Psychological Review
2012, Vol. 119, No. 1, 186–200

0033-295X/11/$12.00 DOI: 10.1037/a0025960

THEORETICAL NOTE

A Stochastic Detection and Retrieval Model for the Study of Metacognition

Yoonhee Jang
University of California, San Diego

Thomas S. Wallsten
University of Maryland, College Park

David E. Huber
University of California, San Diego

We present a signal detection-like model termed the stochastic detection and retrieval model (SDRM) for use in studying metacognition. Focusing on paradigms that relate retrieval (e.g., recall or recognition) and confidence judgments, the SDRM measures (1) variance in the retrieval process, (2) variance in the confidence process, (3) the extent to which different sources of information underlie each response, (4) simple bias (i.e., increasing or decreasing confidence criteria across conditions), and (5) metacognitive bias (i.e., contraction or expansion of the confidence criteria across conditions). In the metacognition literature, gamma correlations have been used to measure the accuracy of confidence judgments. However, gamma cannot distinguish between the first 3 attributes, and it cannot measure either form of bias. In contrast, the SDRM can distinguish among the attributes, and it can measure both forms of bias. In this way, the SDRM can be used to test competing process theories by determining the attribute that best accounts for a change across conditions. To demonstrate the SDRM's usefulness, we investigated judgments of learning (JOLs) followed by cued-recall. Through a series of nested and non-nested model comparisons applied to a new experiment, the SDRM determined that a reduction in variance during the confidence process is the most likely explanation of the delayed-JOL effect, and a stronger relation between information underlying JOLs and recall is the most likely explanation of the testing-JOL effect. Following a brief discussion of implications for JOL theories, we conclude with a broader discussion of how the SDRM can benefit metacognition research.

Keywords: signal detection theory, metacognition, confidence judgment accuracy, criterion variability, judgments of learning

Supplemental materials: http://dx.doi.org/10.1037/a0025960.supp

Research on metacognition typically involves memory paradigms in which participants provide confidence judgments before, during, or after recall or recognition (Nelson & Narens, 1994). For example, in a judgment of learning (JOL) experiment, participants (a) study a list of stimulus–response pairs, (b) at some point view each stimulus and rate their confidence (JOL) that when presented with it again they will be able to recall the associated response, and (c) undergo cued-recall testing. The results of many metacognition experiments were analyzed with the Goodman–Kruskal gamma correlation between rated confidence and memory performance because gamma was believed to have desirable statistical properties (Gonzalez & Nelson, 1996; Nelson, 1984; although see Masson & Rotello, 2009). Gamma, also called resolution, provides an index of how well confidence judgments predict actual memory performance (i.e., metacognitive accuracy) on an item-by-item basis. However, it is a blunt tool to use for testing hypotheses generated from theoretical predictions because it ignores the rich quantitative complexity of the data that may serve to constrain theory. Formal cognitive models are required for this purpose.

Despite the substantial empirical literature on metacognition that has developed over the years (a search of the PsycINFO database using the keyword of *metacognition* yielded nearly 4,500 publications), the field suffers from a dearth of formal models. Notable exceptions are the JOL model proposed by Sikström and Jönsson (2005) and the source activation confusion model (Schunn, Reder, Nhouyvanisvong, Richards, & Stroffolino, 1997). The former invokes slow and fast drifts in memory trace strength to explain the relation between JOLs and subsequent cued-recall performance, and the latter is a network model applied to para-

This article was published Online First November 7, 2011.

Yoonhee Jang and David E. Huber, Department of Psychology, University of California, San Diego; Thomas S. Wallsten, Department of Psychology, University of Maryland, College Park.

This research was supported by National Institute of Mental Health Grant RMH081084A to David E. Huber. We thank Aaron Benjamin and Sverker Sikström for their valuable comments.

Correspondence concerning this article should be addressed to Yoonhee Jang, Department of Psychology, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0109. E-mail: yhjang@ucsd.edu

digms that involve predictions of future recognition of a currently unrecalled item. Augmenting these process models, we propose a measurement model, which provides a formal framework for theory testing. The core of this model assumes two samplings of memory strength per stimulus (e.g., one for retrieval and the other for confidence), allowing separate descriptions of the confidence and retrieval processes. We developed this framework upon realizing that experimental manipulations may affect memory retrieval, use of the confidence scale, or the type of information used for each process, and that current methods of theory formulation and testing cannot distinguish among these possibilities. This model, which we call the stochastic detection and retrieval model (SDRM), is closely linked to signal detection theory (SDT; Green & Swets, 1966; Macmillan & Creelman, 2005) in general and specifically to SDT models of recognition memory. In the spirit of SDT, the SDRM can be applied to a wide variety of memory retrieval and confidence judgments.

The remainder of this article is organized as follows. First, we motivate and develop the SDRM and discuss its relation to SDT. Next, we describe how to use the SDRM as a tool for testing hypotheses about latent processes underlying recall and judgment. This section illustrates the power of the SDRM by applying it to theoretical issues currently under dispute in recall and JOL research. Then, we briefly describe a new experiment to address the issues, apply the SDRM to the data, and draw theoretical conclusions much stronger than those available from the gamma-based analyses typically found in this literature. Finally, the article concludes with discussions about the status of metacognition research and about the SDRM itself.

The SDRM and Its Relation to SDT

The SDRM is similar to SDT in assuming that latent real-valued continuous decision variables underlie memory retrieval and confidence judgments and that observed responses depend on the location of a sampled decision variable relative to one or more criteria. Like SDT, the SDRM is not intended as a model for predicting behavior but is instead a tool for measuring the cognitive processes that lead to the latent decision variables and therefore can be used as a vehicle for testing hypotheses about the underlying processing. Like some of the SDT recognition memory models (see, e.g., Jang, Wixted, & Huber, 2009), the SDRM is a family of multinomial signal-detection-based models and allows for nested model comparisons (e.g., Batchelder & Riefer, 1990).

However, the SDRM differs from SDT in three important ways. We first list these ways and then justify or amplify them.

1. SDT applies only to paradigms that include two (or sometimes more) independently defined categories of stimuli (e.g., studied and non-studied words; presented and non-presented items; benign and non-benign tumors), and therefore SDT does not apply to tasks such as recall memory that involve a single stimulus category. In contrast, the SDRM categorizes trials on the basis of the participant's retrieval behavior (e.g., recalled and not-recalled stimuli).

2. SDT assumes that the decision variable is sampled once per stimulus, and the same sample is used for both binary choice behavior and a confidence rating. In contrast, the SDRM assumes two distinct (possibly related) memory strength samplings, X and Y, with one providing a strength value that is compared to a retrieval threshold and the other proving a strength value that is compared to confidence criteria. The SDRM operationalizes these assumptions in the form of an X–Y bivariate memory-strength distribution.

3. Due to its mathematical structure, SDT cannot identify noise in the confidence rating process (e.g., criterion variance) separate from memory strength variability. In contrast, the SDRM can in some (but not all) paradigms identify the relative contribution of trial-by-trial noise in the confidence and retrieval processes.

To justify Point 1, we assume that information underlying recall exists along a strength continuum, consistent with the fact that people often have partial memories when they cannot fully recall an item or event (e.g., Koriat, Levy-Sadot, Edry, & de Marcas, 2003; Nelson & Narens, 1994). Because there are only targets (and no foils) in a recall memory experiment, the two distributions on this strength continuum, one for recalled and the other for not-recalled memories, result from the internal processes of the observer rather than from independently defined stimulus states (for the recognition memory paradigm, see Clarke, Birdsall, & Tanner, 1959; Galvin, Podd, Drga, & Whitmore, 2003; Pollack, 1959).

Regarding Point 2, some studies found dissociations between confidence and recall accuracy (e.g., Begg, Duft, Lalonde, Melnick, & Sanvito, 1989; Benjamin, Bjork, & Schwartz, 1998; Nelson & Narens, 1994), leading to the claim that a variety of different cues are used for confidence judgments that might or might not be consistent with the factors underlying recall (for detail, see, e.g., Jang & Nelson, 2005; Koriat, 1997). To address this matter, the SDRM assumes that distinct real-valued memory traces underlie confidence judgment and recall tasks. Thus, the SDRM assumes two samplings of memory, one in the service of recall and the other leading to a confidence judgment, and therefore two types of response thresholds (or criteria). When sampling memory for recall, a trace above some threshold leads to recall of that item.[1] When sampling to provide a confidence estimate, the resulting judgment depends on the trace strength relative to the locations of confidence criteria. This perspective is consistent with results in recognition memory research suggesting that different memory strength distributions underlie old/new decisions and confidence judgments (e.g., Busey, Tunnicliff, Loftus, & Loftus, 2000; Tulving, 1981; Van Zandt, 2000). If the two separate samples (or the two different detections) rely on the same underlying information (e.g., encoding strength), and furthermore, if that information remains unchanged between the confidence rating and retrieval, then the samples will be perfectly, or at least very strongly, correlated. However, if some types of information are more important for one process or the other, or if the strength distribution decays or changes in some manner between the confidence rating

[1] It is easy to imagine that multiple memories may exceed the threshold, in which case, the strongest trace would determine the response. Because the strongest trace tends to be the correct one for cued-recall, we limit theorizing in this article to the assumption that the correct trace either is or is not above the threshold and do not distinguish errors of omission due to no trace exceeding threshold from errors of commission due to a misleading trace being the highest above threshold. This could be modeled in the SDRM by including an additional sampling distribution for extra-list items, although we leave this extension to future development.

and retrieval, then these two samples may be somewhat less correlated. The SDRM represents the two samplings by means of a bivariate, instead of a univariate, memory strength distribution, which allows any degree of correlation between the two sampled memory strengths. With this double-detection model, collection of confidence in relation to recall can be a useful method for illuminating the nature of both recall and confidence.

