning (JOLs) are predictions concerning the likelihood of future recall for recently studied material (Arbuckle & Cuddy, 1969; Nelson & Narens, 1990). JOLs have been investigated with a wide variety of study material, including text comprehension (Carroll & Korukina, 1999; Rawson, Dunlosky, & McDonald, 2002), keystroke patterns (Simon & Bjork, 2002), sentence learning (Mazzoni & Cornoldi, 1993), paired associates (Carroll, Nelson, & Kirwan, 1997; Hertzog, Dunlosky, Robinson, & Kidder, 2003; Koriat, 1997; Koriat, Sheffer, & Ma'ayan, 2002; Koriat & Shitzer-Reichert, 2002; Lovelace, 1984), categorization (Kelemen, 2000), memory for sentences (Shaddock & Carroll, 1997) and answers to general knowledge questions (Benjamin, Bjork, & Schwartz, 1998; Shaughnessy & Zechmeister, 1992). If memory-monitoring processes are accurate, the correspondence between JOLs and actual success in remembering will be high. That is, JOLs will accurately predict recall performance when participants are highly sensitive to their actual learning progress. Mazzoni et al. (1990) also suggest that JOLs are dynamic in that they are constantly upgraded and amended as learning progresses. The student in the example described above will therefore make ongoing subjective assessments of her progress and will modify her study strategies in an effort to maximize learning.

On what basis do participants make their JOLs? The two dominant theories that examine the theoretical basis for JOLs are the direct-access view and the inferential approach.

The Direct-Access Approach to Judgments of Learning

The direct-access approach suggests that people actually monitor directly the strength of the memory 'trace' of a studied item (Arbuckle & Cuddy, 1969). This view presumes that individuals are able to constantly gauge the memory strength of a to-be-remembered item and update their study time based on these assessments. Factors that are known to influence JOLs, such as amount of study and extensiveness of encoding, are thought to operate through their influence on the memory trace itself. For example, according to this approach the greater the amount of study time, the stronger the memory trace will be. JOLs are considered to be accurate reports of the strength of the memory trace, such that the stronger the memory trace is, the higher the JOL. A number of problems with this hypothesis have been noted. First, this approach ignores the inferential nature of JOLs and discounts the use of cue heuristics. Cues such as encoding and retrieval fluency, for example, influence JOLs but not actual recall (Benjamin et al., 1998; Hertzog et al., 2003). The importance of cue heuristics on JOLs are described in further detail below (Koriat, 1997). A second, related criticism derives from research indicating an imperfect relationship between JOLs and actual memory performance (e.g., Lovelace, 1984), suggesting that memory monitoring may operate separately from actual learning success. According to the direct-access hypothesis, predictions of future recall should always be accurate, at least to the extent that items judged more likely to be recalled should be recalled better than items with lower ratings (Schwartz, Benjamin, & Bjork, 1997). Yet research has shown that, under some conditions, items given higher recallability ratings are actually recalled more poorly than items with lower ratings (e.g., Benjamin et al., 1998).

An underlying flaw in the direct-access approach is that it assumes that individuals have privileged access to their own prior experiences, and can summon up these events in the same way that one locates and retrieves a document from a filing cabinet. Currently, there is little positive evidence to support this memory-trace view of access to our past. Instead, Whittlesea (2003) and others (e.g., Loftus & Palmer, 1974; Roediger & McDermott, 1995; 1996) propose that remembering is a constructive process based on both fluency of current processing and inferences or attributions made concerning the state of this processing. For example, Leboe & Whittlesea (2002; Experiment 3) found that confidence in recalling the correct target when cued with the stimulus depended upon inferences that participants made as to why a potential target came to mind when they viewed the stimulus. During the training session, participants were presented with stimulus-response pairs that were either related (e.g., LION-TIGER), unrelated (SUMMER-TABLE), or words paired with a string of "XXXX" (e.g., FORK-XXXX). Although overall accuracy was highest for recall of related items, participants were most confident in their response to unrelated targets and the least confident when the target was "XXXX". These authors suggested that response fluency and participants' evaluation of the source of that fluency contributed to their confidence. Specifically, a potential target word that is related to the stimulus comes to mind easily, such as thinking of the word TIGER when presented with the word LION. The target "XXXX" is also produced fluently because one-third of the study items were paired with it, thus making it a relatively frequent target. Yet despite the fluency of producing these responses, participants recognized that these potential targets came to mind not necessarily because they were the correct response, but because it is easy to think of the word TIGER when presented with LION, and "XXXX" comes to mind easily because it was seen frequently during the study phase. In contrast, if upon presentation of a stimulus an unrelated word is produced fluently, participants attributed this fluency to having produced the correct response. More recently, Whittlesea and Leboe (2003; Experiment 3) replicated this finding. In addition to the influence of fluency on confidence of recall, they also found that participants were most impressed by unexpected, or surprising fluency. By performing a median split on response times for each type of context (i.e., related, unrelated, or "XXXX"), Whittlesea and Leboe were able to disentangle the effects of response fluency and the perception of congruency/discrepancy in recall. Participants were more confident in fast responses to unrelated targets compared to fast generation of related targets, despite the fact that participants were actually quicker at generating related targets. Participants interpreted their surprising fluent generation of an unrelated target in response to the stimulus as indicative that they saw the word pair during the study phase, even when they had not. However, according to the direct-access hypothesis there should be no difference in the probability of claiming an incorrect response as actual recall since trace strength should be non-existent for all incorrect responses. The inferential view is preferable then, as it accounts for the reconstructive, evaluative nature of recall, explaining both feelings of remembering that are accurate and inaccurate.

The Inferential Approach to Judgments of Learning

A more promising alternative theory concerning the basis for JOLs is the inferential approach (Schwartz et al., 1997). This view takes into account sources of information available to the participant at the time of making JOLs other than those directly related to the success of learning. For example, there is some evidence that JOLs are partly determined by fluency effects (Benjamin & Bjork, 1996; Benjamin et al., 1998; Bjork, 1999; Hertzog et al., 2003; Matvey, Dunlosky, & Guttentag, 2001). Bjork (1999) suggests that two types of fluency processes influence JOLs. Perceptual fluency refers to speed of perceptual processing of a target item. Considerable research reveals that this influence gives rise to feelings of familiarity. The idea is that people unconsciously attribute fluent perception of a stimulus to prior exposure, leading to a conscious feeling of familiarity (Jacoby, Kelley, & Dywan, 1989; Jacoby & Whitehouse, 1989). Analogously, it appears that perceptual fluency may also influence JOLs by giving rise to the inference that a word is likely to be recalled later. Consistent with this idea, Hertzog et al. (2003) reported that the speed of generating an interactive image between two words (i.e., encoding fluency) was associated with higher JOLs, even though this influence was unrelated to actual recall.

For example, Reder (1987, Experiment 6) found that participants were not only faster to make a response but were also more likely to report that they 'think' they know the answer to a difficult general knowledge question if they had been previously exposed to one or two words from the question sentence (Primed condition), than if they had not been exposed to a word from the question sentence (Unprimed condition). For example, for the question "What term in golf refers to a score of 1 under par on a particular hole?", participants who were primed to this question had been previously exposed to the key words "golf" and "par", under the guise of rating the frequency of these terms. One explanation for why recent exposure to the primed words increased the feeling-of-knowing for the answer is that the difficult questions in the Primed condition were processed more fluently, thereby giving the participant a false feeling-of-knowing.

In contrast to encoding fluency, retrieval fluency involves the speed or certainty with which people can generate a potential response in the context of a memory task. Prior research demonstrates that retrieval fluency can enhance confidence in a response, independent of its accuracy. For example, Kelley and Lindsay (1993) found that prior exposure to responses to general knowledge questions produced higher confidence in those responses. To illustrate, participants in this study who were exposed to the name 'Cody' on a list prior to answering general knowledge questions were more likely to report the answer 'Cody' in response to the question, "What was Buffalo Bill's last name?" and to do so with a high degree of confidence. In this example, relying on retrieval fluency led to a close correspondence between confidence in the answer given and its accuracy. However, Kelley and Lindsay also found that participants exposed to a related-incorrect response (e.g., Hickock) were more likely to report this answer instead, and do so with a high degree of confidence. That is, while retrieval fluency can sometimes serve as a useful cue for judging the correctness of a response, it can also lead people astray when an answer is retrieved fluently for some reason other than its accuracy. Applied to JOLs, retrieval fluency may sometimes correlate with future recall success, but may also lead to erroneous predictions when retrieval fluency occurs for reasons that are not predictive of future remembering (Benjamin et al., 1998).

Thus, the inferential approach can accommodate circumstances in which JOLs will not be predictive of future recall because systematic errors sometimes occur in the use of inferential cues (Schwartz et al., 1997). Koriat's cue-utilization model is a prominent example of the inferential approach to JOLs (Koriat 1997; Koriat & Levy-Sadot, 1999; Koriat et al., 2002).

The Cue-Utilization Model of Judgments of Learning

Koriat (1997) proposed a cue-utilization model for explicating the mechanisms involved in making JOLs. He suggested that JOLs are based on three different types of cues available to the learner: intrinsic, extrinsic, and mnemonic cues. Intrinsic cues are characteristics of the study items that may provide some indicators of their future memorability. These a priori judgments are based on inherent features of a target item that suggest to the learner the ease or difficulty of learning the material. For example, an intrinsic cue in paired-associate tasks is the degree of relatedness between words in a pair. In a typical paired-associates task, participants are presented with a pair of words and are instructed to remember the target (i.e., right-hand word) for a later memory test. Word pairs can be manipulated such that there is a high degree of association between the cue and target word (e.g., HOT-COLD) or a low degree of association (e.g., TRUCK-LAMP). After study, participants are given a cued-recall task in which they have to generate the correct target word when cued with the left-hand (or cue) word. There is evidence that participants are sensitive to the intrinsic cue of relatedness when making JOLs, appreciating that highly associated pairings are easier to learn than less related pairings.

Extrinsic cues refer to study conditions and encoding processes present at the time of acquisition. Examples of study conditions include the number of times a word is presented during a study phase or the number of study items in a list. Encoding processes, such as whether a participant attends to the letter structure, phonology, or the meaning of a word, are extrinsic cues that may also be relied upon when making JOLs.

To illustrate the difference between these two bases of making JOLs, Carroll et al. (1997; Experiment 1) found higher JOLs were reported for paired associates that were less well-learned and related than for items that were overlearned and not related. Participants in this study were required to achieve either only two correct recalls per related item (correct recall of ROCK, given SOIL as a cue twice) or in the overlearned condition, 8 correct recalls per unrelated item (recall DISEASE given ENGINE as a cue eight times). The degree of relatedness between the words is an intrinsic factor of the material to be learned, whereas the amount of learning is extrinsic because it is a condition of the study task. Interestingly, they found that recall was higher for participants in the overlearned condition, even though items from that condition were associated with lower JOLs. This finding indicates that participants believed that an intrinsic cue (relatedness) would be more beneficial than an extrinsic one (amount of learning), although the extrinsic cue was the better predictor of performance. Such evidence that participants discount the use of extrinsic cues when making JOLs is not uncommon (Koriat, 1997; but see Shaddock & Carroll, 1997). Dunlosky and Matvey (2001) also found that the degree of relatedness between paired associates influenced JOLs, even when the relatedness rating was made by a separate group of participants. Thus, the weight of the current evidence suggests that participants rely heavily on intrinsic cues when making their JOLs. It is not clearly understood why participants often do not rely on extrinsic cues when making their JOLs, especially relative to the actual benefit of extrinsic cues on performance. This is an important point and will be examined again later, as it forms part of the motivation for these Experiments.

Mnemonic cues are subjective in nature and may vary considerably between individuals. These cues are used as signals by the learner to infer how well an item has been learned and may be influenced by both intrinsic and extrinsic cues. Examples of commonly-used mnemonic cues for making JOLs include fluency of perception or retrieval (e.g., Benjamin & Bjork, 1996) and memory for whether prior recall attempts were successful (e.g., Mazzoni & Cornoldi, 1993). The use of some mnemonic cues may promote accuracy when making JOLs; others, however, may lead people astray, as in research demonstrating that reliance on processing fluency during study may be unrelated or negatively-correlated with future recall success (Benjamin et al., 1998; Hertzog et al., 2003).

The above discussion provided a brief overview of the main approaches that examine the theoretical basis for JOLs. However, it is important also to understand the relationship between JOLs and actual memory performance. The following discussion looks specifically at the accuracy of JOLs under different study conditions, and the key methodological approaches used to calculate accuracy.

Accuracy of JOLs

Despite evidence that people do not make JOLs based directly on how well they have committed information to memory, previous research indicates that JOLs are usually fairly accurate, although not perfect, in appraising future memory performance (for a review see Koriat & Levy-Sadot, 1999; also Lovelace, 1984). Thus, it appears that inferences about the likelihood of future recall often rely on cues that are related to actual future recall. Nevertheless, after only one exposure to the to-be-remembered information, there is evidence that participants are sometimes overconfident in their JOL predictions (e.g., Lichtenstein, Fischhoff, & Phillips, 1982; Mazzoni & Nelson, 1995). Therefore, although participants tend to be fairly accurate in their JOLs after one study-recall trial, inaccuracies tend to be on the side of overestimating success in future remembering.

A number of factors have been associated with higher JOL accuracy after one study session. For example, when JOLs are delayed until just prior to testing (i.e., JOLs are given at the end of the study session, rather than after each target item), they are more predictive of recall success (Dunlosky & Nelson, 1992, 1994; Hertzog & Hultsch, 2000; Kelemen & Weaver, 1997; Nelson & Dunlosky, 1991; Thiede & Dunlosky, 1994;

Weaver & Kelemen, 1997). Using the method of free-recall for the memory task is also associated with greater JOL accuracy, compared to tests of recognition (Thiede & Dunlosky, 1994). In addition, Leonesio and Nelson (1990) found that the predictive accuracy of JOLs was higher for items that had been overlearned (to a criterion of four correct recalls) compared to items correctly recalled once.

The most dramatic effect on JOL accuracy is the delayed-JOL effect (Nelson & Dunlosky, 1991, 1992). In the typical item-by-item JOL procedure, participants make their JOLs immediately following the presentation of each study item. If participants make JOLs after studying a list of items, but just prior to testing, the accuracy of these predictions increases substantially. For example, in a typical delayed JOL experiment, participants study a list of paired associates and judge how likely they are to recall the target word when presented with the cue word. Unlike item-by-item (or immediate) JOLs that are elicited immediately after presentation of each item, in a delayed-JOL task these judgments are not elicited until a given length of time has passed between study and the JOL (e.g., 10 minutes).

To illustrate, Nelson and Dunlosky (1991) had participants make both item-by-item as well as delayed JOLs. Participants studied a list of paired associates and were required to make immediate JOLs for half of them and delayed JOLs for the other half. JOLs were elicited by prompting participants with the cue word, and then asking how likely they think it is that they will remember the target word later when given the cue word. In the immediate-JOL condition, JOLs were made after presentation of each word pair. For example, participants might be shown the pairing "LAMP-BULB", which would be followed immediately by "LAMP-???" along with the request to predict the likelihood of recalling the target word (BULB) later. In contrast, in the delayed-JOL condition, JOLs were made after presentation of the entire list of associates. For example, participants might be shown the pairing "FISH-SHRIMP", which would be followed after studying the list of associates by "FISH-???" along with the request to predict the likelihood of recalling the target word (SHRIMP) on a future memory test. A cued-recall test of participants' ability to recall target words given the left-hand cue words occurred after a delay of ten minutes.

Two of the more prominent competing views regarding the origin of the delayed-JOL advantage highlight the distinction between the direct-access view and the cue-utilization approach. For example, consistent with the direct-access view Kimball and Metcalfe (2003) have argued that delaying JOLs improves accuracy through improving actual memory performance (i.e., increasing trace strength), not through enhanced memory monitoring. They suggest that accuracy for delayed JOLs is higher than for immediate JOLs because delayed JOLs improve recall through spaced-study opportunities. According to their monitoring-retrieval hypothesis, when cued with a stimulus, a participant tries to recall the target and, if successful, will report a higher JOL than if they are unsuccessful. Thus, for immediate JOLs, participants are provided with only one study opportunity to see the target item and then make a JOL. In contrast, study opportunity is spaced for delayed JOLs because participants are exposed to the target item, and then after some time has passed, are given the opportunity to try to retrieve the target when the JOL is elicited. When differences in spaced study were eliminated by re-exposing participants to the word pairs *after* their initial study-JOL, the advantage of delayed JOLs was eliminated.

Spellman and Bjork (1992) have also argued that delayed JOLs operate through improving memory performance, and that the delayed-JOL effect is not a result of better memory monitoring per se, but due to an improvement in actual memory performance. This view is also consistent with the direct-access view of JOLs discussed above, in that the memory for a stimulus is assumed to be strengthened with repeated successful retrieval attempts. Nelson and Dunlosky (1992) have provided some evidence against this hypothesis, however, ruling out the possibility that actual memory performance is better when participants make delayed JOLs than when they make immediate JOLs. Dunlosky and Nelson (1997) also found no evidence that recognition performance is better after delayed than immediate JOLs, (when JOLs were cued by the stimulus alone), suggesting also that the delayed-JOL effect is not due to memory improvement. Instead, they favour a Monitoring-Dual-Memories (MDM) hypothesis in that people simultaneously access their short-term memory (STM) and long-term memory (LTM) when making JOLs (Nelson & Dunlosky, 1991). When immediate JOLs are made, information in STM about the target item is thought to interfere with information accessible from LTM, although recall is based only on accessing information in LTM. Delayed JOLs are more accurate than immediate JOLs because STM will not be a source of interference when JOLs are made sometime after the initial exposure to the item. Therefore, their argument is consistent with the cue utilization model in that STM interference makes the stimulus cue at the time of an immediate JOL less diagnostic of future recall than when there is no STM interference, such as when JOLs are delayed.

