Subjective learning discounts test type: Evidence from an associative learning and transfer task.

By: Dayna R. Touron, Christopher Hertzog and James Z. Speagle

Touron, D. R., Hertzog, C., & Speagle, J. Z. (2010). Subjective learning discounts test type: Evidence from an associative learning and transfer task. *Experimental Psychology*, *57(5)*, 327-337. DOI: 10.1027/1618-3169/a000039

Abstract:

We evaluated the extent to which memory test format and test transfer influence the dynamics of metacognitive judgments. Participants completed two study-test phases for paired-associates, with or without transferring test type, in one of four conditions: (1) recognition then recall, (2) recall then recognition, (3) recognition throughout, or (4) recall throughout. Global judgments were made prestudy, poststudy, and posttest for each phase; judgments of learning (JOLs) following item study were also collected. Results suggest that metacognitive judgment accuracy varies substantially by memory test type. Whereas underconfidence in JOLs and global predictions increases with recall practice (Koriat's underconfidence-with-practice effect), underconfidence decreases with recognition practice. Moreover, performance changes when transferring test type were not fully anticipated by pretest judgments.

Keywords: learning | memory | metacognition | recognition | judgment | practice | associative learning | metacognitive judgments | recall | psychology | experimental psychology

Article:

Acknowledgement: This research was supported by a National Institute on Aging Grant, NIA R01 AG024485. The authors would like to extend special thanks to the research team at Appalachian State University, particularly Jarrod Hines and Elizabeth Swaim, for their assistance with subject recruitment and data collection

Correspondence concerning this article should be addressed to: Dayna R. TouronDepartment of Psychology296 Eberhart Bldg.PO Box 26170The University of North CarolinaGreensboroNC 27402-6170USA Phone: +1 336 256 0410. Electronic Mail may be sent to: d_touron@uncg.edu.

In recent years, increased attention has been paid to the processes involved in the metacognitive monitoring of memory encoding and retrieval. The accurate monitoring of memory processes can support control behavior which leads to more efficient and effective learning outcomes (see Nelson & Narens, 1990). However, a number of systematic distortions have been found in the

accuracy of memory monitoring for forecasting memory performance (e.g., Benjamin, Bjork, & Schwartz, 1998; Koriat & Bjork, 2005, 2006). Koriat (1997) advocates a perspective that relates metamemory accuracy to the cues that are used to make metacognitive judgments. In some cases, individuals use cues that are not diagnostic of subsequent performance, such as the fluency (speed and ease) of retrieval and encoding (Benjamin et al., 1998; Hertzog, Dunlosky, Kidder, & Robinson, 2003). In other cases, valid cues are ignored or discounted. For example, individuals do not distinguish between forward versus backward associative strength (e.g., in a free-association task, the word “kittens” elicits “cats” but the word “cats” does not as readily elicit “kittens”) when they make judgments of learning (JOLs) for items differing in this characteristic, despite major influences of this cue on recall (Koriat & Bjork, 2005).

It is some consolation, then, that metacognitive judgment accuracy can improve over practice with multiple study-test phases. Such improvements might largely be based on opportunities to benefit from test experience (Finn & Metcalfe, 2007; Hertzog, Dixon, & Hultsch, 1990; Hertzog, Price, & Dunlosky, 2008; Koriat & Bjork, 2006). Despite this, distortions in monitoring can also accompany memory task experience. Koriat and colleagues have uncovered an interesting regularity in the dynamics of monitoring cued-recall accuracy (Koriat, 1997; Koriat, Sheffer, & Ma’ayan, 2002; Koriat, Ma’ayan, Sheffer, & Bjork, 2006). When making JOLs, individuals typically do not anticipate future performance gains due to learning. That is, monitoring learning is underconfident, in the sense that mean JOL confidence becomes lower than the actual level of cued recall. The effect is also observed for global predictions, in which individuals give a single number predicting how many (or what percentage) of items they will remember (Koriat et al., 2002). The absolute accuracy of JOLs (often scaled as the difference score, mean JOL – mean recall) can be relatively good on an initial study-test phase, often exhibiting modest overconfidence. However, it becomes substantially underconfident on a second study-test phase, and thereafter (see also Meeter & Nelson, 2003; Scheck & Nelson, 2005; Serra & Dunlosky, 2005). This underconfidence-with-practice (UWP) effect is particularly intriguing given that practice generally improves the resolution (i.e., relative accuracy or item-by-item discriminability) of JOLs at the same time it negatively affects their absolute accuracy (Connor, Dunlosky, & Hertzog, 1997; Finn & Metcalfe, 2007; Koriat, 1997). Since learning outside the laboratory often involves multiple exposures to material (e.g., reading before class and listening during lecture), understanding the dynamics of subjective learning and their impact on subsequent behavior is particularly important.

It is notable that previous work demonstrating UWP has exclusively relied on cued-recall tests in multiple study-test phases. Metamemory accuracy can vary markedly for different test types. People appear to know that recognition is normatively easier than recall, resulting in higher memory performance with recognition testing than with recall testing (Thiede, 1996). Indeed, test expectancy has been shown to impact learning behavior, such that participants who anticipate a recall test use encoding strategies that benefit both recall and recognition tests (Neely & Balota, 1981). However, participants may not always consider test type differences

when forecasting performance. For example, Mazzoni and Cornoldi (1993) found that JOLs did not vary when participants anticipated a recall test versus a recognition test, even though more study time was allocated when anticipating a recall test.

Manipulation of test type might also moderate the UWP effect. In a skill acquisition task that demands hundreds of associative recognition tests (the noun-pair learning task), we (Hertzog, Touron, & Hines, 2007; Touron & Hertzog, 2004a, 2004b; Touron, Swaim, & Hertzog, 2007) have regularly obtained substantial overconfidence at the end of practice in JOLs for cued recall, a task we administered to measure learning of the associations after extensive experience with the noun pairs, and after associative recognition tests that are part of our noun-pair test procedure. For example, Touron and Hertzog (2004b) found that both young and older adults' associative recognition performance after extended practice was excellent (over 93% correct after 60 repetitions per stimulus pair). When participants were then given a cued-recall test, recall performance was lower than recognition (87% for young and 64% for old). However, mean JOLs for cued recall were overconfident (91% for young and 77% for old), showing no UWP effect. JOLs were not simply reflecting levels of prior recognition performance – both age groups predicted lower recall than the level of recognition performance they achieved at the end of practice. Nevertheless, the extended recognition performance experience in the noun-pair task did not produce UWP upon transfer to a cued-recall test. This outcome implies that the learning history resulting from a particular type of memory test format may influence how memory confidence changes with learning.

