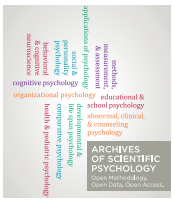
Archives of Scientific Psychology 2013, 1, 23–32
DOI: http://dx.doi.org/10.1037/arc0000003

2169-3269

Archives of Scientific Psychology

http://www.apa.org/pubs/journals/arc



Judgments of Learning Are Influenced by Multiple Cues in Addition to Memory for Past Test Accuracy

Christopher Hertzog and Jarrod C. Hines
Georgia Institute of Technology

Dayna R. Touron
University of North Carolina Greensboro

ABSTRACT

When people try to learn new information (e.g., in a school setting), they often have multiple opportunities to study the material. One of the most important things to know is whether people adjust their study behavior on the basis of past success so as to increase their overall level of learning (e.g., by emphasizing information they have not yet learned). Monitoring their learning is a key part of being able to make those kinds of adjustments. We used a recognition memory task to replicate prior research showing that memory for past test outcomes influences later monitoring, as measured by judgments of learning (JOLs; confidence that the material has been learned), but also to show that subjective confidence in whether the test answer and the amount of time taken to restudy the items also have independent effects on JOLs. We also show that there are individual differences in the effects of test accuracy and test confidence on JOLs, showing that some but not all people use past test experiences to guide monitoring of their new learning. Monitoring learning is therefore a complex process of considering multiple cues, and some people attend to those cues more effectively than others. Improving the quality of monitoring performance and learning could lead to better study behaviors and better learning.

SCIENTIFIC ABSTRACT

An individual's memory of past test performance (MPT) is often cited as the primary cue for judgments of learning (JOLs) following test experience during multitrial learning tasks (Finn & Metcalfe, 2007, 2008). We used an associative recognition task to evaluate MPT-related phenomena, because performance monitoring, as measured by recognition test confidence judgments (CJs), is fallible and varies in accuracy across persons. The current study used multilevel regression models to show the simultaneous and independent influences of multiple cues on Trial 2 JOLs, in addition to performance accuracy (the typical measure of MPT in cued-recall experiments). These cues include recognition CJs, perceived recognition fluency, and Trial 2 study time allocation (an index of reprocessing fluency). Our results expand the scope of MPT-related phenomena in recognition memory testing to show independent effects of recognition test accuracy and CJs on second-trial JOLs, while also demonstrating individual differences in the effects of these cues on JOLs (as manifested in significant random effects for those regression effects in the model). The effect of study time on second-trial JOLs controlling on other variables, including Trial 1 recognition memory accuracy, also demonstrates that second-trial encoding behavior influence JOLs in addition to MPT.

Keywords: judgments of learning, memory for past test, metacognition, multilevel modeling
Supplemental materials: http://dx.doi.org/10.1037/arc0000003.supp
Data Repository: http://dx.doi.org/10.3886/ICPSR34708.v1

This article was published August 5, 2013.

Christopher Hertzog and Jarrod C. Hines, School of Psychology, Georgia Institute of Technology; Dayna R. Touron, Department of Psychology, University of North Carolina Greensboro.

For further discussion on this topic, please visit the *Archives of Scientific Psychology* online public forum at http://arcblog.apa.org.

This research was supported by grants from the National Institute on Aging, one of the National Institutes of Health (R01 AG024485). We would like to extend thanks to the following personnel for their assistance with subject recruitment and data collection: Teri Boutot, Bethany Geist, Devaki Kumarhia, Colin Malone, Melissa McDonald, and Alisha Monteiro. For more information on our research program, consult http://psychology.gatech.edu/CHertzog.

Correspondence concerning this article should be addressed to Christopher Hertzog, School of Psychology, GA Institute of Technology, Atlanta, GA 30332-0170. E-mail: christopher.hertzog@psych.gatech.edu

Metacognitive self-regulation of study involves monitoring current cognitive states and using the products of that monitoring to guide study behavior (e.g., Dunlosky & Metcalfe, 2009; Nelson, 1996). Accurate monitoring of one's current level of learning, often measured by judgments of learning (JOLs), rated confidence in likelihood of remembering a studied item during a subsequent memory test (see Dunlosky & Metcalfe, 2009, for a review), can lead to an optimal selection of items to study and allocation of study time to those items (e.g., Nelson & Narens, 1990; Thiede & Dunlosky, 1999).

Judgments of Learning

Metacognitive research has converged on an accessibility view of JOLs, which states that JOLs are not based on direct access to underlying states of learning, but rather on accessible cues that may or may not be diagnostic of subsequent remembering (Dunlosky & Metcalfe, 2009). Multiple variables have been shown to influence JOLs and JOL resolution (i.e., whether within-person variation in JOLs across items correlates with item variation in memory outcomes; see Gonzalez & Nelson, 1996), including observable stimulus characteristics, learning strategies, and internal mnemonic states (e.g., Koriat, 1997; Nelson, 1996). Typical JOL experiments focus on one or at most a few such cues, seeking to experimentally isolate their influences on JOLs.

Multiple-Cue Utilization Perspective

We argue that additional progress can be gained by adopting a generalization of the accessibility perspective—an explicit *multiple-cue utilization* approach—to account for JOL magnitude and resolution. By this account, a JOL can be considered an outcome of a (possibly informal) decision process in which relevant cues are potentially accessed and then evaluated. The validities of utilized cues for predicting subsequent memory performance (often termed diagnosticity in the judgment and decision making literature), along with the relative weight given to each cue actually accessed by the judge when making the JOL, determine JOL accuracy (resolution). This perspective argues that individuals can (and often do) access multiple cues when constructing JOLs, although ignoring (failing to access) or discounting (failing to utilize) other available cues that may or may not be diagnostic. The task of the researcher is to demonstrate empirically which cues are associated with JOLs, potentially indicating that they are accessed and evaluated during the judgment process. For example, Hertzog, Dunlosky, and Sinclair (2010) showed that JOLs were jointly influenced by an experimentally manipulated stimulus characteristic (associative relatedness of word pairs) and a participant-determined behavior (spontaneous use of mediational strategies).[1]

The experimental literature has generated a great deal of empirical data about different cues that may or may not influence both JOL magnitudes and JOL resolution. JOLs made immediately after studying items for the first time often have rather low resolution that can be attributed in large part to the salience of stimulus features that are mostly irrelevant to subsequent remembering (e.g., perceptual features such as size or loudness; Rhodes & Castel, 2008). Individuals making JOLs also ignore or discount many other cues that are diagnostic of later remembering such as asymmetric direction of associative relatedness of cue-target pairs (e.g., Koriat & Bjork, 2006), instructed encoding strategies varying in normative effectiveness (e.g., Hertzog et al., 2009; Shaughnessy, 1981), and the type of memory test (e.g., Touron, Hertzog, & Speagle, 2010; Weaver & Kelemen, 2003). Immediate JOLs (unlike delayed JOLs or feeling-of-knowing judgments) are also insensitive to the cue's associative set size, which influences the likelihood of implicit retrieval interference at test (Eakin & Hertzog, 2012), but set size apparently is not accessed at the time of encoding when making immediate JOLs. JOLs made during an initial study opportunity are also correlated with encoding fluency (i.e., how quickly an item is encoded; Begg, Duft, Lalonde, Melnick, & Sanvito, 1989; Hertzog, Dunlosky, Robinson, & Kidder, 2003; Undorf & Erdfelder, 2011) and by latencies of retrieving relevant information (e.g., Benjamin, Bjork, & Schwartz, 1998). Even when encoding fluency is not diagnostic of subsequent memory, it appears to influence JOLs, attenuating their resolution (e.g., Hertzog et al., 2003; Robinson, Hertzog, & Dunlosky, 2006).