Consistent with the cue-utilization approach, Dunlosky and Nelson, (1997) also found that JOLs are more accurate when cued by the stimulus alone, rather than the stimulus-response pair, despite the fact that the latter more closely resembles the context at study. In this experiment, participants were prompted to make delayed-JOLs either by the stimulus alone (e.g., FISH-???), or by the stimulus-response pair (e.g., FISH-BOOK). Following the study and JOL phase, participants were then given a recognition test. According to the direct-access view, trace strength should be stronger when the target response is presented both at test and during the JOL phase than when presented during study alone. Consequently, participants should be more accurate in their JOLs in the former condition because increased trace strength should result in better discrimination between recalled vs. unrecalled words. Instead, participants were more accurate in their JOLs when prompted by the stimulus alone, suggesting that participants are relying on cues other than trace strength when making their JOLs.

Very recently, Nelson, Narens, and Dunlosky (2004) revised the standard delayed-JOL methodology to include a measure of recall at the time the JOL is made, thereby refining their analysis of JOL accuracy. They found evidence that the delayed-JOL effect arises from better accuracy at discriminating between a recalled vs. unrecalled item, compared to immediate JOLs in which the relevant discriminations are between recalled items. This also suggests that cues relied upon when making delayed JOLs are more predictive of recall performance.

Calibration as an Indicator of JOL Accuracy

In many of these previous investigations of the factors that influence JOL accuracy, an important method of assessment is the computation of calibration curves. Calibration (i.e., absolute accuracy) refers to the reasonableness of predictions for future recall (Hertzog & Hultsch, 2000; Koriat & Goldsmith, 1996; Lichtenstein et al., 1982). For every given level of JOL, calibration refers to the correspondence between that prediction level and the actual proportion recalled. For instance, perfect calibration occurs if for every given level of JOL (e.g., 40% rating), exactly that proportion of items is actually recalled. Resolution (i.e., relative accuracy) refers to the ability to discriminate between recalled and not recalled items at the time of learning. Thus, high relative accuracy would be associated with the ability to correctly predict items that will be recalled and items that will not be recalled, irrespective of mean JOLs and mean recall. Whereas calibration is typically analyzed by comparing mean JOLs with actual performance, and illustrated in calibration curves, resolution is measured correlationally using the Goodman-Kruskal gamma correlation γ, which is a rank-order index of agreement between JOLs and recall. Although both measures of JOL accuracy are valid for assessing memory-monitoring ability, these Experiments focus on calibration as an indicator of JOL accuracy.

Why is Calibration Important?

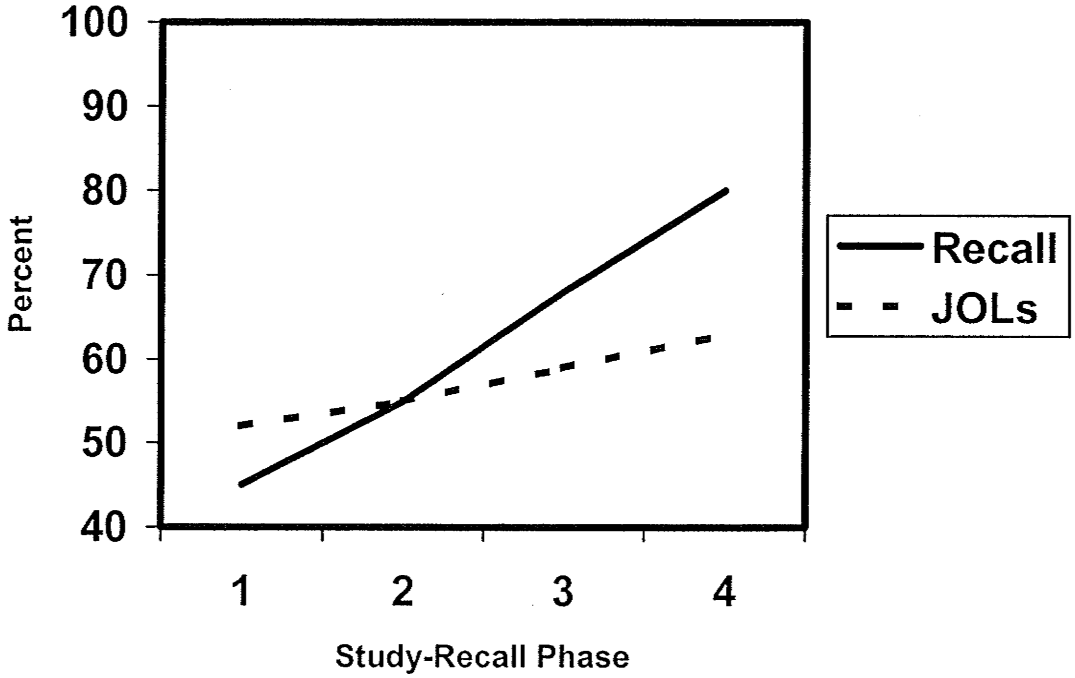
Lichtenstein et al. (1982) posed the following scenario: Consider a situation in which a physician must choose between two possible medical diagnoses. If the patient has condition A, then receiving treatment A would be the most prudent course of action. If, however, the patient has condition B, then treatment B would be the better choice of treatments. Furthermore, assume that treatment A is a better choice overall if the probability that the patient does in fact have condition A is .4 or greater, and the doctor assesses the probability of condition A being present at .45. If this doctor is poorly calibrated because the actual probability of the patient having condition A is .25, the patient would not be receiving the most appropriate and effective treatment. Poorly calibrated judgments can lead to serious negative consequences not only in the medical profession, but also for lawyers, stockbrokers, etc. (Lichtenstein et al., 1982). In everyday life, people are required to make legal, personal, and financial decisions based on predictions of future events. For example, deciding on a type of mortgage (i.e., fixed or variable interest rate), involves the prediction of interest rates years into the future.

The importance of calibration as a dimension of JOL accuracy is also made salient in the Underconfidence-With-Practice effect, a recently recognized phenomenon in the JOL literature that is the primary focus of the present Experiments.

The Underconfidence-With-Practice Effect

Recently, a new form of JOL inaccuracy was identified by Koriat, Sheffer and Ma'ayan (2002). In repeated study-recall cycles, participants show accurate, if slightly overconfident, predictions for the first study-recall cycle, followed by underconfident ratings in subsequent study trials. More specifically, although both recall and JOLs increase with each study-recall cycle, as shown in Figure 1 the increase in recall is significantly more dramatic than the increase in JOLs, resulting in a shift toward underconfidence. Koriat et al. identified this phenomenon as the Underconfidence-With-Practice (UWP) effect.

Figure 1. Typical Mean JOLs and Recall Representative of the Underconfidence-With-Practice Effect on Judgments of Learning.



Thus, calibration between memory monitoring (JOLs) and performance (actual recall) deteriorates with progressive study-recall trials. Koriat and colleagues argue that this shift toward underconfidence represents a deficiency in memory-monitoring effectiveness. Although often not the focus of previous studies involving multiple study-recall sessions, Koriat et al. note that the findings reported provide evidence for good calibration for the first study-recall trial and a shift towards underconfidence during the second presentation (e.g., Koriat, 1997). They also note in their review that the UWP effect is robust regardless of the amount of study-time, levels of incentive for successful recall, associative relatedness of paired associates, and whether or not participants are given accuracy feedback after each recall attempt. Contrary to the latter finding, Thompson (1998) found that providing feedback about the accuracy of answers to general knowledge questions did improve memory-monitoring accuracy from one presentation to the next. However, this finding may be due to the length of time between test sessions which ranged from one to three days, whereas the research discussed in Koriat et al.'s review involved study-recall trials that occurred successively during one experimental session.

Koriat et al. (2002) also noted that the UWP effect generalizes to aggregate as well as item-by-item JOLs. Aggregate JOLs are global predictions concerning the likelihood of future recall for items just presented and are made immediately after presentation of the last study item. Although some researchers have demonstrated a tendency for aggregate JOLs to be underconfident in the first presentation (e.g., Liberman, 2004; Mazzoni & Nelson, 1995), according to Koriat et al. a substantially larger magnitude of underestimation still emerges upon multiple study-recall cycles. In fact, Koriat and colleagues found that participants made lower aggregate JOLs after the fourth study trial than the percentage of words they actually recalled after the previous third trial (73.42% and 83.84% respectively)! Thus, despite a rather small tendency to be underconfident in the first presentation for aggregate JOLs, participants still demonstrate a marked tendency towards greater underconfidence in subsequent study-recall trials. The UWP effect has also been observed across different types of study material. For example, Koriat et al. (2002) observed a UWP effect in memory for motor actions. Very recently, Meeter & Nelson (2003) found a UWP effect for delayed-JOLs as well.

The UWP effect is surprising for several reasons. Absolute accuracy as measured by calibration is usually fairly accurate after one study-recall trial. Thus, participants are accurate, if sometimes slightly overconfident, in their subjective predictions of future performance. Intuitively, calibration should improve with practice when item-by-item JOLs are made because all of the items have not been presented until the end of the first study-recall trial. In the second and subsequent presentations of the study items, participants have had an opportunity to see all the study items and have already made one recall attempt. Consequently one might expect improved absolute accuracy, which is not the case. This effect is also unexpected given the fact that with repeated study-recall cycles participants become better at discriminating between items recalled and not recalled (resolution or relative accuracy improves), whereas calibration (or absolute accuracy) is impaired. This suggests different underlying mechanisms responsible for each of these measures. The UWP effect is also surprising given other evidence that practice retrieving information from memory is beneficial to the learning process (Bjork & Bjork, 1992). Thus, one might assume that calibration would improve not only because of additional study presentations of items but also with repeated recall tests. Although some researchers have found that retrieval practice enhanced JOL accuracy (e.g., Shaughnessy & Zechmeister, 1992), the weight of the evidence strongly suggests that calibration deteriorates with repeated study-recall trials.

So why is the calibration of JOLs impaired with practice, while resolution improves? Several possibilities have already been ruled out. For example, Koriat et al. (2002) concluded that the UWP effect cannot be explained by the distribution of JOL ratings across presentations. If JOLs became more polarized with subsequent study-recall sessions as a result of participants realizing that some items are just too difficult to recall, then items judged as very-unlikely to recall later (or JOLs of around 0%) would increase with each study-recall trial. This would lower the overall mean of JOLs across presentations. However, an examination of frequency distributions in the use of JOLs, collapsing across percentage intervals (i.e., 0-20%, 21-40%, etc.), revealed no increase in the use of the lower numbers across study-recall trials. Thus, the UWP effect is not an artifact of polarized judgments.

Koriat et al. (2002) suggest that the UWP effect is consistent with one of the propositions suggested by the cue-utilization approach. According to this approach, intrinsic cues are more immediately available and, therefore, should play a more dominant role in judgments of learning. As previously discussed, there is some evidence that participants tend to discount extrinsic factors relative to intrinsic cues when making their JOLs (e.g., Carroll et al., 1997). Indeed, much of this evidence reveals people's lack of sensitivity to the benefit of repeated study and recall. Carroll et al. (1997) found that participants undervalued the influence of overlearning material. In this study, participants learned word pairs to a criterion of either 2 correct recalls or 8 correct recalls; items in the latter condition were considered to be 'overlearned'. Although study-recall trials were not consecutive in these experiments, the results of this study suggest that participants are ignoring the value of repeated exposure to study material (an extrinsic cue) in their predictions. Further support comes from the finding that although retrieval experience is advantageous to learning (e.g., Bjork & Bjork, 1992), participants may be underestimating the benefit of this extrinsic cue when making their JOLs.

This emphasis on insensitivity to extrinsic cues is reasonable, although it cannot completely account for the fact that mean JOLs do in fact increase with repeated study-recall trials, (see Figure 1). Nevertheless, failing to appreciate the relevance of extrinsic cues provides a reasonable starting point for understanding why JOLs do not increase with practice, to the extent that they accurately predict recall.

Present Research

The UWP effect has important implications for the study of metamemory and the accuracy of memory-monitoring processes. Judgments of learning success are used to guide subsequent efforts to learn. Thus, efficiency in learning is largely dependent on whether people are accurate in assessing their effectiveness in acquiring knowledge.

Errors in monitoring learning progress will lead either to incomplete acquisition of the studied material or to devoting more time than necessary to the task. Given these implications, it is important for researchers to explore the underlying cause of underestimations of learning success. This phenomenon is also an important avenue of inquiry because the mechanism(s) underlying the UWP effect are currently unknown (Koriat et al., 2002).

The present experiments investigate JOL accuracy using a methodology that more closely matches real-life learning situations than the commonly-used paired associate learning procedure. A list-learning paradigm is used instead of word pairs not only because the list-learning situation is a more ecologically-valid mode of learning, but also because it is a relatively unexplored methodology for studying the UWP effect. Often in real life we are not given specific cues to help us remember information. Consider learning a grocery list; although walking up and down each aisle in the grocery store may aid recall of particular items intended for purchase, relying on cues alone would probably result in coming home without all of the necessary ingredients for that pie you intended to bake! In their discussion of the UWP effect, Koriat et al. (2002) reference only one study that used a list-learning paradigm. In this experiment, the UWP effect did not emerge until the third study-recall cycle, but was significant for that trial as well as the next. Thus, part of the objective for Experiment 1 was to replicate the UWP effect using a list-learning task.

Also, although there is some evidence that delayed JOLs (in paired-associate studies) tend to be more accurate, immediate JOLs were used in this study because it is not possible to 'cue' a single-item target without actually showing the target word again. In paired-associate experiments, for example, a participant would make a delayed JOL for the word SAND when cued by the word BUTTON; however, when only the target word itself was presented at study, a delayed JOL would necessarily have to be prompted by the target word itself, thus increasing the number of presentations of the item. Nevertheless, very recently, the UWP effect has been reported for delayed JOLs using paired-associates (Meeter & Nelson, 2003). Thus, evidence suggests that the critical underlying mechanisms responsible for the UWP effect are not related to the timing of the JOLs.

But what about the basis for JOLs, and their relationship to the UWP effect? The exact nature of this relationship is currently unknown, although as described earlier, there is evidence that people discount extrinsic cues when making their JOLs (e.g., Carroll et al., 1997), and that the difference between mean JOLs and mean recall increases in the direction of underconfidence with repetition for both immediate and delayed JOLs. However, the UWP effect may occur because participants do not fully realize the benefit of the extrinsic cue of repetition. Perhaps people do not appreciate that repetition benefits future recall when they are forced to encode words in the same way during each exposure. Consequently, when individuals are faced with seeing the same items repeatedly, and in the same manner (e.g., reading the word silently), they do not 'clue in' to the fact that re-exposure to the item will significantly aid their recall. A logical hypothesis then, is that in order for people to appreciate that repetition is good for recall, they need to encode the item in a different way during each exposure. Consider a potential graduate student who is preparing for the vocabulary portion of the Graduate Record Exam (GRE). Even though studying the words and their definitions by repeatedly reading them over to herself may in fact aid her recall of the definition, and thus improve her performance on the exam the following day, she may not realize this benefit of study. Instead, she may have the experience that she is not learning anything new, deciding to go out with her friends instead, erroneously believing that further study is a waste of time!

In Experiment 1, participants were given a list of words to alternately study and recall for four cycles and were informed of the need to recall the words for all phases of the study. However, after the first phase, type of encoding was manipulated within-participants. Specifically, for half of the words, participants made separate meaning-based relatedness judgments for each phase after the first; they read the other half of the words silently to themselves. In Experiment 1, there were two extrinsic cues available to participants when making their JOLs: the type of encoding and repetition. Thus, Experiment 1 tested whether the UWP effect occurs because participants discount the extrinsic cue of repetition based on their belief that encountering words the same way multiple times provides little benefit for future recall. That is, participants may become bored with the process of repeatedly studying the same item in the exact same way such that they do not think that they are learning anything new.

This hypothesis has not been previously tested, but research in other cognitive domains suggest that it is a logical starting point for the present investigation. For example, research in the area of novel popout (e.g., Johnston, Hawley, & Farnham, 1993), negative priming (e.g., Leboe, Leboe, & Miliken, 2003), and visual attentional capture (e.g., Yantis & Jonides, 1990), suggest that attention is drawn to novel events. Furthermore, people make inferences concerning their current performance based on their interpretation of this novelty. More specifically in the field of metacognition, Whittlesea (2003) has provided good evidence to suggest that the process of making metacognitive judgments involves using different decision heuristics, depending on the demands of the task. Thus, intuitive theories do play an important role in making subjective judgments. In Experiment 1, the distinct learning experience created by making relatedness judgments may therefore be a cue heuristic that people use when judging how well they think they know a study item. If this assumption is true, then it follows that calibration should be better for target words in which a relatedness judgment is made, because participants may have the intuitive theory that they are learning something new each time they encode a word differently and this will be reflected in higher, (i.e., more accurate), JOLs. That is, for words that require a different relatedness judgment during each repetition, encoding those words differently during each encounter may make people sensitive to the benefit of repetition for future recall. If so, the UWP effect should be minimized, or disappear altogether, for items in the relatedness-judgment condition.