In fact, test type and test transfer differences in the UWP effect might be anticipated from various theoretical approaches to metacognition. Koriat's cue-utilization perspective (1997), as noted above, proposes that the accuracy of metamemory judgments is related to the use of effective cues. Prior research from this perspective indicates that metacognitive judgments focus on intrinsic cues (i.e., stimulus characteristics), such as the degree of semantic relatedness between paired-associates, and may discount extrinsic cues (i.e., learning conditions), such as test type and baseline or chance performance (see Koriat, 1997; Koriat & Bjork, 2006). We might also anticipate that UWP would be impacted by test type based on the memory-for-past-test (MPT) heuristic (Finn & Metcalfe, 2007, 2008), which assumes that JOLs are more informed by prior test performance levels than by monitoring during study or by theory-based expectations (such as those resulting from instructed test type). From this perspective, low accuracy in a recall test might be expected to lead to more underconfident JOLs in future study phases (a typical UWP outcome), whereas high accuracy in a recognition test might be expected to lead to less underconfident JOLs in future study phases.

Differences in subjective learning or UWP across test formats and test transfer could have important implications outside the laboratory. For example, student learning might involve self-assessment primarily via recognition when re-reading and studying class materials, which could create illusory overconfidence relative to an examination format that involves cued recall (such as short answer essays). Test format is likely to vary for a given concept across multiple

occasions such as informal queries in the classroom, short quizzes, student self-assessment, and final examinations; tailoring study to anticipate such variability could be an important component of successful preparation.

Although understanding UWP most critically involves changes in the absolute accuracy of monitoring judgments, it is worth noting that the resolution of monitoring accuracy also varies by test type. Thiede and Dunlosky (1994) obtained higher relative accuracy for JOLs with recall testing compared to recognition testing, regardless of test anticipation or cue type, partially due to correct guessing. This outcome is consistent with meta-analytic findings by Schwartz and Metcalfe (1994) showing that the relative accuracy of feeling-of-knowing judgments depends in large part on the number of test alternatives used, with unreliable judgments in two-choice recognition, but more reliable judgments for recall tests. Weaver and Kelemen (2003) also showed substantially lower relative and absolute accuracy of delayed JOLs in recognition memory tasks, although the test type effect varied for different types of JOLs.

The present work examined two primary questions: (1) how are changes in metacognitive monitoring accuracy impacted by differences in memory test type and (2) how are changes in metacognitive monitoring accuracy impacted when the type of memory test is changed. The design involved persons either experiencing two phases of the same test procedure (recognition or recall) or being transferred from one test format to the other. Although previous work examining changes in monitoring accuracy has generally focused on judgments at study, we also examined judgments at test in the current study as well as global judgments at other periods during the study-test interval. We used multiple different metacognitive judgments in order to clarify the point in the phase of study-test experiences where UWP effects emerged (see Hertzog & Dunlosky, 2004; Hertzog et al., 2008, for broader theoretical discussions of how placement of metacognitive judgments relates to inferences about different kinds of monitoring and their influence on generating new knowledge about effective encoding and retrieval strategies).

First, we compared recognition testing and recall test conditions as to what extent changes in memory accuracy are reflected in changes in item judgments (immediate JOLs at study and confidence judgments [CJs] at test) and global judgments (performance prestudy and poststudy predictions as well as postdictions after the test). Global postdictions and predictions require inferential processes about how to translate item-specific experiences with memory successes and failures (as manifested in CJs) into aggregate representations of the likelihood of success in experimental contexts (Hertzog et al., 2008). Note that the collection of multiple judgments allowed for the examination of changes in monitoring accuracy both within and across two study-test phases. We hypothesized that changes in judgment accuracy would differ by test type, with the typical UWP for recall tests but not recognition tests.

Second, we explored the effects of test type transfer (i.e., recognition to recall or recall to recognition) on item and global judgments. We hypothesized that individuals would fail to adequately adjust metacognitive judgments for transfer in test type. Specifically, we expected

that individuals transferred from recognition tests to recall tests would provide phase 2 JOLs more consistent with their high level of phase 1 success in recognition performance and discount the recall test instruction, whereas individuals transferring from recall tests to recognition tests would provide phase 2 JOLs more consistent with their lower level of phase 1 success and discount the recognition test instruction.

Methods

Design

To fully compare changes in monitoring accuracy with consistent test type versus test type transfer, this experiment involved a 2 (phase 1 test type) × 2 (phase 2 test type) × 2 (phase 1 or 2) mixed factorial design. This resulted in four testing conditions; each participant was tested in one of the following: (1) recognition then recall, (2) recall then recognition, (3) consistent recognition, or (4) consistent recall. Each participant completed two phases, and each phase consisted of a study segment and a test segment. As such, the procedure for each participant included study-test phase 1 followed by study-test phase 2, with the test types as indicated by condition labels.

Participants

One hundred adults between the ages of 18 and 25 years participated and received course credit as compensation for their time. Equal numbers of participants were randomly assigned to each testing condition, as described above.

Procedure

Stimulus presentations and response recordings were controlled by a Visual Basic 6.0 program. All participants completed self-paced computerized instructions. The participants were told they would perform multiple study-test phases for a set of 60 normatively unrelated word pairs (e.g., IVY-BIRD or CAT-MARKET), with 6 s of study per pair. This paired-associates learning task provides a simple analog to associative learning in the classroom, where assessment often requires the correct pairing of terminology with conceptual definitions.

