Journal of Experimental Psychology: Learning, Memory, and Cognition

Metacognitive Awareness and Adaptive Recognition Biases

Diana Selmeczy and Ian G. Dobbins

Online First Publication, July 30, 2012. doi: 10.1037/a0029469

Journal of Experimental Psychology:
Learning, Memory, and Cognition

0278-7393/12/$12.00 DOI: 10.1037/a0029469

Metacognitive Awareness and Adaptive Recognition Biases

Diana Selmeczy and Ian G. Dobbins
Washington University in St. Louis

Prior literature has primarily focused on the negative influences of misleading external sources on memory judgments. This study investigated whether participants can capitalize on generally reliable recommendations in order to improve their net performance; the focus was on potential roles for metacognitive monitoring (i.e., knowledge about one's own memory reliability) and performance feedback. In Experiment 1, participants received explicit external recommendations (*Likely Old* or *Likely New*) that were 75% valid during recognition tests containing deeply and shallowly encoded materials. In Experiment 2, participants received recommendations of differing validity (65% and 85%). Discrimination improved across both experiments when external recommendations were present versus absent. This improvement was influenced by metacognitive monitoring ability measured in the absence of recommendations. Thus, effective incorporation of external recommendations depended in part on how sensitive observers were to gradations of their internal evidence when recommendations were absent. Finally, corrective feedback did not improve participants' ability to use external recommendations, suggesting that metacognitive monitoring ability during recognition is not easily improved via feedback.

Keywords: recognition memory, metacognition, external cues, feedback

Supplemental materials: http://dx.doi.org/10.1037/a0029469.supp

Metacognition, or the ability of observers to introspect about the quality of internal cognitive processes, plays an important role in guiding future cognition and behavior. In the case of memory attribution, metacognition has been studied with judgments of learning (JOL). During such judgments, participants rate their mastery of a newly learned fact or association by predicting their likelihood of later successful retrieval, a skill that may be important for the efficient allocation of study (for a review, see Son & Kornell, 2008). In other words, the extent to which an observer is accurate in his or her assessment of a memory's durability or robustness may constrain how appropriately he or she can allocate study to that particular information, given future goals. Thus, participants with poor metacognitive abilities who inaccurately predict their future performance would be assumed to poorly allocate study opportunities.

Whereas JOLs are prospective judgments, here we focus on immediate retrospective judgments of confidence during item recognition. When providing recognition confidence, participants are not predicting future performance but instead are assessing the robustness of current memory evidence in the current testing situation. Reporting confidence is clearly metacognitive, and the degree to which confidence and accuracy are reliably linked during recognition has implications for applied areas such as eyewitness identification. The fact that increasing recognition confidence usually corresponds to increasing accuracy (Roediger, Wixted, & DeSoto, 2012) suggests that observers have some insight into the nature of the underlying recognition evidence variable or variables, although recognition accuracy and confidence can be decoupled through various manipulations (Busey, Tunnicliff, Loftus, & Loftus, 2000; Chandler, 1994; Dobbins, Kroll, & Liu, 1998). For the current study, we were interested in the degree to which observers can subjectively report gradations in memory evidence through subjective confidence ratings on a trial-by-trial basis, and we take this as our operational definition of metacognitive monitoring during recognition judgment. The reason we are interested in this putative skill is because, as we discuss below, metacognitive monitoring skill may influence the ability of observers to adaptively bias their recognition judgments in the presence of partially diagnostic environmental cues.

Critically, outside the laboratory, recognition does not take place in a vacuum, and external or environmental factors often signal the likely memory status of encountered stimuli. For example, an approaching individual's likely memory status (familiar or novel?) may be signaled by the location (are you in a place where most people tend to be familiar?), time of day (are you likely to encounter familiar people at this time?), or a nearby friend's explicit opinion about whether or not he or she recognizes the approaching individual. Using such external cues is statistically ideal because they convey valuable statistical priors when one is making recognition judgments, and accurate metacognitive monitoring would presumably be important for modulating the degree to which an external cue influenced the current judgment. For example, an observer who accurately deems his or her internal evidence as only weakly favoring a recognition decision should be

Diana Selmeczy and Ian G. Dobbins, Department of Psychology, Washington University in St. Louis.

This research was supported by National Institute of Mental Health Grant MH073982.

Correspondence concerning this article should be addressed to Diana Selmeczy, Department of Psychology, Washington University in St. Louis, Campus Box 1125, One Brookings Drive, St. Louis, MO 63130-4899. E-mail: dselmecz@wustl.edu

much more heavily influenced by an external cue than an observer who accurately deems his or her internal evidence as highly diagnostic in the same situation. In contrast, if an observer is unable to effectively monitor internal evidence levels, he or she will be prone to over- or underreact to the cues in the environment. We will return to the potential dependence between efficient cue use and metacognitive monitoring ability after discussing the basic signal detection model of recognition judgment below.

To more concretely illustrate the utility of external cue use during recognition, it is necessary to more formally model the decision process. The most popular decision model applied to recognition judgments in whole or in part (Wixted & Mickes, 2010; Yonelinas, 2002) is signal detection theory (SDT; Macmillan & Creelman, 2005). Under the model, recognition evidence is assumed to be continuous and normally distributed for the studied and new classes of materials. The distance separating these evidence distributions corresponds to the familiarity accrued through recent study, and, because the evidence for the two classes overlaps, observers must place a decision criterion along the evidence axis in order the categorically rate items as old or new. The original model and several related variants currently in use assume that observers use statistically ideal information when placing this criterion (Glanzer, Hilford, & Maloney, 2009; Pastore, Crawley, Berens, & Skelly, 2003; Swets, Tanner, & Birdsall, 1961; Turner, Van Zandt, & Brown, 2011). Thus, they do not render judgments based on raw stimulus strength or intensity values; instead, they estimate the relative likelihoods that encountered strengths would arise under the hypothesis the item arose from one candidate distribution (e.g., old items in a recognition test) versus the other candidate distribution (e.g., new items in a recognition test), selecting the response reflecting the highest likelihood. More formally, this is captured by taking the ratio of the likelihoods, which are specified by the heights of the old and new item distributions at each point on the axis. In this case, the corresponding decision axis represents the odds of the item arising from the old relative to new item distributions, given its strength value (see Figure 1). The likelihood ratio decision model is statistically ideal because it incorporates all the information necessary to maximize preferred outcomes. For example, in order to maximize the total proportion of correct responses, the criterion should be placed at a likelihood ratio of 1, the point at which either alternative is equally likely given equal numbers of old and new test items. Alternatively, if a payout is imposed such that differential costs and/or benefits accrue for different judgment outcomes, the ideal likelihood ratio criterion can be easily recalculated to determine the location maximizing net winnings under this particular payout (Macmillan & Creelman, 2005).

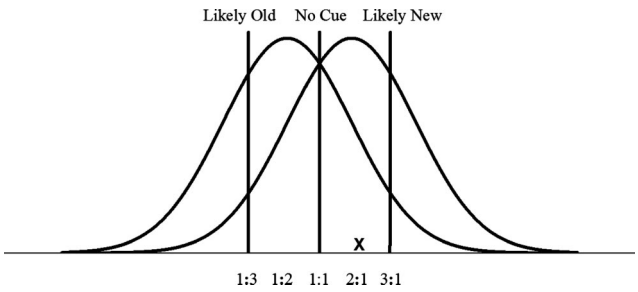

Figure 1. Optimal criteria shifts under external recommendations. The figure depicts how optimal criteria location shifts as a function of external recommendations under a likelihood ratio signal detection theory (SDT) model of recognition memory. The *x*-axis represents likelihood ratios, that is, the probability density of the target distribution divided by that of the lure distribution for each location. Under conditions with no external cue and equal numbers of old and new items, the ideal criterion location is at the intersection of the two distributions. Under conditions with a 75% valid *Likely Old* cue, the ideal observer should shift his or her decision criterion leftward to the point on the decision axis where the likelihood ratio is 1 to 3. Under conditions with a 75% valid *Likely New* cue, the ideal observer should shift his or her decision criterion rightward to the point on the decision axis where the likelihood ratio is 3 to 1. Notice that the response to the recognition strength indicated by an X changes depending on the cuing condition.

If one assumes the likelihood ratio SDT model, then integration of external cues, such as external recommendations, into memory judgments is straightforward provided the reliability of the source is known (cf. Jaeger, Lauris, Selmeczy, & Dobbins, 2012). For example, in the absence of any external cue the observer calculates 2 to 1 odds that a given encountered item was studied (i.e., the likelihood ratio is 2). This should serve as the basis for a reasonably confident "old" judgment. However, if he or she also received an external cue indicating there were 1 in 3 odds that the presented item would be old (i.e., the item is more likely new than old), the judgment should reflect both this prior and the internal evidence noted above. Under Bayes' rule this is done by simply multiplying the two odds ($2/1 \times 1/3 = 2/3$). In this case, despite the internal evidence favoring an old judgment, the ideal judgment is that the item is in fact new.