Experiment 1

Method

Participants

Forty-seven undergraduates from the University of Manitoba enrolled in an introductory psychology course participated in exchange for course credit. Eight participants were eliminated from analysis of the results either due to their failure to follow experimenter instructions or because of computer error in the recording of data. The mean age of the remaining participants was 19.8 years (22 women and 17 men). All participants spoke English as their first language and were under the age of 30. These restrictions were imposed in order to control age and language variability, thereby reducing variability in memory ability across participants.

Apparatus and Stimuli

The study targets in this experiment were 40 concrete nouns varying between 4 and 7 letters in length. Although formal norms for word frequency were not used in constructing these items, all target words were fairly common in everyday usage (e.g., GARDEN, BREAD). From each target word a set of six additional words were constructed, consisting of three words related to the target (e.g., FLOWER for GARDEN; TOAST for BREAD) and three unrelated to the target (e.g., GORILLA for GARDEN; CRAYON for BREAD).

All items were presented on an IBM-compatible computer with a 15-inch monitor. Micro-Experimental Laboratory 2 (MEL2) software was used for presentation of stimuli and recording of participants' responses.

After each presentation of the study list, participants completed a free-recall task by writing responses down on paper.

Procedure

All participants were tested individually in front of a computer. Participants were given separate instructions for Phase 1 (the first study-recall cycle) and Phases 2 to 4 (study-recall cycles 2 to 4). Prior to Phase 1, participants were informed that they would be shown a series of words, which were presented one at a time at the center of the computer screen. They were instructed to read each word silently and try to remember it for a later memory test. The total number of words and the exact timing of the memory test were not explained to participants. Immediately after each word was presented, participants were prompted by instructions on the computer screen to make their JOLs.

Participants were prompted to make JOLs by the appearance of the question, "How likely do you think it is that you will be able to recall this word later?", on the computer screen. Participants were then asked to use the keyboard to type in a two-digit number between 00 indicating 'not at all likely to remember' and 99 indicating 'definitely will remember'. Participants confirmed their answer by typing in the appropriate response, 'Y' for yes and 'N' for no.

JOLs can be reported in a number of ways. The most common types of responses involve reporting a percentage (i.e., how likely do you think it is that you will be able to recall this word later?) with responses ranging from 0% (not at all likely to remember) to 100% (definitely will remember) or in the form of Likert-scale responses (e.g., 0- not at all likely, 3- somewhat likely, 6- very likely), or as a prediction about number of items that were recalled. This latter method is used when participants are asked to make aggregate JOLs after exposure to the last study item. For this experiment, responses in the form of a percentage were chosen for several reasons.[1] First, there is no evidence that reporting JOLs in a form other than a percentage reduces accuracy. Kelemen (2000), for example, found no difference in accuracy of JOLs when they involved percentage ratings of the likelihood of future recall versus predictions about the number of items that would be recalled. As well, reporting JOLs as percentages may be less problematic with respect to subsequent data analysis (see Hertzog & Hultsch, 2000; Koriat, 1997), since they are easier to manipulate statistically. For example, as a percentage, calculating JOL accuracy allows for a more direct comparison, involving a simple computation of the difference between the mean of a participants' JOLs for all study items and the percentage of words correctly recalled. Moreover, although the UWP effect is robust despite differences in how JOLs are reported, the majority of the studies reported in Koriat et al.'s (2002) review of the phenomenon used judgments reported in percentages. Finally, percentage intervals can be easily computed from raw percentages and thus, calibration curves can be generated, allowing efficient graphic representation of the correspondence between JOL magnitude and actual recall success.

Phase 1. Prior to the actual study phase, participants received six practice items, also allowing them practice in making JOLs. The study session consisted of 40 trials. On each of these trials and on the practice trials, target words appeared alone in the center of

[1] Actually, the percentages used in this experiment ranged from 0– 99%, rather than 0– 100%, due to limitations in the MEL2 program. However, participants were clearly instructed to report their JOL as a percentage, and it was further emphasized that 99% referred to 'definitely will remember later', (i.e., 100%). To correct for this, in all Experiments the mean recall was recalculated to a proportion out of 99 for all analyses.

the computer screen for three seconds. After the word disappeared from the screen, participants were prompted for their JOL. They were asked to confirm their judgment by pressing "Y" to indicate the number on the screen was their intended JOL and "N" if they mistyped their response. Whenever "N" was chosen, the screen went back to the JOL prompt and participants re-entered their percentage judgment. After the JOL was entered the next word appeared on the screen. Immediately following the presentation of the final target item, participants were presented with the instruction on the computer screen, "Stop and wait for instructions from the experimenter". At this point, the experimenter provided a sheet of paper on which to perform the recall test. Participants were then given two minutes to recall the words that were just presented, in any order. The outline of the procedure for Experiment 1 is shown in Figure 2.

Figure 2. Experiment 1 Procedure.

Word-Alone Condition			**Relatedness Condition**		
Study 1	bread	**Read + JOL**	**Study 1**	garden	**Read + JOL**
	FREE RECALL			**FREE RECALL**	
Study 2	bread	**Read + JOL**	**Study 2**	crayon garden flower	**Relatedness + JOL**
	FREE RECALL			**FREE RECALL**	
Study 3	bread	**Read + JOL**	**Study 3**	lettuce garden coat	**Relatedness + JOL**
	FREE RECALL			**FREE RECALL**	
Study 4	bread	**Read + JOL**	**Study 4**	button garden seed	**Relatedness + JOL**

Phases 2 to 4. Half of the target words shown in Phase 1 again appeared alone for three seconds each (Word-Alone condition), but the other half of the target words were presented with two additional words: one related to the target and the other unrelated. These two words were presented on either side of the target word, with all three words aligned horizontally at the center of the computer screen (Relatedness condition). Thus, the study session for Phases 2 to 4 consisted of 20 study trials for the Word-Alone condition and 20 trials for the Relatedness condition. For the words in the Relatedness condition, participants were instructed that the middle word was the target word to be studied for a future memory test. In addition, however, participants were instructed to press a specific key if the left word was related to the target word, and a different key if the right word was related. Whether the related word was presented to the left or right of the target was determined at random across trials. Also, words from the Word-Alone and Relatedness conditions were presented in random order. To ensure that differences in recall or JOLs between the two encoding conditions were not due to the particular items chosen, a set of 20 targets were chosen to serve in the Word-Alone condition for half of the participants, with the remaining 20 targets serving as targets in the Relatedness condition. This assignment of words to conditions was reversed for the other half of participants. In other words, the specific target words appearing in the Word-Alone and Relatedness conditions were counterbalanced across participants.

After words from the Word-Alone condition disappeared from the screen and immediately after entering a response for words in the Relatedness condition, participants were prompted to make a JOL. JOL responses were made, following the same procedure as in Phase 1. Participants were also given six practice trials to make relatedness judgments, followed by JOLs, before starting Phase 2. The procedure in Phases 3 and 4 were identical to Phase 2, except that different flanking words formed the basis of participants' relatedness judgments in each of these phases and no practice trials were provided. After each study phase, participants were given a free recall task, following the same procedure as in Phase 1.

Thus, Experiment 1 represents a 4 (Study-Recall Cycles 1-4) X 2 (Word-Alone vs. Relatedness encoding) repeated-measures design. The two primary dependent variables of interest are the percentage of words recalled and mean JOLs obtained for each participant.

Results & Discussion

There were two main points of interest that motivated Experiment 1. First, the Word-Alone condition across the four Phases allowed for a replication of the UWP effect using a list-learning paradigm. Based on the one previous example of the UWP effect using list-learning (Koriat et al., 2002), it was anticipated that a UWP effect would emerge, although the one previous example described by Koriat et al. suggests this effect may not be apparent until Phase 3.

The second point of interest in Experiment 1 is the possible difference in the relationship between JOLs and actual recall between the Word-Alone and Relatedness conditions. As discussed earlier, one possible explanation for the UWP effect is that people are not sensitive to the extrinsic cue of repetition, and that in order for people to appreciate that repetition is good for recall, they need to encode the item in a different way during each exposure. Thus, in Experiment 1, for words that require a different relatedness judgment during each repetition, encoding those words differently during each encounter may make people sensitive to the extrinsic cue of repetition. That is, participants may be more sensitive to the benefit of repetition for future recall in this condition relative to the Word-Alone condition. Consequently, it is expected that calibration across successive study-recall cycles should be better for target words in the Relatedness condition, thereby reducing or possibly eliminating the UWP effect that was expected to occur for the Word-Alone condition. If however, this manipulation did not make the extrinsic cue of repetition salient to participants, or they shifted their reliance to other cues, the UWP effect will remain in the Relatedness condition.

Aside from the predictions concerning JOL accuracy, recall was expected to improve from Phase 1 to Phase 4, given the effect of repeated study and retrieval tasks on actual recall performance. As well, based on the Craik and Lockhart (1972) levels of processing theory that deeper encoding improves recall, I also expected that recall would be better for words in the Relatedness condition. Finally, most studies of the UWP effect demonstrate a small increase in JOLs across study-recall cycles (see Koriat et al., 2002). Therefore, I expected to find a main effect of repeated study and recall on mean JOLs, with JOLs slightly higher for each successive phase of the experiment even for the Word-Alone condition.

Overall Analyses

In an attempt to ensure that any change in the relationship between mean JOLs and mean recall across Phases 2 to 4 was not the product of a small subset of items, mean words recalled and mean JOLs for each of the 40 items was submitted to a repeated measures ANOVA, with encoding condition and phase as within-item factors. This analysis yielded the same main effects and interactions described below when participants were used as the basis of analysis. As expected, these effects were not significantly influenced by differences between items, *F* <1. This is consistent with other research using paired associates, in that immediate JOLs for items did not vary as a function of prior item difficulty (Richards & Nelson, 2004).

The change in relationship on an item-by-item basis between Phases 2 to 4 was also investigated. In only 3 of the 40 items used in this experiment was a violation of the overall shift from higher to lower confidence observed. Thus, rather than serving as a cause of the effects observed in Experiment 1, a select few items actually represented a source of error variance that acted against the effects reported below. Taken together, these analyses clearly demonstrate the validity of the word list used in this study.

In order to avoid recency effects, the last three trials for each subject, in each of Phases 1 through 4, were omitted from further analyses. Furthermore, only data from trials in which correct relatedness judgments were made were included in the analyses. Across all participants, this resulted in 30 trials being omitted out of a total of 2154 relatedness-judgment trials.

Although not taken from formal norms, most participants made no errors in their relatedness judgments. All-correct judgments were made by 33 out of 39 participants; the accuracy rate across all relatedness-judgment trials was 98.6%, indicating that participants were in fact able to correctly identify the word related to the target.

Mean reaction times to make relatedness judgments in Phases 2 to 4 for all words in the Relatedness condition were also computed. Mean reaction times (in *ms*) for each of Phases 2 to 4 were 2327.3, 1951.6, and 1751.7, respectively. Only data from trials in which participants took less than 30 seconds to make their judgments were included in the analyses. Two trials were eliminated for this reason.

Analysis of Recall

A 4 X 2 repeated-measures ANOVA was computed based on the proportion of words recalled for each participant, treating Repetition (Phases 1 to 4) and Encoding condition (Word-Alone vs. Relatedness) as within-participant factors. As stated above, it was hypothesized that there would be a main effect of Repetition, with the proportion of words recalled increasing from Phase 1 to Phase 4. As shown in Table 1, a significant linear increase in recall performance was found; participants recalled more words overall with each exposure to the study list ($F_{\text{linear }(1,38)} = 332.1$, $MSE = 1.3$, $p < .001$). The mean percentage of words recalled in Phases 1 to 4 were 32.5%, 43.5%, 53.0%, and 63.4%, respectively. This finding is consistent with previous studies that have found better recall for items that were presented multiple times, compared to items presented only once (e.g., Shaughnessy and Zechmeister, 1992).

Table 1: Experiment 1: Mean Recall for Phases 1-4 by Encoding Condition (*N* = 39).

Recall (%)

	Phase			
Encoding	1	2	3	4
Word-Alone	32.4 (2.1)	39.1 (2.7)	50.4 (2.6)	60.1 (2.5)
Relatedness	32.5 (2.2)	47.8 (2.3)	55.5 (2.5)	66.6 (2.5)

Note: Mean standard error is in parentheses.

Since relatedness judgments were only presented in Phases 2 to 4, the ANOVA testing the effect of Encoding condition only included data from those phases. As expected, participants recalled more words in the Relatedness Condition than in the Word-Alone condition (see Table 1; $F_{(1,38)} = 11.1$, $MSE = 2.5$, $p < .01$). This finding is also consistent with a large body of research demonstrating that deeper, more meaningful processing provides substantial benefits for future recall (Craik & Lockhart, 1972). Also, it appears that the recall advantage for words in the Relatedness condition occur primarily the first time participants made those judgments. This is further demonstrated by the lack of interaction between Encoding condition and Repetition ($F < 1$). The percentage increase in words recalled from Phase 1 to Phase 2 in the Relatedness and Word-Alone conditions were 15.3 and 6.7 respectively, indicating that significantly more words in the Relatedness condition were recalled in Phase 2 compared to words recalled in the Word-Alone condition. The percentage increase in words recalled from Phase 2 to Phase 4 for Relatedness and Word-Alone conditions were 18.8% and 21.0% respectively, indicating that there is no additional advantage of recall of words in the Relatedness condition after Phase 2.

Analysis of JOLs

A 4 X 2 repeated-measures ANOVA was computed based on mean JOLs for each participant, treating Repetition and Encoding condition as within-participant factors. The two main hypotheses tested here considered the main effects of Repetition and Encoding condition. First, if participants are at all sensitive across the study-recall cycles to the extrinsic cue of practice benefits, mean JOLs should increase from Phase 1 to Phase 4.

Not surprisingly, a significant linear increase was found for JOLs such that participants' judgments of future recall for words increased as a function of the number of times they saw the words ($F_{\text{linear}\ (1,38)} = 12.9$, $MSE = 3.2$, $p < .01$). JOLs for Phases 1-4 for each Encoding condition are shown in Table 2. The mean JOLs for Phases 1 to 4 were 46.5%, 50.5%, 52.8%, and 56.5% respectively.

Table 2: Experiment 1: Mean JOLs for Phases 1-4 by Encoding Conditions ($N = 39$).

JOLs (%)

Phase

Encoding	1	2	3	4
Word-Alone	46.4 (3.1)	49.9 (2.9)	52.0 (3.1)	54.5 (2.9)
Relatedness	46.6 (2.3)	51.1 (2.9)	53.6 (3.1)	58.6 (3.2)

Note: Mean standard error is in parentheses.

If participants are sensitive to the positive contribution of the extrinsic cue of making a judgment based on a word's meaning on future recall, JOLs should be higher for words in the Relatedness condition than in the Word-Alone condition. As in the analysis of recall data, the effect of Encoding condition was tested based only on data from Phases 2 to 4. There was a marginal main effect for Encoding condition; a trend emerged for participants' JOLs to be slightly higher for words in the Relatedness condition than in the Word-Alone condition ($F_{(1,38)} = 3.53$, $MSE = .88$, $p = .068$). JOLs were 1.2% higher in the Relatedness condition than in the Word-Alone condition in Phase 2, 1.6% higher in Phase 3, and 4.1% higher in Phase 4.

Analysis of Calibration

To analyze the relationship between mean JOLs and actual recall, a 4 X 2 X 2 repeated-measures ANOVA was computed, treating Repetition, Encoding condition, and Calibration (Percentage of Words Recalled - Mean JOLs) as within-participant factors. This latter factor permitted computation of the accuracy of JOLs for each of the two study conditions and changes in JOL accuracy as a function of Repetition. In accordance with Koriat et al. (2002), a significant interaction between Repetition and Calibration was expected for the Word-Alone condition, in that there would be a shift from either accurate or overconfident judgments of future recall in Phase 1 to underconfidence in subsequent Phases. In contrast, no significant interaction between Repetition and Calibration was expected for the Relatedness condition. That is, although relatively insensitive to the benefit of repetition for recall when words are presented in the same way multiple times, people ought to appreciate the value of repeated exposure more when words are encoded differently during each exposure. In other words, the prediction was for a significant 3-way interaction between Repetition, Calibration, and Encoding condition. The results failed to support this hypothesis. A significant interaction was found for Proportion of Words Recalled - Mean JOLs and Repetition ($F_{\text{linear}\,(1,38)} = 51.6$, $MSE = 1.9, p < .001$). Mean recall and JOLs in Phases 1-4 for the Word-Alone condition are shown in Figure 3. The mean difference between percentage of words recalled and JOLs gradually shifted from overconfidence in Phase 1 to underconfidence in Phase 4. However, the three-way interaction between Repetition, Encoding condition, and Calibration was not significant ($F_{(1,38)} = 1.55$, $MSE = .90, p = .21$). Thus, the UWP effect emerged not only for the Word-Alone condition but also for the Relatedness condition.

As illustrated in Figure 4, mean recall for words in the Relatedness condition is approximately 8% higher than JOLs in Phase 4. Furthermore, there is a nonsignificant trend towards greater underconfidence in the Relatedness condition.