Importantly, the type of test to be completed (depending on condition and phase) was described in detailed instructions preceding each study-test phase. The instructions also presented sample test items for the test type to be completed. Participants were then asked to provide a Global PreStudy Prediction: The estimated number of word pairs out of sixty total pairs that they would be able to correctly recognize or recall at test.

Following study of each word pair, an item JOL was requested (e.g., “How confident are you that you will be able to remember the previous word-pair 10 min from now? Please estimate your % confidence by typing in a number ranging from 0 to 100.”). Following study of all stimuli,

participants were asked to provide a Global PostStudy Prediction analogous to the global prestudy prediction described above.

Depending on condition and phase, recognition or recall tests followed. For recognition testing, participants were shown a pair and instructed to press a key labeled “Y” if the target pair matched what had been presented in study or a key labeled “N” if the target pair did not match. Thirty matched pairs and thirty unmatched pairs were randomly ordered during recognition testing. Matched pairs were exactly like those presented during study (e.g., IVY-BIRD) and unmatched pairs were generated by randomly selecting the second word from a pair other than the match (e.g., IVY-MARKET). For recall testing, the first word of the pair was presented and participants were required to type in the second word to complete the word pair (e.g., IVY-_____). The first three letters of the word had to be entered correctly for the response to be scored as correct.

Following each word pair test, an item CJ was requested (e.g., “How confident are you that your previous response was correct? Please estimate your % confidence by typing in a number ranging from 0 to 100.”). Following tests for all the 60 word pairs, participants were asked to provide a Global PostTest Postdiction estimating the number of pairs they had correctly recognized or recalled.

Results

We examined the following measures: Memory accuracy, item ratings (CJs and JOLs), and global ratings (predictions and postdictions). Item and global ratings were examined both in terms of response level and in terms of absolute accuracy (i.e., subtractions of accuracy from ratings). Item ratings were additionally examined in terms of relative accuracy (i.e., item-level discriminability as indexed by Goodman-Kruskal gamma correlations).

Memory Accuracy

Memory performance data are given in Figure 1 and Table 1. Because scaling differs for recognition accuracy and recall accuracy – with the functional floor for recognition higher than the floor for recall – differences in outcomes are described but not formally tested. As anticipated, performance was higher when performing recognition than when performing recall, regardless of whether the test type had been performed consistently or transferred. Anticipated performance improvements can also be seen for phase 2. Note that this pattern of outcomes results in a crossover interaction in the transfer conditions, with increasing accuracy when transferring from recall to recognition and decreasing accuracy when transferring from recognition to recall. Metacognitive ratings which accurately reflect memory performance should show a similar pattern with ratings driven by test type and learning and a crossover interaction in the transfer conditions.

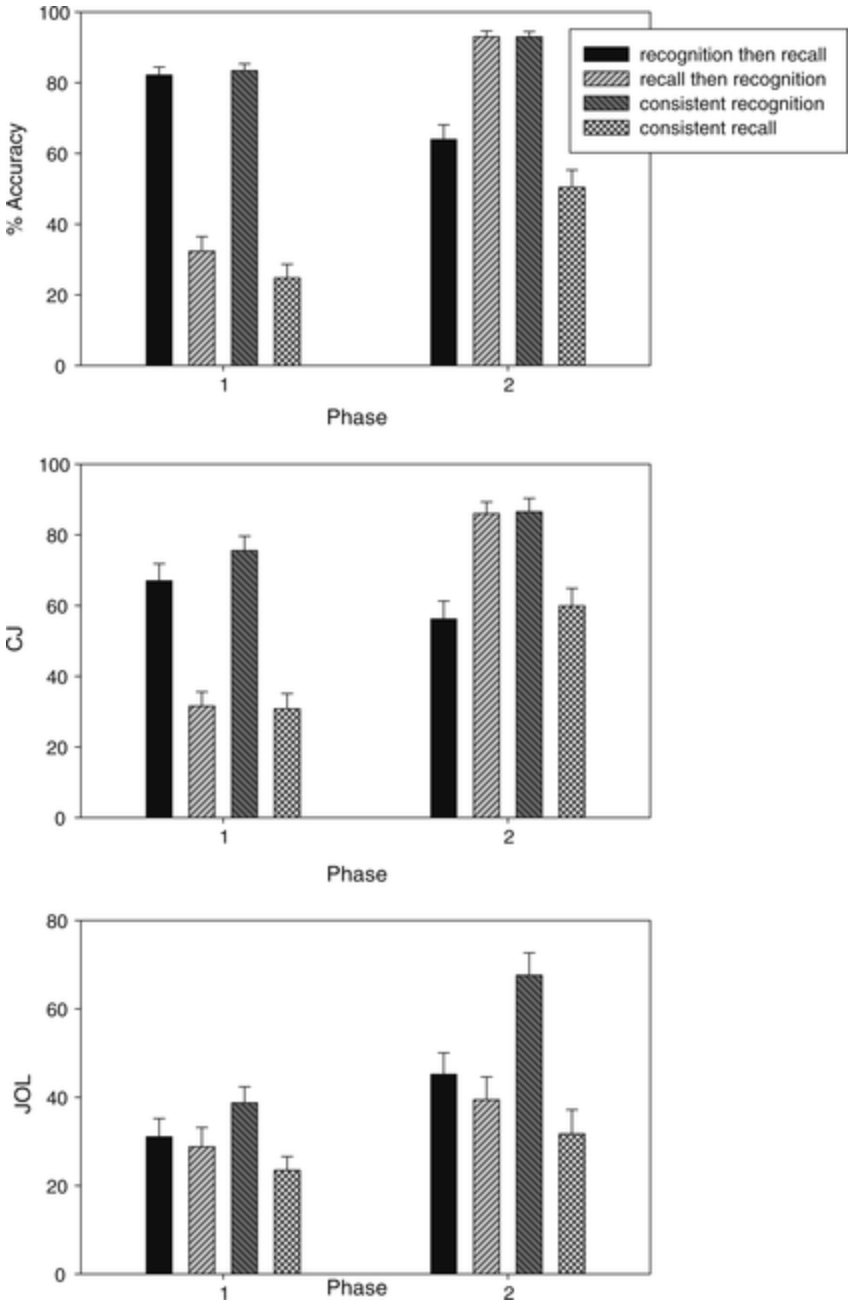


Figure 1. Mean values and standard errors of memory accuracy (top), CJs (middle), and JOLs (bottom) by phase and condition. Note that considerable overlap exists in the top panel. In phase 1, both conditions performing recognition had accuracy near 80%, while both performing recognition had accuracy near 25%; in phase 2, both conditions performing recognition had accuracy near 90%. Overlapping points in the middle panel reflect analogous similarities.