Under the likelihood ratio SDT model, this change in decision outcomes reflects shifts of the decision criterion. If the cue is 75% valid and indicates the upcoming item is likely old, the criterion shifts leftward to a point on the decision axis where the likelihood ratio is 1 to 3. This anticipates the high likelihood that the upcoming item is old, and it captures the Bayesian philosophy that decisions should accord with strongly predictive priors unless the current evidence overwhelmingly suggests otherwise. If the cue instead indicates the item is likely new, the criterion should shift to the point where the likelihood ratio is 3 to 1 (see Figure 1).

By optimally moving the decision criterion under external cuing, as described above, observers maximize their long-term accuracy and elevate their performance relative to situations where external cues are unavailable. For example, the maximum percentage of correct responses possible for an observer with a d' of 1.00 and no environmental cues is 69%. In contrast, if each item is preceded by a predictive cue that is known to be 75% valid, the maximum percentage of correct responses becomes 78%. Thus, the likelihood ratio SDT model concretely illustrates the beneficial nature of integrating external cues and internal evidence during recognition judgments through opportunistic shifting of a decision criterion. Observers who adaptively position the criterion on a trial-by-trial basis in response to external cues stand to gain considerably. Of course, doing so requires not only knowledge of the external cue's validity but also intimate knowledge of the quality of one's internal evidence. If the likelihood ratio variant of the signal detection model were literally accurate, every observer would be capable of optimally incorporating external recommendations provided the validity of those recommendations was known. However, there are several reasons to suspect that observ-

ers do not have explicit access to such statistically ideal information when making recognition judgments. First, it is clearly the case that most observers do not understand likelihood ratios, nor do they have an explicit model of the distributions of target and lure evidence sufficient to calculate likelihood ratio statistics (Hintzman, 1994). Second, they do not shift criteria to an appropriate extent given testwide payoff structures (Healy & Kubovy, 1978), nor do criteria shift sufficiently in response to increases in performance (Stretch & Wixted, 1998). Additionally, in the absence of feedback or explicit warnings, recognition criteria are heavily insensitive to the relative preponderance of old versus new probes present in the test list (e.g., Cox & Dobbins, 2011). Such considerations and findings suggest that the likelihood ratio evidence characterization is best viewed as an ideal solution, with observers varying considerably in how closely they can approximate this ideal approach.

In the current report, we test the idea that one limiting factor in this approximation is the varying degree to which individuals are aware of subtle trial-to-trial gradations in their internal memory evidence, namely, their metacognitive monitoring ability. An individual who is highly insensitive to gradations in internal evidence cannot have a very accurate estimate of the correspondence between internal evidence values and likelihood of success and hence would be prone either to shifting the criterion too little in response to external cues or, potentially just as ineffective, to shifting the criterion too much in the face of an external recommendation. To see why under- or overshifting the criterion is nonoptimal, it helps to graph the relationship between the size of criterion shifts, the internal resolution of the observer, and the validity of the external cue. Figure 2 shows this relationship for two hypothetical observers with d' values of 1.0 and 1.5. The *x*-axis indicates the size of the criterion shift (in standard deviation units – C) with respect to the intersection of evidence distributions. The *y*-axis is the maximum percentage correct given the size of the criterion shift, and the separate lines reflect the different external cue validities. The intercept of each line is the observer's maximum percentage correct in the absence of any criterion shift and, hence, his or her baseline, uncued recognition ability. Three things are apparent from the graphs. First, for any cue validity, the observer can improve performance relative to baseline by shifting the criterion. The maximum of the initially upward trajectory in every curve indicates the maximum possible improvement. Second, when the cue's reliability is higher than the participant's baseline performance, there is a diminishing return as the criterion shift increases. In other words, there is a peak level for shifting, after which the percentages begin to decline and asymptote at the level of the cue's reliability. Because performance is higher than baseline even when overresponding to the cues (i.e., shifting the criterion more than the maximum value indicated on the graph), this still constitutes a beneficial, although nonoptimal, strategy. Finally, when the cue's reliability is lower than the participant's baseline performance, there is a potential cost involved in shifting the criterion too much. As the shift increasingly surpasses the optimal point, performance will decrease to that of the cue, which is below the observer's baseline capabilities. Thus, the figure shows clearly that in order to effectively use external cues, observers must shift the criterion an appropriate degree. To do so requires knowledge about the relative reliability of one's internal evidence in relationship to the reliability of the external cue. In the current report we test the hypothesis that successful use of external

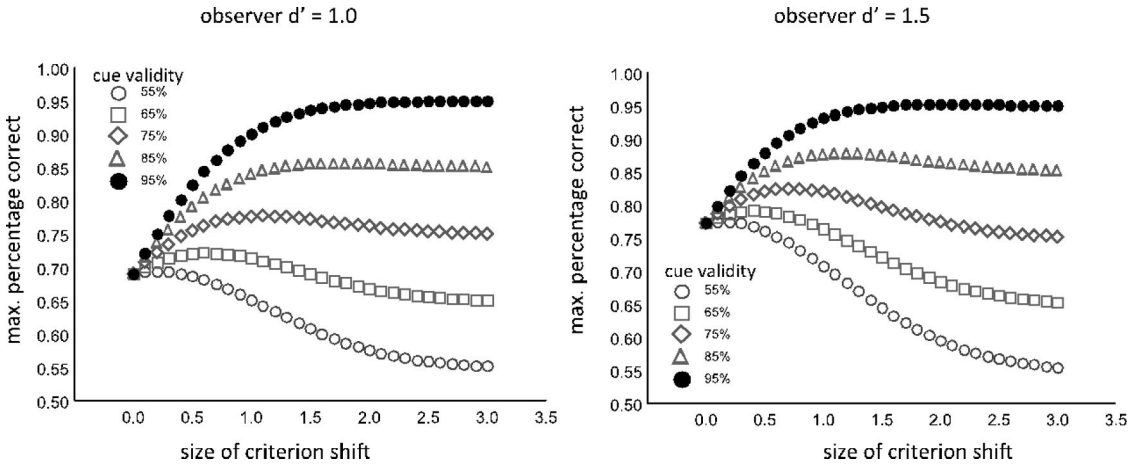

Figure 2. Optimal degree of criterion shift under varying cue validities. The graphs describe the performance benefit that can be achieved through criterion shifts across varying levels of cue validity for two levels of observer d' (1.0 to the left and 1.5 to the right). The *x*-axis represents the size of the criterion shift (in standard deviation units of C), the *y*-axis represents the maximum percentage correct during cued performance, and the lines represent varying cue validities. The maximum height of each line indicates the maximum possible improvement achievable for a given cue validity. When cue validity is higher than the observer's baseline accuracy, performance benefits will always be observed through shifts, even if the size of the criterion shift is nonoptimal. However, in this case, as the size of the shift increases there are diminishing returns in performance benefit. In contrast, when cue validity is lower than the observer's baseline accuracy, any criterion shifts after the optimal point result in performance declines relative to baseline accuracy. A color version of this figure is available online as supplemental material (http://dx.doi.org/10.1037/a0029469.supp).

cues depends in part upon metacognitive monitoring abilities. Before describing the paradigm testing this idea more fully, we briefly discuss one area of confusion that often arises when one talks about elevating accuracy through adaptive movement of a decision criterion.

Although the signal detection measure d' ideally provides a criterion-free estimate of accuracy—that is, d' values are (or should be) insensitive to shifts in the decision criterion—this characterization is true only under a very narrow set of conditions, namely, those in which the criterion remains fixed across the set of trials in question. For example, the two criterion positions illustrated in Figure 1 (under *Likely Old* and *Likely New* cues) yield identical d' values and can be thought of as two observers with identical resolution but different criteria taking the same recognition test. However, if the criterion is moving across the different trials of the test, the measured discrimination performance of the individual depends upon both the relative position of the evidence distributions (d') and the manner of the criterion movement across the individual test trials. As illustrated by Benjamin, Diaz, and Wee (2009; see also Dobbins & Han, 2007), random movement of the criterion across the trials will yield lowered measured performance relative to the hypothetical internal resolution. That is, d' calculated from hits and false alarms under a noisy, random criterion will be lower than that which was theoretically possible given the internal evidence distributions of the observer and a fixed criterion.