To address Point 3, although criterion variability is mathematically non-identifiable under standard applications of SDT, some receiver operating characteristic (ROC) analyses in recognition memory research have considered it. That is, the non-linearity sometimes found in individual z-transformed ROC (zROC) curves (e.g., Heathcote, 2003; Ratcliff, McKoon, & Tindall, 1994) can be interpreted as a consequence of decision noise. Indeed, recent work in recognition memory has challenged the fundamental assumption that criterion placement is a noise-free process and has extended SDT to include decision noise (e.g., Benjamin, Diaz, & Wee, 2009; Mueller & Weidermann, 2008). In suitable paradigms, the SDRM can identify the relative contribution of such variance in both the retrieval and confidence processes by using the approach developed for the stochastic judgment model (SJM; Wallsten & González-Vallejo, 1994), which has been applied to statement verification and probability judgment tasks. The SJM decomposes covert confidence into a knowledge and an error component, and an overt response is treated as a decision based on a degree of covert confidence, which itself depends on the item in question and possibly on momentary fluctuations. Like the SJM, the SDRM assumes that on a trial-by-trial basis, or perhaps as a result of cumulating learning or fatigue, confidence judgment criteria and the recall thresholds may vary.

To summarize and amplify these ideas a bit further, the SDRM allows two samplings of memory traces, one that subserves memory retrieval and another that subserves ratings of confidence. The strength variables at these two different stages may be correlated perfectly, not at all, or anywhere in between. It is convenient, therefore, to represent the memory strength variable by a bivariate X–Y distribution. Memory retrieval is successful when the X sample is sufficiently strong, modeled as exceeding a possibly noisy threshold, and the confidence estimate depends on where the Y sample falls relative to possibly noisy confidence criteria. Because this approach makes no assumptions regarding the timing of the two types of tasks, the model structure is identical for confidence judgments that precede or follow memory retrieval (i.e., prospective or retrospective confidence judgments). Additionally, there is nothing special about the bivariate distribution, which simply allows us to handle the two tasks within a single coherent structure. If the paradigm called for a third task, say another confidence estimate or recall opportunity, we could assume a trivariate distribution. For purposes of this article, we only implement a bivariate distribution. The SDRM is potentially applicable to a wide variety of phenomena in both recall and recognition memory that involve comparisons between different types of judgments. The SDRM assumes two (or more) samplings of memory strength per stimulus but does not specify the representation of those traces or factors that may affect their strength. Like SDT, the SDRM is a theoretical framework within which multiple contrasting assumptions can be tested and compared.

Model Details

To facilitate understanding of the SDRM, consider the example data sets in Figure 1, which summarize results from a JOL-recall experiment that we describe subsequently. Each panel displays the joint JOL-recall outcome response distribution from a different experimental condition. The abscissa shows the six JOL rating scale categories, and the filled and empty bars represent unsuccessful and successful recall performance, respectively. For example, the left-most filled bar in the delayed-JOL panel shows that approximately 40% of the stimuli led to both unsuccessful recall upon testing and a JOL rating of 0% confidence that recall would be successful. The job of the SDRM is to explain these joint response distributions across experimental conditions to determine which latent variables changed between experimental conditions and therefore affected the degree of correspondence between memory retrieval and confidence judgments (e.g., as indexed by gamma). Finally, SDRM-predicted distributions are shown by the circles and are explained below.

In any one condition, the joint response distribution over the 2 × 6, recall-outcome (correct vs. incorrect) by confidence-rating (0%, 20%, . . . , 100%) matrix may be due to factors operating during the process of recall, the process of confidence estimation, or the relation between the two. For simplicity's sake, we first describe the SDRM for a single condition as represented in any one of the panels and then show how it is used to compare two conditions, that is, across panels.

In the SDRM, responses arising from the memory retrieval and confidence judgment processes depend on the sampled memory strength relative to the operative and possibly variable criteria. The top panel of Figure 2 portrays the SDRM representation of this process: Memory strength lies along a continuum X, and an item is recalled if its strength exceeds a memory criterion, C_M. The probability of recall on any given trial, therefore, is the area under the density curve that is above C_M on that trial. As commonly done in SDT, the SDRM makes the simplifying assumption that the memory strength distribution (the broader distribution over X in the figure) is a standard normal distribution with a mean of 0.0 and a standard deviation of 1.0. All other parameters and distributions are scaled relative to the memory strength distribution. We allow variability in C_M, which we illustrate as the taller distribution (M) centered at C_M with standard deviation σ_M.[2] We refer to this noise in the retrieval process as *memory*.[3] Of critical importance is the realization that σ_M indicates the relative contribution of recall criterion variability compared to variability in the memory strength underlying recall, σ_X, which is set to the arbitrary constant 1.0. Thus, *memory* indicates the relative level of noise in the retrieval process as compared to noise in the memory strength distribution.

Consider next the *confidence* component (i.e., noise in the confidence rating process as realized through criterion variability).

[2] Because criterion variability is typically less than memory variability, the criterion distribution is narrow and tall relative to the short and broader memory strength distribution.

[3] Throughout this article, we list the three components of the SDRM that directly affect the gamma correlation in italics. A bias change is a change in the average positions of the criteria and does not affect the correspondence between confidence and retrieval. In contrast, a change in criterion variability for either the *memory* retrieval of *confidence* processes will

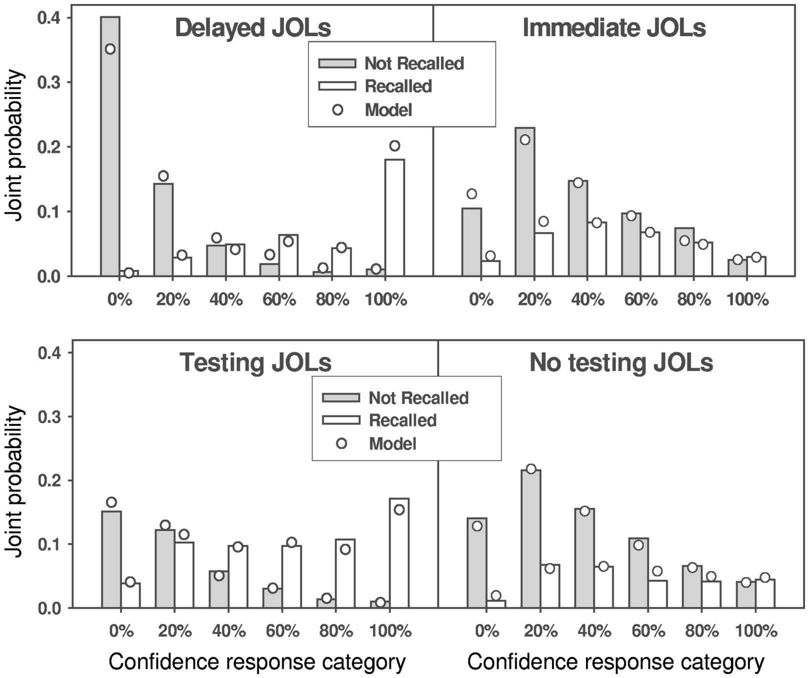

Figure 1. Recall-confidence joint distributions from the reported experiment. The top panels show the delayed (left) and immediate (right) judgment of learning (JOL) conditions, and the bottom panels show the testing (left) and no-testing (right) JOL conditions. The no-testing condition and the immediate condition are functionally identical (both are immediate JOLs without any testing experience), although the data were collected from separate groups of participants. The small dots indicate performance of the stochastic detection and retrieval model with the reported best fitting parameters.

Just as with memory retrieval, the rated confidence that a given item will be or has been recalled depends on memory strength for that item at the time of the JOL rating, as illustrated in the middle panel of Figure 2 (the broader distribution over *Y*). As with the trace strength distribution at recall, we assume that this distribution is normal with a mean of 0.0 and a standard deviation of 1.0. The participant provides a rating that corresponds to the location of the strength variable relative to the decision criteria on that trial: for example, *0% sure* if the variable is below C_1, *20% sure* if it is between C_1 and C_2, *40% sure* if it is between C_2 and C_3, . . . , or *100% sure* if the variable is above C_5. Just as with the memory criterion, we assume that the confidence criteria are variable, and that all are normally distributed with standard deviations σ_C as illustrated by the taller distributions (*C*) centered on the criteria. Similar to the memory criterion, σ_C indicates the relative contribution of confidence criterion variability as compared to variability in the memory strength underlying confidence, which is set to the arbitrary constant 1.0. The criteria may vary in a linked fashion, or they may vary independently of each other. The two possibilities lead to distinct models. To our knowledge, neither is a special case of the other. We cover both cases below.

Finally, memory strength sampled from the *Y* distribution during the confidence stage may or may not be correlated with that sampled from the *X* distribution at recall. The bottom panel of Figure 2 shows the *X*–*Y* bivariate distribution with positive covariance for illustrative purposes. The strength of the correlation between *X* and *Y* depends on the extent of overlap in the memory information sampled at the two stages. At one extreme, the same information underlies both judgments, and the distributions are perfectly correlated with unit covariance; at the other, the sampling is independent in the two cases, and the distributions are unrelated (i.e., 0.0 covariance). We represent this range of possibilities with a linear correlation parameter ρ between *X* and *Y* and refer to this component of the SDRM as *correlation*. Thus, in the SDRM, the memory strength distribution is bivariate normal with both variances set to 1.0, with both means set to 0.0, and with a single ρ parameter determining the relation between *X* and *Y*.

In summary, the degree of correspondence between confidence ratings and memory retrieval, such as traditionally measured with a gamma correlation, is determined by the level of noise in the retrieval process (i.e., *memory* threshold variability), noise in the confidence process (i.e., *confidence* criteria variability), or the *correlation* between the information underlying each process.

To implement the SDRM, we consider the linked-confidence-criteria version first because of its simpler mathematics. The assumption here is that the confidence criteria move up or down on *Y* in lockstep, depending on the randomly sampled criterion noise term. Still just focusing on a single condition of a JOL-recall experiment, consider the joint probability that an item is both successfully recalled and receives JOL rating J_i, where the rating

affect the gamma correlation as will a change in the *correlation* between the information used for each process.