Table 1 has been omitted from this formatted document.

Item Ratings

CJs

Figure 1 also shows the mean CJs for test conditions over phases. Note that the pattern of CJs was indeed similar to the pattern observed for mean levels of memory performance, with condition differences in each phase driven by test type and practice-based improvements, as well as the crossover interaction with transfer. Accordingly, the absolute accuracy of CJs (computed as a simple subtraction of accuracy from ratings) was quite good (Figure 2), demonstrating that metacognitive ratings can track performance differences/changes by test type, training, and transfer. The relative accuracy of CJs was also generally high (Table 2), particularly in the second phase. As noted earlier, test type differences in relative accuracy may be driven in large part by the possibility of correct guessing for recognition tests but not for recall tests (Schwartz & Metcalfe, 1994; Thiede & Dunlosky, 1994; Weaver & Kelemen, 2003).

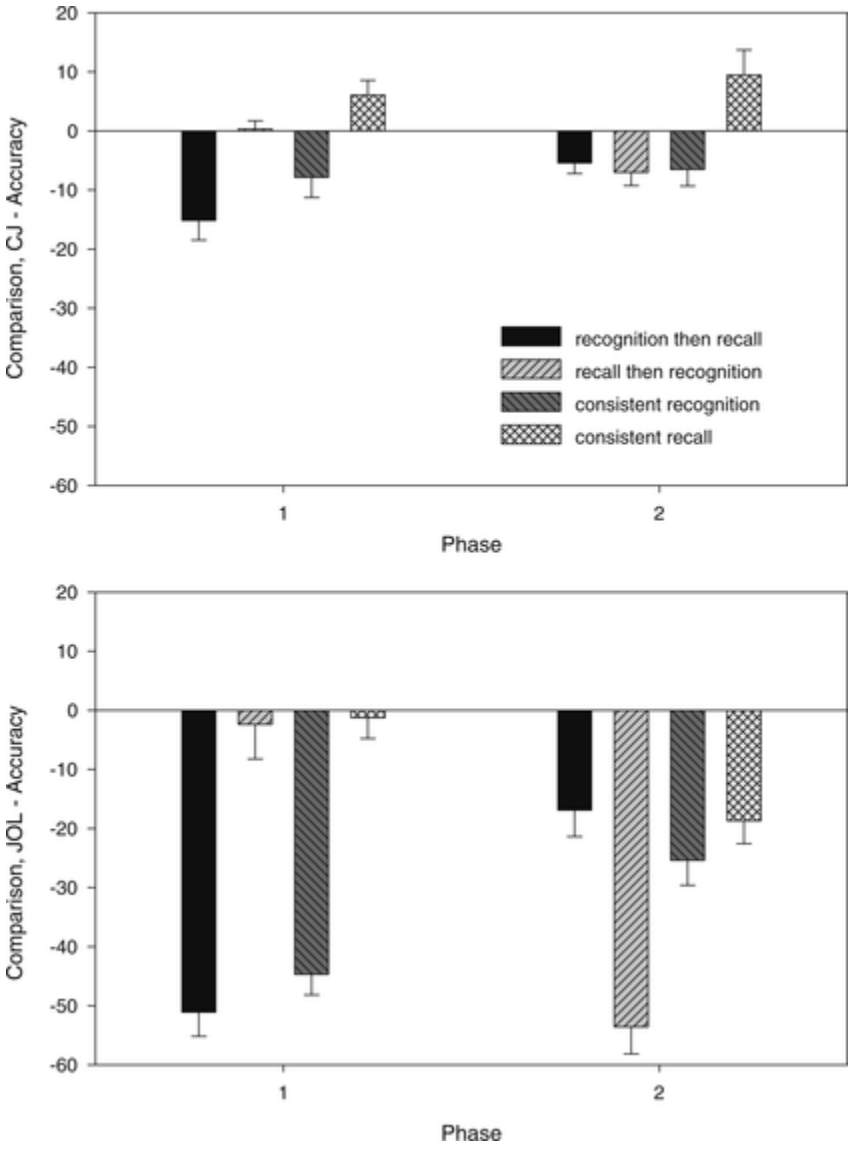


Figure 2. Mean values and standard errors for absolute accuracy of CJs (accuracy; top) and JOLs (accuracy; bottom) by phase and condition.

Table 2 is omitted from this formatted document.