JOLs and the Memory-For-Past-Test Heuristic

Studies of multiple study-test trials with cued recall of lists of paired-associate (PA) items show that JOL resolution increases until asymptotic performance begins to restrict item variance in recall (e.g., Koriat, 1997; Koriat & Bjork, 2005). This increase in resolution typically requires intervening memory tests rather than just additional study opportunities (Koriat & Bjork, 2005). A plausible mechanism for this effect is that individuals remember past test performance for an item and use that information when making a JOL on the next study opportunity. Indeed, memory for past test (MPT) has been shown to have a strong influence on subsequent JOLs (Ariel & Dunlosky, 2011; Finn & Metcalfe, 2007, 2008). Reliance on MPT might help to account for the stability bias in JOLs (failure to predict increases in performance after an additional study opportunity; e.g., Kornell, Rhodes, Castel, & Tauber, 2011) as well as the underconfidence-with-practice effect (increasing discrepancy between the aggregate level of JOL confidence and the list-wise probability of recall; Finn & Metcalfe, 2007).

However, People do not exclusively rely on MPT as a heuristic for making JOLs. Finn and Metcalfe (2008) found that underconfidence with practice occurred even when no intervening test was given between JOLs, and used this and other findings to argue for the likely influence of other cues they had not measured on second-trial JOLs. Ariel and Dunlosky (2011) used prestudy JOLs, which, in one condition, were accompanied by prompts about past recall success for that item. They reasoned that if MPT was the only cue that was considered when making a second-trial JOL, then these prestudy JOLs should be as accurate as standard JOLs. They were not. Their two experiments indicated that second-trial JOLs were influenced by MPT, but also by new learning and item forgetting. Tauber and Rhodes (2012) used multilevel regression models to show that second-trial recall predicted prior JOLs independently of first-trial recall, demonstrating that new learning was indeed influencing JOLs. However, although these studies demonstrated the likely existence of influences other than MPT on second-trial JOLs, they did not actually measure any of the other cues potentially affecting second-trial JOLs. Hence, they could neither empirically demonstrate other cues' influence nor estimate the magnitude of those effects.

Goals of the Present Study

This study was motivated by two broad goals. One was to extend the concept of MPT to embrace aspects of test performance other than cued- recall success, as has previously been studied. We used an associative recognition memory task to broaden the range of MPT-related phenomena beyond those typically seen in cued-recall tasks. A second goal was to explicitly evaluate whether cues besides MPT

[1] Goals, agendas, expectations, individuals' cognitive resources, and task-imposed processing constraints are also potential influences on which and how many cues are evaluated and utilized when making JOLs.

influence JOLs at the second study opportunity. Embracing the multiple-cue perspective, we argue that at least two other variables in addition to those related to MPT should predict second-trial JOLs: prior JOL confidence and second-trial study time.

Expanding MPT-Related Phenomena in an Associative Recognition Task

Recognition memory tests often use item-level confidence judgments (CJs)—rated confidence in the accuracy of memory test responses—to address multiple research questions, including those regarding influences of response criteria on responses and influences of recollection versus familiarity on remembering (e.g., Yonelinas & Parks, 2007). Metacognitive research has used CJs to evaluate the accuracy of performance monitoring, including retrieval and memory test decisions (e.g., Kelley & Sahakyan, 2003; see Dunlosky & Metcalfe, 2009, for a review) Cued recall for lists with associatively unrelated PA items are typically characterized by high omission error responses ("I can't remember") and low intrusion error rates (e.g., Dunlosky, Hertzog, & Powell-Moman, 2005). Probably as a consequence, item-level CJs after each cued-recall test are highly accurate even in the absence of performance feedback (Dunlosky & Hertzog, 2000; Finn & Metcalfe, 2007, Experiment 3), with recall-CJ resolution often approaching perfect 1.0 correlations, except when to-be-remembered materials are deliberately manipulated to create false memories (e.g., Jacoby & Rhodes, 2006; Kelley & Sahakyan, 2003; Roediger & McDermott, 1993).

In recognition memory tests, however, monitoring successful remembering is more ambiguous and error-prone, and other variables such as response latency may carry important information about future accessibility of underlying memory representations (e.g., Ratcliff, Thapar, & McKoon, 2011). Associative recognition tests often compare recognition of intact paired-associated (unchanged from original study) with false alarms for rearranged pairs (in which words for originally different studied items are paired together at test. Associative recognition task CJs manifest imperfect performance monitoring accuracy (Hines, Touron, & Hertzog, 2009); e.g., older adults are prone to high-confidence recognition memory errors (e.g., Shing, Werkle-Bergner, Li, & Lindenberger, 2009). Recollective experiences during the recognition test may include more or less access to vivid details about the original encoding, the associative mediators that were originally produced (if any), or other cues that could influence both CJs and subsequent JOLs (e.g., Cohn & Moscovitch, 2007; Hicks & Marsh, 2001; Parks, 2007).

Because associative recognition test CJs are imperfectly correlated with recognition memory success, using a recognition memory task allowed us to separate cues associated with memory test responses from subjective confidence in the accuracy of those responses—both variables can be used to predict second trial JOLs. Indeed, in the present study we tested whether CJs predicted second-trial JOLs independent of actual recognition memory accuracy. To the extent that CJs fully capture the subjective experience of remembering during the recognition test, recognition memory success should not predict JOLs independent of CJs.

There is good reason to believe that neither recognition memory accuracy nor CJs will capture all the available cues about recognition memory performance that will later be accessed when making JOLs. For instance, both variables may not capture retrieval fluency effects during recognition memory that influence JOLs (e.g., Benjamin, Bjork, & Schwartz, 1998). Recognition memory reaction time (RTs) are one measure of retrieval fluency in recognition tasks. They typically decrease monotonically with the degree of item learning and, therefore, may serve as a valid cue for future correct retrieval (e.g., Cerella, Onyper, & Hoyer, 2006; Logan, 1988; Ratcliff & Starns, 2009), perhaps because they are related to gradations in memory strength above a threshold required for successful recognition (Bower, 2000). Retrieval fluency effects have typically been ignored in existing work on MPT, which has focused exclusively on recall accuracy. We hypothesized that recognition test response latencies would also predict subsequent JOLs independent of prior recognition accuracy and CJs.

Time monitoring in memory experiments is often rather inaccurate, with individuals underestimating the time required for successful retrieval, possibly because of the degree of attentional control devoted to the retrieval search (e.g., Craik & Hay, 1999; Hertzog, Touron, & Hines, 2007). Thus, subjective fluency (measured by participant-estimated retrieval times) could influence JOLs, whether or not actual retrieval fluency does so. Indeed, from a metacognitive perspective, one might argue that perceived retrieval fluency should have a stronger influence on later JOLs than actual retrieval fluency (see Hines et al., 2009; Robinson, Hertzog, & Dunlosky, 2006).