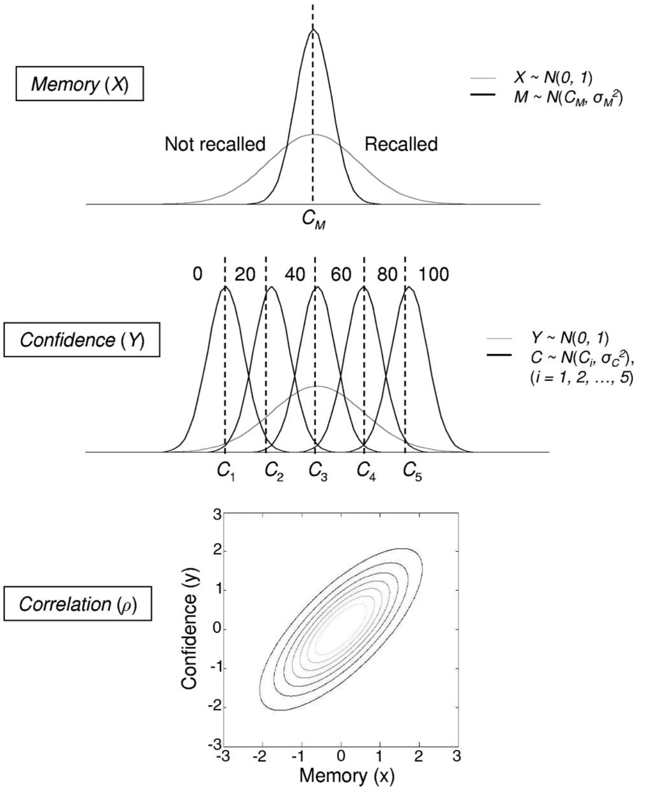

Figure 2. Three components of the stochastic detection and retrieval model that can affect judgment of learning accuracy. The top panel represents the *memory* component: the distribution (X) of memory strength underlying recall responses, which is a standard normal distribution with a mean of 0.0 and a standard deviation of 1.0, and the distribution (M) of the criterion that separates recalled from not-recalled items, which is a normal distribution with a mean of C_M and a standard deviation of σ_M. The middle panel represents the *confidence* component: the distribution (Y) of memory strength underlying confidence judgments, which is a standard normal distribution with a mean of 0.0 and a standard deviation of 1.0, and the five criteria distributions (C) that separate the six confidence judgments (i.e., from *0% sure* to *100% sure*), which are normal distributions with a mean of C_i (i = 1–5) and a standard deviation of σ_C. The bottom panel represents the *correlation* (ρ) component, which determines the relation between X and Y through a bivariate normal distribution (ρ = .75 in this example).

is neither the lowest nor the highest allowed (i.e., $0 < i < n$, where n is the number of confidence criteria and $n + 1$ is the number of JOL categories). This probability is given by

$$p(J_\mathrm{i}, \mathit{recalled}) = \iint h(x,y,\rho)N(x|C_M,\sigma_M)[N(y|C_{i+1},\sigma_C) - N(y|C_i,\sigma_C)]dxdy, \quad (1)$$

where $N(x|C_M,\sigma_M)$ is the cumulative probability that the sampled recall criterion is less than the sampled memory strength x, given criterion mean C_M and standard deviation σ_M, and therefore recall is successful. The difference term within the brackets is the probability that the sampled memory strength y falls between confidence thresholds C_i and C_{i+1}, given unbiased random error around the confidence thresholds with standard deviation σ_C. The bivariate normal density term in Equation 1 is the usual

$$h(x,y,\rho) = \frac{1}{2\pi\sqrt{1-\rho^2}}\, exp\left(\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right),$$

with correlation parameter ρ. This equation assumes that the distributions of memory strengths underlying both recall and confidence have a mean of 0.0 and a standard deviation of 1.0.

The probability that a response is not correctly recalled and receives confidence estimate J_i is given by

$$p(J_\mathrm{i}, \mathit{not\ recalled}) = \iint h(x,y,\rho)[1 - N(x|C_M,\sigma_M)] \times [N(y|C_{i+1},\sigma_C) - N(y|C_i,\sigma_C)]dxdy. \quad (2)$$

Equation 2 differs from Equation 1 only in the memory term, within the first set of brackets, which estimates the probability that the recall criterion falls above the sampled memory strength x.

Applying Equations 1 and 2 to all confidence judgments, J_0, J_1, . . . , J_n, with suitable changes in the confidence term for the end-value judgments J_0 and J_n, provides a mutually exclusive and exhaustive partitioning (MEE) of the unit area under the bivariate memory trace strength distribution $h(x, y, \rho)$.

The situation is slightly more complicated when the modeled criteria are allowed to vary independently of each other, with the potential to become out of order (e.g., Rosner & Kochanski, 2009; Treisman & Faulkner, 1985). Because criterion order is not preserved, Equations 1 and 2 no longer lead to a MEE of the area under $h(x, y, \rho)$, and a normalization procedure is required. Complete equations and details for both linked and independent criteria are available in Section A of the supplemental materials.

In the modeling that follows, we implemented the SDRM with both linked and independent criteria and found that the two versions yield qualitatively similar results, although the independent criteria model tends to provide a better fit to recall data. Therefore, the results we report below reflect the independent criteria version of the SDRM.

Using the SDRM

To reproduce observed data distributions, as in Figure 1, the parameter space of the SDRM is searched using maximum likelihood estimation (MLE) procedures. For a single condition, the 11 degrees of freedom contained in the 12 response categories only lightly constrain the SDRM, which has nine free parameters (σ_M, σ_C, ρ, C_M, and 5 C_is). The SDRM, therefore, readily captures the data in any single condition. For example, it accounted for 99.92% of the variance of the empirical proportions for the delayed-JOL condition (top left panel of Figure 1). Importantly, fitting the model to the data of a single condition is of no help in testing theoretical predictions. For example, the almost perfect fit to this single condition provides no information as to why JOL accuracy, as measured by gamma, is better in the delayed (γ = .93) than the immediate (γ = .32; top right panel of Figure 1) JOL condition.

The goal, however, is not merely to fit the SDRM to data from different experimental conditions but to use the model across two or more conditions to test hypotheses about how the experimental conditions affect cognitive and metacognitive processing. This goal is accomplished by constraining some parameters to equal each other across conditions and allowing others to vary. For example, if a particular theory specifies that two conditions differ only in processing at the time of recall, then a version of the SDRM that allows *memory* to differ while constraining all the other parameters to be equal across conditions should be nonsignificantly different from the full model, which allows all the parameters to vary. Or if the theory implies that an independent variable affects processing only at the time of the confidence rating, then a model that allows only *confidence* to vary across conditions should not differ significantly from the full model. Likewise, if the theory suggests that different information underlies recall and confidence judgments, then a model that allows only *correlation* to vary across conditions should fit as well as the full model.

These examples invoke the logic of testing both nested models (i.e., comparing a superordinate model to one subsumed under it by virtue of constraining some parameter estimates) and non-nested models (i.e., comparing two models, neither one of which is a special case of the other, such as comparing a *memory*-constrained model to a *confidence*-constrained model). Both nested and non-nested tests are necessary in the service of theory development and testing: the former to establish the descriptive validity of the SDRM and the latter to compare specific theories of underlying recall and confidence judgment processes.

Simple Bias and Metacognitive Bias

The average positions of the confidence criteria (the C_is) and the memory threshold (C_M) affect the marginal distributions of confidence ratings and memory performance, respectively, but these variables do not affect the correspondence between confidence and memory retrieval.[4] Thus, these variables index bias. More specifically, the confidence criteria can shift upward, representing a tendency to give lower confidence ratings, or downward, representing a tendency to give higher confidence ratings. We refer to such a shift as simple bias. In application to JOLs across different conditions, we did not observe a change in simple bias, and greater constraint was imposed by using the same criteria values across conditions. However, a different kind of bias shift tends to occur with a change in the correspondence between confidence and memory retrieval. When the correspondence is high (i.e., a high gamma correlation), participants tend to use the extremes of the confidence scale giving both high and low confidence ratings (e.g., Dunlosky & Nelson, 1994; Koriat & Goldsmith, 1996; Koriat, Sheffer, & Ma'ayan, 2002); in contrast, when the correspondence is low (i.e., a low gamma correlation), participants tend to use the middle of the confidence scale. These data patterns correspond to a contraction or expansion of the confidence criteria, respectively, and we refer to such a coordinated shift as metacognitive bias. We measured the extent of metacognitive bias by setting the criteria of one condition equal to the criteria of another condition multiplied by a metacognitive bias parameter, $\beta > 0$. When $\beta > 1$, the confidence criteria in the condition to which β applies are spread further apart than in the other condition, and the judgment distribution tends to be more of an inverted-U shape. When $0 < \beta < 1$, the criteria are closer together, and the distribution tends to be more of a U-shape.

Current Accounts of the Underlying Processes of Recall and JOLs

To illustrate the power of the SDRM, we next apply it to new data within the JOL paradigm for the purpose of distinguishing among competing theoretical accounts. This section first summarizes the relevant JOL phenomena and then explores the different theoretical explorations in terms of the SDRM's components.

Delayed-JOL Effect

Typically, JOL accuracy as measured by gamma is much better when recall confidence is judged at least 30 s after study than when it takes place immediately. Nelson and Dunlosky (1991) first observed this phenomenon and called it the delayed-JOL effect. To explain it, they proposed the monitoring-dual memories (MDM) hypothesis, which asserts that when assessing the likelihood of a subsequent successful retrieval, one monitors information retrieved from both short- and long-term memory (STM and LTM). STM information retrieved during immediate JOLs is strong and effectively adds noise to the prediction of subsequent recall because it is not available at the time of recall. STM information is much weaker for delayed JOLs and therefore does not interfere with LTM information, which more reliably predicts recall success.

Spellman and Bjork (1992) offered a different account of the delayed-JOL effect. They assumed that individuals covertly attempt recall when providing JOLs, which results in retrieval practice, and therefore items retrieved in the service of a JOL are more easily retrieved in the final recall. However, this retrieval practice is not particularly effective with immediate JOLs because the words were viewed just a few seconds ago. Nelson, Narens, and Dunlosky (2004) termed this the self-fulfilling-prophecy (SFP) hypothesis, and Kimball and Metcalfe (2003) called it the memory hypothesis.