JOLs

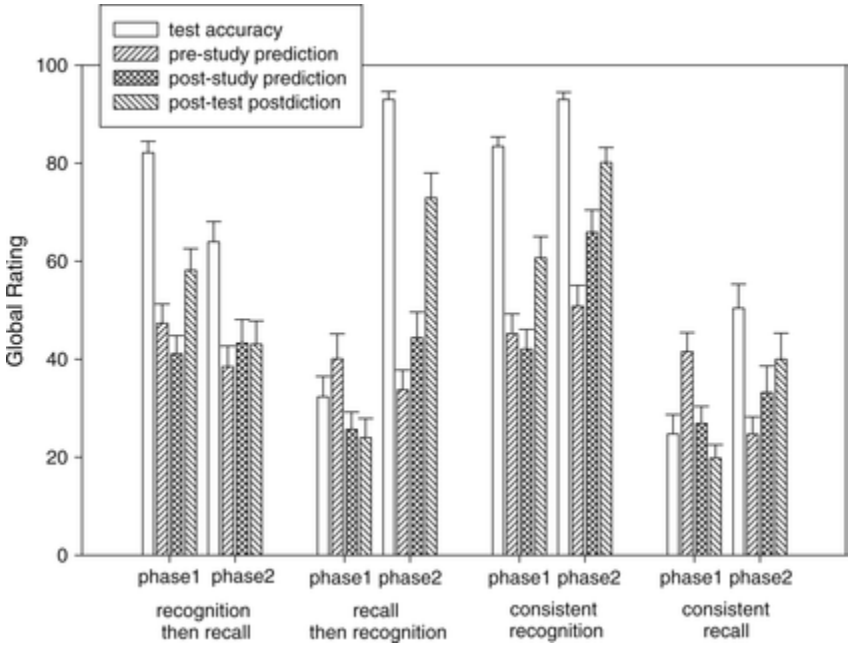
Mean JOLs and standard errors by phase and test type are shown in Figure 1. The correspondence between performance and CJ was not similarly reflected in JOLs. At phase 1, the test type difference, comparing those in the recall test conditions to those in the recognition test conditions, was small but reliable, $F(1, 99) = 5.16$, $MSE = 372$, $p = .03$. Note that the average JOL for the recognition conditions was below 40% – lower than a benchmark of 50% based on random guessing. Improvements from phase 1 to phase 2 were anticipated across conditions, $F(1, 96) = 62.09$, $MSE = 193$, $p < .01$. It appears that the largest improvements were anticipated by participants performing consistent recognition. Indeed, phase 2 JOLs for consistent recognition were reliably higher than for each of the other groups (all $p < .01$), which did not differ (all $p > .05$), as indicated by examining all focused (pairwise) condition comparisons. Note that decreases in memory performance when transferring from recognition to recall testing were not reflected in JOL ratings, and participants who transferred from recognition to recall appeared to anticipate improved performance at phase 2 instead of the negative transfer effect. Individuals' JOLs apparently did not consider the fact that recognition is substantially easier than recall. Note, however, that increases in JOL ratings were less pronounced for participants transferring from recognition to recall, compared to those performing consistent recognition, indicating some sensitivity to transfer, $F(1, 48) = 6.90$, $MSE = 190$, $p = .01$.

The absolute accuracy of JOLs is given in Figure 2. At phase 1, groups performing recall showed good absolute accuracy while groups performing recognition showed pronounced underconfidence – an effect driven by differences in accuracy given the JOL similarity noted above, $F(1, 98) = 115.07$, $MSE = 456$, $p < .01$. Furthermore, groups performing recall at phase 1 showed decreased absolute JOL accuracy at phase 2 (i.e., increased underconfidence – the UWP effect), while groups performing recognition at phase 1 showed increased absolute JOL accuracy at phase 2 (i.e., decreased underconfidence), $F(1, 98) = 8.98$, $MSE = 613$, $p < .01$. Earlier recall performance appeared to profoundly reduce absolute JOL accuracy by participants in the recall to recognition transfer condition. Focused comparisons at phase 2 (examining all pairwise condition comparisons) indicate that participants in the recall-then-recognition condition were more underconfident than other conditions (all $p < .01$), which did not differ in the absolute accuracy of their JOLs (all $p > .05$). This outcome appears to be driven by a failure to anticipate the substantial performance improvements afforded by the transfer to a recognition task. Relative accuracy was poorer for JOLs compared to CJs (Table 2). As for CJs, relative JOL accuracy was higher for phase 2.

In summary, item ratings indicate good absolute and relative accuracy for CJs but poorer monitoring for JOLs. These outcomes confirm previous findings that metacognitive monitoring during test, as a more proximal indicator of performance, is more accurate (also see Dunlosky & Hertzog, 2000; Thiede & Dunlosky, 1994). While the conditions with recall tests in phase 1 showed a typical UWP outcome for JOLs, the conditions with recognition tests in phase 1 showed an opposite outcome of initially pronounced underconfidence which partially resolved in phase 2.