In addition to the influence of these metacognitive Trial 1 cues on Trial 2 JOLs, the Trial 2 study experience immediately preceding Trial 2 JOLs could also have an influence that should be also independent of MPT. Study time at first exposure to new information is a multi-determined variable, being influenced by the nature of the study strategy, the difficulty of implementing a selected strategy, and additional rehearsal or elaborative processing after completion of a study strategy (e.g., Koriat, Ma'ayan, & Nussinson, 2006). In contrast, study time on a restudy trial following a memory test can reflect (a) relative study emphasis—in particular, a decision to devote less study time to items already remembered, which should be entrained by MPT (e.g., Mazzoni & Cornoldi, 1993); or (b) variations in reprocessing fluency (e.g., Masson, 1993), which might arise in part because individuals can choose to reproduce or retrieve outcomes from a study strategy or vary encoding strategies (e.g., either (a) retrieve and rehearse the same study mediators generated during the first study opportunity or (b) choose to generate new mediators; see Pyc & Rawson, 2012). Hines, Touron, and Hertzog (2009) showed that that study times following a test trial are influenced by past associative-recognition test accuracy and recognition CJs. Controlling on past test accuracy, an independent effect of restudy time on JOLs made immediately after restudy would indicate emergent influences of encoding on JOLs that are not accounted for by a MPT heuristic.

Testing Multiple-Cue Hypotheses About Restudied Item JOLs

We used multilevel regression models (Singer, 1998; Snijders & Bosker, 1999) to assess the relative contributions of multiple independent variables (that carry the effects of different metacognitive cues) on JOLs made during a second study-test trial in our associative recognition task. These types of analyses are especially well-suited to evaluating the simultaneous unique predictive power of multiple cues on JOLs (see Hoffman & Rovine, 2007), assessing whether a particular cue is predictive of subsequent JOLs controlling on other measured cues (via the partial regression coefficient associated with each cue). Multilevel modeling also affords the opportunity to examine in a single model the roles of both item-level and person-level variation in cues in predicting subsequent JOLs. For example, one can simultaneously examine (a) whether items that were encoded or retrieved more quickly also predict higher JOLs (item-level association); and (b) whether people with faster average encoding or recognition times also exhibit greater average JOL confidence. It is also possible to estimate specific item differences in JOLs and to control for these effects in evaluating the influences of other variables on JOLs. This

approach insures that any effects are not an epiphenomenon of unusual item characteristics impacting the JOLs.

Through the use of multilevel modeling we show that, although variables related to past test performance do have a powerful influence on subsequent JOLs, other cues related to the learning experience also influence them.

In a preliminary study we analyzed unpublished data from Hines et al.'s (2009) experiment assessing age differences in metacognitive predictors of study time allocation in an experiment with two study-test trials. We had collected Trial 2 JOLs in that study but did not report on them in the 2009 article. Multilevel regression models measured item-level and person-level influences on Trial 2 JOLs. We obtained reliable prediction of Trial 2 JOLs by JOLs at Trial 1, indicating stable item-related variance in JOLs. Expected patterns of MPT were also found, with first-trial recognition accuracy predicting Trial 2 JOLs. We also found reliable random effects in this regression coefficient, indicating that the MPT effect varied across individuals. Critically for a multiple-cue perspective, Trial 2 study time predicted Trial 2 JOLs independent of Trial 1 recognition success.

We conducted a new experiment that replicated and extended those preliminary findings, testing several hypotheses about variables that may influence the construction of Trial 2 JOLs. First, we hypothesized that recognition CJs and estimated recognition RTs (i.e., subjective retrieval fluency) would also predict Trial 2 JOLs independent of recognition memory accuracy. Second, we hypothesized across-trial consistency in JOLs (prediction of Trial 2 JOL for an item by the Trial 1 JOL for that same item) that would not be eliminated by the MPT effects. Third, and most critically, we hypothesized that study time at Trial 2 (reflecting reprocessing fluency at encoding) would be associated with Trial 2 JOLs, independent of the Trial 1 variables. This outcome would confirm prior reports of unknown variables other than MPT influencing second-trial JOLs, while demonstrating that one of these previously unidentified variables was the fluency of second-trial encoding.

Method

Design

The experiment used two study-test trials in a within-subjects design.

Participants

Fifty-one undergraduates between the ages of 18 and 25 years (M = 19.22) participated and received course credit for doing so.

Materials and Procedure

The stimulus materials were identical to those reported by Hines et al. (2009), consisting of 120 concrete nouns that were used to construct 60 normatively unrelated paired associate items (see Table S2 in Supplement A for the list of paired associate items). The general procedure involved two consecutive study-test trials using an associative recognition test with intact and rearranged pairs. During study, the task presented centrally fixated word pairs under intentional memorization instructions. Study was self-paced at both trials, with participants terminating item-level study by pressing the spacebar. During each study trial, 60 associatively unrelated word pairs (e.g., IVY–BIRD) were presented individually in a randomized order. After studying each item, participants gave a JOL rating their ability to remember the item approximately 10 minutes later on a continuous scale from 0% to 100% confidence. Study was terminated by a keypress used to record the elapsed study time.

Self-paced testing presented both intact (e.g., an item as presented during study) and rearranged (e.g., IVY–BIRD might have been tested when the originally studied items were IVY–STAR and BARREL–BIRD) trials. Each item was tested in intact form and used to form one rearranged pair, resulting in 120 recognition test trials. Upon the presentation of each stimulus, participants indicated by keypress ("yes" or "no") whether the item was an intact pair that had been studied previously. Following each test trial, participants provided reaction time (RT) estimates and CJs. Like JOLs, CJs were made using an integer between 0 and 100% confidence. Participants estimated their recognition RT by using a continuous visual analog scale. Participants controlled the scale by moving the computer mouse to slide an indicator from the left (0 s) to the right (10 s). As the mouse moved, the computer also displayed a numerical value of the estimated RT below the slider, rounded to the nearest .1 s. The order of collecting CJs and RT estimates was counterbalanced across participants who were randomly assigned to the counterbalancing conditions.

Results

Table 1 contains the means and standard errors of variables entered into our multilevel models, as well as Goodman-Kruskal gamma correlations (see Gonzalez & Nelson, 1996) often used in the metacognitive literature to evaluate increases in JOL resolution across trials (e.g., Koriat, 1997). These data should help the reader assess the relative impact of each variable in the regressions and are informative concerning results that are typically emphasized in studies of JOLs in multitrial learning experiments.