There is an ongoing debate whether the MDM or the SFP hypothesis provides a better explanation of the delayed-JOL effect. The MDM hypothesis postulates that the difference between immediate and delayed JOL performance is due to changes within the confidence process (i.e., metamemory improvement after a delay), whereas the SFP hypothesis postulates that it is due to changes in the recall process (i.e., memory improvement after a delay). Empirical comparisons of the two accounts have not been decisive (e.g., Kimball & Metcalfe, 2003; Nelson et al., 2004). The SFP hypothesis is supported if recall following delayed JOLs is greater than recall following immediate JOLs or nothing (control). However, summarizing the relevant studies, Sikström and Jönsson (2005) concluded that recall was not systematically enhanced after delayed JOLs. Kimball and Metcalfe (2003) and Nelson et al. (2004) tested these accounts independently by re-exposing the items following JOLs. However, this manipulation itself could affect gamma correlations because the JOLs and recall were

[4] However, average criterion placement can affect the gamma correlation by changing the proportion of dyads that produce ties. For instance, if adjacent confidence criteria are placed far apart, it becomes more likely that any two JOL judgments will yield the same JOL value and such ties are eliminated from the calculation of gamma.

achieved before and after the re-exposure, respectively (Sikström & Jönsson, 2005).

The SDRM provides a means for formulating competing predictions from the two theories. Within the SDRM, the MDM (or metamemory) hypothesis corresponds to changes in confidence judgments because it assumes that the process underlying the JOL responses has been changed. In contrast to this account, from the SDRM perspective, the SFP (or memory) hypothesis appeals directly to the recall process because it predicts decreases in retrieval variability following delayed JOLs. In other words, if the MDM account is correct, the immediate and delayed conditions differ in the confidence judgment stage, and if the SFP account is correct, they differ in the recall stage.

How do these different predictions translate to different parameter constraints in the SDRM? The MDM and SFP hypotheses are verbal theories, and neither provides formal quantitative specification of the joint distribution between JOLs and recall. Therefore, it is difficult to state with certainty how the confidence and retrieval processes should be affected under each account. Nevertheless, we posit the most natural mapping when relating each account to the components of the SDRM. Because the MDM hypothesis assumes that STM introduces noise to the JOL process, this implies that σ_C varies between the immediate and delayed conditions. In contrast, the SFP hypothesis assumes that recall is more reliable due to previous retrieval attempts with delayed JOLs, which implies that σ_M varies between the immediate and delayed conditions. When interpreting changes in these parameters, it is important to note that the criterion variability parameters are expressed as the ratio of criterion variability relative to memory strength (*X* or *Y*) variability, which is set to 1.0 for convenience. Thus, observed differences in estimates of either of these two parameters could be due to changes in criterion variability or to changes in the relevant memory strength variable, *X* or *Y*. The important conceptual point is that one account assumes that the relative noise level occurring during the judgment process is affected, whereas the other account assumes that the relative noise level occurring during the recall process is affected.

Testing-JOL Effect

JOL accuracy is also affected by practice. In what we call the testing-JOL effect, JOL accuracy as measured by gamma improves when participants cycle through the study, JOL rating, and test phases more than once (e.g., Finn & Metcalfe, 2007; Koriat, 1997; Koriat et al., 2002). Critically, this cycling involved repeated testing of the same items, which provides not only practice with JOL ratings in general but also a chance to develop item-specific JOL knowledge.

During multi-trial learning, people tend to accurately distinguish previously recalled and not-recalled items and can monitor their knowledge of the outcomes of previous tests (Bisanz, Vesonder, & Voss, 1978; Gardiner & Klee, 1976; Klee & Gardiner, 1976; Robinson & Kulp, 1970). Consequently, they learn items efficiently on subsequent study trials. Such results suggest that retrieval practice can play an important role in metacognitive judgments. Indeed, the JOL to a repeated item is more strongly correlated with recall on the previous test of that item than with recall on the subsequent test of that item (Finn & Metcalfe, 2007; King, Zechmeister, & Shaughnessy, 1980; Koriat, 1997; Lovelace, 1984), which suggests that JOLs are based on information pertaining to the outcome of the previous recall. In other words, JOLs constitute at least in part postdiction based on previous retrievals. Finn and Metcalf (2007) referred to this account as the memory for past test (MPT) hypothesis.

In terms of the SDRM, the MPT hypothesis corresponds to different information sources underlying the JOLs in the first and second study-JOL-test cycles. This is because the MPT hypothesis assumes that second cycle JOLs depend on the information from the previous recall test, whereas this information is not available during the initial cycle. Thus, the correlation between memory sampling at test and at JOL should be different between the first and second cycles.

Experiment: JOL Accuracy and JOL-Recall Joint Distributions

Having now cast these theories in terms of restrictions on the SDRM across experimental conditions, we briefly describe an experiment that yielded data to which we applied the model. The experiment had additional purposes as well. Because the testing-JOL effect is new and not yet fully explored, one purpose was to establish its cause more precisely. Using S, J, and T to refer to study, JOL rating, and test, respectively, previous studies have compared JOL accuracy when cycling through SJT once (the usual procedure) to that when cycling through a second time with the same items. Based on these previous studies, it is unclear whether the improvement on the second SJT cycle is due to the entire prior cycle, to only one component of the prior SJT cycle, or to some combination: for example, the prior S (with JT being irrelevant), the prior SJ (with T being irrelevant), and so on. Another purpose of the experiment was to look at the delayed- and testing-JOL effects jointly, which has not yet been done. Thus, this experiment crossed the immediate-delay variable with type of prior practice (i.e., none, which serves as the control condition, S, SJ, ST, and SJT).

The experimental method is described completely in Section B of the supplemental materials. Here, we provide a very brief overview. The experimental design was a 5 × 2 mixed factorial with type of practice (control, S, SJ, ST, and SJT) preceding a full SJT cycle manipulated between subjects (45 participants per condition) and JOL timing (immediate and delayed) manipulated within subjects. Concrete unrelated noun–noun pairs (Paivio, Yuille, & Madigan, 1968) were used: 24 pairs for each of the immediate and delayed JOLs. Participants were instructed to study word pairs and to indicate their JOL for a pair (i.e., to predict the future recall probability: 0%, 20%, 40%, 60%, 80%, and 100%) whenever the cue word appeared alone. Immediate JOLs were elicited right after the offset of each pair, and the delayed JOLs were elicited after all the pairs had been studied. In both cases, JOL responses were self-paced. Finally, during the recall phase, participants typed the target word when cued by the first word of the pair.

Before applying the SDRM, we compared gamma values across experimental conditions to provide a point of contact with other studies. In short, we replicated the delayed-JOL effect and identified the causal event in the testing-JOL effect. Specifically, the delayed-JOL effect showed up under each of the five practice conditions, a testament to its robustness. In contrast, the testing-

JOL effect was evident only in the immediate conditions (ST and SJT), apparently due to a ceiling effect for the delayed conditions. In addition, this experiment is the first to demonstrate the aspect (S, J, or T) that induces a testing-JOL effect—because the ST and SJT conditions revealed a testing-JOL effect, whereas the S and SJ conditions did not, this demonstrates that the testing-JOL effect depends on a prior recall test. The complete results are reported in Section B of the supplemental materials, and here we focus mainly on the JOL-recall joint distribution data (to which the SDRM is applied).

As noted previously, gamma is a blunt tool for theory testing and is insensitive to response distributions, such as those displayed in Figure 1. The top panels of this figure show the immediate and delayed JOL distributions for the control condition participants, and the bottom panels show the no-testing and testing immediate JOL distributions for the SJT condition participants. We note that the no-testing condition and the immediate part of the control condition are functionally identical (these are items from the first study list to which participants gave immediate JOLs), although the data were collected from separate groups of participants. These four distributions exemplify the delayed- and testing-JOL effects (top two and bottom two, respectively). For completeness, the remaining distributions are available in Section C of the supplemental materials. Next, we apply the SDRM to the two distributions for each effect to determine whether the observed increases in JOL accuracy are due to reduced noise in the recall process, reduced noise in the JOL process, the use of more consistent information for the two processes, or a change only in metacognitive bias. Arriving at the most descriptive sub-models for each effect provides a means for examining the MDM and SFP hypotheses of the delayed-JOL effect and the MPT hypothesis of the testing-JOL effect.

Applying the SDRM

In this section, we demonstrate how the SDRM uses the full response distribution data to determine the best explanation for the difference in JOL accuracy across conditions.

Model Hierarchy

Each descriptive theory implies a different set of constraints on the parameter values of the SDRM, which in turn defines a different sub-model under the unconstrained, or full SDRM. Comparing the goodness-of-fit (GOF) statistics of these models provides a means for evaluating and comparing the descriptive validity of the competing theories. Figure 3 shows the model hierarchy that guides this process for the two experimental conditions: delay versus immediate (under no testing), and testing versus no testing (under immediate).

M1, the full model, shows a full set of parameters across the two conditions. The second subscript, 1 or 2, for each parameter denotes the respective conditions. As described earlier, the means and standard deviations of the bivariate *X*–*Y* memory strength distribution are set at (0, 0) and (1, 1), respectively, and are therefore not shown.