To see why this is the case, consider a massive random movement from trial to trial of plus or minus 4 standard deviation units (with respect to the intersection of the distributions) for an observer with an internal d' of 1.0. On trials in which the movement is leftward, the observer will perfectly identify all of the old items presented on those trials (hit rate = 1) but will also incorrectly classify all of the new items as "old" (false alarm rate = 1), because of the extreme leftward location of the criterion. However, this only constitutes half of the total test trials. For the remaining half, there are extreme rightward criterion shifts, which will yield zero hit and false alarm rates because of the extreme rightward criterion position. Thus, the measured net hit rate for the entire test in which there were both leftward and rightward movements is .5(1) + .5(0) and the measured false alarm rate is .5(1) + .5(0). Because the net hit and false alarm rates are both .5, this leads to an accuracy estimate, d' of essentially 0, far below the internal resolution of 1.0. Of course, less extreme random variation will yield lower performance costs, but the point is the same; random movements from trial to trial lower measured performance. These movements are "random" because they are completely uncorrelated with the test status of the items on each trial. In other words, one cannot predict the location of the criterion given the knowledge that an item is studied or novel.

Now consider if the criterion movement is not random but is actually informed by a reliable external cue or recommendation on each trial. Taking the extreme case again, consider a cue that is 99% valid. Here, during an old recommendation by the cue (which occurs on half of the trials), the observer shifts the criterion extremely leftward (because of the high odds in favor of an old item, given the high cue validity). For this half of the trials, the appropriate SDT equations yield a hit rate of .9997 and a false alarm rate of .9927. The fact that both values are essentially 1 reflects the extreme leftward location of the optimal criterion position given the high cue validity. On the remaining half of the trials in which the cue recommends a new response, the hit and false alarm rates are approximately .0073 and .0003, respectively. These values are extremely low because the criterion location is now positioned to the extreme right. Critically, however, unlike the random movement case above, the cue recommendations and hence the criterion shifts are not independent of the actual status of the items, because the cues are highly valid. Thus, the testwide hit rate is the proportion of times the cue recommends an old response times the hit rate under that scenario [.99(.9997)] plus the proportion of times the cue invalidly recommends a new response times the hit rate under that scenario [.01(.0073)], or about .99. The false alarm rate is the proportion of times the cue invalidly recommended an old response times the false alarm rate under that scenario [.01(.9927)] plus the proportion of times the cue validly recommended a new response times the rate at which the observer nonetheless incorrectly classified the item as new [.99(.0003)], or about .01. These proportions yield a d' estimate of ~4.6, which is well in excess of the observer's internal resolution in this example.

The gains that are possible are less extreme as cue validity decreases, but the general point holds. If the observer positions the criterion adaptively on every trial of a test, measured performance will exceed internal resolution. This simply reflects the fact that two sources of reliable information (cues + internal evidence) optimally combined will necessarily yield higher performance than one source (internal evidence) in isolation. Although the observer is expected to make a high rate of errors on invalidly cued trials, these occur extremely infrequently as the cue becomes increasingly valid; thus, they play a more minor role in the overall rates for the entire test. For example, in the case of a 99% valid cue and a test consisting of 100 old and 100 new items, there will be exactly 2 trials in which the cue provides an incorrect recommendation, whereas the cue will provide the correct answer on 198 trials. Clearly then, the costs of inappropriate criterion shifts for 2 trials will be largely outweighed by the gains of appropriate criterion shifts on 198 trials, as illustrated in the example above.[1]

Thus, the SDT model and Bayesian reasoning dictate that if external cues are available and reliable, observers should use them to adaptively bias their decisions by combining both the cue and the internal evidence (i.e., by appropriately shifting decision criteria). However, to the degree that observers differ in the ability to assess the quality of their own internal memory evidence, there will be variation in how effectively they can incorporate the external cues. As mentioned previously, if one does not know the quality of one's own memory evidence, it is impossible to consistently incorporate external recommendations in order to bolster performance. Couched in the language of the likelihood ratio SDT framework, an observer with poor metacognitive monitoring is one who does not have reliable estimates of the relative likelihoods that memory strength signals originated from old versus new item

[1] For simplicity, both examples of criterion movement used an observer with a fixed d' of 1.0. However, the relative costs (of random movements) and benefits (of informed movements) are also dependent upon the accuracy of the observer. For observers with increasing internal resolution, both the costs and the benefits of criterion movement are smaller than those for observers whose internal resolution is poor.

classes and who hence cannot shift the criterion to the adaptive positions in response to external recommendations.

Below we test this general idea in two experiments, in which we compare trials where valid external anticipatory cues are available and trials where they are not (uncued). Within the paradigms we calculate a measure of metacognitive monitoring during the uncued trials based on the hypothesis that observers with higher metacognitive monitoring ability will also demonstrate a greater ability to utilize the external cues when available. In Experiment 1, we manipulated levels of processing during encoding in order to assess whether participants are able to benefit from external cues under varying levels of memory evidence. Under the signal detection model, the optimal criterion shift in units of standard deviation (C measure) should be smaller when internal evidence is accurate than when it is inaccurate. For example, with a 75% cue, an observer should shift the criterion 1.10 units under situations in which internal d' is 1.0 and should shift it 0.73 units under conditions in which d' is 1.5. This simply reflects the idea that as internal resolution increases, external cues should become less influential. In Experiment 2, we tested the complementary idea by manipulating the level of cue validity in order to assess whether participants are able to differentially benefit from varying cue validities. For a fixed level of internal resolution, one should react more strongly as the validity of the external cue increases.

Additionally, we consider whether the presence or absence of feedback helps in this ability, given that prior work has suggested that trial-by-trial feedback may be necessary for accurate representations of statistical likelihoods (Turner et al., 2011) and that feedback results in more appropriate criterion placement (Estes & Maddox, 1995; Kantner & Lindsay, 2010; Rhodes & Jacoby, 2007; Verde & Rotello, 2007). This prior work implies that feedback may be crucial for participants to instantiate a criterion shift, presumably because feedback serves to help participants realize that a manipulation is present and they need to change their pattern of responding. In our experiments, feedback may not be pertinent to inform participants of the need to shift their criteria, because participants already have explicit knowledge of cue validity and we encourage them to incorporate cues. Instead, feedback may serve to help participants fine-tune this process. As is evident in Figure 1, the size of the appropriate criterion shift is determined by relative balance of internal versus external reliability, with the potential for observers to shift too much or too little given their internal resolution and the external cue's validity (see Figure 2). If they were under- or overresponsive to the external cue, feedback may help appropriately lessen or increase the size of the shifts. Thus, we hypothesized that feedback might result in more appropriate criterion shifts in response to the external recommendations and, hence, a somewhat greater improvement in performance when one is comparing uncued to cued recognition accuracy.

Experiment 1

Method

Participants. Experiment 1 included 37 Washington University students (average age = 20.9 years, 23 female) who were paid $20 for participation. Three participants were removed due to low performance ($d' < 0.19$), leaving 34 participants for analyses. Although we removed participants with near-chance performance under the assumption that they were unlikely to be engaged during the task, all results still hold when low performers are included in the analysis. All participants provided informed consent in accordance with the university's institutional review board.

Materials and procedure. Testing was self-paced, with observers entering their responses via keyboard and presentation and timing controlled via Matlab's Psychophysics Toolbox (Version 3.0.8; Brainard, 1997; Pelli, 1997). For each participant, words were randomly selected from a 1,216-item pool, with an average of 7.09 letters and 2.34 syllables and a Kučera–Francis frequency of 8.85.

We used a 2 × 2 × 2 mixed design with repeated factors of levels of processing (deep vs. shallow targets present during test) and cue condition (cued vs. uncued) and a between-subjects factor of feedback (present vs. absent). Participants completed four study/test cycles, with two tests preceded by deep encoding and two tests preceded by shallow encoding. The order of deep and shallow tests sequentially alternated, with half the participants beginning with the shallow test condition and half of them beginning with the deep test condition (100 study items and 200 test items for each cycle). During shallow encoding, participants indicated whether the first and last letter of each presented word were in alphabetical order, whereas during deep encoding they performed an abstract/concrete rating. Recognition testing immediately followed each study phase, with participants indicating whether randomly intermixed old and new items were studied ("old") or novel ("new"; 100 old items, 100 new items). On 120 of the test trials (60 old, 60 new) a probabilistic mnemonic cue, *Likely Old* or *Likely New,* was presented 1 s before the probe word appeared. These cues were correct 75% of the time, with participants correctly informed that "cues will be correct 75% of the time. This means about 7 out of 10 times the cue will give you the correct answer and should be useful for your recognition judgment." In addition to the cued trials, there were 80 (40 old, 40 new) uncued trials intermixed in the test phase, with participants notified that some portion of the probes would be presented without anticipatory cues. After each old/new recognition decision, participants provided confidence on a 6-point scale ranging from 50% (guessing) to 100% (certain) in 10% intervals. Corrective feedback immediately followed for half the participants.

Results and Discussion

The order in which the two levels of processing conditions were administered did not influence accuracy (d') or criteria (C), nor did it interact with other factors. Given this, we collapsed across test order in the analyses below. Hit rates of 1 and false alarm rates of 0 were corrected with the formulas suggested by Macmillan and Creelman (2005), $1 - 1/(2N)$ for hits and $1/2N$ for false alarms, where N is the number of trials.