One outcome that can be observed in Table 1 is that the mean level of JOL confidence, collected on a scale of 0%–100%, was well below the average level of associative recognition memory performance. Supplemental results available with this article (see Supplement A) demonstrated that JOLs are highly underconfident of recognition memory performance at first test, and that this underconfidence is substantially but not fully repaired by test experience (also see Touron

Table 1
Means and Standard Errors for Measured Variables for Both Learning Trials

Variable	Trial 1		Trial 2	
	Mean	*SE*	Mean	*SE*
Study time(s)	7.36	0.65	1.62	0.29
Judgment of learning	40.76	2.47	81.70	1.88
Recognition accuracy	0.82	0.02	0.93	0.01
Confidence judgment	85.53	1.41	93.64	0.91
Recognition RT(s)	1.65	0.04	1.09	0.03
Estimated recognition RT(s)	1.49	0.12	1.26	0.11
Judgment of learning × Recognition accuracy gamma	0.24	0.07	0.30	0.08
Trial 1 recognition accuracy × Trial 2 Judgment of learning gamma	0.68	0.04		

Note. Estimated RT was measured in seconds, as were all other temporal measures. Judgments of Learning and Confidence Judgments were scaled from 0 to 100% on a continuous scale, and recognition accuracy represents the proportion of correct recognition responses for each experimental group.

et al., 2010).[2] Hence, the underconfidence with practice effect emphasized to date in the JOL literature (e.g., Finn & Metcalfe, 2007) is specific to cued recall. JOL resolution for associative recognition also did not increase across trials, unlike well-replicated findings with cued recall tasks. These results confirm previous findings that associative recognition behaves differently from cued recall (see Touron, Hertzog, & Speagle, 2010), and reinforce the possibility that multiple cues could influence second-trial JOLs.

Multilevel Regression Model for Trial-2 JOLs

Our main goal in this article was to use regression models to test hypotheses relevant to multiple-cue influences on Trial 2 JOLs. Multilevel regression models were constructed using SAS PROC MIXED (Littell, Milliken, Stroup, & Wolfinger, 2000).[3] The item-level predictor variables included in the initial regression equation were the specific item identifier, Trial 1 study time, JOL, recognition accuracy, CJ, recognition RT, perceived recognition RT, and Trial 2 study time. We employed the item identifier ("Cue") as a nuisance factor to control for item differences in average JOLs. This procedure insured that any Level 1 regression effects were not an artifact of variance associated with items being normatively perceived as more easily or less easily learned. The person-level predictor variables were the person means for all item-level predictors.

We used within-person and between-person centering of predictors to account for within- and between-person effects (see Enders & Tofighi, 2007; Singer, 1998). Within-person centering scales each predictor variable as its deviation from that person's mean for that cue. For example, if an individual has on average 70% confidence in recognition CJs, that person's CJ of 50 would be rescaled to −20, whereas a CJ of 90% confidence would be rescaled to +20. Each person's predictor variables are centered to the person's own mean. So for a different individual with 30% mean CJs, item CJs of 70% and 90% would be rescaled to +40% and +60%, respectively. Within-person centering allows one to evaluate whether within-person variation in other cues predicts within-person variation in JOLs, or what would be termed a Level 1 regression effect in multilevel regression models.

With between-person scaling, a new variable is created for each item that is the person's mean on the predictor variable, centered against the overall sample mean for the predictor. This variable is a constant within a set of items for each person, but differs between persons. So if the sample mean CJ was 50%, the two individuals with mean CJs of 70% and 30% would have centered values of +20% and −20%, respectively, for each and every one of their items. This between-person centered variable captures individual differences in means on the predictor variables. Done in this way, the two rescaled variables are orthogonal to each other, and can both be used as predictors in the regression model (Singer, 1998).

Using both types of variables allows the simultaneous estimation of both item-level (i.e., Level 1 effects) effects and individual differences in predictors on JOLs (Level 2 effects). This practice can address within- and between-level questions, such as: "Does a higher JOL for an item at Trial 1 predict a higher JOL for that same item at Trial 2?" and "Do people who had higher mean JOLs at Trial 1 also have higher mean JOLs at Trial 2?" If these within-person and between-person influences are not independently identified and jointly evaluated they can contaminate one-another (e.g., Snijders & Bosker, 1999).

We also evaluated random effects for the regression coefficients predicting JOLs from recognition memory accuracy and CJs in order to determine whether there were reliable individual differences in these MPT effects (see Gelman & Hill, 2007). To do so, we specified a hierarchical series of models with maximum likelihood (ML) estimation, first adding a random intercept (individual differences in mean JOLs for Trial 2), next adding a random effect for recognition memory accuracy, and finally adding a random effect for CJs. This procedure allowed us to compute differences in fit (−2 Log Likelihood, or −2LL) for the ML solution, which are likelihood ratios whose differences are distributed as central χ^2 variates under the null hypothesis that the row of random effects added to the covariance matrix was generated from a vector of 0 population variances and covariances (Pinheiro & Bates, 2000).

Multilevel regression models involve estimation and interpretation of raw score regression coefficients; in our analyses item-level predictors can have over 1,000 *df* for their error terms. To avoid consideration of trivially small effect sizes, we set the Type I error criterion at $\alpha = .01$, although we report actual *p* values for all fixed-effect *F* tests.

We had counterbalanced the order of RT estimates and CJs after each recognition test, with persons randomly assigned to counterbalancing conditions. A preliminary model (see Table S3 in Supplement A) tested for counterbalancing effects (Order) and interactions. No effect reached the .01 criterion.[4]

The initial model reported here included the item-level predictor variables of Trial 1 study time, JOL, recognition accuracy, CJ, recognition RT, perceived recognition RT, and Trial 2 study time allocation. It also included person-level predictor variables for these same variables.

First, we evaluated goodness-of-fit tests for the sequence of hierarchically nested models. Table 2 reports the sequence of likelihood ratio tests, starting with an initial unconditioned model (that estimated the JOL intercept without the inclusion of any predictors (see Snijders and Bosker, 1999). This unconditioned model can be used to generate pseudo-R^2 values for each model by subtracting the residual variance associated with the final model from that associated with a model containing predictor variables. Model 2 added a random effect on the intercept, generating a pseudo-R^2 that is also an estimate of the intraclass correlation, with about 19% of the total variance attributable to consistent individual differences in mean JOL confidence.

We next entered the desired fixed effects in Model 3. For Model 3 it is possible to compute the contribution of fixed effects to the pseudo-R^2. The estimated random variance in intercepts dropped from 169.62 in Model 2 to 94.76 when fixed effects were added (between-subjects effects accounted for some of the individual differences in JOLs estimated in Model 2). Using these values, we estimated that 29% of the total variance in item-level Trial 2 JOLs was accounted for by the fixed effects portion of Model 3; including random effects, the model accounted for 46% of the JOL variance.

[2] Subjective confidence and objective probability of success are deliberately reported in different scales (% confidence and proportion correct). The implicit assumption that the two types of measures can be aligned assumes equivalence of subjective probability and frequentist probability metrics that may not be justified (see Keren, 1991). The rating scale used can influence the calibration of those ratings (see Weber & Brewer, 2003). Individuals have been shown to discount chance recognition performance when making JOLs and CJs (Touron et al., 2010). Nevertheless, our focus is on subjective probability and not on judgment calibration or underconfidence. Hence, the use of a full (0%–100%) scale that has an intuitively meaningful range of confidence for raters was considered most appropriate for the current study.

[3] Our raw data (Supplement B), as well as SAS PROC MIXED (Supplement C) and SPSS MIXED (Supplement D) syntax code and the outputs they generate (Supplements E and F, respectively) are provided so that our analyses can be evaluated and reproduced.

[4] There was a trend ($p = .046$) for an Order × within-person RT estimates interaction, indicating that RT estimates had a greater influence on Trial 2 JOLs if they preceded CJs than if they followed them.