In order to further investigate the shift towards underconfidence in JOLs from Phases 2 to 4, a one-way repeated-measures ANOVA was computed for each of Phases 1 to 4 and for each Encoding condition. As shown in Table 3, the results for Phase 1 in which all words were only read, revealed overconfidence with mean JOLs 14% higher than the actual percentage of words recalled, ($F_{(1,38)} = 18.2$, $MSE = 4.0$, $p < .001$). In Phase 2, participants were still overconfident in their JOLs for words in the Word-Alone condition, ($F_{(1,38)} = 11.4$, $MSE = 1.8$, $p <.01$) but there was no difference between JOLs and mean words recalled in the Relatedness condition ($F_{(1,38)} = 1.34$, $MSE = 1.2$, $p = .254$). In Phase 3, there were no differences between JOLs and mean words recalled for either Encoding conditions ($F < 1$). By the fourth Phase, participants were underconfident in their estimations of future recall for both Word-Alone, ($F_{(1,38)} = 4.25$, $MSE = 1.8$, $p <.05$), and Relatedness, ($F_{(1,38)} = 7.98$, $MSE = 1.9$, $p <.01$) conditions. Thus, in both Encoding conditions, participants started out as overconfident and shifted towards underconfidence by the fourth study-recall trial.

Table 3: Experiment 1: Mean Difference Between JOLs and Recall for Phases 1-4 by Encoding Condition ($N = 39$).

		Phase			
Encoding	%	1	2	3	4
Word-Alone	JOL – Recall	14.0	10.8	1.6	- 5.2
Relatedness	JOL – Recall	14.1	3.3	- 1.9	- 8.0

Note: Positive numbers indicate overconfidence; negative numbers indicate underconfidence.

Figure 3. Experiment 1: Mean Recall and JOLs for Word-Alone Condition as a Function of Study-Recall Phase.

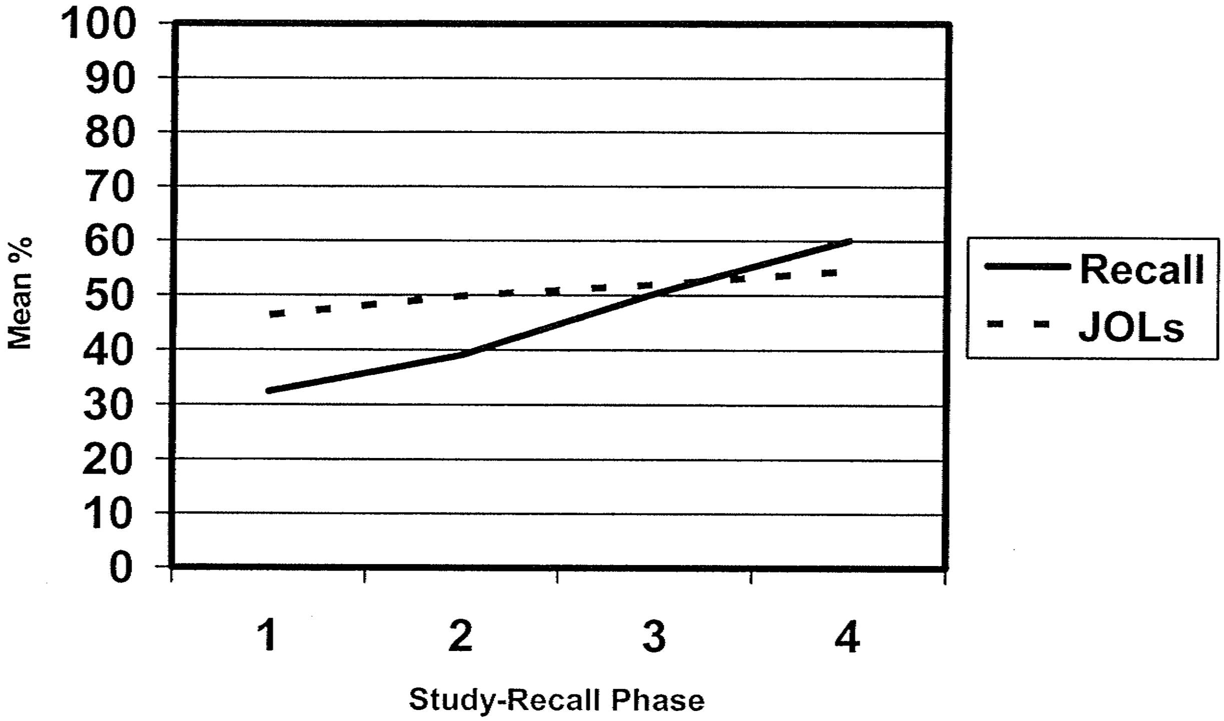
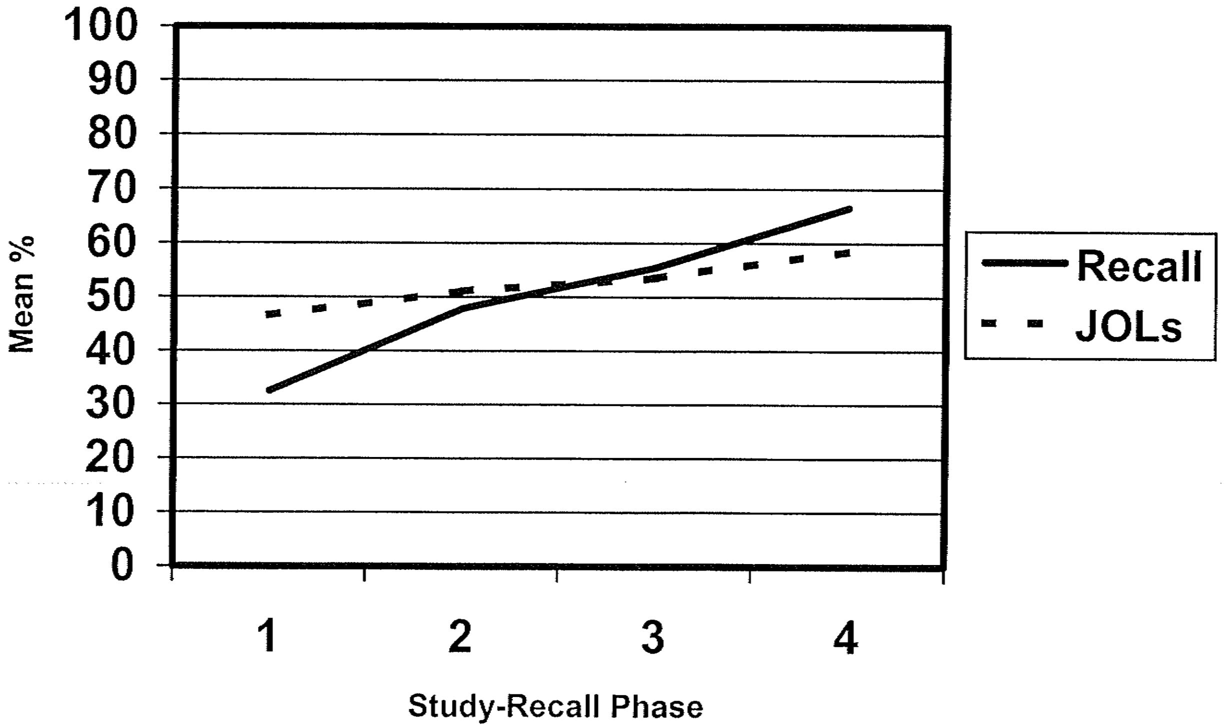


Figure 4. Experiment 1: Mean Recall and JOLs for Relatedness Condition as a Function of Study-Recall Phase.



This is consistent with Koriat et al.'s (2002) earlier finding in which underconfidence in estimating future recall of single-word items emerged during a later study-recall phase, compared to when paired-associates are used. Whereas Koriat et al. report that underconfidence emerges in the second study-recall phase when paired-associates are used, the one study described therein that used single-word target items did not reveal underconfidence until the third study-recall phase. The reason why underconfidence-with-practice still exists using a list-learning procedure, albeit after more study-recall cycles than in paired-associate tasks, is currently unknown and is worthy of future investigation.

Overall, the research hypotheses were partially supported by the results of Experiment 1. Specifically, these findings replicate previous research indicating that people tend to be underconfident in their estimations of future recall when presented with the same study items multiple times (Koriat et al., 2002). Contrary to expectations, however, the UWP effect was apparent in the Relatedness condition as well. The hypothesis that having people encode words differently by making relatedness judgments would minimize the UWP effect was not supported. Experiment 1 suggests that people's failure to recognize the benefit of making relatedness judgments is insufficient for explaining the UWP effect. One reason for this may be that encoding words differently is not sufficient to make the benefit of repetition salient to participants. The mean reaction time to make a relatedness judgment was only 2.01 seconds; participants may have considered the judgment too easy to provide any new learning relative to the initial presentation. In addition, participants became faster at making the relatedness judgment from Phase 2 (2.33 seconds), to Phase 4 (1.75 seconds), indicating that they may have found the task increasingly easier. As a consequence, this manipulation may not have caused participants to fully realize the benefit of repetition on their future recall of words in this condition. In effect, participants may have underestimated future recall in Phase 4 for both the Word-Alone and Relatedness conditions because they considered additional study of words in either condition not much more valuable for recall than the original exposure to these words in Phase 1. It is possible that the UWP effect is an unfortunate consequence of people's knowledge about the relationship between effort and learning. That is, people may have the belief that encountering a word multiple times is not much more valuable for future recall than encountering a word once, unless effortful processing is involved. Although making a relatedness judgment clearly involves more effort than reading a word silently to oneself, the results of Experiment 1 suggest that a greater amount of effort may be required.

If this interpretation of the results of Experiment 1 is accurate, it might be possible to cause people to appreciate the benefit of repetition more if the encoding process was made more effortful. According to Robert Bjork and colleagues (Bjork, 1999; Bjork & Bjork, 1992), learning conditions that pose difficulties for the learner actually promote better recall performance. Consequently, the more effortful processing is, the better recall should be. Experiment 2 explored the possibility that effortful processing may make the benefit of repetition salient. Participants in Experiment 2 were required to perform effortful tasks while encoding half of the target words. The first two study sessions in Experiment 2 were identical to that of Experiment 1; participants only read the word in the first study phase and then performed a relatedness judgment for half of the words in the second study phase. In the third study trial, half of the target words in the relatedness judgment condition were presented backwards and in alternating upper- and lower-case letters, making decoding the meaning of the target somewhat more challenging. In the fourth study trial, targets in the relatedness judgment condition were presented as anagrams, requiring participants to unscramble the letters of the target before performing the relatedness judgment. Thus, for half of the target words, each study trial involved different *and* effortful encoding. The prediction is that the UWP effect will not occur for words that are processed with effort, if participants are sensitive to the benefit of effortful processing.

Experiment 2

Method

Participants

Forty-one undergraduates from the University of Manitoba enrolled in an introductory psychology course participated in exchange for course credit. Five participants were eliminated from analysis of the results due to their failure to follow experimenter instructions. The mean age of the remaining participants was 19.3 years (26 women and 10 men). All participants spoke English as their first language and were under the age of 30 for the same reason as in Experiment 1.

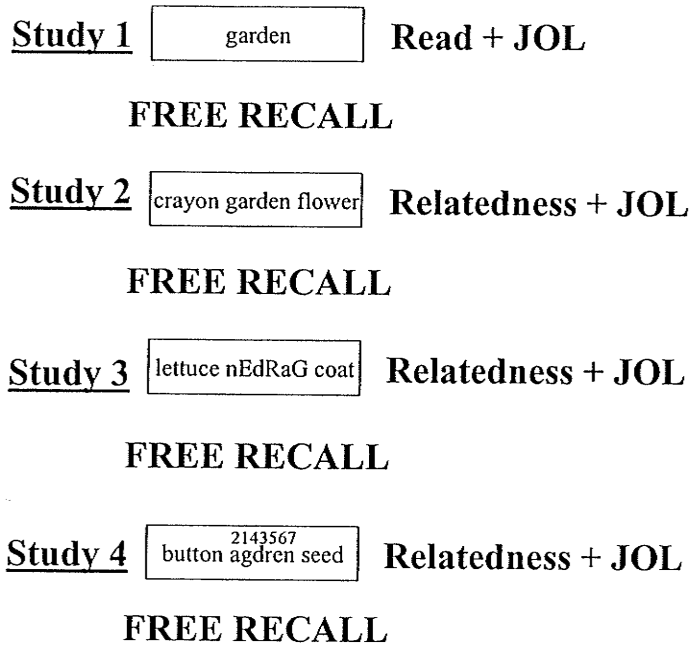
Apparatus and Stimuli

The materials used in this experiment were identical to that of Experiment 1.

Procedure

The procedure of Experiment 2 was identical to that of Experiment 1 except for the presentation of target words in the Relatedness condition for Phases 3 and 4, (see Figure 5). In Phase 3, target words presented for a relatedness judgment were presented backwards and in alternating upper- and lower-case letters. For example, the target word 'GARDEN' appeared as 'nEdRaG'. Participants were instructed to first decipher the word by reading it backwards, and then make a judgment about whether that word was related to the word on the right or the word on the left. The flanking words appeared normally with no alternating typeface. In Phase 4, target words presented for a relatedness judgment were presented as an anagram, appearing in a 2-1-4-3-5-6-7 format, such that the first two letters of the word were interchanged, the second two letters were interchanged, and the remaining letters appeared in their usual order. For example, the target word 'GARDEN' appeared as 'AGDREN'. A number key appeared above the target word indicating the order in which the letters went to solve the anagram. Thus, for the target word 'GARDEN', the numbers '2 1 4 3 5 6 7' appeared directly above the target word. As in Phases 2 and 3, the target word appeared between one related and one unrelated word from the set. These flanking words appeared in their regular, unscrambled form. Participants were instructed to first solve the anagram and then make the relatedness judgment. My assumption was that making a relatedness judgment when the target word appeared in regular, backward and alternating case, or anagram form required more effortful encoding than reading the word silently to oneself. For example, previous research indicates that anagrams involve effortful processing (Allen & Jacoby, 1990; Jacoby, 1991; Jacoby & Dallas, 1981; Jacoby & Hollingshead, 1990).

Figure 5. Experiment 2 Procedure.



Study 1 garden **Read + JOL**

FREE RECALL

Study 2 crayon garden flower **Relatedness + JOL**

FREE RECALL

Study 3 lettuce nEdRaG coat **Relatedness + JOL**

FREE RECALL

Study 4 button $\overset{2143567}{\text{agdren}}$ seed **Relatedness + JOL**

FREE RECALL

Note: Word-Alone condition same as in Experiment 1.

As one possible explanation for the UWP effect is that if people are not sensitive to the extrinsic cue of repetition, then perhaps in order for people to appreciate that repetition is beneficial for recall they need to encode the item differently *and* with effort during each exposure. Thus, it is hypothesized that calibration across study-recall phases should be better for target words which required effort to encode, thereby reducing or eliminating the UWP effect for target words in the Effortful encoding condition. If, instead, participants do not rely on this cue, or this manipulation of encoding difficulty does not make the extrinsic cue of repetition salient to them, the UWP effect will emerge nonetheless.

Experiment 2 represents a 4 (Study-Recall Cycles 1 through 4) X 2 (Word-Alone vs. Effortful Encoding) repeated-measures design. In addition to the hypotheses discussed in Experiment 1 concerning the benefit of repeated study-retrieval trials on recall, the prediction here was that recall would improve as a function of task difficulty. Mazzoni & Nelson (1995), for example, reported higher recall for words studied as anagrams than for words appearing alone. Bjork (1999; & Bjork, 1992) also suggests that recall improves as a function of study difficulty.

Results & Discussion

Overall Analyses

As in Experiment 1, the last three trials for each subject, in each of Phases 1 through 4, were omitted from further analyses resulting in a total of 37 target words per phase. Only data from trials in which correct relatedness judgments were made (Phases 2-4) were included in the analyses. Overall, relatedness judgment errors led to the elimination of 1.1% of trials from the Effortful encoding condition.

Mean reaction times to make relatedness judgments for all words in the Effortful condition for each of Phases 2 to 4 were computed. Only data from trials in which participants took less than 30 seconds to make their judgments were included in the analyses. Across participants, this resulted in one trial omitted out of a total of 1934 trials. Mean reaction times for Phases 2 to 4 were 1.80, 4.78, and 4.49 seconds respectively. In contrast to Experiment 1, reaction times for Experiment 2 indicate that participants did not find the relatedness judgment task easier across Phases, and that this task required greater effort than in Experiment 1.

The analysis of Recall, JOLs, and Calibration proceeded in a manner identical to that of Experiment 1. The only difference is that data from the Effortful encoding condition occupied the role served by data from the Relatedness condition in Experiment 1.

Analysis of Recall

Table 4: Experiment 2: Mean Recall for Phases 1-4 by Encoding Condition (*N* = 36).

Recall (%)

	Phase			
Encoding	1	2	3	4
Word-Alone	34.0 (2.4)	38.2 (2.9)	49.3 (3.2)	54.7 (2.9)
Effortful	31.4 (2.1)	44.6 (2.5)	55.8 (2.3)	64.4 (2.3)

Note: Mean standard error is in parentheses.

A 4 X 2 repeated-measures ANOVA was computed treating Repetition (Phases 1 to 4) and Encoding condition (Word-Alone vs. Effortful) as within-participant factors. As was found in Experiment 1, a significant main effect of Repetition on recall was found ($F_{\text{linear}\,(1,35)} = 146.0$, $MSE = 2.1$, $p < .001$). As shown in Table 4, participants recalled more words with each successive Phase. Mean recall in each successive phase was 32.7%, 41.4%,52.6%, and 59.6%, respectively. This replicates the effect of repetition on recall observed in Experiment 1 and conforms well with previous research demonstrating that repeated exposure to a study item benefits future recall (e.g., Shaughnessy and Zechmeister, 1992).