Does accuracy improve with provision of cues? Descriptive statistics demonstrating performance and confidence as a function of cuing condition are presented in Tables 1 and 2.

To assess potential gains in accuracy (d') as a function of cue condition, we used a 2 × 2 × 2 mixed analysis of variance (ANOVA) with repeated-measures factors of levels of processing (deep vs. shallow targets present during test) and cue condition (cued vs. uncued) and a between-subjects factor of feedback (present or absent). Results revealed a main effect of levels of process-

Table 1
Accuracy (d′) and Average Response Rates (Hits and False Alarms) Under Uncued and Cued Conditions With Standard Deviations in Parentheses

	Uncued			Cued		
Condition	*d′*	Hits	False alarms	*d′*	Hits	False alarms
Experiment 1						
Shallow	0.91 (0.39)	0.65 (0.13)	0.32 (0.10)	1.17 (0.37)	0.69 (0.10)	0.26 (0.08)
Deep	2.09 (0.66)	0.87 (0.20)	0.22 (0.10)	2.32 (0.62)	0.89 (0.07)	0.18 (0.08)
Experiment 2						
65% predictive	1.20 (0.40)	0.73 (0.09)	0.29 (0.10)	1.30 (0.48)	0.73 (0.11)	0.27 (0.10)
85% predictive				1.65 (0.39)	0.79 (0.10)	0.21 (0.06)

ing, $F(1, 32) = 174.32$, $MSE = 0.26$, $p < .001$, reflecting higher accuracy for deep than shallow tests. There was also a main effect of cue condition, $F(1, 32) = 36.65$, $MSE = 0.06$, $p < .001$, indicating that participants significantly improved performance on cued versus uncued trials (see Table 1). There was no main effect of feedback, $F(1, 32) = 0.32$, $MSE = 0.74$, $p = .57$, suggesting that feedback did not influence overall accuracy. The interaction between feedback and cue condition was also not significant, $F(1, 32) = 1.09$, $MSE = 0.06$, $p = .30$, indicating that feedback did not have an appreciable effect on cuing benefit. This result was unexpected, as we hypothesized that participants who received feedback may gain additional information about their performance, resulting in more ideal criterion shifts and hence greater benefits from external recommendations. None of the remaining two-way interactions were significant, and the three-way interaction also failed to reach significance.

Overall, these analyses demonstrate that participants increased their accuracy on cued trials relative to uncued trials for both deeply and shallowly encoded items, and this improvement in performance was not altered by the provision of feedback. Thus, although participants are effectively incorporating the cues into their judgments, the mechanism by which this occurs does not appear to require or benefit from feedback-based learning. We further consider the inefficacy of feedback in the discussion.

Reactivity to cues. Because the accuracy analysis demonstrates that observers are improving when cues are in the environment, it is clearly the case that these cues are being used to adjust decision standards (see Table 3 for descriptive statistics). Nonetheless, we wanted to verify that observers were shifting criteria more vigorously during shallow tests than during deep tests, because this pattern should result if the cues are being considered in light of the internal recognition evidence. That is, the cues should have more influence when the internal evidence is less discriminable (shallow tests) than when it is more discriminable (deep tests). Using C as our criterion measure, we conducted a $2 \times 2 \times 2$ mixed ANOVA with repeated measures of levels of processing (deep vs. shallow) and cue type (*Likely Old* vs. *Likely New*) and a between-subjects measure of feedback (present or absent). Results revealed a main effect of levels of processing, $F(1, 32) = 36.55$, $MSE = 0.08$, $p < .001$. However, we do not interpret this finding directly because the interpretation of criterion differences across conditions in which accuracy is also vastly different is highly dependent on the particular measure of criterion employed (Pastore et al., 2003). As anticipated, there was main effect of cue type, $F(1, 32) = 70.47$, $MSE = 0.23$, $p < .001$, suggesting that observers responded more liberally under *Likely Old* than *Likely New* cues. The main effect of feedback was not significant, $F(1, 32) = 36.55$, $MSE = 0.18$, $p = .80$, indicating that feedback did not influence overall criterion placement. Critically, a significant interaction was found between levels of processing and cue type, $F(1, 32) = 21.34$, $MSE = 0.06$, $p < .001$, demonstrating that the difference in criterion across *Likely Old* and *Likely New* cue trials was greater for shallow tests than deep tests. The remaining two-way interactions were not significant, and the three-way interaction among levels of processing, cue type, and feedback was also not significant, $F(1, 32) = 0.094$, $MSE = 0.06$, $p = .76$. These results suggest that participants' absolute shifts in criteria are greater under conditions where internal memory resolution is lower,[2] which confirms that the cues' influence is larger for the condition with poorer internal recognition evidence. Furthermore, the size of these criteria shifts are not feedback dependent, again demonstrating that the use of cues does not seem to benefit from feedback-based learning.

Individual differences in efficacy of cue use. Although on average accuracy benefited from cuing, there were large individual differences in the degree of improvement. As noted in the introduction, the effective use of external cues may critically depend upon metacognitive monitoring. To examine the role of monitoring we used the gamma index, which captures the correspondence between changes in subjective confidence and changes in accuracy at the trial-by-trial level for each participant (Nelson, 1984). Because gamma has a restricted range, we used the logit transformation of gamma (G*) to improve its scale properties (Benjamin & Diaz, 2008).

If metacognition plays a role in cue utilization skill and is not a simple alternative measure of uncued observer accuracy, hierarchical regression analysis should demonstrate that it makes a significant and unique contribution to cued performance when uncued performance has been appropriately partialed from the data. Therefore, we examined if metacognitive awareness explains

[2] Although comparing criteria under different levels of accuracy can be problematic, we are interpreting the interaction only in terms of the absolute difference in criteria between the two cuing conditions for deep and shallow test items. In other words, we are interpreting only the absolute shift between *Likely Old* and *Likely New* cues and are not making direct comparisons about relative criterion locations across different accuracies.

Table 2
Average Confidence for Uncued and Cued Conditions With Standard Deviations in Parentheses

	Uncued				Cued			
Condition	Hits	Misses	Correction rejections	False alarms	Hits	Misses	Correction rejections	False alarms
Experiment 1								
Shallow	82.7 (9.7)	72.4 (11.2)	75.2 (11.6)	72.8 (10.8)	80.6 (8.9)	72.2 (11.1)	77.4 (11.1)	71.6 (11.2)
Deep	94.5 (5.0)	76.8 (10.4)	82.9 (9.9)	78.6 (11.4)	93.8 (5.6)	76.2 (11.3)	84.2 (9.7)	75.4 (11.6)
Experiment 2								
65% predictive	85.2 (9.5)	73.8 (10.9)	78.11 (11.0)	74.0 (10.6)	84.7 (9.9)	72.9 (9.7)	78.6 (10.5)	73.0 (10.4)
85% predictive					86.5 (9.2)	74.2 (10.4)	80.9 (10.5)	74.4 (10.0)

any unique variance in cued performance that is nonoverlapping with uncued recognition skill. Using a hierarchical regression analysis, we examined the contribution of metacognition to cued accuracy (d') by entering feedback condition (dummy coded) and uncued recognition accuracy as predictors in Step 1. Next, we examined whether metacognitive monitoring made a contribution beyond these factors by entering each participant's G* as an additional predictor in Step 2. Critically, G* was calculated from uncued performance and is therefore a measure of metacognition in the complete absence of cues. Table 4 shows simple correlations and Table 5 shows the results of the two hierarchical regressions that were separately conducted for the shallow and deep tests. For the shallow test, in Step 1 uncued accuracy was a significant predictor, $B = 0.60$, $t(30) = 5.02$, $p < .001$, of cued accuracy, but feedback group was not, $B = -0.04$, $t(30) = -0.47$, $p = .64$. It is not surprising that participants with high accuracy in the uncued condition would also have high accuracy under cuing. However, the clear absence of any contribution of the feedback variable serves again to underscore the fact that the provision of feedback has no appreciable influence on the manner in which participants use the cues in this paradigm (see ANOVA results above). Entering G* in Step 2 explained an additional 7.37% variance, $F(1, 30) = 4.72$, $p = .04$. When the hierarchical regression was repeated for the deep test list condition, a similar pattern emerged. In Step 1, uncued accuracy was a significant predictor of cued accuracy, $B = 0.77$, $t(29) = 8.40$, $p < .001$, and feedback group was not, $B = -0.16$, $t(29) = -1.28$, $p = .21$, and during Step, 2 G* accounted for an additional 5.08% of unique variance in cued performance, $F(1, 29) = 6.00$, $p = .02$.