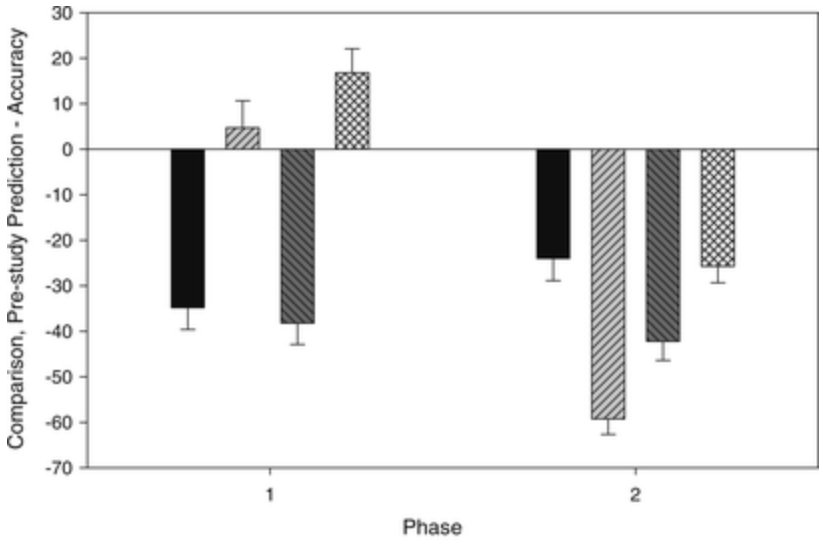
Global Ratings

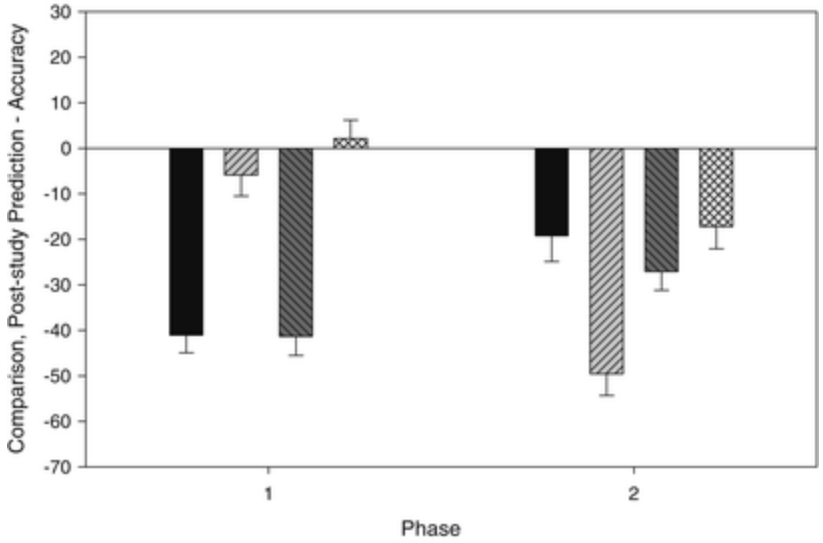
Global metacognitive ratings are given in Figure 3. Mean ratings for each condition are presented chronologically with test accuracy to allow for the comparison of changes in monitoring accuracy during and across phases. In phase 1, prestudy predictions were around or below 50% and did not vary reliably by test type (those performing recall compared to recognition), $F(1, 98) = 1.69$, $MSE = 436$, $p > .05$. For phase 1, note that although ratings universally drop from prestudy to poststudy, they increase from poststudy to posttest for groups performing recognition but decrease from poststudy to posttest for groups performing recall. These data will be examined below in the context of their absolute accuracy.



The absolute accuracy of global metacognitive ratings is given in Figure 4. Patterns of accuracy by test type and transfer were qualitatively similar for prestudy and poststudy predictions and quite similar to patterns of absolute accuracy for JOLs, as follows (noted differences in these planned comparisons were reliable at $p < .05$). At phase 1, groups performing recall showed good absolute accuracy for predictions while groups performing recognition were substantially underconfident. At phase 2, groups performing recall at phase 1 showed increased underconfidence in predictions, while groups performing recognition at phase 1 showed stable or

decreased underconfidence. In the group which transferred from recall to recognition, phase 1 recall again seems to have markedly reduced phase 2 predictions relative to accuracy. As with CJs, posttest postdictions were less deviant from accuracy. However, phase 1 postdictions showed good absolute accuracy for recall but underconfidence for recognition, consistent with other global ratings and JOLs.


Comparison, Pre-study Prediction - Accuracy
30
20
10
0
-10
-20
-30
-40
-50
-60
-70
1
2
Phase


Comparison, Post-study Prediction - Accuracy
30
20
10
0
-10
-20
-30
-40
-50
-60
-70
1
2
Phase

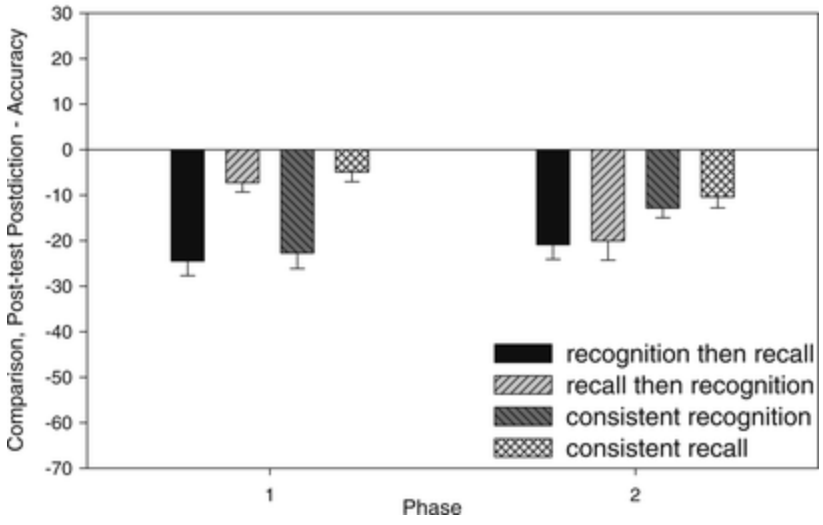

Comparison, Post-test Postdiction - Accuracy
30
20
10
0
-10
-20
-30
-40
-50
-60
-70
1
2
Phase
recognition then recall
recall then recognition
consistent recognition
consistent recall

Figure 4. Mean values and standard errors for absolute accuracy of PRS (accuracy; top), PSS (accuracy; middle), and PST (accuracy; bottom) by phase and condition.

These outcomes aid interpretation of the mean global ratings in terms of how test type and transfer impact changes in monitoring accuracy. Participants who performed recall in phase 1 show stable or increased underconfidence (the typical UWP effect) within phase 1 as well as across phases. In contrast, participants who performed recognition in phase 1 show decreased underconfidence within phase 1 as well as across phases. Likewise, decreased ratings from phase 1 poststudy to posttest by participants performing recall reflects increased underconfidence (i.e., poorer absolute accuracy; again the typical UWP effect), whereas increased ratings from phase 1 poststudy to posttest by participants performing recognition reflects decreased underconfidence (i.e., better absolute accuracy).

However, comparison of phase 1 postdictions and phase 2 predictions reveals both test type and transfer effects. While participants performing consistent recognition show decreased ratings and those performing consistent recall do not show substantially increased ratings from phase 1 posttest to phase 2 prestudy, participants transferring from recall to recognition do seem to anticipate better phase 2 performance and participants transferring from recognition to recall do seem to anticipate poorer phase 2 performance. Nevertheless, in neither case does the anticipated change account for the impact of test type on performance. It is also notable that the decrease in ratings from prestudy to poststudy in phase 1 did not occur in phase 2, where ratings increased from prestudy to poststudy and also to posttest (except for in the recognition to recall group), suggesting that the resolution of recall UWP begins within the second study-test phase rather than following it.

In summary, absolute accuracy of global ratings was consistent with findings for item ratings. The UWP effect occurred within as well as between phases for those who performed recall tests in phase 1, while participants who performed recognition tests in phase 2 showed an opposite effect – decreased underconfidence – both within and across phases. While we obtained some evidence that participants considered test type and test transfer in comparisons between phase 1 posttest and phase 2 pretest ratings, such considerations appear to be inadequate.