As in Experiment 1, participants were expected to recall more words in the Effortful condition than in the Word-Alone condition. As relatedness judgments were only presented in Phases 2 to 4, only data from these phases was included in a 3 (Phases 2 to 4) X 2 (Effortful vs. Word-Alone) repeated-measures ANOVA. As hypothesized, participants recalled more words that were presented in the Effortful condition than in the Word-Alone condition ($F_{(1,35)} = 17.8$, $MSE = 1.7$, $p < .001$). Overall, participants recalled 7.5% more words in the Effortful condition than in the Word-Alone condition. These results further support previous research indicating that effortful learning benefits future remembering (Bjork, 1999; Bjork & Bjork, 1992).

Contrary to expectations, there was no interaction between Encoding condition and Repetition ($F < 1$). This indicates that the benefit of recall for words in the Effortful condition occurred primarily the first time participants made the relatedness judgments (i.e., Phase 2). In the Effortful condition, 13.2% more words were recalled after the second presentation of the study list than after the initial presentation compared to an increase of only 4.2% for the Word-Alone condition. After Phase 2, however, the increase in words recalled for the Effortful encoding condition was 11.2% and 8.6% compared to an increase of 11.1% and 5.4% for the Word-Alone condition.

Thus, the main effect of Encoding condition is consistent with previous evidence that deep, meaning-based and more effortful processing benefits future remembering (e.g., Craik & Lockhart, 1972).

Analysis of JOLs

A 4 (Phases 1 to 4) X 2 (Effortful vs. Word-Alone) repeated-measures ANOVA was conducted treating Repetition and Encoding condition as within-participant factors. As was found in Experiment 1, there was a significant main effect for Repetition, in that participants' JOLs increased with each successive Phase (see Table 5, $F_{\text{linear}\ (1,34)} = 39.4$, $MSE = 3.0, p < .001$). This is also consistent with previous studies showing that JOLs do increase with each exposure to the targets (Koriat, 1997; Koriat et al., 2002). The mean JOL in Phase 1 was 39.0%, 43.9% in Phase 2, 51.0% in Phase 3, and 55.9% in Phase 4.

Table 5: Experiment 2: Mean JOLs for Phases 1-4 by Encoding Condition ($N = 39$).

JOLs (%)

	Phase			
Encoding	1	2	3	4
Word-Alone	38.9 (3.1)	43.5 (3.2)	50.4 (3.3)	54.7 (3.3)
Effortful	39.1 (2.9)	44.2 (3.1)	51.7 (3.5)	57.1 (3.2)

Note: Mean standard error is in parentheses.

As in Experiment 1, and in the analysis of recall results described above, the effect of Encoding condition was based on data from Phases 2 to 4. Results were similar to those found in Experiment 1; participants' JOLs were not significantly higher for words in the Effortful condition than in the Word-Alone condition ($F_{(1,34)} = 2.16$, $MSE =$

.53, $p = .15$). This finding suggests that participants were relatively insensitive to the benefit for future recall that is gained by effortful processing.

Analysis of Calibration

A 4 (Phase 1-4) X 2 (Effortful vs. Word-Alone) X 2 (Proportion of Words Recalled – Mean JOLs) repeated-measures ANOVA was conducted. This analysis yielded a significant Proportion of Words Recalled – Mean JOLs X Repetition interaction, $F_{(3,102)} = 5.42$, $MSE = 1.4$, $p < .01$. This interaction reflected a UWP effect in that participants were overconfident in their estimations of future recall in Phase1, shifting toward underconfidence by Phase 4. These results are consistent with previous research on the UWP effect (Koriat et al., 2002), and with the results of Experiment 1.

To determine whether the Encoding condition had an effect on Repetition and Calibration, a 3 (Phases 2 to 4) X 2 (Effortful vs. Word-Alone) X 2 (Proportion of Words Recalled vs. Mean JOLs) repeated-measures ANOVA was computed. A significant interaction was found for the latter two factors ($F_{(1,34)} = 14.7$, $MSE = .66$, $p < .01$). Participants were more underconfident in the Effortful Encoding condition than in the Word-Alone condition. Mean recall and JOLs for the Word-Alone condition across Phases 1-4 are illustrated in Figure 6. Figure 7 shows the mean recall and JOLs from Phases 1-4 for words in the Effortful encoding condition. In addition, the three-way interaction was not significant, $F < 1$, contrary to the hypothesis that the UWP effect would be eliminated for the Effortful encoding condition.

Figure 6. Experiment 2: Mean Recall and JOLs for Word-Alone Condition as a Function of Study-Recall Phase.

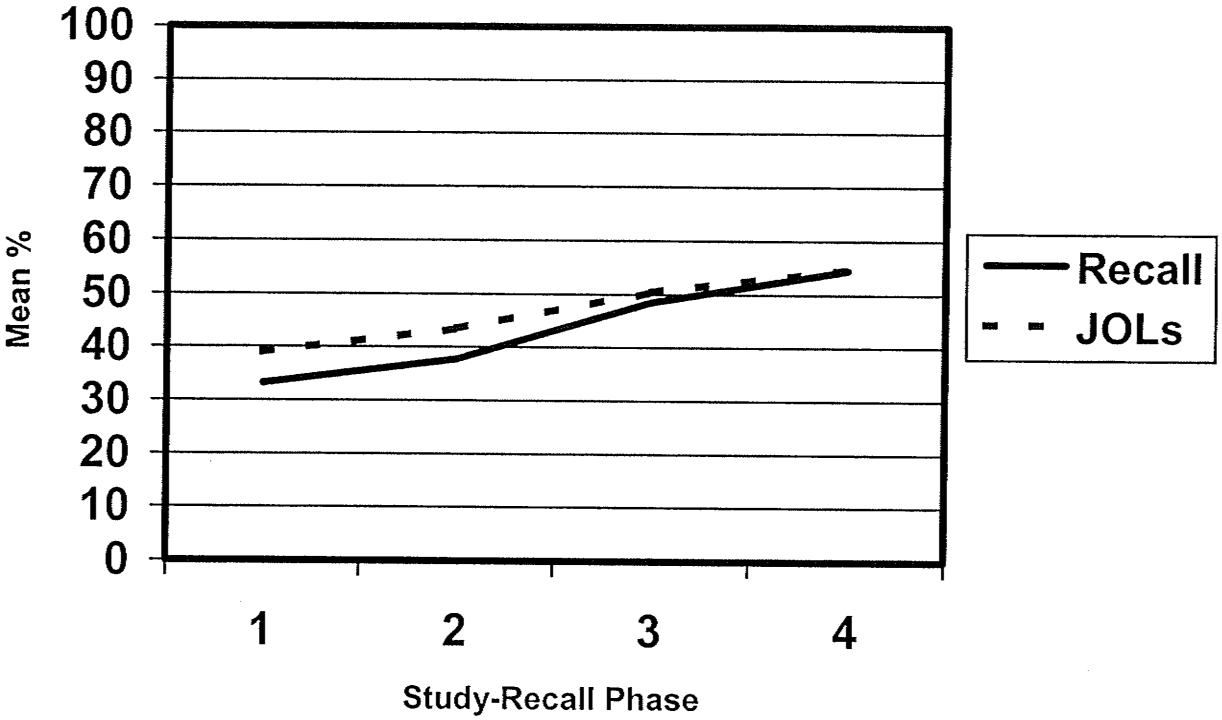


Figure 7. Experiment 2: Mean Recall and JOLs for Effortful Encoding Condition as a Function of Study-Recall Phase.

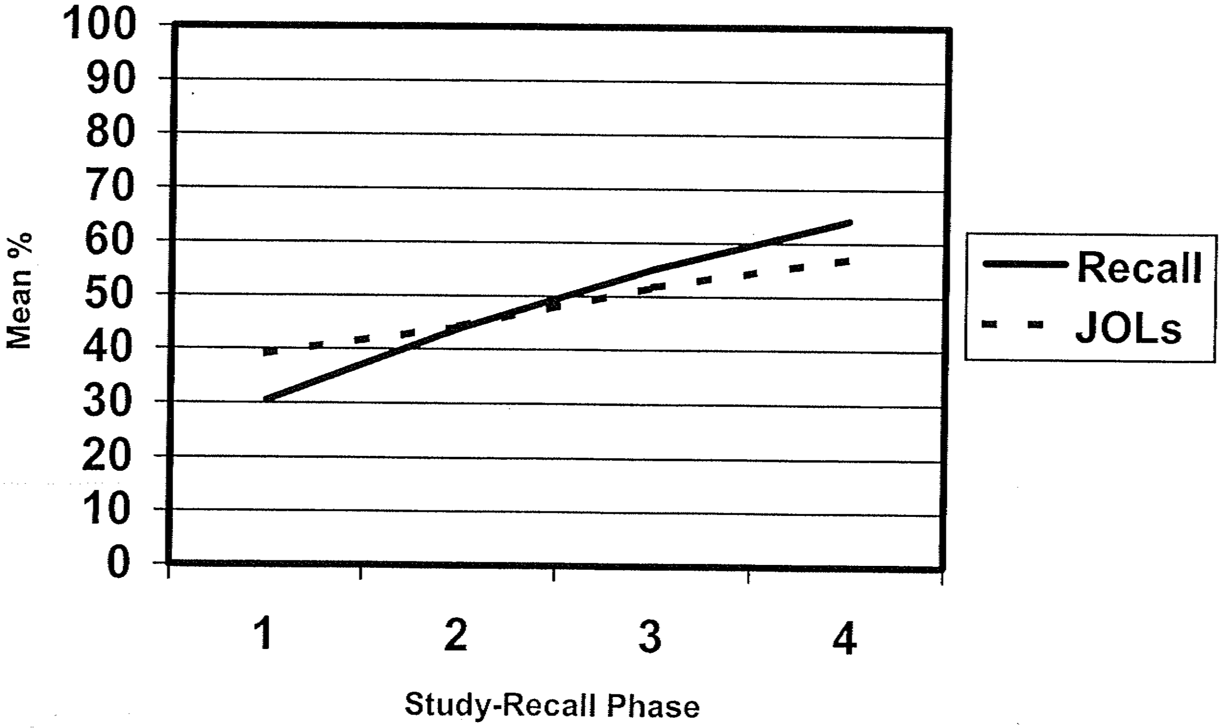


Table 6: Experiment 2: Mean Difference Between JOLs and Recall for Phases 1-4 by Encoding Condition ($N = 36$).

		Phase			
Encoding	%	1	2	3	4
Word-Alone	JOL – Recall	4.9	5.3	1.1	0.0
Effortful	JOL – Recall	7.7	- 0.4	- 4.1	-7.3

Note: Positive numbers indicate overconfidence; negative numbers indicate underconfidence.

To further clarify the relationship between Calibration and Repetition for each Encoding condition, a one-way repeated-measures ANOVA was computed for Phases 1 to 4 for each Encoding condition. As shown in Table 6, participants were somewhat overconfident in their JOLs after the initial presentation of the study list, with mean JOLs about 6.3% higher than the percentage of word recalled in Phase 1, $F_{(1,35)} = 3.97$, $MSE = 4.0$, $p = .054$. In the second and third study-recall phase, participants were again fairly accurate in their JOLs for both Encoding conditions with no significant differences between mean words recalled and mean JOLs ($p > .05$). By the fourth study-recall cycle, participants demonstrated a marginally significant trend toward underconfidence in the Effortful condition with participants mean JOLs 7.3% lower than the percentage of words actually recalled, $F_{(1,35)} = 3.50$, $MSE = 2.5$, $p = .07$, but no difference was found in the Word-Alone condition, $F < 1$.

The results of Experiment 2 closely mirror those of Experiment 1 for the Relatedness conditions. The UWP effect appeared in both the Distinctive learning condition (Experiment 1) as well as the Effortful encoding condition (Experiment 2). Thus, the hypothesis that the UWP effect would be minimized in the Effortful encoding condition was not supported by the results of Experiment 2. Instead, it appears that meaningful and effortful processing improves recall to a greater extent than it effects people's estimations of future recall.

Another possibility concerning why participants discount the use of extrinsic cues in their JOLs is that manipulating both encoding condition and task difficulty is still not sufficient for making the benefit of these cues salient at the time of their judgments. Instead, perhaps participants need to be explicitly informed of this benefit. In keeping with the theory that people discount extrinsic cues when making their JOLs (Koriat, 1997), and extending the work of Experiments 2 and 3 in which the extrinsic cues of repetition and encoding condition were available implicitly for participants to use, Experiment 3 examined whether or not participants were sensitive to the benefit of repetition if *explicitly* told about this benefit. Experiment 3 explored the possibility that if participants are explicitly told that repetition improves performance, perhaps this knowledge will then influence their JOLs such that there is a closer correspondence between JOLs and recall performance. If so, participants should be fairly accurate in their judgments throughout all the phases and, therefore, the UWP effect will be minimized or disappear. In Experiment 3, information concerning the benefit of the extrinsic cue of repetition was manipulated between-participants. In using the same procedure as in Experiment 1, this allowed a direct comparison between participants who are either (a) instructed that repetition is beneficial, (b) instructed that repetition has no effect or (c) are not provided any explicit instructions (Experiment 1). If informing participants is successful, no UWP effect should occur for participants in the first condition as their JOLs should more accurately reflect the benefit of repetition. It is expected that similar results may emerge for both of the latter groups if participants have an a priori belief that repetition is not beneficial, with the UWP effect present in both.

Experiment 3

Method

Participants

Fifty-five undergraduates from the University of Manitoba enrolled in an introductory psychology course participated in exchange for course credit according to the same restrictions as in Experiments 1 and 2. Five participants failed to follow instructions and were omitted from the analyses. Two additional participants were omitted because their accuracy rates were approximately at chance level, (error rate across Phases 2-4 = 41.4% and 43.2%). Thirty-five women and 13 men, mean age = 20.4 years, were included in the analyses.

Apparatus and Stimuli

The same apparatus and word list used in the previous two experiments was also used in Experiment 3.

Procedure

The procedure of Experiment 3 was identical to that of Experiment 1 except for the following modifications. First, prior to the start of Phase 1, participants were read a short paragraph informing them that either (a) repetition benefits recall or (b) repetition does not benefit recall. Second, at the end of the last recall test, participants were given a manipulation check. Specifically, participants were given a short survey to measure the effectiveness of providing explicit information about whether or not repeated exposure to information provides benefits for future recall.

Half of the participants (Repetition-Benefit condition) were read the following paragraph prior to the start of the first study-recall session:

> Successful learning depends, in part, on how well people think they have studied something and how likely they are to remember it later. For example, when studying for an exam, knowing what helps you remember information later is important in determining whether you do the right things to maximize your score on the exam. Previous research shows very clearly that studying the same item multiple times will help you to remember it later. So, for example, if you are required to remember a list of words, the more times you look at the word, the more likely you are to remember that word later. We are interested in learning more about how people learn when the same material is presented multiple times.

Prior to the start of Phase 4, participants in this condition were briefly reminded that:

> Once again, previous research shows very clearly that studying the same item multiple times will help you to remember it later. We are interested in learning more about how people learn when the same material is presented multiple times.

The other half of participants (Repetition-No Benefit condition) were read the following paragraph prior to the start of Phase 1:

> Successful learning depends, in part, on how well people think they have studied something and how likely they are to remember it later. For example, when studying for an exam, knowing what helps you remember information later is important in determining whether you do the right things to maximize your score on the exam. Previous research shows very clearly that studying the same item multiple times will not help you to remember it later. So, for example, if you are required to remember a list of words, looking at the word multiple times does not help you remember that word later. We are interested in learning more about how people learn when the same material is presented multiple times.

Prior to the start of Phase 4, participants in this condition were briefly reminded that:

> Previous research shows very clearly that studying the same item multiple times will not help you to remember it later. We are interested in learning more about how people learn when the same material is presented multiple times.

Participants in both groups then completed the following survey after the final recall test:

Question 1: "In *general*, how does seeing a word multiple times relate to its memorability?" Participants responded by circling their answer on an 11-point scale ranging from "0"-makes it much harder to remember, "5"-neither easier nor harder to remember, to"10"-makes it much easier to remember.

Question 2: "How do *you* think that seeing a word multiple times related to your recall of the word?" Participants responded by circling their response on an 11-point scale ranging from "0"-made it much harder to remember, "5"-neither easier nor harder to remember, to "10"-made it much easier to remember.

This check served two purposes. Question 1 relates directly to the Repetition-Benefit vs. Repetition-No Benefit manipulation, in that it is hypothesized that participants in the Repetition-Benefit condition would be more likely to report that seeing a word multiple times is beneficial, whereas participants in the Repetition-No Benefit condition were expected to report less benefit. It is important to note that although mean ratings for the benefit of repetition were expected to be lower for the Repetition-No Benefit condition, than for the Repetition-Benefit condition, they were still expected to be on the positive side of the scale (i.e., rated "5" or higher). Since the recall task provides some feedback to participants concerning the increasing number of words recalled after each successive study phase, it would have been surprising if participants denied any positive influence of repetition on recall. Nonetheless, participants in the Repetition-Benefit condition should report significantly higher ratings for the effect of repetition on memorability than participants in the Repetition-No Benefit condition.

Question 2 also relates to the manipulation of Repetition Benefit vs. No Benefit. It was expected that participants will rate the benefit of repetition for their own recall higher in the Repetition-Benefit condition than participants in the Repetition-No Benefit condition. However, Question 2 allows for a comparison between what participants may believe to be true *for most people* (Question 1) based on the instructions they received and what they themselves believe the effects of repetition were on their *own* performance. It is possible that explicit instructions will convince people that repetition is beneficial in an abstract, general sense, while remaining relatively insensitive to the contribution of repetition to recall during the course of the experiment.