The regression analyses demonstrate that cued performance is linked with uncued accuracy but, more important, that after controlling for uncued performance, metacognitive monitoring is a significant predictor of cued performance gains. Again, the provision of feedback had little influence on the effective use of the external cues. These results hold for both shallowly and deeply encoded items. For two reasons, the failure of feedback to improve cue utilization is unlikely simply the result of low power. First, the null effect of feedback group replicated across both deep and shallow tests. Second, the numerical direction of the effect is opposite that predicted under the idea that feedback would improve performance. That is, the beta weights in both regressions were negative, indicating that cued performance was actually slightly numerically worse for the feedback group (with other factors controlled). This suggests that even large increases in sample size would still fail to show an actual benefit for the provision of feedback. In summary, Experiment 1 demonstrates that participants are able to benefit from external cues for both deep and shallow test items, individual differences in cue utilization skills are in part related to metacognitive monitoring ability, and feedback does not improve cue utilization skill.

Table 3
Criteria (C) and Accuracy (d′) Under Likely Old and Likely New Cues With Standard Deviation in Parentheses

	Likely old		Likely new	
Condition	C	d'	C	d'
Experiment 1				
Shallow	−0.37 (0.40)	0.83 (0.47)	0.51 (0.34)	0.83 (0.50)
Deep	−0.47 (0.38)	1.96 (0.60)	0.01 (0.35)	2.16 (0.61)
Experiment 2				
65% predictive	−0.27 (0.34)	1.14 (0.52)	0.27 (0.26)	1.26 (0.55)
85% predictive	−0.42 (0.43)	1.17 (0.50)	0.41 (0.09)	1.17 (0.50)

Experiment 2

Method

For Experiment 2 we wanted to replicate our results from Experiment 1 and examine whether participants are able to effectively differentiate between cues of differing validity (65% and 85% predictive). We again also examined whether cued performance is, in part, dependent on metacognitive monitoring.

Table 4
Simple Correlations (R) for Shallow and Deep Encoding

Variable	Cued d'	Uncued d'	Feedback	G*
Shallow encoding				
Cued d'	—			
Uncued d'	.69***	—		
Feedback	−.17	−.17	—	
G*	.68***	.70***	−.26	—
Deep encoding				
Cued d'	—			
Uncued d'	0.83***	—		
Feedback	−.10	.02	—	
G*	.58***	.42*	−.28	—

*$p < .05$. ***$p < .001$.

Table 5
Hierarchical Regression Analysis Predicting Cued Accuracy (d′) for Shallow and Deep Encoding

Variable	*B*	*SE B*	β	R^2	ΔR^2
Shallow encoding					
Step 1					
Uncued recognition (*d′*)	0.60	0.12	.67***		
Feedback	−0.04	0.09	−.06	.47***	
Step 2					
Uncued recognition (*d′*)	0.37	0.18	.41*		
Feedback	−0.01	0.12	−.01		
Metacognitive ability (G*)	0.95	0.43	.39*	.55***	.07*
Deep encoding					
Step 1					
Uncued recognition (*d′*)	0.77	0.09	.84***		
Feedback	−0.16	0.12	−.12	.71***	
Step 2					
Uncued recognition (*d′*)	0.67	0.09	.73***		
Feedback	−0.06	0.10	−.05		
Metacognitive ability (G*)	0.69	0.28	.26*	.76***	.05*

* $p < .05$. *** $p < .001$.

Participants. Experiment 2 included 38 Washington University students (average age = 21.5 years, 18 female) who were paid $20 for participation. Three participants were removed due to near-chance performance ($d' < 0.19$), leaving 35 participants for analyses. Once again, the exclusion of low performers does not change our overall findings. All participants provided informed consent in accordance with the university's institutional review board.

Materials and procedure. Testing was self-paced, with observers entering their responses via keyboard and presentation and timing controlled via Matlab's Psychophysics Toolbox (Version 3.0.8; Brainard, 1997; Pelli, 1997). For each participant, words were randomly selected from a 1,216-item pool, with an average of 7.09 letters and 2.34 syllables and a Kučera–Francis frequency of 8.85.

We used a 3 × 2 mixed design with a repeated-measures factor of cue condition (uncued, 65% predictive cue, 85% predictive cue) and a between-subjects factor of feedback (present vs. absent). Participants completed four study/test cycles (100 study items each) during which the encoding task was syllable counting (1, 2, 3, or more syllables?). During each recognition test a total of 160 (80 old, 80 new) words were preceded by a probabilistic mnemonic cue (*Likely Old* or *Likely New*) 1 s before the word probe appeared. Cue predictability varied for this experiment, where half the cues were 65% predictive (40 old, 40 new) and half the cues were 85% predictive (40 old, 40 new). Participants were clearly informed of the two different cue validities. The 65% predictive cues were presented in a smaller blue font with the number 65 appearing next to the cue. The 85% predictive cues were presented in a larger yellow font with the number 85 appearing next to the cue. Instructions stated, "Cues that are 65% correct will give you the correct answer about 6 out of 10 times. Cues that are 85% correct will give you the correct answer about 8 out of 10 times. Use the cues to help increase your performance." In addition to the cued trials, there were 40 (20 old, 20 new) uncued trials intermixed in the test phase, with participants notified that some portion of the probes would be presented without anticipatory cues. Following each recognition decision participants rated confidence on a 6-point scale ranging from 50% (guessing) to 100% (certain) in 10% intervals. Corrective feedback followed for half the participants.

Results and Discussion

Does accuracy improve with provision of cues? To assess gains in accuracy (*d′*), we used a 3 × 2 mixed ANOVA with a repeated-measures factor of cue condition (uncued, 65% predictive cue, 85% predictive cue) and a between-subjects factor of feedback (present or absent). Results revealed a significant main effect of cue condition, $F(2, 66) = 44.13$, $MSE = 0.05$, $p < .001$; no significant effect of feedback, $F(1, 33) = 1.18$, $MSE = 0.45$, $p = .28$; and no significant interaction between cue condition and feedback, $F(2, 68) = 0.54$, $MSE = 0.05$, $p = .58$. Follow-up post hoc tests on the main effect of cue condition demonstrated that relative to uncued trials, there was a significant increase in performance on 85% predictive cued trials ($MSE = 0.04$, $p < .001$) and a numeric though unreliable improvement on 65% predictive cued trials ($MSE = 0.05$, $p = .17$; see Table 1). Overall, these results demonstrate that participants can benefit from the use of cues, even when two differing levels of cue validity are intermixed. It is not necessarily surprising that performance does not significantly improve with the 65% predictive cues, as these cues are not highly accurate. However, replicating the results from Experiment 1, we find that when cues are highly predictive, participants are able to improve their performance. Furthermore, we again see that cuing benefits do not increase with the provision of corrective feedback.

Reactivity to cues. We wanted to examine if observers were shifting criteria more vigorously during highly predictive cues (85%), because this pattern should result if participants appropriately consider the relative predictability of the two cue levels. That is, as the 85% cues are accurate more often, we would expect participants respond more vigorously to these more valid cues. With criterion measure C as our dependent variable, we ran a 2 × 2 × 2 mixed ANOVA with repeated measures of cue type (*Likely Old* vs. *Likely New*) and cue condition (65% predictive cue, 85% predictive cue) and a between-subjects factor of feedback (absent or present). Results revealed a main effect of cue type, $F(1, 33) = 70.68$, $MSE = 0.23$, $p < .001$, indicating that observers responded more liberally under *Likely Old* than *Likely New* cues. The main effects of cue condition, $F(1, 33) = 0.01$, $MSE = 0.02$, $p = .92$, and feedback, $F(1, 33) = 0.03$, $MSE = 0.18$, $p = .97$, were not significant. Importantly, there was a significant interaction between cue type and cue condition, $F(1, 33) = 29.67$, $MSE = 0.02$, $p < .001$, showing a greater difference in criterion locations for 85% predictive cues than 65% predictive cues (see Table 3). As before, the three-way interaction among cue type, cue condition, and feedback was not significant, $F(1, 33) = 0.007$, $MSE = 0.02$, $p = .94$. These results suggest that participants are in fact more influenced by highly predictive cues than less predictive cues, and this relationship is not affected by feedback.