Table 2
Goodness-of-Fit Statistics, LR χ^2 Tests of Random Effects, and Psuedo-R^2 for Hierarchically Nested Models Predicting Trial 2 JOLs

Model	−2LL	BIC	Residual	$\Delta\chi^2$	Δdf	R^2
1. Unconditioned	28146	28163	578.42	—	—	—
2. Random intercept	27250	27262	408.81	896.0**	1	.293
3. Add fixed effects	26439	26738	315.17	811.7**	1	.455
4. Add random accuracy	26375	26682	302.95	63.5**	2	.476
5. Add random CJ	26325	26643	295.17	50.7**	3	.490
6. Drop random accuracy	26343	26650	299.49	18.3**	3	.482

Note. R^2 is the total pseudo-R^2 for all effects (fixed and random). 2LL = −2 Log Likelihood (fit function); BIC = Bayesian Information Criterion; LR = likelihood ratio; JOL = judgment of learning; CJ = confidence judgment.
** $p < .001$.

Model 4 added a random effect for recognition memory accuracy, which most closely aligns to the typical effect studied for MPT in free recall. Model 5 added a random effect of CJs to see if there were random effects in this regression relationship as well. The likelihood ratio (LR) χ^2 tests revealed that each addition improved the model fit, so random effects for the JOL intercept, the memory accuracy slope, and the CJ slope were retained in the final model. This final model accounted for about 49% of the total variance in JOLs. We also evaluated Model 6, which deleted the random effects for recognition memory accuracy, leaving the other two random sources. Given the expected positive correlation of CJs and recognition memory accuracy, it was possible that CJs alone could account for random MPT effects. The loss of fit was significant, causing us to retain all three random effects.

Table 3 reports the fixed effects regression estimates and Table 4 reports the associated *F* tests for two models: Model 3 with random effects for JOL intercepts only and the final model, Model 5. The major difference between these two models is the within-subject fixed effect for estimated RTs is highly significant in Model 3 but was eliminated with random effects for two variables closely related to estimated RTs, recognition accuracy and CJs, were added to the model. In addition, *p* values for several person-centered predictors were reduced in Model 5 relative to Model 3, suggesting that these effects may have been an artifact of not modeling individual differences in the effects of key MPT predictors. We focus the rest of our treatment on the outcomes from Model 5.

Table 3
Multi-Level Regression Predicting Trial 2 Judgments of Learning

Effect	Model with random intercept		Model with all random effects	
	Estimate	*SE*	Estimate	*SE*
Intercept	74.85	4.67	79.23	4.48
Trial 1 study time (within subjects)	−0.02	0.05	−0.03	0.05
Trial 1 study time (between subjects)	−0.50	0.33	−0.01	0.32
Trial 1 JOL (within subjects)	0.11	0.02	0.11	0.02
Trial 1 JOL (between subjects)	0.24	0.10	0.18	0.09
Trial 1 recognition accuracy (within subjects)	12.79	1.08	12.12	1.59
Trial 1 recognition accuracy (between subjects)	18.68	16.57	8.73	15.33
Trial 1 CJ (within subjects)	0.16	0.02	0.15	0.03
Trial 1 CJ (between subjects)	0.45	0.21	0.24	0.19
Trial 1 recognition RT (within subjects)	0.26	0.34	0.02	0.34
Trial 1 recognition RT (between subjects)	−1.64	3.30	−0.48	3.09
Trial 1 recognition RT estimate (within subjects)	−1.71	0.53	−1.06	0.54
Trial 1 recognition RT estimate (between subjects)	−0.95	1.83	0.70	1.72
Trial 2 study time (within subjects)	−0.61	0.15	−0.69	0.15
Trial 2 study time (between subjects)	0.76	0.66	0.58	0.62

Note. JOL = judgment of learning; CJ = confidence judgment. All temporal measures are scaled in seconds. Effects associated with specific word pair cues were not included in this table but can be generated using code found in Supplements C and D.

On average, participants' Trial 2 JOL confidence was approximately 79%. The random intercept variance = 109 (see Table 5). Assuming a normal distribution of individual differences in intercepts, about 84% of the participants had mean Trial 2 JOLs in the interval between 69% and 89% confidence.

The systematic item differences in mean JOLs, captured by the Cue variable in Table 4, did not reach the Type I criterion. Nevertheless, covarying on the specific item variable assured that the remaining within-person regression effects could not be attributed to unusual characteristics of a few items. There were indications of stability in JOLs at the item-level across the two study trials, both at the item-level and at the person-level. A 1% difference in an item's Trial 1 JOL predicted a 0.11% difference in Trial 2 JOL for that item (see Table 3).

The substantively critical effects involved the expanded MPT hypothesis that recognition accuracy and recognition CJs would both influence Trial 2 JOLs. The analysis revealed reliable item-level effects of Trial 1 recognition memory accuracy on Trial 2 JOLs, consistent with the MPT hypothesis. Participants were approximately 12% more confident in their learning given an accurate prior item-level recognition test. However, the model also supported the argument that subjective confidence in recognition outcomes influenced later JOLs, controlling on recognition accuracy. Trial 1 CJs also predicted Trial 2 JOLs, with a 1% increase in an item-level CJ confidence predicted a 0.15% increase in subsequent JOL confidence for that same item. Given that Trial 1 JOLs were also included as predictors in the regression model, both of these effects can be interpreted as indicating that changes in JOLs from Trial 1 to Trial 2 were predicted both by recognition memory accuracy and by recognition memory CJs.

These fixed effects represent the average influence across all persons of recognition accuracy and CJs on Trial 2 JOLs. The significant random effects associated with these predictors indicate that the effects differ across persons. Table 5 reports the estimated covariance matrix of the random effects. Assuming a normal distribution, 95% of the within-person recognition accuracy effects lie in the interval 12.11 ± 14.23. Thus, a few individuals show no effect of recognition accuracy on Trial 2 JOLs, whereas a few at the other extreme show a greater than 25% confidence difference for previously correct versus incorrect items. Likewise, 95% of the within-person CJ effects lie in

Table 4
F-Tests for Multi-Level Regression Fixed Effects

Effect	Model with random intercept			Model with all random effects		
	df	*F*	*p*	*df*	*F*	*p*
Trial 1 Study Time (Within Subjects)	3009.00	0.17	0.683	2952.00	0.24	0.626
Trial 1 Study Time (Between Subjects)	51.00	2.24	0.141	57.50	0.05	0.820
Trial 1 JOL (Within Subjects)	3009.00	41.46	<.001	2979.00	40.06	<.001
Trial 1 JOL (Between Subjects)	51.00	6.52	0.014	54.80	4.22	0.045
Trial 1 Recognition Accuracy (Within Subjects)	3009.00	139.16	<.001	36.60	58.37	<.001
Trial 1 Recognition Accuracy (Between Subjects)	51.00	1.27	0.265	51.40	0.32	0.572
Trial 1 CJ (Within Subjects)	3009.00	88.80	<.001	46.90	24.72	<.001
Trial 1 CJ (Between Subjects)	51.00	4.66	0.036	50.00	1.59	0.213
Trial 1 RT (Within Subjects)	3009.00	0.62	0.431	2846.00	0.00	0.963
Trial 1 RT (Between Subjects)	51.00	0.25	0.621	54.30	0.02	0.878
Trial 1 RT Estimate (Within Subjects)	3009.00	10.55	0.001	2590.00	3.79	0.052
Trial 1 RT Estimate (Between Subjects)	51.00	0.27	0.608	54.30	0.17	0.684
Trial 2 Study Time (Within Subjects)	3009.00	16.12	<.001	2980.00	20.79	<.001
Trial 2 Study Time (Between Subjects)	51.00	1.31	0.257	51.90	0.87	0.355
Cue	3009.00	1.43	0.017	2971.00	1.35	0.039

Note. JOL = judgment of learning; CJ = confidence judgment; all time-based measures are scaled in seconds.

the interval 0.15 ± 0.30, indicating that some individuals actually manifest a negative CJ-JOL relationship.[5]

Given the presence of the two random effects, one possibility is that the two metacognitive cues play off against one another, with some persons weighting CJs more when making JOLs, and other persons weighting recognition accuracy more. If this were the case, we would expect a negative covariance between the two sources of random effects. Table 5 shows this is not the case; the estimated covariance is positive but not different from zero, $z = 0.65$, $p > .50$. It appears that the effects for the two cues across persons are essentially independent of one another.