Results & Discussion

Experiment 3 represents a 4 (Study-Recall Cycles 1 through 4) X 2 (Word-Alone condition vs. Relatedness condition) X 2 (Repetition-Benefit condition vs. Repetition-No Benefit condition) mixed design. As in Experiments 1 and 2, the first two variables represented within-participants factors. The manipulation of instruction (Repetition-Benefit vs. Repetition-No Benefit) is a between-participants factor and is the unique focus of interest for Experiment 3. In addition to the hypotheses discussed in Experiments 1 and 2 concerning the benefit of repeated study-retrieval sessions on recall, it was hypothesized that there would be a closer correspondence between JOLs and recall (i.e., improved calibration) for participants in the Repetition-Benefit condition, as explicit instructions may make them more sensitive to the extrinsic cue of repetition.

Other than the introduction of this between-participants factor, the analysis of proportion of words recalled, mean JOLs, and calibration proceeded in a manner identical to that of Experiment 1. The most critical difference is that the analysis of calibration involved submitting the data to a 2 X (4 X 2 X 2) mixed ANOVA, treating Repetition (Phases 1 through 4), Encoding condition (Word-Alone vs. Relatedness), and Calibration (Proportion of Words Recalled vs. Mean JOLs) as within-participant factors and Instruction condition (Repetition Benefit vs. Repetition-No Benefit) as a between-participants factor. As observed in Experiments 1 and 2, a significant interaction between Repetition and Calibration was expected for both the Word-Alone and Relatedness condition, revealing a shift from overconfident predictions of future recall in Phase 1 toward less or underconfidence in Phase 2. However, this relationship was expected only for the Repetition-No Benefit condition. Participants in the Repetition-Benefit condition were expected to be more sensitive to the extrinsic cue of repetition because they were informed of this benefit, and therefore less or no shift toward underconfidence was hypothesized across study-recall cycles. In other words, the UWP effect was expected to be minimized in the Repetition-Benefit condition, relative to the Repetition-No Benefit condition. This effect of Instruction condition was expected to be reflected in a significant 3-way interaction between Repetition, Calibration, and Instruction condition.

Overall Analyses

As in Experiments 1 and 2, the last three trials for each subject, in each of Phases 1 through 4, were omitted from further analyses. Only data from trials in which participants took less than 30 seconds to make their judgments were included in the analyses. Across all participants, one trial was eliminated for this reason. Furthermore, only data from trials in which correct relatedness judgments were made were included in the analyses. This resulted in the elimination of 0.4% of trials from further analyses.

Analyses of Instruction Condition

In order to assess whether participants believed that repetition is helpful for memorability in general, (Manipulation Check - Question 1), and whether participants believed that repetition was helpful for their own performance (Manipulation Check - Question 2), these two questions were submitted to an independent samples t-test with Instruction (Repetition-Benefit vs. Repetition-No Benefit) as the between-participants factor. Not surprisingly, for Question 1, participants in the Repetition-Benefit condition reported significantly higher ratings for the benefit of seeing a word multiple times for recall in general, $t_{(46)} = 3.46$, $SE = 0.44$, $p < .001$ (one-tailed). As predicted, participants in both conditions thought that repetition aided memorability, with mean rated benefit of repetition greater than 5 in both cases. However, participants in the Repetition-Benefit condition reported a mean rating of 8.19, $SD = 1.4$, whereas participants in the Repetition No-Benefit condition reported an average rating of 6.67, $SD = 1.6$. It was expected that ratings for both conditions would be on the 'positive' side of the scale since the recall task provided some feedback to participants concerning the increasing number of words recalled after each study-recall phase.

For Question 2, participants in the Repetition-Benefit condition were also more likely to report that seeing a word multiple times made it easier for them to remember in the current experiment, compared to the Repetition-No Benefit condition, $t_{(46)} = 2.23$, $SE = 0.40$, $p < .05$ (one-tailed). As was found for Question 1, for Question 2 participants in both conditions believed that repetition was helpful for their own performance with participants in the Repetition-Benefit condition reporting a mean rating of 7.85, $SD = 1.2$ and participants in the Repetition-No Benefit condition reported an average rating of 6.95, $SD = 1.6$.

Thus, these results indicate that the manipulation of Benefit vs. No Benefit was successful, in that participants in the Benefit condition reported significantly higher ratings for the benefit of repetition on recall in both the general sense, (i.e., what is true for most people), and also for their own performance in this experiment.

Analyses of Recall

A 4 X 2 X 2 mixed-design Analysis of Variance (ANOVA) was computed based on the proportion of words recalled for each participant, treating Repetition (Phases 1 to 4) and Encoding condition (Word-Alone vs. Relatedness) as within-participant factors and Instruction (Repetition Benefit vs. No-Benefit) as the between-participants factor. As shown in Table 7, a significant linear increase was found in the number of words recalled with each successive phase, $F_{\text{linear }(1,46)} = 345.6$, $MSE = 1.6$, $p < .001$. As expected, there was no difference in recall for the two Instruction conditions, $F < 1$. The mean percentage of words recalled across Instruction conditions for Phases 1 to 4 were 26.1%, 37.5%, 48.7%, and 57.8%, respectively.

Table 7: Experiment 3: Mean Recall for Phases 1-4 as a Function of Instruction and Encoding Conditions ($N = 48$).

Recall (%)

		Phase			
Instruction	Encoding	1	2	3	4
Benefit	Word-Alone	30.0 (2.0)	36.0 (2.5)	46.1 (3.6)	54.4 (3.4)
	Relatedness	24.8 (2.3)	37.3 (2.5)	51.5 (3.0)	62.1 (2.9)
No Benefit	Word-Alone	25.6 (2.3)	34.5 (2.8)	44.9 (4.1)	54.8 (3.8)
	Relatedness	24.0 (2.6)	42.0 (2.9)	52.4 (3.4)	59.7 (3.3)

Note: Mean standard error is in parentheses.

As in Experiments 2 and 3, a main effect for Encoding condition was also expected, in that participants were expected to remember a higher proportion of words in the Relatedness condition than in the Word-Alone condition. As was found in Experiment 1, participants recalled more words in the Relatedness condition than in the Word-Alone condition (see Table 7; $F_{(1,46)} = 8.91$, $MSE = 2.6$, $p < .01$).

Analysis of JOLs

A 4 X 2 X 2 mixed-design ANOVA was computed based on mean JOLs for each participant, treating Repetition and Encoding condition as within-participant factors and Instruction as the between-participants factor. As reported in Table 8, a significant linear increase was found for JOLs such that across Instruction conditions, participants' judgments of future recall for words increased as a function of the number of times they saw the words, $F_{\text{linear }(1,46)} = 12.6$, $MSE = 2.9$, $p < .01$. A significant main effect was also found for the Instruction condition such that participants in the Benefit condition reported significantly higher JOLs overall than did participants in the No-Benefit condition, $F_{(1,46)} = 6.86$, $p < .05$. The mean JOLs for Phases 1 to 4 for the Repetition-Benefit condition were 51.9%, 55.7%, 55.6%, and 60.2%, respectively. Mean JOLs for Phases 1 to 4 for the No-Benefit condition were 40.5%, 41.8%, 43.7%, and 50.3%, respectively. As in Experiments 1 and 2, to further analyze the effect of Encoding condition the analysis was based only on data from Phases 2 to 4. There was a main effect for Encoding condition such that participants' JOLs were higher for words in the Relatedness condition than in the Word-Alone condition ($F_{(1,46)} = 4.46$, $MSE = .74$, $p < .05$). This effect of Relatedness condition was the same for both Instruction conditions (Encoding condition X Instruction interaction, $F < 1$).

Table 8: Experiment 3: Mean JOLs for Phases 1-4 as a Function of Instruction and Encoding Conditions ($N = 48$).

JOLs (%)

		Phase			
Instruction	Encoding	1	2	3	4
Benefit	Word-Alone	52.5 (3.3)	54.8 (3.4)	55.2 (3.5)	59.8 (3.6)
	Relatedness	51.4 (3.2)	56.6 (3.4)	56.0 (3.7)	60.7 (3.9)
No Benefit	Word-Alone	40.9 (3.8)	39.3 (3.9)	42.3 (4.0)	49.4 (4.1)
	Relatedness	40.1 (3.6)	44.2 (3.9)	45.0 (4.2)	51.1 (4.4)

Note: Mean standard error is in parentheses.

Analysis of Calibration

To analyze the relationship between mean JOLs and actual recall for each Instruction condition, a 2 X (4 X 2 X 2) mixed-design ANOVA was computed, treating Repetition, Encoding condition, and Calibration (Proportion of Words Recalled - Mean JOLs) as within-participant factors and Instruction (Repetition-Benefit vs. Repetition-No Benefit) as the between-participants factor. First, a significant interaction was found between Calibration and Repetition, $F_{\text{linear}\,(1,46)} = 103.9$, $MSE = 1.4$, $p < .001$. Although this finding is consistent with Experiments 1 and 2, contrary to my hypothesis, this effect did not interact with Instruction condition, $F < 1$. However, there was a significant interaction between Calibration and Instruction condition, $F_{(1,46)} = 5.74, p < .05$, resulting from significantly higher confidence in estimates of future recall across all study phases for the Repetition-Benefit condition. Table 9 reports mean recall and JOLs for Phases 1-4 for the Benefit condition, collapsing across Encoding condition. Note the bottom row; participants in this condition were highly overconfident in Phases 1-3, and exhibited overconfidence for all Phases.

Table 9: Experiment 3: Mean Recall and JOLs in Benefit Condition for Phases 1-4 Across Encoding Condition ($N = 27$).

		Phase			
	Encoding	1	2	3	4
Recall (%)	Mean	27.4 (1.7)	36.7 (2.1)	48.8 (2.5)	58.3 (2.8)
JOLs (%)	Mean	51.9 (3.2)	55.7 (3.1)	55.6 (3.5)	60.2 (3.7)
	JOLs - Recall	24.5	19.0	6.8	1.9

Note: Mean standard error is in parentheses.

In contrast, participants in the No-Benefit condition exhibited the UWP effect. As shown in Table 10 (bottom row), participants in this condition were overconfident in Phases 1 and 2, and shifted toward underconfidence in Phases 3 and 4.

Table 10: Experiment 3: Mean Recall and JOLs in No-Benefit Condition for Phases 1-4 Across Encoding Condition ($N = 21$).

		Phase			
	Encoding	1	2	3	4
Recall (%)	Mean	24.8 (1.9)	38.3 (2.4)	48.7 (2.8)	57.3 (3.2)
JOLs (%)	Mean	40.5 (3.6)	41.8 (3.5)	43.7 (4.0)	50.3 (4.2)
	JOLs - Recall	15.7	3.5	- 5.0	- 7.0

Note: Mean standard error is in parentheses.

In order to clarify the change in calibration across the course of the experiment, mean JOLs and mean percentage of words recalled for each Instruction condition was further subjected to a one-way ANOVA for each Encoding condition and for each Phase. For Phase 1, Proportion of Words Recalled - Mean JOLs was significant for both the Repetition-Benefit condition, $F_{(1,26)} = 48.18$, *MSE* = 3.3, $p < .001$, as well as the Repetition-No Benefit condition, $F_{(1,20)} = 15.14$, *MSE* = 3.3, $p < .01$. Participants in both conditions demonstrated overconfidence with mean JOLs 24.5% higher than recall in the Benefit condition and 15.7% higher in the No-Benefit condition. Mean recall and JOLs across Phases 1-4 for Benefit condition are illustrated in Figure 8; note the very shallow slope of JOLs across the Phases. JOLs in this condition were high in comparison with the JOLs made by the No-Benefit group, and remained that way across all Phases. Figure 9 shows mean recall and JOLs across all Phases for the No-Benefit condition. Notice that Figure 9 closely resembles that of the 'typical' UWP effect as illustrated in Figure 1 in the Introduction. In Phase 1, JOLs in the Benefit condition were 11.4% higher than JOLs in the No-Benefit condition. This is interesting because it appears that explicitly telling participants about the benefit of repetition on recall actually increases overall confidence,

Figure 8. Experiment 3: Mean Recall and JOLs for Benefit Condition as a Function of Study-Recall Phase.

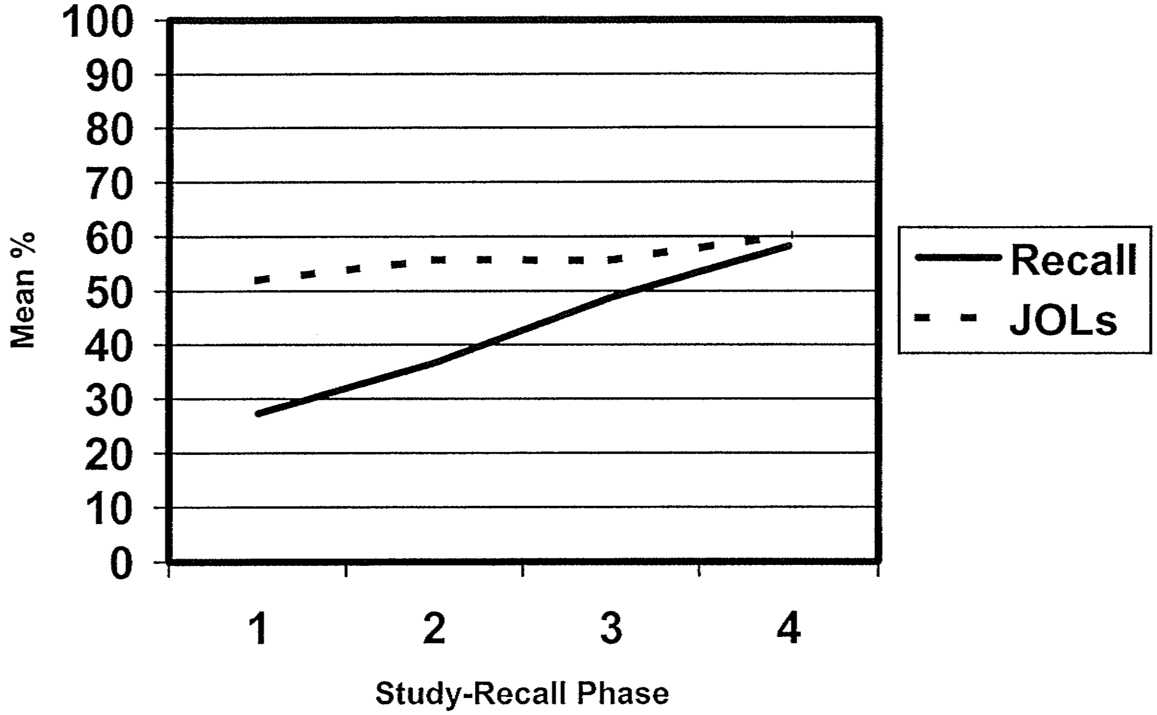
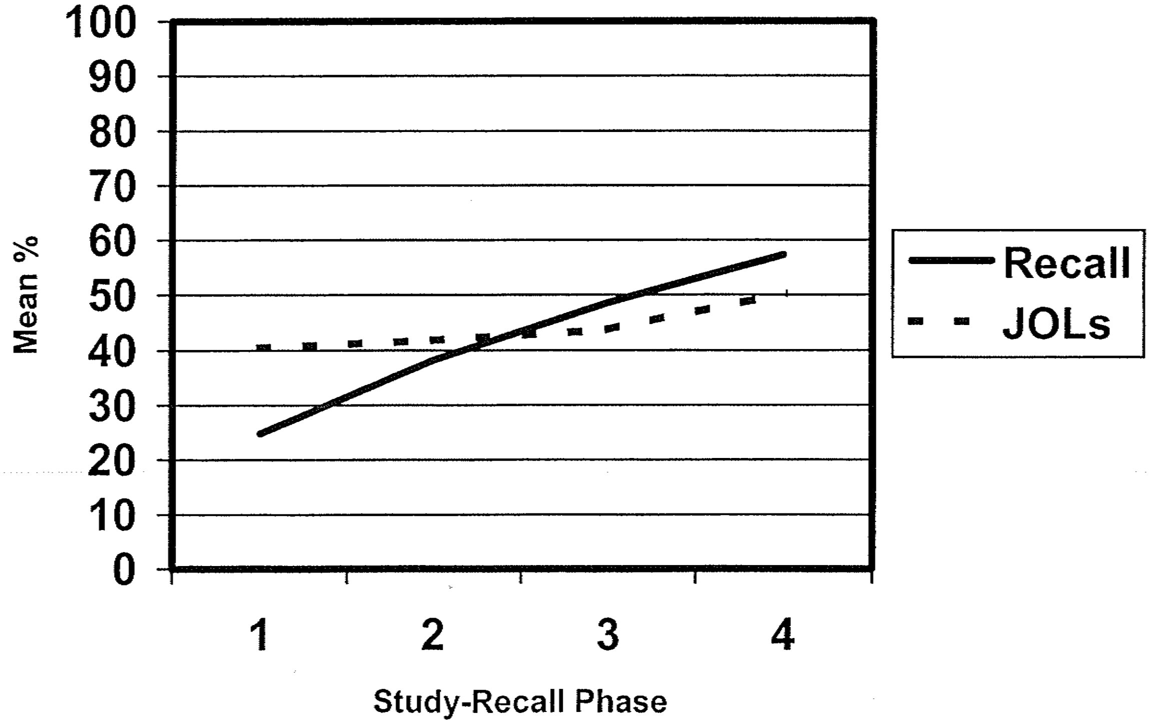


Figure 9. Experiment 3: Mean Recall and JOLs for No-Benefit Condition as a Function of Study-Recall Phase.



even before any repetition occurred in the experiment! This point will be discussed in further detail later.