Individual differences in efficacy of cue use. Although participants as a whole increased their cued performance relative to uncued performance, there were once again large individual differences in cuing benefits. We wanted to replicate results from

Experiment 1 and demonstrate that metacognitive monitoring contributes unique variance to cued performance above and beyond uncued performance. To examine this, we ran a separate hierarchical regression analysis on 65% predictive cued performance and 85% predictive cued performance, with uncued recognition accuracy (d') and feedback as predictors in Step 1 and metacognitive monitoring (G*) as a predictor in Step 2. Table 6 shows the simple correlations and Table 7 shows the results for the two separate hierarchical regression analyses conducted for 65% and 85% predictive cues. Because the two different cue validities were intermixed with uncued trials, both analyses use the same measure for uncued recognition and for metacognition (which is again determined from uncued confidence reports). In Step 1 for 65% predictive cued performance, uncued accuracy was a significant predictor, $B = 1.04$, $t(32) = 9.70$, $p < .001$, and feedback was not, $B = -0.10$, $t(32) = -1.16$, $p = .25$. After controlling for uncued performance and feedback, metacognitive ability explained an additional 7.70% of the variance in cued performance, $F(1, 31) = 14.02$, $p < .001$. Similar results were found when using 85% predictive cued performance, where again in Step 1 uncued performance was a significant predictor, $B = 0.59$, $t(32) = 4.50$, $p < .001$, and feedback was not, $B = -0.18$, $t(32) = -1.73$, $p = .09$. In Step 2, metacognitive ability explained an additional 7.74% of the variance in cued performance, $F(1, 31) = 4.90$, $p = .03$.These results demonstrate that although uncued recognition skill is related to cued performance, there is additional unique variance explained by metacognitive monitoring. These results hold for both 65% and 85% predictive cues and are not affected by feedback. Thus, replicating results from Experiment 1, we found that metacognitive monitoring is a significant predictor in cued performance above and beyond uncued accuracy, and corrective feedback does not seem to influence this relationship.

The current design also afforded the opportunity to ask a slightly different question about the reliability of cue utilization skill across observers. If this represents a moderately stable skill that varies across individuals, one would expect that an individual who is good at extracting useful information from cues that are 65% valid would also be fairly good at doing so when cues are 85% valid. Likewise, someone who is poor under one condition should be poor under the other. We tested this idea in two ways, first using the equal variance signal detection model and second using regression in a somewhat similar fashion to the analyses reported above.[3] Under the signal detection model, given a baseline d' reflecting the observer's skill in the absence of external cues and an external cue with a fixed validity (e.g., 75%), one can work out the expected measured d' if the observer ideally integrates the cues and internal evidence. For example, consider an observer with a baseline d' of 1.0 and an external cue that is 75% valid. Under the signal detection model, the observer should shift the criterion to a location on the evidence axis representing a likelihood ratio of 3 to 1 whenever the cue indicates *Likely New* and of 1 to 3 whenever the cue indicates *Likely Old* (Macmillan & Creelman, 2005). If the baseline d' value is 1.0, ideal criterion placement in response to the cue will yield a d' value of 1.52, which is the maximum possible accuracy achievable when combining the cue and the evidence under the signal detection model. Figure 3 shows the relationship between baseline d' and the maximum possible d' under cuing for the two cue validities used in Experiment 2 across a range of baseline d' values. Returning to the question of individual differences in cue-use skills, one can use these ideal values as benchmarks by taking the maximum d' possible under the two cue validities (given that participant's baseline d') and dividing each of these ideal values by the d' scores that participant actually achieved under the two cuing conditions. A participant who is near ideal in one instance should be near ideal in the other if he or she possesses a stable cue integration skill. When we conducted this analysis on the current data, the two proportional scores were highly correlated ($r = .70$, $p < .001$) suggesting a stable skill across the cuing conditions.

The above analysis is wholly contained within the signal detection measurement model. It does not require calculating gamma and so avoids some of the criticisms of that particular statistic. However, given that the equal variance model is generally assumed to be only a rough approximation of recognition decision making (e.g., Wixted & Mickes, 2010; Yonelinas, 2002), it is useful to consider cue utilization reliability using a different ap-

Table 6
*Simple Correlations (*R*) for 65% and 85% Predictive Cues*

Variable	Cued d'	Uncued d'	Feedback	G*
65% predictive cues				
Cued d'	—			
Uncued d'	.86***	—		
Feedback	−.17	−.07	—	
G*	.62***	.42*	−.10	—
85% predictive cues				
Cued d'	—			
Uncued d'	.62***	—		
Feedback	−.28	−.07	—	
G*	.52**	.42*	−.10	—

* $p < .05$. ** $p < .01$. *** $p < .001$.

Table 7
*Hierarchical Regression Analysis Predicting Cued Accuracy (*d′*) for 65% Predictive Cues and 85% Predictive Cues*

Variable	*B*	*SE B*	β	R^2	ΔR^2
65% predictive cues					
Step 1					
Uncued recognition (d')	1.04	0.12	.85***		
Feedback	−0.10	0.08	−.10	.75***	
Step 2					
Uncued recognition (d')	0.88	0.08	.73***		
Feedback	−0.07	0.07	−.08		
Metacognitive ability (G*)	0.79	0.21	.31***	.82***	.07***
85% predictive cues					
Step 1					
Uncued recognition (d')	0.59	0.13	.60***		
Feedback	−0.18	0.10	−.23	.43***	
Step 2					
Uncued recognition (d')	0.46	0.14	.47**		
Feedback	−0.16	0.13	−.21		
Metacognitive ability (G*)	0.64	0.29	.31*	.51***	.08*

* $p < .05$. ** $p < .01$. *** $p < .001$.

[3] We thank an anonymous reviewer for suggesting this analysis.

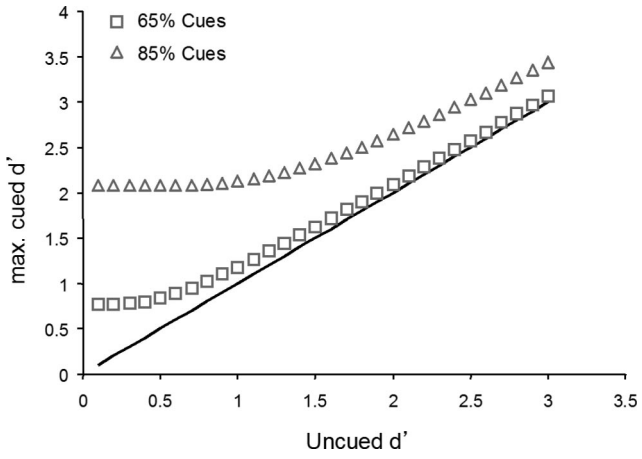

Figure 3. Maximum d' under 65% and 85% predictive cues with optimal criterion shifts. The *x*-axis represents uncued or baseline d' and the *y*-axis represents the maximum cued d' possible under 65% predictive cues (squares) and 85% predictive cues (triangles) under an equal variance signal detection model. A color version of this figure is available online as supplemental material (http://dx.doi.org/10.1037/a0029469.supp).

proach in order to make sure the conclusions are not model specific. To do so we returned to multiple regression. If an observer has a stable cue utilization skill, there should be a relationship between cued performance under the 65% validity condition and the 85% validity condition that is statistically independent of the baseline performance of the observer. That is, there should be some process that allows an observer to more or less effectively incorporate cues that is not wholly determined by whether or not the observer is simply more or less accurate in the absence of external cues. To assess this, we simply took each participant's baseline d' and his or her effective d' under the 85% valid cue condition and used these to predict performance under the 65% valid cue condition. The results of this hierarchical regression are shown in Table 8. Critically, after controlling for uncued recognition in Step 1, cued accuracy under 85% valid cues is a significant predictor and explains an additional 10.1% of the variance, $F(1, 32) = 20.4$, $p < .001$, in cued accuracy under 65% valid cues. Supporting the signal detection analysis above and the gamma statistic findings, performance in the 85% cue validity condition is predictive of performance in the 65% cue validity condition even when baseline accuracy has been partialed out. In other words, there is a characteristic of observers, apart from baseline recognition ability, that results in a stable relationship across cuing conditions.

Overall, the data converge with those of Experiment 1 in suggesting that the ability to effectively use external cues is partially dependent upon metacognitive monitoring skill demonstrated in the absence of external cues during baseline recognition. Experiment 2 replicated this basic finding across cues of widely differing validity (65% and 85%), and it provided three demonstrations that converge in suggesting that this reflects a stable skill, at least across intermixed cuing conditions within a single experiment. First, in the analyses using baseline accuracy and gamma to predict cued performance (see Table 7), gamma resulted in a highly similar increase in variance accounted for across the two cuing conditions (and similar regression coefficients). Thus, its explanatory power was stable across the two cuing conditions, which suggests it was tracking a stable characteristic of the observers. Additionally, the gamma statistic is calculated entirely from uncued, baseline recognition performance and thus forms an index of metacognitive monitoring that is statistically independent of actual cued performance. Second, the signal detection analysis that computed the proportion of optimal d' achieved under the two cuing conditions revealed a highly significant correlation in these values across the two cuing conditions. Finally, the regression analysis that used baseline and 85% valid cued performance to predict 65% valid cued performance revealed a stable relationship between cued performance under the two different cue validities even when baseline accuracy was statistically controlled. Thus, the three approaches for estimating reliability of metacognitive monitoring skill all converged on the same result. This suggests that the ability to use external cues to elevate performance during recognition is a skill that varies across individuals and that similarly contributes to performance under different cue validities.