Critically, Trial 2 item study time, capturing reprocessing fluency, predicted Trial 2 JOLs, showing the influence of emergent aspects of Trial 2 encoding experiences on JOLs independent of MPT effects and all other predictors. An increase of 1s in Trial 2 study time for an item was related to a decrease of approximately 0.69% JOL confidence for that item.

We had hypothesized an effect of subjective fluency, as measured by estimated RTs, on Trial 2 JOLs. As can be seen in Table 3, the within-person regression of JOLs on estimated RTs was not reliable in Model 5, although it had been significant in Model 3. One potential issue is that estimated RTs correlate moderately with actual RTs, consistent with the idea that accurate monitoring of recognition Test RTs generates a correlation of subjective and objective latency. Indeed, when we dropped actual RT variables from the model, the effect of within-person estimated RTs changed to −1.04, $SE = 0.44$, but failed to reach the specified level of 99% confidence to reject the null hypothesis, $p = .017$.

Table 5
Covariance Matrix of Random Effects For Model 5 (Standard Errors in Parentheses)

	Intercept	Trial 1 Accuracy	Trial 1 CJs
Intercept	108.69 (25.37)	—	—
Trial 1 Accuracy	−7.85 (17.63)	52.68 (24.22)	—
Trial 1 CJs	1.17 (0.38)	0.19 (0.29)	0.024 (.008)

Note. Only the lower triangular portion of the symmetric matrix is shown.

Discussion

This study provides unambiguous evidence that JOLs in a multitrial associative recognition task are influenced by multiple sources of information, not just recall of past memory performance accuracy. Although the results indicated a direct influence of past test accuracy on Trial 2 JOLs, consistent with the operation of a memory-for-past-test heuristic (Finn & Metcalfe, 2007, 2008), they also show that recognition CJs influence Trial 2 JOLs independent of prior test performance. The independent effect of CJs supports a conception of the MPT heuristic as a broader subjective experience of past memory access that is determined by multiple cues about underlying memory representations, not just retrieved memories of past test accuracy.

Our findings also confirm inferences from recent research (e.g., Ariel & Dunlosky, 2011; Tauber & Rhodes, 2012) that other cues besides those affecting a MPT heuristic must be influencing JOLs. Critically, however, we showed that Trial 2 self-paced study time predicted JOLs independent of the MPT-related variables. We view this effect as reprocessing fluency influencing on JOLs, in part because the effect was significant when controlling on study time, memory accuracy, and CJs at Trial 1 (variables that should carry any effects of item-variance in normative learning difficulty). The analysis controls for the influence of MPT-related variables on study time, implying that the observed Trial 2 study time effect is due to emergent influences associated with Trial 2 study behavior, such as actual or perceived encoding fluency (see Robinson et al., 2006). When other published studies have specifically examined time to complete encoding operations (e.g., interactive images), similar negative correlations have been found for encoding times and JOLs (e.g., Hertzog et al., 2003). The effect seen here is specific to reprocessing fluency because it was not observed for Trial 1 study time–JOL relationships in this experiment. Given that overall study time subsumes initial encoding fluency and other influences (such as additional item rehearsal), initial

[5] One uninteresting possible interpretation of the random effect for recognition accuracy is that a few persons close to or at ceiling in recognition memory performance could not contribute within-person variation in recognition outcomes to predict Trial 2 JOLs. We checked this possibility by excluding 10 persons with recognition accuracy above 95%. This exclusion had minimal effects on estimated random effects for recognition accuracy or CJs.

study time may not always correlate with JOLs (see Koriat et al., 2006; Nelson, 1993).

In addition to showing that multiple cues have an effect on JOLs (what are termed fixed effects in multilevel regression models), we also demonstrated reliable random effects (individual differences) in the MPT effects associated with recognition accuracy and CJs. Thus, there are individual differences in these variables' effects on JOLs, which is a new and interesting outcome. Finn and Metcalfe (2007) reported near-perfect gamma correlations of a rating of past cued-recall test performance on an item (taken at the time of the next JOL) and the actual memory outcome for that item, implying highly veridical assessment of past test performance by all individuals when they are prompted to make a JOL (see Tauber & Rhodes, 2012). The random effects found in this study would not have been anticipated from those studies—another possible manifestation of differences between recognition and cued-recall MPT phenomena. It is possible that all individuals in our experiments access MPT-related cues when making second-trial JOL, but that some participants discount them when constructing the JOL, perhaps due to individual differences in the amount of recollective detail that was accessed (see Ariel & Dunlosky, 2011). Indeed, examining alternate measures of MPT, such as remember/know/new judgments, could help to clarify MPT effects (given that these judgments may correlate more highly with test performance than CJs; McCabe, Geraci, Boman, Sensenig, & Rhodes, 2011). Regardless, the reasons why some individuals would show weaker item-level MPT effects than other individuals is unknown, and further research is warranted to understand these individual differences.

Tauber and Rhodes (2012) found that earlier immediate JOLs influenced JOLs made in second and third study-test trials in their associative memory task. Our experiment also shows that Trial 1 JOLs influence Trial 2 JOLs, even when controlling on consistent item differences in mean JOLs. Indeed, the only between-person predictor to approach significance, controlling on the random effects, was for mean Trial 1 JOLs predicting mean Trial 2 JOLs. Apparently there are persistent individual differences in mean JOLs across two study trials, which could reflect overall confidence that an individual has about his or her own memory performance (see Kleitman & Stankov, 2007, for arguments about confidence judgments as a facet of personality, and a stable dimension of individual differences).

We do not claim to have captured all relevant cues operating on JOLs. Indeed, the current model accounts for just about half of the item-level variance in Trial 2 JOLs. Of course, most experimental studies do not attempt to account for all item variance, assuming large amounts of error variance that are often hidden by aggregation over items prior to data analysis. Nevertheless, it is likely the case that variables not included in this study would have explained additional variance in Trial 2 JOLs. As reviewed earlier in this article, many different cues have been demonstrated as having valid and invalid influences on JOLs, and many of these are likely operate in multiple-trial learning. Other unmeasured cues could certainly have stronger influences on JOLs than Trial 2 study time. For example, Hertzog, Sinclair, and Dunlosky (2010) showed that successful, uninstructed use of mediational strategies during a single study-test trial (i.e., spontaneous use of interactive imagery to create the new association) influenced JOLs independently of a manipulated item characteristic, associative relatedness. We did not measure strategy use in the present experiments, but it is likely that spontaneous mediator use occurred (e.g., Naveh-Benjamin, Brav, & Levy, 2007; Pyc & Rawson, 2012; Richardson, 1998). It could even be the case that a major influence on study time at Trial 2 is the difference between retrieving a mediator formed at Trial 1 versus constructing a new mediator. Remembering the mediator used at Trial 1 may lead to reduced study time and higher confidence that the mediator will facilitate subsequent item recognition.