Discussion

Over two study-test phases, substantial differences in subjective learning were obtained for recognition versus recall testing, as demonstrated for both item and global judgments. Whereas consistent recall testing, as has been previously examined, produced UWP, a separate pattern emerged for recognition testing. Rather than being overconfident in phase 1 and underconfident in phase 2, recognition testing led to extreme underconfidence in phase 1 which diminished in phase 2. Given the mean levels of JOLs and global predictions, it appears that participants performing recognition did not consider that they had a substantial probability of correct guessing in a forced-response recognition task.

A number of explanations for these findings can be considered. For the most part, metacognitive ratings appear to be driven by test type for the previous phase more so than by expectations of improved or reduced performance with learning or test type transfer. These patterns implicate a first possible explanation – that theory-based extrinsic cues are discounted or insufficiently considered when making metacognitive judgments (Koriat, 1997; Koriat & Bjork, 2006). In this study, available extrinsic cues during both phases include anticipated test type (recall vs. recognition) as well as related beliefs regarding test difficulty. Thiede (1996) separated these factors and found that JOLs, self-regulation behaviors, and test performance were governed more so by test type compared to test difficulty, suggesting that participants either do not monitor or discount relative difficulty in favor of a priori beliefs about test type. Having obtained similar JOLs and predictions suggests that test type and difficulty expectations were not considered by participants in the current phase 1 with test type varied between subjects. Further research could separate these factors to determine the extent to which the discounting of a priori expectations about test type as well as previous test difficulty contribute to the insufficient sensitivity to test type transfer we observed in the current phase 2.

Second, the outcomes could be explained by the possibility that subjective confidence in future success is not necessarily isomorphic with the objective aggregate likelihood of success (see Gigerenzer, 2000). JOLs and CJs require a subjective probability of success, scaled in % confidence, for each item. By contrast, performance in the recognition and recall memory tasks is scaled as % of items remembered, aggregated over the whole list. When asked to make JOLs for recognition judgments, individuals may use a subjective probability scale that does not align in an absolute sense with probability of correct recognition. Metacognitive researchers examining bias through absolute accuracy of item-level judgments implicitly assume these two probability scales align in interval-scale fashion, even though this is a major methodological assumption that is not necessarily correct (e.g., Keren, 1991). Arguing against this explanation, however, are the facts that (1) subjective confidence in CJs – in contrast to JOLs – aligns well with recognition probabilities and (2) our global predictions were scaled as the expected number (not the expected percentage) of correct items. Despite the differences in scaling, JOLs and global predictions behave quite similarly and show discounting of the cue of test type. We conclude that scaling effects cannot fully account for the results (see also Dougherty, Scheck, Nelson, & Narens, 2005).

Future research might inform participants about the normative difficulty of the recognition and recall tests to see if this alters JOL rating behavior. Such a study could also manipulate whether restricting options on the subjective confidence scale (e.g., only allowing recognition test JOLs of 50–100% confidence to align with the probability of successful guessing) would influence the absolute accuracy of judgments in the recognition test. Future research should also more broadly consider whether, the extent to which, and how individuals correct monitoring judgments for guessing when anticipating recognition tests versus recall tests, and how such corrections might impact indices of monitoring accuracy.

Third, outcomes are somewhat consistent with the MPT heuristic, advanced by Finn & Metcalfe (2007), which posits that JOLs are entrained by prior performance levels rather than by monitoring item study outcomes. By this account, participants increase phase 2 JOLs when they remember getting the item right on the phase 1 test, as is more likely with recognition, but decrease JOLs when they remember getting the item wrong on the phase 1 test, as more likely with recall. Certainly, postdictions were more accurate than predictions, indicating that people learn from test experience about their memory performance, even without performance feedback (see also Hertzog et al., 2008).

To further consider the influence of the MPT heuristic on our findings, we directly compared accuracy data with global and item judgments (see Table 1). Marked changes in confidence occur during the study and test intervals as well as between the test and subsequent study. For example, confidence declined from prestudy predictions to JOLs in each condition during phase 1 and increased in each condition during phase 2. For both phases and in each condition, global postdictions were below both previous response CJs and actual performance. Whereas comparison of absolute accuracy indicates that recall UWP resolves across phases, underconfidence also resolves during the study phase of the second study-test phase interval. Such outcomes do not directly follow from the MPT heuristic. Finn and Metcalfe (2007) also concluded that the MPT heuristic was not a sufficient account of UWP effects in consistent recall tasks, for different reasons.

One method that has been used to demonstrate the MPT heuristic is the comparison of JOLs for items which were not recalled on a first test but recalled on a second test (FR or forgot-recalled items) to JOLs for items which were recalled on both a first and second test (RR items; see Finn & Metcalfe, 2007, 2008). If the MPT heuristic is being employed, JOLs in a second phase should be lower for items that were not recalled in the first test (FR items), regardless of test 2 recall (hence the comparison to RR items). We adopted this approach and compared the difference in phase 2 JOLS between FR and RR items for participants by condition (see Table 3). As predicted by the MPT heuristic, this difference was greater than zero in each condition (all $p < .01$). Furthermore, the size of the difference varied by condition. Whereas the difference was minimal for participants who completed a recognition test in phase 1, it was larger for those who completed a recall test in phase 1. It appears that recall test performance has a greater influence on subsequent ratings compared to recognition test performance. Further research should more closely examine potential sources of this discrepancy.

Table 3 is omitted from this formatted document.

In addition, comparisons of transfer and consistent conditions for the second study-test phase suggest that individuals raised their phase 2 predictions and JOLs well above phase 1 postdictions when transferred to recognition memory, relative to those who continue recall testing. Conversely, individuals transferred to a recall test lower global predictions and JOLs more compared to those who will continue recognition. Thus, both item judgments and global

judgments do appear to reflect changes in test type transfer in addition to any influence of past performance.

Individuals' metacognitive judgments were influenced by test type (e.g., Thiede, 1996), but not enough to produce high levels of absolute accuracy in JOLs and global predictions. Such findings are consistent with the argument that judgment accuracy is improved by task experience, but not in a manner completely consistent with the MPT heuristic. To the extent that previous performance influences subsequent metacognitive ratings, it appears to do so using what might be fallible monitoring of that previous performance (see Hines, Touron, & Hertzog, 2009). Moreover, other factors seem to be in play as well.