General Discussion

Our study examined the ability of observers to integrate reliable external recommendations into recognition judgments. Before discussing our main results, we want to make note of another literature, referred to as memory conformity, which also examines how observers are influenced by external sources of information. Memory conformity studies generally have a confederate intentionally provide misinformation on a subset of trials, while the participant is led to believe that he or she studied the same material as the confederate. The overall finding from this memory conformity research is that people's decisions are in fact influenced by others' responses (Allan & Gabbert, 2008; Axmacher, Gossen, Elger, & Fell, 2010; Betz & Skowronski, 1996; Meade & Roediger, 2002; Reysen, 2005; Roediger, Meade, & Bergman, 2001; Schneider & Watkins, 1996; Walther et al., 2002; Wright, Gabbert, Memon, & London, 2008; Wright, Mathews, & Skagerberg, 2005; Wright, Self, & Justice, 2000). These prior studies focus on the negative aspect of conformity, mainly that performance is decreased when participants are given inaccurate external information. Implications from memory conformity research are especially important when the goal is to minimize external influences, such as eyewitness testimony situations where the goal of the legal system is to preserve the original fidelity of the observer's remembrances (not to maximize discrimination accuracy). However, most recognition decisions are not made in the context of the legal system or in the

Table 8
*Hierarchical Regression Analysis Predicting 65% Cued Accuracy (*d′*)*

Variable	*B*	*SE B*	β	R^2	ΔR^2
Step 1					
Uncued recognition (d')	1.04	0.11	0.86***	0.74***	
Step 2					
Uncued recognition (d')	0.74	0.11	0.61***		
85% cued recognition (d')	0.50	0.11	0.40***	0.84***	0.10***

***$p < .001$.

context of deceptive others. Generally, one's goal is to maximize accuracy, and, in the presence of useful sources of external information, this goal is achieved by judiciously integrating external influences with internal memory evidence. Thus, our study specifically examined how observers are able to benefit from explicitly reliable external cues and found that participants are in fact able to boost performance under such conditions.

In particular, we were interested in determining whether metacognitive monitoring plays a role in observers' ability to capitalize on external cues. We hypothesized that the ability to assess the quality of one's own internal memory evidence will influence one's ability to properly weight external memory cues. As described in the introduction, under SDT, in order to successfully capitalize on external cues with a known validity, observers must adaptively shift their decision criteria under the different cuing conditions. An observer with more accurate insight about his or her internal memory representation (i.e., reliable approximations of a likelihood ratio evidence variable) can more appropriately determine ideal criteria placement and hence benefit more from external cues. Under this perspective, criterion shifts are not a nuisance phenomenon but are reflective of an adaptive decision process that capitalizes on environmental context cues during recognition judgment. Additionally, despite growing research examining recognition memory, prior research in the field has surprisingly not yet investigated the role of metacognitive monitoring in the adaptive placement of recognition criteria.

In the current report, across varying encoding conditions and cue validities, we demonstrated that, after controlling for uncued recognition skill, observers with greater metacognitive monitoring do in fact benefit more from reliable external cues. Thus, our results suggest that at least one factor governing successful adaptive criterion placement is metacognitive monitoring ability. Although metacognitive monitoring and uncued recognition accuracy have shared variance, our hierarchical regression analyses demonstrate that metacognitive monitoring has a unique contribution to cued performance beyond that of basic recognition skill. Although metacognitive monitoring is in part related to the benefit achieved from cues, future work should assess how other general abilities, such as working memory capacity, inhibitory control, or intelligence, may influence cuing benefit and the relationship of these abilities to metacognitive monitoring skills. Future work examining adaptive criterion shifting may also benefit from jointly assessing individual differences in metacognition along with other variables that may influence criteria placement such as personality variables (e.g., agreeableness or conscientiousness) and development. For example, given prior aging research suggesting behavioral inhibition deficits in healthy older adults (Hasher & Zacks, 1988), it may be the case that older adults overrely on external recommendations because they are unable to use recovered memory evidence to countermand the expectations instilled by the recommendations. One related example by Rogers, Jacoby, and Sommers (2012) examined the use of auditory context cues, where older adults exhibited more false hearing than younger adults did due to a greater reliance on auditory context cues. In their study Rogers et al. created congruent and incongruent contexts using a cue–target training procedure. During the test phase, participants were presented with previously learned semantically related word pairs (e.g., *BARN–HAY*), where the cue word was clearly aurally presented and the target word was masked in noise. For congruent trials the target word was the same as the trained target (e.g., *HAY*), and for incongruent trials the target word was phonologically similar (e.g., *PAY*). Despite a warning that the targets would not always match previously learned targets and titration for age-related hearing differences, older adults were more likely to false alarm and favor the previously learned context (e.g., *HAY*) than younger adults were. However, when the prior context and current target matched, older adults outperformed younger adults. Thus, it appears that at least for the cause of auditory cues, older adults tend to more heavily rely on prior external context. This resulted in greater accuracy when the context was facilitative but produced greater false hearing when the context was incongruent.

In addition to assessing metacognitive monitoring, we examined whether corrective feedback influenced observers' ability to benefit from external recommendations. Somewhat surprisingly, corrective feedback did not improve the extent to which participants benefited from external cuing, nor did it influence the degree to which metacognitive monitoring predicted cued performance. These results may seem puzzling because feedback could potentially inform participants about the adequacy of their criterion placement strategies and hence help them respond more ideally. Prior studies have demonstrated that feedback is sometimes critical for observers to realize that a shift of the criterion may be appropriate or useful (Estes & Maddox, 1995; Kantner & Lindsay, 2010; Rhodes & Jacoby, 2007; Verde & Rotello, 2007). The key difference between prior work using feedback and the current study is that in the former, feedback was typically used to alert the participant to some experimental manipulation that should ideally induce a criterion shift. For example, in Rhodes and Jacoby (2007) the probability of encountering studied items was correlated with screen locations such that words presented on one side were more likely to be targets and should result in more liberal responding than words presented on the other side of the screen. In Verde and Rotello (2007), the strength of old items (manipulated via repetition) was considerably higher on the first than the second half of the recognition test, and thus the optimal criterion placement was different in the two halves of the test. The key commonality across these studies is that feedback appeared critical for the participants to realize that responding similarly to the two locations (Rhodes & Jacoby, 2007) or similarly in the two test halves (Verde & Rotello, 2007) was not ideal, because the overall distributions of targets or the average target strength differed across locations or test periods. In the current study, however, the question was not whether observers would realize that the external cues were potentially useful, as this information was already clearly provided. Instead, the key question was whether the feedback would increase the ability of the observers to optimally integrate the cues into their judgments. In this context, feedback could help participants tune the size of criterion shifts in order to increase the odds of correct responding. Alternatively, the feedback on baseline, uncued trials might be used to calibrate the metacognitive judgments, which in turn would improve the use of cues because it would reflect an improvement in the observers' understanding of the link between internal evidence and the likelihood of successful judgments. Indeed, a recent model by Turner et al. (2011) specifically assumed that correct feedback plays a critical role in improving the observer's estimates of internal evidence distributions; thus, it should result both in an improvement in gamma during uncued trials and in an increased ability to incorporate external cues, given the

improved knowledge of the relationship between strength and likelihood of success. Unfortunately, neither possibility was supported by the current data, because the feedback group did not demonstrate increased cued recognition performance and had numerically lowered metacognitive monitoring scores on uncued trials relative to the no feedback group. In Experiment 1, in the feedback group, the average G* was 0.25 for shallow test items and 0.49 for deep test items; in the no feedback group, G* was 0.33 for shallow test items and 0.62 for deep test items. In Experiment 2, the average G* was 0.76 in the feedback group and 0.79 in the no feedback group.

This lack of improvement from feedback perhaps suggests limits on the plasticity of metacognitive monitoring ability during recognition. Prior research in the field of metacognition in an educational setting has in fact indicated that metacognitive abilities are often resistant to change (e.g., Bol, Hacker, O'Shea, & Allen, 2005; Koriat, 1997; Nietfeld, Cao, & Osborne, 2005), although the results are somewhat mixed, with some studies finding successful interventions (for a review, see Hacker, Bol, & Keener, 2008). However, improvements in the ability to monitor learning or mastery of complex materials may not transfer very well to basic judgments of recognition confidence. It is also the case that experimental manipulations of feedback may play a fairly minor role in judgments, such as recognition, with which observers have had a lifetime of experience (for a discussion, see Turner et al., 2011). However, given that metacognitive awareness and cue utilization skills are clearly far from perfect in the current participants (although quite good in some participants), this would have to be viewed as a premature cessation of the learning process.