One implication of the multiple-cue perspective argued here is a need to be concerned about how the mixture of measured and unmeasured cues might alter estimated regression effects. As an anonymous reviewer reminded us, self-generated validity effects could be operating (Feldman & Lynch, 1988) such that simply drawing attention to a cue by measuring it may alter its influence on metacognitive judgments. For instance, preliminary analyses of counterbalancing effects in these data (see Supplement A) found a trend toward an interaction of counterbalancing order of judgments (CJs followed by RT estimates, or vice versa) on the effect of RT estimates on Trial 2 JOLs. Thus, one cannot assume that estimated cue effects in one experiment will generalize to other experiments with different sets of measured variables or instructions that may affect the salience of different cues at the time of the metacognitive judgment. Spontaneous reliance on cues differs in principle from experimenter-guided attention to cues, as may the metacognitive consequences of explicit cue measurement. This concern underscores the value of replicating effects across variations in task contexts, as argued decades ago by Brunswik (1955) when he promoted the concept of representative design. In that vein, this study reinforces previous findings regarding predictive value of past memory performance (and, implicitly, memory performance monitoring) for subsequent JOLs, and also demonstrating that MPT effects differ between recall and recognition task contexts because of imperfect performance monitoring.

This study illustrates how multiple cues can be explicitly measured and then entered into a multilevel regression model to evaluate whether cues operate independently of one another to influence JOLs. These cues can either be experimentally manipulated or merely passively observed, but they must be empirically measured. We readily acknowledge that this study does not definitively address a number of critical questions about the nature of the multiple-cue utilization process. For example, what are the limits on the number of cues that can be considered on any given trial? What evidence can be generated that multiple cues are used on any given item, rather than different cues being accessed on different subsets of items? Would instructions to give different weights to specific different cues alter their relative impact on JOLs?

It is common practice at present for metacognitive researchers to manipulate one or two cues in isolation and then evaluate their influence on metacognitive judgments without consideration of other cues that may also be concomitantly influencing those judgments. Experimental isolation is of course highly useful and even critical for identifying candidate influences on metacognitive judgments. However, we argue that such approaches are usually insufficient as a means of testing multiple-cue theories about metacognitive judgments because it is often impractical or even impossible to manipulate all relevant cues in a single experiment (especially those cues that derive from participant behaviors and strategies, such as encoding fluency effects). As shown here, demonstrations of potent MPT accuracy effects as in earlier research do not rule out other influences on Trial 2 JOLs, and the typical experiment cannot fully control participant-generated access to relevant cues. Evaluating the influence of multiple cues on metacognitive judgments requires multilevel models like the ones used here to test whether measured cues have a unique influence on JOL magnitudes controlling on other measured cues, whether experimentally manipulated or passively observed.

References

Ariel, R., & Dunlosky, J. (2011). The sensitivity of judgments-of-learning resolution to past test performance, new learning, and forgetting. *Memory & Cognition, 39,* 171–184. doi:10.3758/s13421-010-0002-y

Begg, I., Duft, S., Lalonde, P., Melnick, R., & Sanvito, J. (1989). Memory predictions are based on ease of processing. *Journal of Memory and Language, 28,* 610–632. doi:10.1016/0749-596X(89)90016-8

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. doi:10.1037/0096-3445.127.1.55

Bower, G. (2000). A brief history of memory research. In E. Tulving & F. I. M. Craik (Eds.), *The Oxford handbook of memory* (pp. 3–32). New York, NY: Oxford University Press.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review, 62,* 193–217. doi:10.1037/h0047470

Cerella, J., Onyper, S. V., & Hoyer, W. J. (2006). The associative memory basis of cognitive skill learning. *Psychology and Aging, 21,* 483–498. doi:10.1037/0882-7974.21.3.483

Cohn, M., & Moscovitch, M. (2007). Dissociating measures of associative memory: Evidence and theoretical implications. *Journal of Memory and Language, 57,* 437–454. doi:10.1016/j.jml.2007.06.006

Craik, F. I. M., & Hay, J. F. (1999). Aging and judgments of duration: Effects of task complexity and method of estimation. *Perception & Psychophysics, 61,* 549–560. doi:10.3758/BF03211972

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. doi:10.1037/0882-7974.15.3.462

Dunlosky, J., Hertzog, C., & Powell-Moman, A. (2005). The contribution of mediator-based deficiencies to age differences in associative learning. *Developmental Psychology, 41,* 389–400. doi:10.1037/0012-1649.41.2.389

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition*. Thousand Oaks, CA: Sage.

Eakin, D. K., & Hertzog, C. (2012). Immediate judgments of learning are insensitive to implicit interference effects at retrieval. *Memory & Cognition, 40,* 8–18. doi:10.3758/s13421-011-0138-4

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multi-level models: A new look at an old issue. *Psychological Methods, 12,* 121–138. doi:10.1037/1082-989X.12.2.121

Feldman, J. M., & Lynch, J. G., Jr. (1988). Self-generated validity and other effects of measurement on belief, attitude, intention, and behavior. *Journal of Applied Psychology, 73,* 421–435. doi:10.1037/0021-9010.73.3.421

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. doi:10.1037/0278-7393.33.1.238

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

Gelman, A., & Hill, J. (2007). *Data Analysis using regression and multi-level/hierarchical models*. Cambridge, NY: Cambridge University Press.

Gonzalez, R., & Nelson, T. O. (1996). Measuring ordinal association in situations that contain tied scores. *Psychological Bulletin, 119,* 159–165. doi:10.1037/0033-2909.119.1.159

Hertzog, C., Dunlosky, J., Robinson, A. E., & Kidder, D. P. (2003). Encoding fluency is used for judgments about learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 22–34. doi:10.1037/0278-7393.29.1.22

Hertzog, C., Dunlosky, J., & Sinclair, S. (2010). Episodic feeling-of-knowing resolution derives from the quality of original encoding. *Memory & Cognition, 38,* 771–784. doi:10.3758/MC.38.6.771

Hertzog, C., Price, J., Burpee, A., Frentzel, W. J., Feldstein, S., & Dunlosky, J. (2009). Why do people show minimal knowledge updating with task experience: Inferential deficit or experimental artifact? *Quarterly Journal of Experimental Psychology, 62,* 155–173. doi:10.1080/17470210701855520

Hertzog, C., Sinclair, S. M., & Dunlosky, J. (2010). Age differences in the monitoring of learning: Cross-sectional evidence of spared resolution across the life-span. *Developmental Psychology, 46,* 939–948. doi:10.1037/a0019812

Hertzog, C., Touron, D. R., & Hines, J. C. (2007). Does a time monitoring deficit influence older adults' delayed retrieval shift during skill acquisition? *Psychology and Aging, 22,* 607–624. doi:10.1037/0882-7974.22.3.607