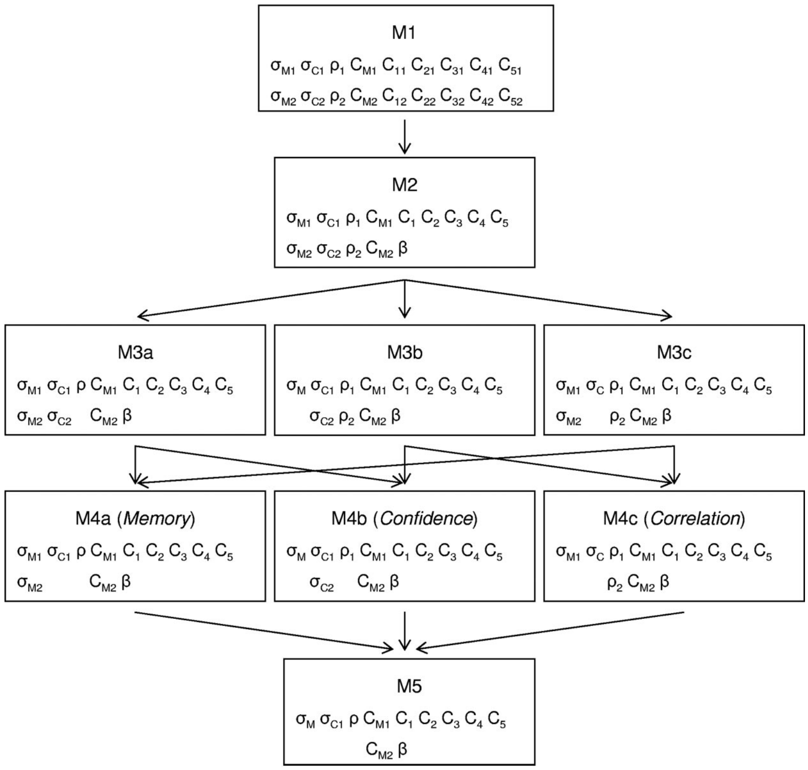
M1
$\sigma_{M1}\ \sigma_{C1}\ \rho_1\ C_{M1}\ C_{11}\ C_{21}\ C_{31}\ C_{41}\ C_{51}$
$\sigma_{M2}\ \sigma_{C2}\ \rho_2\ C_{M2}\ C_{12}\ C_{22}\ C_{32}\ C_{42}\ C_{52}$
M2
$\sigma_{M1}\ \sigma_{C1}\ \rho_1\ C_{M1}\ C_1\ C_2\ C_3\ C_4\ C_5$
$\sigma_{M2}\ \sigma_{C2}\ \rho_2\ C_{M2}\ \beta$
M3a
$\sigma_{M1}\ \sigma_{C1}\ \rho\ C_{M1}\ C_1\ C_2\ C_3\ C_4\ C_5$
$\sigma_{M2}\ \sigma_{C2}\ C_{M2}\ \beta$
M3b
$\sigma_{M}\ \sigma_{C1}\ \rho_1\ C_{M1}\ C_1\ C_2\ C_3\ C_4\ C_5$
$\sigma_{C2}\ \rho_2\ C_{M2}\ \beta$
M3c
$\sigma_{M1}\ \sigma_{C}\ \rho_1\ C_{M1}\ C_1\ C_2\ C_3\ C_4\ C_5$
$\sigma_{M2}\ \rho_2\ C_{M2}\ \beta$
M4a (*Memory*)
$\sigma_{M1}\ \sigma_{C1}\ \rho\ C_{M1}\ C_1\ C_2\ C_3\ C_4\ C_5$
$\sigma_{M2}\ C_{M2}\ \beta$
M4b (*Confidence*)
$\sigma_{M}\ \sigma_{C1}\ \rho_1\ C_{M1}\ C_1\ C_2\ C_3\ C_4\ C_5$
$\sigma_{C2}\ C_{M2}\ \beta$
M4c (*Correlation*)
$\sigma_{M1}\ \sigma_{C}\ \rho_1\ C_{M1}\ C_1\ C_2\ C_3\ C_4\ C_5$
$\rho_2\ C_{M2}\ \beta$
M5
$\sigma_{M}\ \sigma_{C1}\ \rho\ C_{M1}\ C_1\ C_2\ C_3\ C_4\ C_5$
$C_{M2}\ \beta$

Figure 3. Model hierarchy for nested model comparisons. Directed arrows between models indicate subset relations between models.

M1 accounts for changes in metacognitive bias by using completely independent confidence criteria for each of the two conditions. With so many free parameters (i.e., 10 confidence criteria), it is difficult to identify the cause of the change in JOL accuracy. Therefore, M2, nested under M1, is a much more constrained model that still allows for a change in metacognitive bias; in this model, the five criteria in Condition 2 differ from the five in Condition 1 only by a multiplicative factor β.

Models M3a, M3b, and M3c, all equivalently nested under M2, are obtained via parameter restrictions at different processing stages. Thus, M3a represents the case in which Conditions 1 and 2 entail the same relation between the two memory strength samplings *X* and *Y*, represented as $\rho_1 = \rho_2 = \rho$. M3b assumes that the level of noise in the recall process is the same for both conditions, represented as $\sigma_{M1} = \sigma_{M2} = \sigma_M$. Finally, M3c assumes that the level of noise in the JOL rating process is the same in both conditions, represented as $\sigma_{C1} = \sigma_{C2} = \sigma_C$.

Models M4a, M4b, and M4c are each nested under the two models above them, as indicated by the directed arrows, and combine the corresponding two sets of restrictions. Thus, M4a represents theories that assume that Conditions 1 and 2 differ only in the variance of underlying memory recall processes. M4b represents theories that assume the two conditions differ only in the variance underlying JOL ratings. M4c represents theories that assume the conditions differ only in the relation between the memory processes at recall and at JOL ratings.

Finally, M5 is nested under all three M4 models. M5 is a necessary comparison because a change in metacognitive bias can by itself cause an apparent change in JOL accuracy due to poor resolution when using a small number of discrete confidence ratings. This can be seen by considering an extreme situation in which the metacognitive bias parameter, β, is set to infinity (i.e., all of the confidence criteria are placed at either positive or negative infinity), in which case all of the memory strength samples fall between the same two criteria (i.e., the same JOL rating is given for all items), and JOL accuracy is at chance. Thus, as metacognitive bias becomes more liberal (i.e., a small value of β places the criteria closer to the center, which results in a U-shaped confidence distribution), there is more opportunity for differences in the JOL ratings to indicate differences between recalled and not-recalled items. Comparing M5 to the M4 models tests whether metacognitive bias alone explains the differences in JOL accuracy.

Model Recovery Probabilities

Our primary interests are in comparisons across the non-nested models of M4a, M4b, and M4c for theory testing. It is important that we investigate whether model mimicry can occur among these non-nested models—if one model can mimic the behavior of another, but not vice versa, then that model is more flexible and can fit a greater range of data patterns without necessarily being more accurate. We examined model mimicry by means of parametric bootstrap simulations[5] (for detail, see Navarro, Pitt, & Myung, 2004; Wagenmakers, Ratcliff, Gomez, & Iverson, 2004), which allowed us to determine whether the true model can be recovered when it is in competition with other models. The details are reported in the Appendix.

We also developed and estimated a new model comparison statistic, which we term the Bayesian recovery probability (BRP). Using Bayes rule and an assumption of equal priors, the BRP uses the results of the Monte Carlo simulations to estimate the probabilities that each model is the correct one given that it emerged as the model that best fit the observed data. The BRP therefore provides a measure of certainty for the model selection process that factors in the functional form of the models. Based on our delayed- and testing-JOL effect data, these values ranged from .772 to .921, as reported in the Appendix: that is, assuming that one of the three models is correct, it is unlikely that an error was made when concluding that the best fitting one was the true model.

Using the SDRM to Test Theoretical Accounts

Delayed-JOL effect. To compare theoretical explanations of the delayed-JOL effect, we fit the nine versions of the SDRM shown in Figure 3 to the response distributions of the delayed versus immediate data set for Conditions Control, S, and SJ (i.e., the conditions that do not show the testing-JOL effect). The results for each fit were very similar, and we report here only those for the control condition: For completeness, the best fits of Conditions S and SJ are available in Section C of the supplemental materials.

The results of fitting the SDRM to the control condition delayed versus immediate data set are shown in Figure 4. The Models M1–M5 are those illustrated in Figure 3, and the chi-square values superimposed on the directed arrows were obtained via maximum-likelihood tests comparing each nested model to the one above it, or in the case of M1 to a model that used the observed data to define a multinomial distribution. The chi-square degrees of freedom (*df*s) equals the difference in the number of free parameters when comparing two nested models, or between the number of observed data types (i.e., 24 combinations of JOL and recall levels across experimental conditions) and M1. Values significant at $\alpha = .05$ are indicated by solid arrows, and non-significant values are indicated by dotted lines.

Note that M1 is a significantly worse model than one based on the data (i.e., a model with a free parameter for each observed frequency), and M2 is a significantly worse model than M1, as measured by chi-square GOF difference. However, considering the huge *N*s involved in this experiment (i.e., 45 participants × 24 words × 2 conditions = 2,160), this does not necessarily mean that these models are fitting poorly in an absolute sense. Indeed, these models captured the data pattern extremely well, explaining 95% and 83% of the variance of the empirical proportions, respectively.

M2 is the first substantive model that includes some constraint between the conditions of interest, and we take this model as the point of departure in drawing theoretical conclusions regarding the underlying explanation behind changes in JOL accuracy. In moving down Figure 4 from M2, the paths of interest are the ones in which the lower model does not differ significantly from the one above it, implying that the extra free parameters of the upper model are unnecessary. The logic of this process is based on Type

[5] The procedure of Wagenmakers, Ratcliff, Gomez, and Iverson (2004) included non-parametric sampling of the observed data set for each Monte Carlo simulation as well parametric data generation. However, for signal-detection models of recognition memory, there was little or no difference between Monte Carlo simulations that did or did not include non-parametric sampling (Jang, Wixted, & Huber, 2011).

1 error rates when deciding whether a nested model is significantly worse than the one above it. The associated Type 2 error rate is unknown, although as noted above, power was very high for this experiment. As seen in Figure 4, this nested model comparison process identified Models M3a and M3b as non-significantly different from M2. Continuing down the figure, M4b is non-significantly different from both M3a and M3b. Model M5, however, differed significantly from M4b.

On the basis of comparing successively nested models, M4b clearly appears to provide the best description of the data as represented by M2. Moreover, Models M4a and M4c differ significantly from M2, whereas M4b does not. This point is seen by summing the chi-square values (and their *df*s) on the distinct paths connecting M2 to each M4 model. The results are shown in the top row of Table 1. Thus, M4b, which allows noise in the confidence judgment process to change between the two conditions, provided the best fit among all the models that included some form of constraint between the conditions of interest. The top panels of Figure 1 show the fit of M4b compared to the data: The circles show the predicted joint response distributions based on the M4b MLE parameters. The M4b response proportions accounted for 82% of the variance of the empirical proportions.

The empirical data column on the delayed versus immediate side of Table 2 reports the best fitting parameter values based on M4b. The simulated data column shows the parameter averages and standard deviations from fitting the 1,000 Monte Carlo simulations that generated artificial data based on the best fitting parameters listed in the empirical data column. A comparison between the parameter estimates based on the empirical versus the simulated data demonstrates that parameter recovery was fairly reliable.