In Phase 2, participants in the Benefit condition demonstrated overconfidence in both the Word-Alone condition, $F_{(1,26)} = 24.88$, $MSE = 1.8$, $p < .001$, and Relatedness condition, $F_{(1,26)} = 32.57$, $MSE = 1.5$, $p < .001$, mean overconfidence = 18.8% and 19.3% respectively. In the No-Benefit condition, participants showed a nonsignificant trend towards underconfidence in the Word-Alone condition with mean JOLs 4.8% higher than actual recall, $F < 1$. No significant difference was found between JOLs and recall in the Relatedness condition, $F < 1$.

In Phase 3, participants in the Benefit condition continued to show slight overconfidence in the Word-Alone condition, $F_{(1,26)} = 3.77$, $MSE = 2.7$, $p = .063$, and a nonsignificant trend towards overconfidence in the Relatedness condition, $F < 1$, (mean overconfidence 9.1% and 4.5%, respectively). In the No-Benefit condition, however, despite nonsignificant differences between mean JOLs and recall for both Word-Alone, F < 1, and Relatedness conditions, $F_{(1,20)} = 2.44$, $MSE = 2.7$, $p = .13$, the means for this group suggest a trend towards underconfidence with JOLs 2.6% below recall in the Word-Alone condition and 7.4% below recall in the Relatedness condition.

By Phase 4, there were no significant differences between JOLs and recall in the Benefit condition for both Word-Alone and Relatedness conditions, $F < 1$. However, a closer look at the means shows a tendency towards underconfidence in the Relatedness condition with mean JOLs 1.4% below that of recall. In the No-Benefit condition, a marginally significant difference between JOLs and recall was found for the Word-Alone condition, $F_{(1,20)} = 2.48$, $p = .13$, in the direction of underconfidence (JOL – recall = -

4.6%), and significant underconfidence for the Relatedness condition, $F_{(1,20)} = 5.8$, $MSE = 1.5$, $p < .05$, (JOL – recall = - 8.6%).

Overall, the above pattern of results for the No-Benefit condition is similar to that of Experiments 1 and 2. That is, participants in the No-Benefit condition demonstrated a pattern of overconfidence in Phase 1 that gradually shifted towards less confidence by Phase 4. This finding suggests that participants who are not told anything about the effect of repetition on recall may actually hold the *same* expectations and beliefs about their future performance as participants who *are* explicitly told that repetition does not aid recall. Given that the mean rating for the Manipulation Check Question 1 was greater than 5, indicating that participants who were told that repetition does not aid recall did in fact report that they believed it was at least of some benefit[2], this suggests that perhaps participants are somewhat sensitive to the benefit of repetition on recall when explicitly asked, but that they are either not sensitive to this benefit during the actual experiment or else their JOLs are not accurately reflecting this belief.

Participants in the Benefit condition, on the other hand, demonstrated strong overconfidence even during the first study-recall phase. This overconfidence continued until Phase 3, although the means demonstrate slight overconfidence for the Word-Alone condition even in Phase 4. This is an interesting observation because it suggests that explicitly telling participants that repetition aids recall serves to increase *overall* confidence, rather than improve calibration. If participants were sensitive to the benefit of repetition on recall, there should be no difference between Benefit and No-Benefit conditions at Phase 1, but better calibration in Phases 2 to 4 in the Benefit condition, as a

[2] Although participants in the Repetition-No Benefit condition were significantly less impressed with this benefit than participants in the Repetition-Benefit condition, as indicated by their lower ratings.

result of increasingly higher JOLs. Instead, the slope of increase in participants JOLs is virtually the same for the Benefit and No-Benefit conditions. In other words, there is no evidence that participants in the Benefit condition used the instructions provided to improve calibration between their JOLs and the true benefit of repetition for recall. This being so, it is interesting to note that, combining data for both Encoding conditions, participants in the Benefit condition shifted from massive overconfidence in Phase 1 (mean JOLs – mean recall = 24.5 %) to much more accurate JOLs by Phase 4 (1.9%). This shift toward greater accuracy is misleading, arising primarily as a result of successive increases in recall across phases, with relatively little change in JOLs as a function of repetition. As shown in Figure 8, across Encoding conditions, mean JOLs for the Benefit condition increased from 51.9% to 60.2% in Phase 4, while the percentage of words recalled increased from 27.4% in Phase 1 to 58.3% in Phase 4.

To summarize, the results of Experiment 3 reveal several interesting findings. First, the pattern of results for participants in the No-Benefit condition is similar to that of the previous two Experiments in which participants were told nothing about the relationship between repetition and recall performance. Either people hold an implicit belief that repetition is not beneficial for future remembering (which is unlikely given the outcome of the Manipulation check), or else perhaps people do not effectively incorporate this belief into their subjective judgments. Secondly, telling participants that repetition aids future recall increases overall confidence, (i.e., higher JOLs), but does not improve actual JOL accuracy.

General Discussion

Memory monitoring is an important factor in acquiring new skills and learning new information. It involves the capacity to regulate the effectiveness of study strategies and make judgments regarding how well the new information has been processed and acquired for future use. For example, students must decide what material they have already learned, and what material they need to spend their time on to maximize their performance on exams (Maki, 1998).

When learning new information, there is evidence that people may use effective study strategies when acquiring information for the first time. For example, they may devote study time to items that are judged to be learnable, thereby diverting resources away from items judged to be very well learned or very difficult to learn (Mazzoni et al., 1990). This is especially true when the amount of study time provided is insufficient to learn all the material (Son & Metcalfe, 2000). For example, Son and Metcalfe (2000) found that people will allocate more study time to items judged easy to learn, thereby maximizing their performance. In addition, Son and Metcalfe also reported that when the study list is shorter, increasing the amount of time participants can spend studying each item, people will spend more time learning difficult items. Thus, when people are presented with new information to learn, they tend to use fairly effective strategies to maximize their performance. Despite their often intelligent use of learning resources, however, Koriat et al. (2002) identified a major source of error in memory monitoring in that people tend to become underconfident when faced with multiple study sessions with the same material. In some ways, this situation might be more representative of actual study experiences. That is, when people are presented with a situation involving the learning of new information, they often look at the material more than once (i.e., multiple study sessions). Investigating why this UWP effect occurs and, as a consequence, discovering possible strategies for reducing its occurrence was the primary motivation for the present Experiments.

The results of Experiments 1 and 2 suggest that although distinctiveness of learning and effortful processing did improve recall performance, and to a lesser extent influenced JOLs, overall participants in these Experiments exhibited a tendency to overestimate their likelihood of future recall in the first two study-recall phases and become less confident in their assessments of learning with repeated exposure to study items. In Experiments 1 and 2, participants were shown a list of words, presented one at a time on a computer screen. Immediately following each word, participants were prompted to make their JOL. After the word list was shown, participants were given a free-recall task. This procedure was repeated three more times, for a total of four study-recall cycles. In Experiment 1, participants were required to make a relatedness judgment for half of the words in Phases 2-4, thereby creating a distinctive learning environment. In Experiment 2, participants had to first decode the target word and then make a relatedness judgment in Phases 2-4. Thus, studying words in the Relatedness condition involved effortful processing. In both conditions, participants were overconfident in Phases 1-2, shifting towards less confidence by Phase 4. Interestingly, participants were more underconfident for words involving distinctive learning or effortful encoding, due to the advantage of these factors on recall performance. In fact, the UWP effect was only consistently observed in the Relatedness conditions. This is surprising, given that Koriat et al. (2002) found a UWP effect by the third study-recall phase using a single-word list. One reason for why a significant UWP effect was not found in every Word-Alone condition may be due to poorer recall of words in that condition due to the presence of better-remembered words in the Relatedness condition. That is, stimulus words that were presented alone may be susceptible to higher rates of forgetting due to interference from deeper encoded items. In contrast, Koriat et al. presented all of the stimulus words alone. Relative to the experiment reported by Koriat, if memory for words presented alone was worse in Phases 2-4 of the Experiments reported here, then a UWP effect would be more difficult to observe.

Why are participants not sensitive to the benefit of repetition even when learning conditions are distinctive or effortful? Koriat (1997; Koriat et al., 2002) suggested that people discount extrinsic cues when making JOLs. Extrinsic cues refer to aspects of the study conditions, and encoding processes engaged in during learning. In Experiments 1 and 2, extrinsic cues available to the participants were number of study phases, distinctiveness of learning, and effortful encoding. The hypothesis was that if the UWP effect occurs because people hold intuitive beliefs that they are not learning anything new when looking at a word multiple times, then this phenomenon may be eliminated in conditions under which learning is distinct and/or effortful. Although participants in these Experiments did in fact have higher JOLs for words in which relatedness judgments were made, JOLs were less influenced by this manipulation than actual recall was, resulting in exacerbation of the UWP effect for this condition, rather than elimination. While contrary to my hypothesis, these findings are consistent with Koriat's cue-utilization view; participants in this study discounted the extrinsic cues of learning distinctiveness and effortful encoding.

Thus, one possibility is that people do not fully appreciate that these factors do improve performance. Dunlosky and Matvey (2001) argue that cues will influence JOLs only to the extent that people believe they are indicative of their performance. These authors suggest that the extrinsic-intrinsic distinction proposed by Koriat (1997) may not be a useful theoretical distinction in terms of understanding cues people utilize in making their JOLs. Based on his analysis of previous JOL studies, Koriat had concluded that although extrinsic cues may be useful diagnostic indicators for making metacognitive judgments, people often discounted extrinsic cues, instead relying on intrinsic factors. In contrast, Dunlosky and Matvey (2001) found that study conditions, an extrinsic cue according to Koriat's theory, did influence JOLs whereas the relatedness of paired-associates, an intrinsic cue, was sometimes discounted. Specifically, these researchers found that serial position and order effects influenced JOLs, such that degree of relatedness between stimulus-response pairs was less influential on JOLs than the order in which the items were presented. In their Experiment 2, for example, recall was greater for related than for unrelated items presented in the first block, and yet the JOLs discounted this when related items were presented first. Nevertheless, the findings from Experiments 1 and 2 in this thesis suggest that people do not take into account the benefit of learning distinctiveness or effortfulness in their calculations of JOLs. As a result, a significant UWP effect emerged for the Relatedness condition in Experiment 1 and a marginal UWP effect occurred in Experiment 2.

Alternatively, perhaps poor calibration between recall performance and JOLs is not due to an inherent flaw in one's intuition or a lack of a priori beliefs concerning the benefit of repetition for recall performance, but instead is a product of people's inability to use their knowledge at the time of making JOLs. That is, perhaps people do at some level fully appreciate that repetition, learning distinctiveness, and effortful encoding are beneficial for remembering, but these inferences are not accurately represented in the probability judgments people make. For example, people may 'know' that repetition is beneficial, but 'forget' to include it in their subjective calculations. If this is true, then information explicitly provided to participants at the time of making metacognitive judgments may help compensate for this poor reasoning ability. In Experiment 3, participants were explicitly told either that repetition improves recall, or that repetition has no effect on recall. Results of this experiment indicate that telling participants that repetition aids recall served to increase overall confidence, in that JOLs in the Repetition Benefit condition were higher than those in the Repetition No-Benefit condition even during the first study phase.

Why then did telling participants that repetition is beneficial for recall increase JOLs equally across all phases instead of causing participants to report increasingly high JOLs from Phase 1 to Phase 4? One possibility is that, instead of using the information given to them as a basis for making their subjective judgments, participants used it as a social cue. de Carvalho Filho and Yuzawa (2001) found that social cues given during the judgment phase can influence JOLs. In their study, participants made JOLs for both easy stimulus-response pairs that were considered a priori to have a high degree of relatedness (e.g., MAGAZINE-NEWSPAPER), and difficult stimulus-response pairs that were unrelated (e.g., ANIMAL-CLOCK). During the JOL phase, participants were exposed to information concerning the fictitious performance of previous 'participants'. In the High Cue condition, participants were given information on the bottom of their computer screen indicating that previous college students recalled a mean of 87% of the easy paired-associates and 77% of the difficult pairs correctly. In the Low Cue condition, participants were told that college students recalled a mean of only 57% of easy paired-associates and 47% of difficult pairs correctly. A control group performed the tasks with no additional information given at the time of making JOLs. After making their judgments, participants were given a recall test followed by a metacognitive assessment. This assessment consisted of four tasks designed to assess overall metacognitive ability including prediction accuracy, as well as strategy selection and production. de Carvalho Filho and Yuzawa found that participants with low metacognition scores who were in the High Cue condition had higher overall JOLs than participants with low metacognitive ability who were in the control group. The lowest JOLs were given by participants with low metacognition scores who were in the Low Cue condition.

These findings indicate that participants who are relatively poor in metacognitive ability are easily influenced by social cues when making JOLs. Since overall recall in this study was high, ranging from 51-55% for difficult word pairs to 89-92% for easy pairs, participants with low metacognitive scores who were in the High Cue condition actually had the highest relative accuracy, due to the influence of increasing their JOLs to more closely approximate that of actual recall. In this thesis, the results of Experiment 3 show that in Phase 1, the JOLs are closer to actual recall for the No-Benefit condition, as a result of the high degree of overconfidence exhibited by the Benefit condition. In contrast, by Phase 4 JOLs were closer to actual recall for the Benefit condition, because high recall by this phase was closely matched by the (already high) JOLs. It is important to note that this result should not be interpreted as an accuracy advantage of the No-

Benefit condition, but as an artifact of the extreme JOLs made by participants in the Benefit condition. Consequently, what appears to be an elimination of the UWP effect by Phase 4 in the Benefit condition is a product of improved recall performance, not accurate JOLs. It is likely that if participants completed additional study-recall phases (i.e., a total of 6 phases rather than 4), the UWP effect would emerge for the Benefit condition as well.

A possible explanation for the overall high degree of confidence demonstrated by the Repetition Benefit group in Experiment 3 is that these participants were using the information given to them regarding the benefit of repetition as a social cue, leading them to make their JOLs in an atmosphere of heightened confidence. The outcome was that JOLs for the Benefit condition were higher in general, and were relatively uninfluenced by the content of the instructions that clearly emphasized the contribution of "repetition" to enhancing recall. In other words, explicitly telling participants that repetition is beneficial may have served as a general cue that recall should be high after studying a list of target words, not just high after repeatedly seeing the targets.

Another possibility for why JOLs were higher overall for the Repetition Benefit condition relates back to the idea that people are poor at reasoning with probabilities. In addition to the influence of social cues on reasoning judgments, there is evidence that people's JOLs are also influenced by the specific instructions given to them when making metacognitive judgments. Very recently, Liberman (2004) found that participants made very different confidence judgments depending on the instructions they were given at the time of making their JOLs. In this study, underconfidence in global assessments of confidence was due to people's failure to account for correct guessing. In a series of experiments, participants were given random pairings of company names and were asked to circle the name of the business that had the highest sales. After each pair, participants were then asked how confident they were that they had chosen the correct answer. After all 40 pairs were presented, participants made global JOLs concerning the overall percentage of questions they think they answered correctly. In the Unrestricted condition, participants were not given any additional instructions regarding their global JOL. In the Restricted condition, participants were further told that their estimate of correct responses should be greater than 50%, because random guessing should result in 50% accuracy. In the Reminder condition, participants were informed that if they answered randomly, about 50% of the responses should be correct. The key difference between the latter two conditions was that in the Restricted group participants were explicitly told that they should not give a response lower than 50%, whereas in the Reminder condition participants were only 'reminded' that random guessing would result in about 50% correct responses, (i.e., no lower response boundary was set). Although no differences in actual performance was found across conditions, participants in the Restricted condition had higher confidence than participants in both the Unrestricted and Reminder conditions. Explicitly telling participants to use ratings greater than 50%, and providing the logic why they should do so, increased confidence in this study. Interestingly, participants in the Unrestricted condition, who were not given any explicit instructions regarding how they should make their judgments, reported a global confidence judgment of only 54.4% in one experiment, despite the fact that guessing alone should result in approximately 50% accuracy! Also, although accuracy was not the focus of this study, the mean performance and global confidence assessments across the conditions show that participants in the Unrestricted condition, who were not given any instructions when making their global confidence judgments, demonstrated underconfidence, whereas participants in the Restricted condition were overconfident. Results for the Reminder condition were mixed, in that participants in this condition were underconfident in one experiment but overconfident in another. Thus, explicitly giving participants logical advice concerning how they should make their judgments actually influenced their JOLs. It is worth noting that it is likely that participants already 'knew' that random guessing should result in approximately 50% accuracy and that in accounting for this, their judgments should exceed 50%. Explicitly telling participants information that supports their intuitive metacognitive beliefs served to reinforce those beliefs, resulting in higher confidence ratings. Furthermore, it is also likely that information consistent with a priori beliefs are more likely to influence JOLs than information that is not consistent. In Experiment 3, participants in the Repetition Benefit condition were given information that corresponded with the belief that the more times you see a word, the more likely you are to recall that word later. Since it is unlikely that people hold equally strong beliefs that seeing a word multiple times does not enhance recall, JOLs in the Repetition No-Benefit condition were not as influenced by the information given to them. This is further evidenced by the similarity in JOLs between participants in the No-Benefit condition in Experiment 3 and participants in Experiments 1 and 2 who were given no information.