Finally, the individual differences we found in the ability of observers to capitalize on external recommendations also have implications for signal detection models that assume a likelihood ratio decision axis. To date, there has been little consideration of the significance of the metacognitive literature for this particular class of decision models. However, this literature and the current findings suggest that there are likely important constraints on the ability of observers to assess the kind of information assumed under the likelihood ratio SDT model. Thus, even if one assumes that observers may be performing something roughly analogous to a likelihood ratio decision process (cf. Hintzman, 1994), they may do so in only a very limited or heuristic sense. One limitation revealed in the current study is poor metacognitive monitoring, and an interesting follow-up question to the current work would be to examine whether this metacognitive limitation is domain specific or fairly general. The latter would reveal that individuals who experienced difficulty incorporating external cues during recognition judgment would also have difficulty incorporating analogous cues during, say, perceptual judgments (e.g., gender discriminations), and in both cases this would be in part mediated by generally poor metacognitive monitoring of internal evidence during uncued performance.

References

Allan, K., & Gabbert, F. (2008). I still think it was a banana: Memorable "lies" and forgettable "truths." *Acta Psychologica, 127,* 299–308. doi:10.1016/j.actpsy.2007.06.001

Axmacher, N., Gossen, A., Elger, C. E., & Fell, J. (2010). Graded effects of social conformity on recognition memory. *PLoS ONE, 5,* e9270. doi:10.1371/journal.pone.0009270

Benjamin, A. S., & Diaz, M. (2008). Measurement of relative metamnemonic accuracy. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of metamemory and memory* (pp. 73–94). New York, NY: Psychology Press.

Benjamin, A. S., Diaz, M., & Wee, S. (2009). Signal detection with criterion noise: Applications to recognition memory. *Psychological Review, 116,* 84–115. doi:10.1037/a0014351

Betz, A., & Skowronski, J. (1996). Shared realities: Social influence and stimulus memory. *Social Cognition, 14,* 113–140. doi:10.1521/soco.1996.14.2.113

Bol, L., Hacker, D. J., O'Shea, P., & Allen, D. (2005). The influence of overt practice, achievement level, and explanatory style on calibration accuracy and performance. *Journal of Experimental Education, 73,* 269–290. doi:10.3200/JEXE.73.4.269-290

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10,* 433–436. doi:10.1163/156856897X00357

Busey, T. A., Tunnicliff, J., Loftus, G. R., & Loftus, E. F. (2000). Accounts of the confidence–accuracy relation in recognition memory. *Psychonomic Bulletin & Review, 7,* 26–48. doi:10.3758/BF03210724

Chandler, C. C. (1994). Studying related pictures can reduce accuracy, but increase confidence, in a modified recognition test. *Memory & Cognition, 22,* 273–280. doi:10.3758/BF03200854

Cox, J. C., & Dobbins, I. G. (2011). The striking similarities between standard, distractor-free, and target-free recognition. *Memory & Cognition, 39,* 925–940. doi:10.3758/s13421-011-0090-3

Dobbins, I. G., & Han, S. (2007). What constitutes a model of item-based memory decision making? *Psychology of learning and motivation: Vol. 48. Strategic and nonstrategic influences on memory attribution* (pp. 95–144). London, England: Elsevier.

Dobbins, I. G., Kroll, N. E., & Liu, Q. (1998). Confidence–accuracy inversions in scene recognition: A remember–know analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 1306–1315. doi:10.1037/0278-7393.24.5.1306

Estes, W. K., & Maddox, W. (1995). Interactions of stimulus attributes, base rates, and feedback in recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 1075–1095. doi:10.1037/0278-7393.21.5.1075

Glanzer, M., Hilford, A., & Maloney, L. T. (2009). Likelihood ratio decisions in memory: Three implied regularities. *Psychonomic Bulletin & Review, 16,* 431–455. doi:10.3758/PBR.16.3.431

Hacker, D. J., Bol, L., & Keener, M. C. (2008). Metacognition in education. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of metamemory and memory* (pp. 429–455). New York, NY: Psychology Press.

Hasher, L., & Zacks, R. T. (1988). Working memory, comprehension, and aging: A review and a new view. In G. Bower (Ed.), *Psychology of learning and motivation: Vol. 22. Advances in research and theory* (pp. 193–225). San Diego, CA: Academic Press.

Healy, A. F., & Kubovy, M. (1978). The effects of payoffs and prior probabilities on indices of performance and cutoff location in recognition memory. *Memory & Cognition, 6,* 544–553. doi:10.3758/BF03198243

Hintzman, D. L. (1994). On explaining the mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 201–205. doi:10.1037/0278-7393.20.1.201

Jaeger, A., Lauris, P., Selmeczy, D., & Dobbins, I. G. (2012). The costs and benefits of memory conformity. *Memory & Cognition, 40,* 101–112. doi:10.3758/s13421-011-0130-z

Kantner, J., & Lindsay, D. S. (2010). Can corrective feedback improve recognition memory? *Memory & Cognition, 38,* 389–406. doi:10.3758/MC.38.4.389

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. doi:10.1037/0096-3445.126.4.349

Macmillan, N., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, NJ: Erlbaum.

Meade, M. L., & Roediger, H. L. (2002). Explorations in the social contagion of memory. *Memory & Cognition, 30,* 995–1009. doi: 10.3758/BF03194318

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95,* 109–133. doi:10.1037/0033-2909.95.1.109

Nietfeld, J. L., Cao, L., & Osborne, J. W. (2005). Metacognitive monitoring accuracy and student performance in the postsecondary classroom. *Journal of Experimental Education, 74,* 7–28.

Pastore, R. E., Crawley, E. J., Berens, M. S., & Skelly, M. A. (2003). "Nonparametric" A' and other modern misconceptions about signal detection theory. *Psychonomic Bulletin & Review, 10,* 556–569. doi: 10.3758/BF03196517

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10,* 437–442. doi: 10.1163/156856897X00366

Reysen, M. B. (2005). The effects of conformity on recognition judgments. *Memory, 13,* 87–94. doi:10.1080/09658210344000602

Rhodes, M. G., & Jacoby, L. L. (2007). On the dynamic nature of response criterion in recognition memory: Effects of base rate, awareness, and feedback. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 305–320. doi:10.1037/0278-7393.33.2.305

Roediger, H. L., Meade, M. L., & Bergman, E. T. (2001). Social contagion of memory. *Psychonomic Bulletin & Review, 8,* 365–371. doi:10.3758/BF03196174

Roediger, H. L., Wixted, J., & DeSoto, A. K. (2012). The curious complexity between confidence and accuracy in reports from memory. In L. Nadel & W. P. Sinnott-Armstrong (Eds.), *Memory and law.* New York, NY: Oxford University Press.

Rogers, C. S., Jacoby, L. L., & Sommers, M. S. (2012). Frequent false hearing by older adults: The role of age differences in metacognition. *Psychology and Aging, 27,* 33–45. doi:10.1037/a0026231

Schneider, D., & Watkins, J. M. (1996). Response conformity in recognition testing. *Psychonomic Bulletin & Review, 3,* 481–485. doi:10.3758/BF03214550

Son, L. K., & Kornell, N. (2008). Research on the allocation of study time: Key studies from 1890 to the present (and beyond). In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of metamemory and memory* (pp. 333–351). New York, NY: Psychology Press.

Stretch, V., & Wixted, J. T. (1998). On the difference between strength-based and frequency-based mirror effects in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 1379–1396. doi:10.1037/0278-7393.24.6.1379

Swets, J. A., Tanner, W. P., Jr., & Birdsall, T. G. (1961). Decision processes in perception. *Psychological Review, 68,* 301–340. doi: 10.1037/h0040547

Turner, B. M., Van Zandt, T., & Brown, S. (2011). A dynamic stimulus-driven model of signal detection. *Psychological Review, 118,* 583–613. doi:10.1037/a0025191

Verde, M. F., & Rotello, C. M. (2007). Memory strength and the decision process in recognition memory. *Memory & Cognition, 35,* 254–262. doi:10.3758/BF03193446

Walther, E., Bless, H., Strack, F., Rackstraw, P., Wagner, D., & Werth, L. (2002). Conformity effects in memory as a function of group size, dissenters and uncertainty. *Applied Cognitive Psychology, 16,* 793–810. doi:10.1002/acp.828

Wixted, J. T., & Mickes, L. (2010). A continuous dual-process model of remember/know judgments. *Psychological Review, 117,* 1025–1054. doi:10.1037/a0020874

Wright, D. B., Gabbert, F., Memon, A., & London, K. (2008). Changing the criterion for memory conformity in free recall and recognition. *Memory, 16,* 137–148. doi:10.1080/09658210701836174

Wright, D. B., Mathews, S. A., & Skagerberg, E. (2005). Social recognition memory: The effect of other people's responses for previously seen and unseen items. *Journal of Experimental Psychology: Applied, 11,* 200–209. doi:10.1037/1076-898X.11.3.200

Wright, D. B., Self, G., & Justice, C. (2000). Memory conformity: Exploring misinformation effects when presented by another person. *British Journal of Psychology, 91,* 189–202. doi:10.1348/000712600161781

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46,* 441–517. doi:10.1006/jmla.2002.2864

Received April 5, 2012
Revision received June 20, 2012
Accepted June 22, 2012 ■