A limitation of this study is that recognition and recall tests do not have the same baseline of chance performance. The assumed base rate for guessing on our recall test is essentially zero, because there were few intrusion errors. Future research should compare recognition tests with varying numbers of response alternatives to characterize insensitivity to baseline performance more fully. It could also consider variants of forced recall tests (e.g., Koriat & Goldsmith, 1996) that might render the recall and recognition tests more comparable. Finally, manipulations that would reverse the performance advantage of recognition over recall would be useful to determine whether it is location in a range of the possible scale, rather than the nature of the memory test, that influences UWP differences (see Dougherty et al., 2005). Nevertheless, the present results are important because they open a new line of inquiry regarding test type effects and UWP.

In summary, these outcomes further demonstrate that systematic distortions occur in the monitoring and forecasting of memory performance. Effects in the data show that sensitivity to the cue of test type, although present, is insufficient to maintain high levels of judgment accuracy when one is transferred to a new test type. Participants also largely discounted the impact of learning on test performance. Finally, participants' failure to properly evaluate the differences in test type was sufficiently large that it overrode the UWP effect – which apparently is not universal, and could even be specific to the associative cued recall in which it was discovered. Whereas recall ratings became underconfident with practice, recognition ratings were highly underconfident initially, but this underconfidence decreased with recognition task practice.

References

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.

Connor, L. T., Dunlosky, J., & Hertzog, C. (1997). Age-related differences in absolute but not relative metamemory accuracy. *Psychology and Aging*, *12*, 50– 71.

Dougherty, M. R., Scheck, P., Nelson, T. O., & Narens, L. (2005). Using the past to predict the future. *Memory and Cognition, 33*, 1096– 1115.

Dunlosky, J., & Hertzog, C. (2000). Updating knowledge about encoding strategies: A componential analysis of learning about strategy effectiveness from task experience.*Psychology and Aging, 15*, 462– 474.

Finn, B., & Metcalfe, J. (2007). The role of memory for past test in the underconfidence with practice effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition,33*, 238– 244.

Finn, B., & Metcalfe, J. (2008). Judgments of learning are influenced by memory for past test. *Journal of Memory and Language, 58*, 19– 34.

Gigerenzer, G. (2000). *Adaptive thinking: Rationality in the real world*. Oxford, England: Oxford University Press.

Hertzog, C., Dixon, R. A., & Hultsch, D. F. (1990). Relationships between metamemory, memory predictions, and memory task performance in adults. *Psychology and Aging, 5*, 215– 227.

Hertzog, C., & Dunlosky, J. (2004). Aging, metacognition, and cognitive control. In B. H.Ross (Ed.) , *Psychology of learning and motivation* (pp. 215– 251). San Diego: CA: Academic Press.

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

Hertzog, C., Price, J., & Dunlosky, J. (2008). How is knowledge generated about memory encoding strategy effectiveness?*Learning and Individual Differences, 18*, 430– 455.

Hertzog, C., Touron, D. R., & Hines, J. (2007). Does a time monitoring deficit influence older adults' delayed retrieval shift during skill acquisition?*Psychology and Aging, 22*, 607– 624.

Hines, J., Touron, D. R., & Hertzog, C. (2009). Metacognitive influences on study time allocation in an associative recognition task: An analysis of adult age differences. *Psychology and Aging, 24*, 462– 475.

Keren, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica, 77*, 217– 273.

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., & Bjork, R. A. (2005). Illusions of competence in monitoring one's knowledge during study. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *31*(2), 187– 194.

Koriat, A., & Bjork, R. A. (2006). Mending metacognitive illusions: A comparison of mnemonic-based and theory-based procedures. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *32*, 1133– 1145.

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

Koriat, A., Ma'ayan, H., Sheffer, L., & Bjork, R. A. (2006). Exploring a mnemonic debiasing account of the underconfidence-with-practice effect. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *32*(3), 595– 608.

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.

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.

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

Neely, J. H., & Balota, D. A. (1981). Test-expectancy and semantic organization effects in recall and recognition. *Memory & Cognition*, *9*, 283– 300.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G.Bower (Ed.) , *The psychology of learning and motivation: Advances in research and theory* (*Vol. 26*, pp. 125– 173). San Diego, CA: Academic Press.

Scheck, P., & Nelson, T. O. (2005). Lack of pervasiveness of the underconfidence-with-practice effect: Boundary conditions and an explanation via anchoring. *Journal of Experimental Psychology: General*, *134*, 124– 128.

Schwartz, B. L., & Metcalfe, J. (1994). Methodological problems and pitfalls in the study of human metacognition. In J.Metcalfe, & A. P.Shimamura (Eds.) , *Metacognition*. Cambridge, MA: MIT Press.

Serra, M. J., & Dunlosky, J. (2005). Does retrieval fluency contribute to the underconfidence-with-practice effect?*Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 1258– 1266.

Thiede, K. W. (1996). The relative importance of anticipated test format and anticipated test difficulty on performance. *The Quarterly Journal of Experimental Psychology*, *49A*, 901– 918.

Thiede, K. W., & Dunlosky, J. (1994). Delaying students' metacognitive monitoring Improves their accuracy in predicting their recognition performance. *Journal of Educational Psychology*, *86*(2), 20– 302.

Touron, D. R., & Hertzog, C. (2004a). Distinguishing age differences in knowledge, strategy use, and confidence during strategic skill acquisition. *Psychology and Aging*, *19*(3), 452– 466.

Touron, D. R., & Hertzog, C. (2004b). Strategy shift affordance and strategy choice in young and older adults. *Memory and Cognition*, *32*, 298– 310.

Touron, D. R., Swaim, E., & Hertzog, C. (2007). Moderation of older adults' retrieval reluctance through task instructions and monetary incentives. *Journal of Gerontology: Psychological Sciences*, *62B*, 149– 155.

Weaver, C. A., & Kelemen, W. L. (2003). Processing similarity does not improve metamemory: Evidence against transfer-appropriate monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 1058– 1065.