Table 1
Goodness-of-Fit Difference Between M2 and M4

		Memory	*Confidence*	*Correlation*
Data set	*df*	(M4a)	(M4b)	(M4c)
Delayed vs. Immediate	2	23.46	1.93 $p = .38$	32.81
Testing vs. No testing	2	53.69	54.36	0.39 $p = .82$

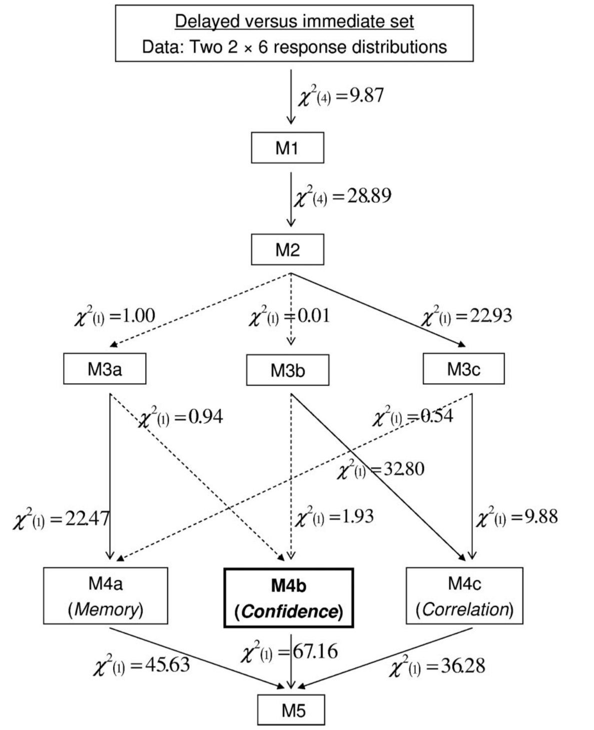

Figure 4. Chi-square statistics comparing nested models for the delayed versus immediate data set. Chi-square values significant at $\alpha = .05$ are shown via solid arrows. Non-significant values are shown via dashed arrows. Among the models that include some form of constraint between the conditions (i.e., M2–M5), the bold box (M4b, *confidence*) provides the best account of the delayed-judgment-of-learning effect.

Note from the delayed-JOL parameter estimates in Table 2 that the position of the memory criterion (C_M) was lower for the delayed condition (0.23) than the immediate condition (0.54). This difference reflects different levels of proportion correct recall for the two conditions (in this experiment, $M = 0.32$, $SE = 0.04$ for immediate; and $M = 0.37$, $SE = 0.04$ for delayed), although this change cannot explain the difference in JOL accuracy.[6] Note next that the estimate of noise in the confidence judgment process (σ_C) was greater for immediate JOLs (2.27) than for delayed JOLs (0.05), a result that supports the MDM hypothesis of less noise from STM with delay. Finally, note that the metacognitive bias parameter (β) is lower for the delayed condition (0.45) than the immediate condition (1.00), which indicates that the confidence scale was used in a more liberal fashion for delayed JOLs. However, metacognitive bias alone does not provide a good explanation of the data, as can be seen by comparing M5 to M4b and noting the substantial reduction in GOF.

Testing-JOL effect. To account for the testing-JOL effect, we fit the nine versions of the SDRM shown in Figure 3 to the response distributions of the testing versus no-testing data set (i.e., the second SJT cycle vs. the first SJT cycle in Condition SJT), in the same manner as for the delayed versus immediate data set. As illustrated in Figure 5, the nested model analysis clearly identified M4c as the best account of the data in terms of non-significant paths from one model to another for the paths between Models M2 and M4c as well as the significant path between M4c and M5. The summed values, as shown in Table 1 (bottom row), confirm this conclusion by showing that Models M4a and M4b differed significantly from M2, whereas M4c did not. Thus, comparing across the three models, M4c, which allows different correlations between the two memory strength samplings in the testing and no-testing conditions, provided the best explanation of the testing-JOL effect. The circles of Figure 1 (bottom panels) show the predicted joint response distributions based on the M4c best fitting parameters. The M4c response proportions account for 82% of the variance of the empirical proportions.

[6] Accurate JOLs occur when higher confidence ratings reliably predict a higher probability of recall, which is a function of the joint probability distribution rather than the marginal distribution represented by the overall probability of recall. For further discussion on this issue, see, for example, Kimball and Metcalfe (2003), Nelson and Dunlosky (1992), and Sikström and Jönsson (2005).

Table 2
Best-Fitting SDRM Parameter Estimates of Empirical Data and Simulated Data

	Delayed vs. Immediate (M4b: *Confidence*)				Testing vs. No testing (M4c: *Correlation*)			
	Empirical data		Simulated data		Empirical data		Simulated data	
Parameter	Delayed	Immediate	Delayed	Immediate	Testing	No testing	Testing	No testing
C_1	−1.89		−1.93 (0.33)		−1.02		−1.08 (0.13)	
C_2	0.27		0.32 (0.21)		−0.18		−0.17 (0.05)	
C_3	1.41		1.52 (0.26)		0.33		0.37 (0.10)	
C_4	2.43		2.62 (0.21)		0.81		0.85 (0.05)	
C_5	3.51		3.64 (0.24)		1.32		1.36 (0.05)	
C_M	0.23	0.54	0.28 (0.05)	0.45 (0.05)	−0.19	0.60	−0.19 (0.05)	0.61 (0.07)
β	0.45	1.00[a]	0.43 (0.35)	1.00[a]	0.83	1.00[a]	0.74 (0.13)	1.00[a]
σ_M	0.45		0.52 (0.10)		0.05		0.06 (0.12)	
σ_C	0.05	2.27	0.11 (0.22)	2.40 (0.36)	0.05		0.06 (0.14)	
ρ	0.91		0.91 (0.02)		0.68	0.35	0.71 (0.07)	0.33 (0.07)

Note. Parameter standard deviations are in parentheses. The values on the simulated data column are the parameter averages from the 1,000 Monte Carlo simulations. SDRM = stochastic detection and retrieval model.
[a] The parameter value was set to 1.00.

Similar to the delayed-JOL effect results, Table 2 shows there was a decrease in C_M for the testing condition (−0.19 vs. 0.60 for the no-testing condition) because as expected, recall accuracy was higher in that condition (in this experiment, $M = 0.27$, $SE = 0.03$ for no-testing JOLs, and $M = 0.62$, $SE = 0.04$ for testing JOLs). There was also a decrease in metacognitive bias (β) for the testing condition (0.83 vs. 1.00 for the no-testing condition), although this effect by itself was unable to account for the change in JOL accuracy. As shown in Table 2 for the testing-JOL effect, the estimate of the relation (ρ) between the memory strength underlying recall and the memory strength underlying confidence was greater after testing (.68) than prior to test experience (.35). This result supports the MPT hypothesis (i.e., more diagnostic JOL information with test experience).

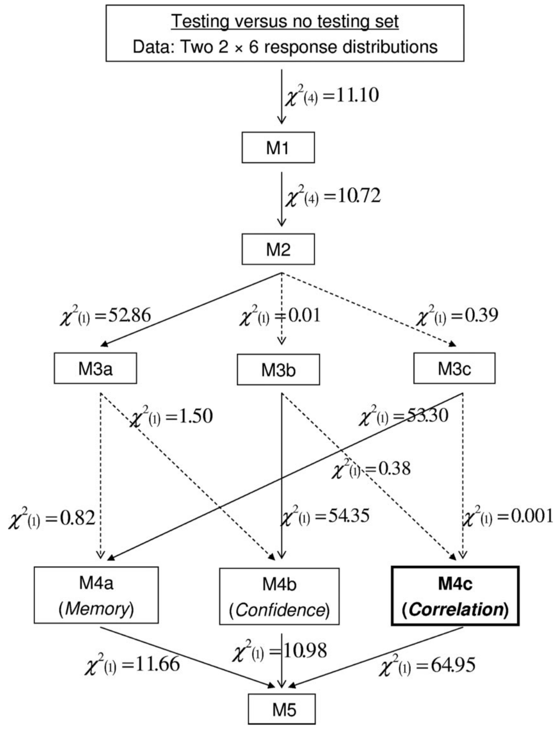

Figure 5. Chi-square statistics comparing nested models for the testing versus no-testing data set. Chi-square values significant at α = .05 are shown via solid arrows. Non-significant values are shown via dashed arrows. Among the models that include some form of constraint between the conditions (i.e., M2–M5), the bold box (M4c, *correlation*) provides the best account of the testing-judgment-of-learning effect.

Finally, recall that the immediate and no-testing control conditions are functionally the same condition except that they involved different groups of participants. Therefore, it may seem surprising that many of the parameters in Table 2 differ when comparing these conditions. However, this does not reflect poor reliability when applying the SDRM, as demonstrated by the small parameter standard deviations seen in Table 2, and as demonstrated by successful model recovery. Instead, this occurred because the best model for the delayed-JOL effect was a different, non-nested model than the best model for the testing-JOL effect (i.e., M4b and M4c, respectively). When the same model is applied to the delayed- and testing-JOL effects, the nominally identical control condition for each effect does in fact produce nearly identical best fitting parameter values. This highlights the fact that the SDRM is not a single model; instead, it is a framework for comparing different measurement models that may correspond to different process models of interest.

General Discussion

The SDRM as a New Method for Contrasting Theories

Based on detection theory, the SDRM identifies sources of variance in related memory and judgment tasks and compares confidence criteria placement across tasks. We created this model to examine confidence accuracy of memory performance and to provide parsimonious explanations of theoretical issues.

Focusing on JOL research, we demonstrated that the SDRM is useful for distinguishing among competing theories. In the case of the delayed-JOL effect, the SDRM found that the increase in JOL accuracy was due to less noise in the confidence judgment process with delayed than immediate JOLs. This result is consistent with the MDM (metamemory) hypothesis but not with the SFP (memory) hypothesis. However, neither hypothesis is sufficiently specified to make quantitative predictions regarding the joint JOL and recall distributions. Based on qualitative descriptions of these theories, we assumed that the MDM hypothesis corresponds to a change in variance within the JOL process, whereas the SFP hypothesis corresponds to a change in variance within the retrieval process. Proponents of the SFP hypothesis may disagree with this assumption, in which case the SFP hypothesis may be compatible with our results. Regardless, application of the SDRM has quantitatively specified the source of the delayed-JOL effect, which should aid further development of either account.