Other factors may play a role in accuracy of judgments as well. Pallier, Wilkinson, Danthiir, Kleitman, Knezevic, Stankov, & Roberts (2002) identified a "confidence trait" that is weakly related to cognitive ability but is a significant determinant of confidence accuracy. This is a particularly interesting study in that they also found that some personality traits related to proactiveness and activity also correlated with this trait. Their main conclusion was that individual differences among participants, including personality traits as well as overall cognitive and metacognitive abilities, are important considerations for understanding the underlying mechanisms involved in miscalibration. Although beyond the scope of the present Experiments, it would be interesting and perhaps fruitful for future calibration researchers to include some of these variables in their analyses.

The above discussion highlights the potential influence that experimental instructions can have on participants' subjective reasoning. What information, then, do participants need to improve their absolute accuracy? One possible solution may be to give participants' feedback that will be a useful cue in making their JOLs. For example, there is some evidence that providing performance feedback can improve overall calibration accuracy, and that this benefit is transferable across different tasks (e.g., Lichtenstein et al., 1982). In contrast, Koriat (1997) did not find that feedback improved calibration across study-recall cycles. However, participants in Koriat's study were given feedback as to the correctness of their response after each item, not after each recall phase. Perhaps participants in this study were unable to use this cue effectively because item-by-item feedback is not as predictive of future recall as feedback concerning overall performance. One way to investigate this possibility would be to give global feedback to participants after each recall. That is, immediately following recall participants would be told what percentage of items were recalled correctly. Variations of this experimental design could include conditions in which participants are given global feedback only, or in conjunction with item-by-item feedback. Alternatively, in the second and subsequent JOL phases participants could be given information about the previous JOL they made for that item at the time they are making their current JOL; during this stage participants could also be given feedback as to whether or not they correctly recalled this item in the previous recall phase. I expect that calibration accuracy would improve under these circumstances, as participants' JOLs should be more diagnostic of future recall when previous performance is used as a cue.

Implications and Future Research: The UWP Enigma

Investigations into memory monitoring in general, and judgments of learning in particular, have important developmental and educational implications. For example, Plude et al. (1998) approach memory monitoring from a developmental perspective. They emphasize the importance of studying memory-monitoring processes not only in young adults but also in children and in the elderly. There is some evidence, for instance, that children's' JOLs (see Koriat & Shitzer-Reichert, 2002) and older adults' JOLs (see Hertzog & Hultsch, 2000) operate in ways similar to young adults, but there has been relatively little research conducted with these groups. Future research on the UWP effect could focus on whether or not this error in memory monitoring occurs in children and older adults.

As well, Cavanaugh and Morton (1988) highlight the importance of memory self-efficacy in older adults and its relationship with memory ability. Future research on JOLs in older adults could examine the relationship between memory self-efficacy and JOLs in the elderly. For example, lower memory self-efficacy in older adults may produce lower confidence in JOLs generally, even without multiple exposures to the same material. However, given that older adults generally have poorer recall than younger adults (e.g., Hultsch, Hertzog, Dixon, & Small, 1998), it would be interesting to explore whether older adults' JOLs are more predictive of actual recall with repeated study-recall cycles, or if the UWP effect would be found in this population as well.

A number of interesting research questions arise from the current investigation into the UWP effect. For example, would the UWP effect emerge for study-test trials that occurred over an extended period of time? Carroll et al., (1997) reported that JOLs were not very accurate at predicting future recall when the retention period was over several weeks. Specifically, Carroll et al. found that participants were overconfident in their predictions for recall after a long delay such that the JOLs made in anticipation of a six week delay prior to recall were the same as those made in anticipation of only a two week delay. Not surprisingly, but not consistent with equivalent JOLs for the two groups, recall was better after two weeks than after six weeks. It would be interesting to investigate whether this relative overconfidence for the long delay group would shift toward underconfidence with repeated study-recall sessions with retention intervals extending over six weeks or more.

Another research question posed by the current investigation concerns the relationship between JOLs and other memory-monitoring processes. For example, Leonesio and Nelson (1990) reported that other memory-monitoring measures such as ease-of-learning (EOL), feelings-of-knowing (FOK), and JOLs are only weakly correlated with each other. EOL judgments are made prior to acquisition, and involve inferential, a priori assessments concerning the ease or difficulty in learning the items (see Nelson & Narens, 1990). FOK judgments are also metacognitive judgments, made during either the acquisition or retrieval phases of the learning process. FOK judgments involve subjective assessments about whether a currently not retrieved item will likely be recalled later (see Nelson, 1988; Nelson & Narens, 1990). What, if any, underlying mechanisms are shared by these measures and what are unique? Would a UWP effect emerge for these memory measures as well? More recently, Dunlosky, Kubat-Silman, and Hertzog (2003) identified another monitoring measure: quality-of-encoding (QUE). QUEs are subjective judgments concerning how well an item has been encoded. Their evidence suggests some age-related impairment in memory-monitoring effectiveness based on this measure. Might there be a UWP effect for QUEs? Would this effect be particularly strong in older adults?

Currently, there is very little known about the UWP effect and why it occurs. Koriat et al. (2002) speculate that the UWP effect emerges because people rely on intrinsic cues when making JOLs and discount the benefit of extrinsic cues. Experiments 1-3 in this thesis provide a contribution toward a greater understanding of the conditions under which the UWP effect occurs.

References

Allen, S. W., & Jacoby, L. L. (1990). Reinstating study context produces unconscious influences of memory. *Memory & Cognition, 18,* 270-278.

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

Benjamin, A. S., & Bjork, R. A. (1996). Retrieval fluency as a metacognitive index. In Reder, L. M. (Ed). *Implicit Memory & Metacognition.* Mahwah, NJ: Lawrence Erlbaum Assoc.

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General, 127,* 55-68.

Bjork, R. A. (1999). Assessing our own competence: Heuristics and illusions. In D. Gopher, & A. Koriat (Eds.), *Attention and performance XVII: Cognitive regulation of performance: Interaction of theory and application* (pp. 435-459). Cambridge, MA: The MIT Press.

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *Essays in honor of William K. Estes. Vol. 1: From learning theory to connectionist theory. Vol. 2: From learning processes to cognitive processes* (pp. 35-67). Hillsdale, England: Lawrence Erlbaum Assoc.

Carroll, M., & Korukina, S. (1999). The effect of text coherence and modality on metamemory judgments. *Memory, 7,* 309-322.

Carroll, M., Nelson, T. O., & Kirwan, A. (1997). Tradeoff of semantic relatedness and degree of overlearning: Differential effects on metamemory and on long-term retention. *Acta Psychologica, 95,* 239-253.

Cavanaugh, J. C., & Morton, K. R. (1988). Older adults' attributions about everyday memory. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory: Current research & issues, vol. 1. : Memory in everyday life* (pp. 209-214). Oxford, England: John Wiley & Sons.

Craik, F. I. & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning & Verbal Behavior, 11,* 671-684.

de Carvalho Filho, M. K., & Yuzawa, M. (2001). The effect of social influences and general metacognitive knowledge on metamemory judgments. *Contemporary Educational Psychology, 26,* 571-587.

Dunlosky, J., Kubat-Silman, A. K., & Hertzog, C. (2003). Effects of aging on the magnitude and accuracy of quality-of-encoding judgments. *American Journal of Psychology, 116,* 431-454.

Dunlosky, J., & Matvey, G. (2001). Empirical analysis of the intrinsic-extrinsic distinction of judgments of learning (JOLs): Effects of relatedness and serial position on JOLs. *Journal of Experimental Psychology: Learning, Memory & Cognition, 27,* 1180-1191.

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

Dunlosky, J., & Nelson, T. O. (1994). Does the sensitivity of judgments of learning (JOLs) to the effects of various study activities depend on when the JOLs occur? *Journal of Memory & Language, 33,* 545-565.

Dunlosky, J., & Nelson, T. O. (1997). Similarity between the cue for judgments of learning (JOL) and the cue for test is not the primary determinant of JOL accuracy. *Journal of Memory & Language, 36,* 34-49.

Hertzog, C., Dunlosky, J., Robinson, A. E., & Kidder, D. P. (2003). Encoding fluency is a cue used for judgments about learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 29,* 22-34.

Hertzog, C., & Hultsch, D. F. (2000). Metacognition in adulthood and old age. In F. I. M. Craik, & T. A. Salthouse (Eds.), *The handbook of aging & cognition, 2nd Ed.* (pp. 417-466). Mahwah, NJ: Lawrence Erlbaum Assoc.

Hultsch, D. F., Hertzog, C., Dixon, R. A., & Small, B. J. (1998). *Memory change in the aged.* New York: Cambridge University Press.

Jacoby, L. L. & Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General, 110,* 306-340.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory & Language, 30,* 513-541.

Jacoby, L. L., & Hollingshead, A. (1990). Toward a generate/recognize model of performance on direct and indirect tests of memory. *Journal of Memory & Language, 29,* 433-454.

Jacoby, L. L., Kelley, C. M., & Dywan, J. (1989). Memory attributions. In H. L. Roediger & F. I. M. Craik (Eds.), *Varieties of memory and consciousness: Essays in honour of Endel Tulving* (pp. 391-442). Hillsdale, NJ: Lawrence Erlbaum Assoc.

Jacoby, L. L., & Whitehouse, K. (1989). An illusion of memory: False recognition influenced by unconscious perception. *Journal of Experimental Psychology: General, 118,* 126-135.

Johnston, W. A., Hawley, K. J., & Farnham, J. M. (1993). Novel popout: Empirical boundaries and tentative theory. *Journal of Experimental Psychology: Human Perception and Performance, 19,* 140-153.

Kelemen, W. L. (2000). Metamemory cues and monitoring accuracy: Judging what you know and what you will know. *Journal of Educational Psychology, 92,* 800-810.

Kelemen, W. L., & Weaver, C. A. (1997). Enhanced metamemory at delays: Why do judgments of learning improve over time? *Journal of Experimental Psychology: Learning, Memory, & Cognition, 23,* 1394-1409.

Kelley, C. M., & Lindsay, D. S. (1993). Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions. *Journal of Memory & Language, 32,* 1-24.

Kimball, D. R., & Metcalfe, J. (2003). Delaying judgments of learning affects memory, not metamemory. *Memory & Cognition, 31,* 918-929.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126,* 349-370.

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review, 103,* 490-517.

Koriat, A., & Goldsmith, M. (1998). The role of metacognitive processes in the regulation of memory performance. In G. Mazzoni & T. O. Nelson, (Eds.), *Metacognition and cognitive neuropsychology: Monitoring and control processes* (pp. 97-118). Mahwah, NJ: Lawrence Erlbaum Assoc.

Koriat, A., Goldsmith, M., & Pansky, A. (2000). Toward a psychology of memory accuracy. *Annual Review of Psychology, 51,* 481-537.

Koriat, A., & Levy-Sadot, R. (1999). Processes underlying metacognitive judgments: Information-based and experience-based monitoring of one's own knowledge. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 483-502). New York: Guilford Press.

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

Koriat, A., & Shitzer-Reichert, R. (2002). Metacognitive judgments and their accuracy. In P. Chambres, M. Izaute, & P. J. Marescaux (Eds.), *Metacognition: Process function, and use* (pp. 1-17). Dordrecht, Netherlands: Kluwer Academic Publishers.

Leboe, J. P., Leboe, L. C., & Miliken, B. (2003). Another look at the effect of a surprising intervening event on negative priming. *Canadian Journal of Experimental Psychology, 57,* 115-124.

Leboe, J. P. & Whittlesea, B. W. A. (2002). The inferential basis of familiarity and recall: Evidence for a common underlying process. *Journal of Memory and Language, 46,* 804-829.

Leonesio, R. J., & Nelson, T. O. (1990). Do different metamemory judgments tap the same underlying aspects of memory? *Journal of Experimental Psychology, 16,* 464-470.

Liberman, V. (2004). Commentary: Local and global judgments of confidence. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 30,* 2004.

Lichtenstein, S., Fischhoff, B. & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp.306-334). Cambridge, England: Cambridge University Press.

Loftus, E. F. & Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior, 13,* 585-589.

Lovelace, E. A. (1984). Metamemory: Monitoring future recallability during study. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 10,* 756-766.

Maki, R. H. (1998). Test predictions over text material. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 117-144). Mahwah, NJ: Lawrence Erlbaum Assoc.

Matvey, G., Dunlosky, J., & Guttentag, R. (2001). Fluency of retrieval at study affects Judgments of learning (JOLs): An analytic or nonanalytic basis for JOLs? *Memory & Cognition, 29,* 222-233.

Mazzoni, G., & Cornoldi, C. (1993). Strategies in study time allocation: Why is study time sometimes not effective? *Journal of Experimental Psychology: General, 122,* 47-60.

Mazzoni, G., Cornoldi, C., & Marchitelli, G. (1990). Do memorability ratings affect study-time allocation? *Memory & Cognition, 18,* 196-204.

Mazzoni, G., & Nelson, T. O. (1995). Judgments of learning are affected by the kind of encoding in ways that cannot be attributed to the level of recall. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 21,* 1263-1274.

Meeter, M., & Nelson, T. O. (2003). Multiple study trials and judgments of learning. *Acta Psychologica, 113,* 123-132.

Nelson, T. O. (1988). Predictive accuracy of the feeling of knowing across different criterion tasks and across different subject populations and individuals. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory: Current research & Issues, vol. 1. : Memory in everyday life* (pp. 190-196). Oxford: John Wiley & Sons.

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

Nelson, T. O., & Dunlosky, J. (1992). Commentary: How shall we explain the delayed-judgment-of-learning effect? *Psychological Science, 3,* 317-318.

Nelson, T. O., Dunlosky, J., Graf, A., & Narens, L. (1994). Utilization of metacognitive judgments in the allocation of study during multitrial learning. *Psychological Science, 5,* 207-213.

Nelson, T. O., & Leonesio, R. J. (1988). Allocation of self-paced study time and the "Labor-in-Vain effect". *Journal of Experimental Psychology: Learning, Memory, & Cognition, 14,* 676-686.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory: Vol. 26.* San Diego, CA: Academic Press.

Nelson, T. O., Narens, L., & Dunlosky, J. (2004). A revised methodology for research on metamemory: Pre-judgment recall and monitoring (PRAM). *Psychological Methods, 9,* 53-69.

Pallier, G., Wilkinson, R., Danthiir, V., Kleitman, S., Knezevic, G., Stankov, L., & Roberts, R. D. (2002). The role of individual differences in the accuracy of confidence judgments. *The Journal of General Psychology, 129,* 257-299.

Paris, S. G. (2002). When is metacognition helpful, debilitating, or benign? In P. Chambres, M. Izaute, & P. J. Marescaux, (Eds.), *Metacognition: Process, function and use* (pp. 105-120). Norwell, MA: Kluwer Academic Publishers.

Plude, D. J., Nelson, T. O., & Scholnick, E. K. (1998). Analytical research on developmental aspects of metamemory. *European Journal of Psychology of Education, 13,* 29-42.

Rawson, K. A., Dunlosky, J., & McDonald, S. L. (2002). Influences of metamemory on performance predictions for text. *The Quarterly Journal of Experimental Psychology, 55a,* 505-524.

Reder, L. M. (1987). Strategy selection in question answering. *Cognitive Psychology, 19,* 90-138.

Richards, R. M., & Nelson, T. O. (2004). Effect of the difficulty of prior items on the magnitude of judgments of learning for subsequent items. *American Journal of Psychology, 117,* 81-91.

Roediger, H. L. & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 803-814.

Roediger, H. L. & McDermott, K. B. (1996). False perceptions of false memories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 814-816.

Schwartz, B. L., Benjamin, A. S., & Bjork, R. A. (1997). The inferential and experiential bases of metamemory. *American Psychological Society, 6,* 132-137.

Shaddock, A., & Carroll, M. (1997). Influences on metamemory judgments. *Australian Journal of Psychology, 49,* 21-27.

Shaughnessy, J. J., & Zechmeister, E. B. (1992). Memory-monitoring accuracy as influenced by the distribution of retrieval practice. *Bulletin of the Psychonomic Society, 30,* 125-128.

Simon, D. A., & Bjork, R. A. (2002). Models of performance in learning multisegment movement tasks: Consequences for acquisition, retention, and judgments of learning. *Journal of Experimental Psychology: Applied, 8,* 222-232.

Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning, Memory & Cognition, 26,* 204-221.

Spellman, B. A., & Bjork, R. A. (1992). Commentary: When predictions create reality: Judgments of learning may alter what they are intended to assess. *Psychological Science, 3,* 315-316.

Thiede, K. W., & Dunlosky, J. (1994). Delaying students' metacognitive monitoring improves their accuracy in predicting their recognition performance. *Journal of Educational Psychology, 86,* 290-302.

Thompson, W. B. (1998). Metamemory accuracy: Effects of feedback and the stability of individual differences. *American Journal of Psychology, 111,* 33-42.

Weaver, C. A., & Kelemen, W. L. (1997). Judgments of learning at delays: Shifts in response patterns or increased metamemory accuracy? *American Psychological Society, 8,* 318-321.

Whittlesea, B. W. A. (2003). On the construction of behavior and subjective experience: The production and evaluation of performance. In J. S. Bowers & C. J. Marsolek, (Eds.), *Rethinking implicit memory* (pp. 239-260). New York: Oxford University Press.

Whittlesea, B. W. A. & Leboe, J. P. (2003). Two fluency heuristics (and how to tell them apart). *Journal of Memory and Language, 49,* 62-79.

Yantis, S. & Jonides, J. (1990). Abrupt visual onsets and selective attention: Voluntary Versus automatic allocation. *Journal of Experimental Psychology: Human Perception and Performance, 16,* 121-134.