Hicks, J. L., & Marsh, R. L. (2001). False recognition occurs more frequently during source identification than during old-new recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 375–383. doi:10.1037/0278-7393.27.2.375

Hines, J. C., 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. doi:10.1037/a0014417

Hoffman, L., & Rovine, M. J. (2007). Multilevel models for the experimental psychologist: Foundations and illustrative examples. *Behavior Research Methods, 39,* 101–117. doi:10.3758/BF03192848

Jacoby, L. L., & Rhodes, M. G. (2006). False remembering in the aged. *Current Directions in Psychological Science, 15,* 49–53. doi:10.1111/j.0963-7214.2006.00405.x

Kelley, C. M., & Sahakyan, L. (2003). Memory, monitoring, and control in the attainment of memory accuracy. *Journal of Memory & Language, 48,* 704–721. doi:10.1016/S0749-596X(02)00504-1

Keren, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica, 77,* 217–273. doi:10.1016/0001-6918(91)90036-Y

Kleitman, S., & Stankov, L. (2007). Self-confidence and metacognitive processes. *Learning and Individual Differences, 17,* 161–173. doi:10.1016/j.lindif.2007.03.004

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

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,* 187–194. doi:10.1037/0278-7393.31.2.187

Koriat, A., & Bjork, R. A. (2006). Illusions of competence during study can be remedied by manipulations that enhance learners' sensitivity to retrieval conditions at test. *Memory & Cognition, 34,* 959–972. doi:10.3758/BF03193244

Koriat, A., Ma'ayan, H., & Nussinson, R. (2006). The intricate relationships between monitoring and control in metacognition: Lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology: General, 135,* 36–69. doi:10.1037/0096-3445.135.1.36

Kornell, N., Rhodes, M. G., Castel, A. D., & Tauber, S. K. (2011). The ease-of-processing heuristic on the stability bias: Dissociating memory, memory beliefs, and memory judgments. *Psychological Science, 22,* 787–794. doi:10.1177/0956797611407929

Littell, R. C., Milliken, G. A., Stroup, W. W., & Wolfinger, R. D. (2000). *SAS system for mixed models* (4th. ed.). Cary, NC: SAS Institute, Inc.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review, 95,* 492–527. doi:10.1037/0033-295X.95.4.492

Masson, M. E. (1993). Episodically enhanced comprehension fluency. *Canadian Journal of Experimental Psychology, 47,* 428–465. doi:10.1037/h0078822

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. doi:10.1037/0096-3445.122.1.47

McCabe, D. P., Geraci, L., Boman, J. K., Sensenig, A. E., & Rhodes, M. G. (2011). On the validity of remember–know judgments: Evidence from think aloud protocols. *Consciousness and Cognition, 20,* 1625–1633. doi:10.1016/j.concog.2011.08.012

Naveh-Benjamin, M., Brav, T. K., & Levi, D. (2007). The associative memory deficit of older adults: The role of efficient strategy utilization. *Psychology and Aging, 22,* 202–208. doi:10.1037/0882-7974.22.1.202

Nelson, T. O. (1993). Judgments of learning and the allocation of study time. *Journal of Experimental Psychology: General, 122,* 269–273. doi:10.1037/0096-3445.122.2.269

Nelson, T. O. (1996). Consciousness and metacognition. *American Psychologist, 51,* 102–116. doi:10.1037/0003-066X.51.2.102

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. doi:10.1016/S0079-7421(08)60053-5

Parks, C. M. (2007). The role of noncriterial recollection in estimating recollection and familiarity. *Journal of Memory and Language, 57,* 81–100. doi:10.1016/j.jml.2007.03.003

Pinheiro, J., & Bates, D. (2000). *Mixed-effects models in S and S-plus*. New York, NY: Springer-Verlag. doi:10.1007/978-1-4419-0318-1

Pyc, M. A., & Rawson, K. A. (2012). Why is retrieval practice beneficial for memory? An evaluation of the mediator shift hypothesis. Journal of Experimental Psychology. *Learning, Memory & Cognition, 38,* 737–746. DOI; 10.1037/a0026166.

Ratcliff, R., & Starns, J. J. (2009). Modeling confidence and response time in recognition memory. *Psychological Review, 116,* 59–83. doi:10.1037/a0014086

Ratcliff, R., Thapar, A., & McKoon, G. (2011). Effects of aging and IQ on associative memory. *Journal of Experimental Psychology: General, 140,* 464–487. doi:10.1037/a0023810

Rhodes, M. G., & Castel, A. D. (2008). Memory predictions are influenced by perceptual information: Evidence for metacognitive illusions. *Journal of Experimental Psychology: General, 137,* 615–625. doi:10.1037/a0013684

Richardson, J. T. E. (1998). The availability and effectiveness of reported mediators in associative learning: A historical review and an experimental investigation. *Psychonomic Bulletin & Review, 5,* 597–614. doi:10.3758/BF03208837

Robinson, A. E., Hertzog, C., & Dunlosky, J. (2006). Aging, encoding fluency, and metacognitive monitoring. *Aging, Neuropsychology, and Cognition, 13,* 458–478. doi:10.1080/13825580600572983

Roediger, H. L., & McDermott, K. B. (1993). Implicit memory in normal human subjects. In F. Boller & J. Grafman (Eds.), *Handbook of neuropsychology* (Vol. 8, pp. 63–131). Amsterdam, The Netherlands: Elsevier.

Shaughnessy, J. J. (1981). Memory monitoring accuracy and modification of rehearsal strategies. *Journal of Verbal Learning & Verbal Behavior, 20,* 216–230. doi:10.1016/S0022-5371(81)90389-3

Shing, Y. L., Werkle-Bergner, M., Li, S.-C., & Lindenberger, U. (2009). Committing memory errors with high confidence: Older adults do but children don't. *Memory, 17,* 169–179. doi:10.1080/09658210802190596

Singer, J. D. (1998). Using SAS proc mixed to fit multi-level models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics, 23,* 323–355.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multi-level analysis: An introduction to basic and advanced multi-level modelling*. London, UK: Sage.

Tauber, S. K., & Rhodes, M. G. (2012). Multiple bases for young and older adults' judgments of learning in multitrial learning. *Psychology and Aging, 27,* 474–483. doi:10.1037/a0025246

Thiede, K. W., & Dunlosky, J. (1999). Toward a general model of self-regulated study: An analysis of selection of items for study and self-paced study time. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 1024–1037. doi:10.1037/0278-7393.25.4.1024

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,* 327–337. doi:10.1027/1618-3169/a000039

Undorf, M., & Erdfelder, E. (2011). Judgments of learning reflect encoding fluency: Conclusive evidence for the ease-of-processing hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37,* 1264–1269. doi:10.1037/a0023719

Weaver, C. A., III, & 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. doi:10.1037/0278-7393.29.6.1058

Weber, N., & Brewer, N. (2003). The effect of judgment type and confidence scale on confidence-accuracy calibration in face recognition. *Journal of Applied Psychology, 88,* 490–499. doi:10.1037/0021-9010.88.3.490

Yonelinas, A. P., & Parks, C. M. (2007). Receiver operating characteristics (ROCs) in recognition memory: A review. *Psychological Bulletin, 133,* 800–832. doi:10.1037/0033-2909.133.5.800

Received December 19, 2012
Revision received May 23, 2013
Accepted June 5, 2013 ■