In the case of the testing-JOL effect, the SDRM found that the increase in JOL accuracy was due to a closer correspondence between the information underlying JOLs and the memory strength supporting recall after testing experience, supporting the MPT hypothesis. The SDRM also revealed different metacognitive biases across conditions for both the delayed- and testing-JOL effects (which was not considered by any of the theories), although changes in metacognitive bias alone were unable to explain the increase in JOL accuracy with delay or testing experience. Thus, the SDRM provides far stronger evidence for the supported theories than is possible under traditional gamma-based analyses.

Theoretical Mechanisms and Models of Recall and JOLs

The most important theoretical point of the SDRM is that one can (we would argue must) distinguish retrieval from confidence judgment processes, considering that either can change the accuracy of confidence judgments. In support of this distinction, it has been observed that recall and confidence judgments are dissociated in patient populations (e.g., Shimamura & Squire, 1986), in midazolam-induced amnesia (e.g., Merritt, Hirshman, Hsu, & Berrigan, 2005), and with alcohol intoxication (e.g., Nelson, McSpadden, Fromme, & Marlatt, 1986). Nonetheless, the consequences of this distinction have not been well considered. Full consideration of this distinction leads to the realization that the observed relation between confidence and retrieval can be affected by noise in the confidence process, noise in the retrieval process, or the type of information used for each process. Unlike simple correlation measures (such as gamma), the SDRM can separately measure each of these influences.

A comparison between the SDRM and the formal model of recall and JOLs proposed by Sikström and Jönsson (2005) is warranted. Their model provided good fits of JOL distributions, although the data were not separated into recalled versus not-recalled JOL values (i.e., the model was fit to the marginal JOL distribution rather than the joint distribution with recall). Specifically, this model assumes that confidence judgments only depend on memory strength, which decays at different rates. Aside from producing power-law forgetting, this model captures the delayed-JOL effect due to the passage of time and intervening events rather than the use of different processes. Intuitively, it seems that Sikström and Jönsson's model corresponds to a change in *correlation* of the SDRM because it assumes that a different kind of information underlies immediate JOLs (fast memory traces) compared to delayed JOLs (slow memory traces). But intuitions can be misleading, and it is difficult to conclusively determine whether this is the case because their model was formulated in terms of prediction accuracy rather than a detection process with decision noise. More specifically, the equations of Sikström and Jönsson's model explicitly produce a gamma correlation rather than constructing the gamma correlation out of the joint probability distribution of recalled and not-recalled at each confidence level.

Beyond Gamma

We now return to our original motivation for developing the SDRM, which was to have a more sensitive way to compare the predictions of competing theories than non-parametric measures such as gamma. Nelson (1984) argued that gamma provides a measure of confidence judgment accuracy that is free of the assumptions entailed by SDT (for further discussion on this issue, see Nelson, 1986; Swets, 1986a, 1986b). This proved to be a persuasive argument, and gamma has been used in almost all metacognition articles published since the mid-1980s.

Providing evidence against the preeminence of gamma, Masson and Rotello (2009) demonstrated that the computation of gamma is not free of distributional assumptions and that the empirically determined value of gamma systematically deviates from its actual value under realistic conditions (also see Rotello, Masson, & Verde, 2008). The root of the problem is that calculation of gamma excludes ties, which are pairs of test items that produced equivalent JOL ratings or were equivalently recalled or not-recalled. When ties constitute a substantial fraction of the data, the underlying distributions are modified in unsystematic ways, and the results can be very misleading. In contrast, the SDRM uses all of the data, and because the distributional assumptions are explicit, they can be altered if necessary.

Conclusion

The SDRM emphasizes the role of stochastic variability during the confidence and retrieval processes. Assuming that confidence and retrieval can each be described by a criterial process, it is clear that a lack of correspondence between confidence and retrieval can arise from decision noise in one criterial process or the other. Furthermore, the two criterial processes may be based on different kinds of information. The SDRM, as a general method, distinguishes among different explanations for a change in confidence accuracy. Although the current application was in the study of cued-recall and JOLs, the SDRM can be used in any area of metacognition (which involves both recall and recognition). The SDRM specifies a family of measurement models depending on which components are allowed to vary across experimental conditions. Comparing models within the family provides a means for differentiating between verbal theories that advocate different explanations for a change in confidence accuracy. Furthermore, the SDRM can measure both simple bias shifts and changes in metacognitive bias.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Caski (Eds.), *Proceeding of the* *second international symposium on information theory* (pp. 267–281). Budapest, Hungary: Akademiai Kiado.

Batchelder, W. H., & Riefer, D. M. (1990). Multinomial processing models of source monitoring. *Psychological Review, 97,* 548–564. doi:10.1037/0033-295X.97.4.548

Begg, I., Duft, S., Lalonde, P., Melnick, R., & Sanvito, J. (1989). Memory predictions are based on easy 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

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

Bisanz, G. L., Vesonder, G. T., & Voss, J. F. (1978). Knowledge of one's own responding and the relation of such knowledge to learning. *Journal of Experimental Child Psychology, 25,* 116–128. doi:10.1016/0022-0965(78)90042-5

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

Clarke, F. R., Birdsall, T. G., & Tanner, W. P., Jr. (1959). Two types of ROC curves and definitions of parameters. *The Journal of the Acoustical Society of America, 31,* 629–630. doi:10.1121/1.1907764

Dunlosky, J., & Nelson, T. O. (1994). Does the sensitivity of judgments of learning (JOLs) to the effects of various study activities depend on when the JOLs occur? *Journal of Memory and Language, 33,* 545–565. doi:10.1006/jmla.1994.1026

Finn, B., & Metcalfe, J. (2007). The role of memory for past test in underconfidence with practice effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 238–244. doi:10.1037/0278-7393.33.1.238

Galvin, S. J., Podd, J. V., Drga, V., & Whitmore, J. (2003). Type 2 tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions. *Psychonomic Bulletin & Review, 10,* 843–876. doi:10.3758/BF03196546

Gardiner, J. M., & Klee, H. (1976). Memory for remembered events: An assessment of output monitoring in free recall. *Journal of Verbal Learning and Verbal Behavior, 15,* 227–233. doi:10.1016/0022-5371(76)90021-9

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

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* New York, NY: Wiley.

Heathcote, A. (2003). Item recognition memory and the receiver operating characteristic. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 1210–1230. doi:10.1037/0278-7393.29.6.1210

Jang, Y., & Nelson, T. O. (2005). How many dimensions underlie judgments of learning and recall? Evidence from state-trace methodology. *Journal of Experimental Psychology: General, 134,* 308–326. doi:10.1037/0096-3445.134.3.308

Jang, Y., Wixted, J. T., & Huber, D. E. (2009). Testing signal-detection models of yes/no and two-alternative forced-choice recognition memory. *Journal of Experimental Psychology: General, 138,* 291–306. doi:10.1037/a0015525

Jang, Y., Wixted, J. T., & Huber, D. E. (2011). The diagnosticity of individual data for model selection: Comparing signal-detection models of recognition memory. *Psychonomic Bulletin & Review, 18,* 751–757. doi:10.3758/s13423-011-0096-7

Kimball, D. R., & Metcalfe, J. (2003). Delaying judgments of learning affects memory, not metamemory. *Memory & Cognition, 31,* 918–929. doi:10.3758/BF03196445

King, J. F., Zechmeister, E. B., & Shaughnessy, J. J. (1980). Judgments of knowing: The influence of retrieval practice. *The American Journal of Psychology, 93,* 329–343. doi:10.2307/1422236

Klee, H., & Gardiner, J. M. (1976). Memory for remembered events: Contrasting recall and recognition. *Journal of Verbal Learning and Verbal Behavior, 15,* 471–478. doi:10.1016/S0022-5371(76)90042-6

Koriat, A. (1997). Monitoring one's knowledge during study: A cue-utilization framework to judgments of learning. *Journal of Experimental Psychology: General, 126,* 349–370. doi:10.1037/0096-3445.126.4.349

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review, 103,* 490–517. doi:10.1037/0033-295X.103.3.490

Koriat, A., Levy-Sadot, R., Edry, E., & de Marcas, S. (2003). What do we know about what we cannot remember? Accessing the semantic attributes of words that cannot be recalled. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 1095–1105. doi:10.1037/0278-7393.29.6.1095

Koriat, A., Sheffer, L., & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: Judgments of learning exhibit increased underconfidence with practice. *Journal of Experimental Psychology: General, 131,* 147–162. doi:10.1037/0096-3445.131.2.147

Lovelace, E. A. (1984). Metamemory: Monitoring future recallability during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 756–766. doi:10.1037/0278-7393.10.4.756

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). New York, NY: Cambridge University Press.

Masson, M. E. J., & Rotello, C. M. (2009). Sources of bias in the Goodman–Kruskal gamma coefficient measure of association: Implications for studies of metacognitive processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 509–527. doi:10.1037/a0014876

Merritt, P., Hirshman, E., Hsu, J., & Berrigan, M. (2005). Metamemory without the memory: Are people aware of midazolam-induced amnesia? *Psychopharmacology, 177,* 336–343. doi:10.1007/s00213-004-1958-8

Mueller, S. T., & Weidermann, C. T. (2008). Decision noise: An explanation for observed violations of signal detection theory. *Psychonomic Bulletin & Review, 15,* 465–494. doi:10.3758/PBR.15.3.465

Navarro, D. J., Pitt, M. A., & Myung, J. I. (2004). Assessing the distinguishability of models and the informativeness of data. *Cognitive Psychology, 49,* 47–84. doi:10.1016/j.cogpsych.2003.11.001

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

Nelson, T. O. (1986). ROC curves and measures of discrimination accuracy: A reply to Swets. *Psychological Bulletin, 100,* 128–132. doi:10.1037/0033-2909.100.1.128

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect." *Psychological Science, 2,* 267–270. doi:10.1111/j.1467-9280.1991.tb00147.x

Nelson, T. O., & Dunlosky, J. (1992). How shall we explain the delayed-judgment-of-learning effect? *Psychological Science, 3,* 317–318. doi:10.1111/j.1467-9280.1992.tb00681.x

Nelson, T. O., McSpadden, M., Fromme, K., & Marlatt, G. T. (1986). Effects of alcohol intoxication on metamemory and on retrieval from long-term memory. *Journal of Experimental Psychology: General, 115,* 247–254. doi:10.1037/0096-3445.115.3.247

Nelson, T. O., & Narens, L. (1994). Why investigate metacognition? In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 1–25). Cambridge, MA: MIT Press.

Nelson, T. O., Narens, L., & Dunlosky, J. (2004). A revised methodology for research on metamemory: Pre-judgment recall and monitoring (PRAM). *Psychological Methods, 9,* 53–69. doi:10.1037/1082-989X.9.1.53

Paivio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns [Monograph]. *Journal of Experimental Psychology, 76*(1, Pt. 2), 1–25. doi:10.1037/h0025327

Pollack, I. (1959). On indices of signal and response discriminability. *The Journal of the Acoustical Society of America, 31,* 1031. doi:10.1121/1.1907802

Ratcliff, R., McKoon, G., & Tindall, M. (1994). Empirical generality of data from recognition memory receiver-operating characteristic functions and implications for the global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 763–785. doi:10.1037/0278-7393.20.4.763

Robinson, J. A., & Kulp, R. A. (1970). Knowledge of prior recall. *Journal of Verbal Learning and Verbal Behavior, 9,* 84–86. doi:10.1016/S0022-5371(70)80012-3

Rosner, B. S., & Kochanski, G. (2009). The law of categorical judgment (corrected) and the interpretation of changes in psychophysical performance. *Psychological Review, 116,* 116–128. doi:10.1037/a0014463

Rotello, C. M., Masson, M. E. J., & Verde, M. F. (2008). Type I error rates and power analyses for single-point sensitivity measures. *Perception & Psychophysics, 70,* 389–401. doi:10.3758/PP.70.2.389

Schunn, C. D., Reder, L. M., Nhouyvanisvong, A., Richards, D. R., & Stroffolino, P. J. (1997). To calculate or not to calculate: A source activation confusion model of problem familiarity's role in strategy selection. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 3–29. doi:10.1037/0278-7393.23.1.3

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461–464. doi:10.1214/aos/1176344136

Shimamura, A. P., & Squire, L. R. (1986). Metamemory and memory: A study of the feeling-of-knowing phenomenon in amnesic patients. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12,* 452–460. doi:10.1037/0278-7393.12.3.452

Sikström, S., & Jönsson, F. (2005). A model for stochastic drift in memory strength to account for judgments of learning. *Psychological Review, 112,* 932–950. doi:10.1037/0033-295X.112.4.932

Spellman, B. A., & Bjork, R. A. (1992). When predictions create reality: Judgments of learning may alter what they are intended to assess. *Psychological Science, 3,* 315–316. doi:10.1111/j.1467-9280.1992.tb00680.x

Swets, J. A. (1986a). Form of empirical ROCs in discrimination and diagnostic tasks: Implications for theory and measurement of performance. *Psychological Bulletin, 99,* 181–198. doi:10.1037/0033-2909.99.2.181

Swets, J. A. (1986b). Indices of discrimination or diagnostic accuracy: Their ROCs and implied models. *Psychological Bulletin, 99,* 100–117. doi:10.1037/0033-2909.99.1.100

Treisman, M., & Faulkner, A. (1985). Can decision criteria interchange locations? Some positive evidence. *Journal of Experimental Psychology: Human Perception and Performance, 11,* 187–208. doi:10.1037/0096-1523.11.2.187

Tulving, E. (1981). Similarity relations in recognition. *Journal of Verbal Learning and Verbal Behavior, 20,* 479–496. doi:10.1016/S0022-5371(81)90129-8

Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 582–600. doi:10.1037/0278-7393.26.3.582

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & Iverson, G. J. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology, 48,* 28–50. doi:10.1016/j.jmp.2003.11.004

Wallsten, T. W., & González-Vallejo, C. (1994). Statement verification: A stochastic model of judgment and response. *Psychological Review, 101,* 490–504. doi:10.1037/0033-295X.101.3.490

Appendix

Model Flexibility and Bayesian Recovery Probability (BRP)

Traditional flexibility correction procedures, such as Akaike's information criterion (Akaike, 1973) and Bayesian information criterion (Schwarz, 1978) cannot apply here because the three versions of M4 are at the same level in the hierarchy and therefore have the same number of parameters. Model mimicry methods, such as the parametric bootstrap cross-fitting method (Wagenmakers et al., 2004), also cannot be applied in this instance because they rely on pairwise comparisons, which may yield intransitive results when comparing three or more models. Therefore, we developed a new method, or the BRP, that avoids all these difficulties by estimating the probability of recovering the true model given that model and its competitors.

The procedure entails using each of the M4 models to generate an artificial data set, which is followed by application of all of the candidate models to each artificial data set to determine which model yields the best fit. By carrying this process out thousands of times, one can determine whether differential model flexibility is a problem. We first determined the proportion of times each model yielded the best fit to the 1,000 data sets generated for each model, which are shown in Part A of Table A1.

These entries in Part A show the percentage of model *i* providing the best fit given that model *j* generated the data when the competition is among models *i*, *j*, *k*. What we want, however, when considering empirical data for which the true model is not known, is the reverse probability, the BRP, which is the probability that model *i* is the correct model given that it emerged as the best fitting among candidate models *i*, *j*, *k*. We obtain the BRP by assuming equal prior probabilities among the three models and applying Bayes rule to the probabilities in Part A. The results are shown in Part B. For example, when considering the delayed versus immediate paradigm, the probability that M4a is the true model is .797 given that it best fit the data among candidate models, M4a, M4b, and M4c. Continuing along this row where M4a provides the best fit, the probability that M4b is the true model is .150, and the probability that M4c is the true model is .053. It is apparent that the models are approximately equal in flexibility.

(*Appendix continues*)

Table A1
Model Recovery: (A) Percentage of the Best Fit to Simulated Data (N = 1,000 per Cell) and (B) Probability of Each Model Given the Best Model Fit Under the Equal Priors of Bayes Rule (Bayesian Recovery Probability)

			True model		
			Memory	*Confidence*	*Correlation*
	Data set	Fitted model	(M4a)	(M4b)	(M4c)
(A)	Delayed	*Memory* (M4a)	84.0	15.8	5.6
	vs.	*Confidence* (M4b)	3.5	78.3	3.2
	Immediate	*Correlation* (M4c)	12.5	5.9	91.2
	Testing	*Memory* (M4a)	77.4	19.9	2.9
	vs.	*Confidence* (M4b)	6.5	77.3	4.8
	No testing	*Correlation* (M4c)	16.1	2.8	92.3
(B)	Delayed	*Memory* (M4a)	0.797	0.150	0.053
	vs.	*Confidence* (M4b)	0.041	0.921	0.038
	Immediate	*Correlation* (M4c)	0.114	0.054	0.832
	Testing	*Memory* (M4a)	0.772	0.199	0.029
	vs.	*Confidence* (M4b)	0.073	0.872	0.054
	No testing	*Correlation* (M4c)	0.145	0.025	0.830

Received December 13, 2010
Revision received August 25, 2011
Accepted August 26, 2011 ■

A correction to this article appears on page 221.

Correction to Jang, Wallsten, and Huber (2011)

In the article “A stochastic detection and retrieval model for the study of metacognition,” by Yoonhee Jang, Thomas S. Wallsten, and David E. Huber (*Psychological Review*, Online First Publication. November 7, 2011. doi:10.1037/a0025960), incorrect equations were published. The corrected forms of Equations (1) and (2) in this article are as follows:

$$p(J_i, \text{recalled}) = \int\int h(x, y, \rho)N(x|C_M, \sigma_M)[N(y|C_i, \sigma_C) - N(y|C_{i+1},\sigma_C)]dxdy \qquad (1)$$

$$p(J_i, \text{not recalled}) = \int\int h(x, y, \rho)[1 - N(x|C_M, \sigma_M)][N(y|C_\mathrm{i}, \sigma_c) - N(y|C_{i+1}, \sigma_C)]dxdy \qquad (2)$$

The corrected forms of Equations in the Supplemental Material are as follows:

When $i = 0$,

$$p(J_i, \text{recalled}) = \int\int h(x, y, \rho)N(x|C_M, \sigma_M)[1 - N(y|C_{i+1}, \sigma_C)]dxdy,$$

$$p(J_i, \text{not recalled}) = \int\int h(x, y, \rho)[1 - N(x|C_M, \sigma_M)][1 - N(y|C_{i+1}, \sigma_C)]dxdy,$$

when $0 < i < n$,

Linked version of the SDRM:

$$p(J_i, \text{recalled}) = \int\int h(x, y, \rho)N(x|C_M, \sigma_M)[N(y|C_i, \sigma_C) - N(y|C_{i+1}, \sigma_C)]dxdy,$$

$$p(J_i, \text{not recalled}) = \int\int h(x, y, \rho)[1 - N(x|C_M, \sigma_M)][N(y|C_i, \sigma_C) - N(y|C_{i+1}, \sigma_C)]dxdy,$$

Independent version of the SDRM:

$$p(J_i, \text{recalled}) \propto \int\int h(x, y, \rho)N(x|C_M, \sigma_M)N(y|C_i, \sigma_C)[1 - N(y|C_{i+1}, \sigma_C)]dxdy,$$

$$p(J_i, \text{not recalled}) \propto \int\int h(x, y, \rho)[1 - N(x|C_M, \sigma_M)]N(y|C_i, \sigma_C)[1 - N(y|C_{i+1}, \sigma_C)]dxdy,$$

and when $i = n$,

$$p(J_i, \text{recalled}) = \int\int h(x, y, \rho)N(x|C_M, \sigma_M)N(y|C_i, \sigma_C)dxdy,$$

$$p(J_i, \text{not recalled}) = \int\int h(x, y, \rho)[1 - N(x|C_M, \sigma_M)]N(y|C_i, \sigma_c)dxdy.$$

DOI: 10.1037/a0026619