Journal of Memory and Language 62 (2010) 183–203

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Recollection and confidence in two-alternative forced choice episodic recognition

Andrew Heathcote *, Beatrice Bora, Emily Freeman

School of Psychology, The University of Newcastle, Australia

ARTICLE INFO

Article history:
Received 9 July 2009
revision received 30 October 2009
Available online 8 December 2009

Keywords:
Recognition memory
Memory modeling
Face recognition memory
Recognition memory confidence
Remember–know

ABSTRACT

We used quantitative modeling of remember–know responses (Tulving, 1985) to investigate the processes underlying recognition memory for faces in the choice-similarity paradigm (Tulving, 1981). Similarity between recognition test choices produces opposite effects on confidence and accuracy in this paradigm. We extended existing models of this double dissociation to account for remember–know responses, by adding a variable recollection criterion to Clark's (1997) single-process model and by adding graded recollection strength to Dobbins, Kroll, and Liu's (1998) dual-process model. Both models provided an accurate and comprehensive account of objective and subjective judgments in an experiment we conducted on memory for faces, and of data from Dobbins et al. on memory for natural scenes. Model selection techniques were used to refine these accounts, providing insight into the psychological processes proposed by each approach and into their implications for the relationship between recollection and confidence in two-alternative forced choice recognition.

Introduction

In this paper we investigate the relationship between confidence and recollection in episodic recognition memory. Episodic recognition requires a decision about whether a test item appeared in a previous study episode. Recollection is a process that retrieves from memory details that were associated with a test item during the study episode. For example, on first encountering a person you may be struck by their hooked nose, which reminds you of a classical Roman statue. When you later meet this person again you recollect thinking about their Roman nose on your first encounter. That recollection serves to confirm that you have previously met the person, allowing you to confidently greet them on your second meeting.

The idea that episodic recognition is in part based on recollection is commonly called dual-process theory (Yonelinas, 2002). Mandler (1980) illustrated dual-process theory with a story about encountering a man on a bus. The man seemed highly familiar, but it was only after a conscious search process, during which memory was probed with various potential contexts, that the man was recognized as the butcher from the supermarket. Mandler goes on to propose that, although the context-free familiarity occurring in the "butcher-on-the-bus" example is unusual, it illustrates the two processes that generally underpin episodic recognition, and that: "In the normal course of events, the two separate processes occur conjointly; recognition involves the additive effects of familiarity and retrieval." (p. 253). Rather than a literally additive relationship, Mandler proposed an either/or process for using these two sources of mnemonic information. If recollection occurs, the recognition decision is based only on recollection. When recollection fails, the decision is based on familiarity, as familiarity is always available. This proposal has been adopted by all subsequent dual-process theories, with the exception of Wixted (2007).

Participants are often asked to rate their confidence in recognition decisions, and are usually more confident in

* Corresponding author. School of Psychology, Psychology Building, The University of Newcastle, University Avenue, Callaghan 2308, Australia. Fax: +61 2 49216906.
E-mail address: andrew.heathcote@newcastle.edu.au (A. Heathcote).

0749-596X/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jml.2009.11.003

conditions where they are also more accurate. In extending dual-process theory to explain recognition confidence, Yonelinas (1994) proposed that recollection results in highly confident recognition. In contrast to Mandler (1980), he assumed familiarity is continuous and normally distributed, with a higher mean for studied (old) than unstudied (new) test items, but with equal variance in both cases. Confidence based on familiarity is determined in the manner of signal detection theory (Macmillan & Creelman, 2005), by placing a set of criteria on the familiarity dimension. Values of familiarity above the largest criterion result in a highly confident old decision even in the absence of recollection. Less confident decisions are given when familiarity falls between intermediate criteria. Yonelinas showed that this model provided an accurate account of Receiver Operating Characteristic (ROC) data obtained by having participants rate their confidence in single test item recognition decisions.

We investigate the extension of the dual-process approach to confidence in two-alternative forced choice (2AFC) episodic recognition decisions. We focus on 2AFC responding because, in contrast to the usual finding in single item recognition, confidence and accuracy can become dissociated (i.e., less accurate responses are given with greater confidence) when the two choice alternatives are sufficiently similar. In particular, Tulving (1981) found that episodic recognition of natural scene stimuli was more accurate but less confident when the choice alternatives were highly similar compared to when they were less similar. His finding is surprising, not only because the confidence–accuracy dissociation differs from the usual finding of a positive relationship between confidence and accuracy, but also because similarity is usually found to be detrimental rather than helpful for recognition accuracy (e.g., Wickelgren, 1977).

Dobbins, Kroll, and Liu (1998) used Tulving's (1985) "remember–know" (RK) methodology to investigate the cause of Tulving's (1981) surprising findings. The RK methodology relies on participant's ability to report whether episodic recognition decisions are based on recollection or familiarity. In an experiment using natural scene stimuli Dobbins et al. found that recognition decisions were less often classified as based on recollection when choice-similarity was high than when it was low. They suggested that high choice-similarity tends to result in recollection of both the target (i.e., the previously studied test alternative) and the distracter (i.e., false recollection). Participants respond to the conflicting recollections by basing their recognition decision on familiarity. As familiarity, on average, produces lower confidence decisions, confidence tends to be less when choice-similarity is high than when it is low. However, accuracy is improved because errors caused by false recollection are less common when choice-similarity is high. That is, because false recollection is correlated with correct recollection when choice-similarity is higher, participants are less often misled by false recollection than when choice-similarity is lower.

In summary, Dobbins et al. (1998) proposed a dual-process explanation of both Tulving's (1981) confidence–accuracy dissociation and of the effect of choice-similarity on RK responses. Recently, Heathcote, Freeman, Etherington, Tonkin, and Bora (2009) replicated the confidence–accuracy dissociation with face stimuli. Here we use the RK methodology to investigate the cause of the confidence–accuracy inversion with face stimuli. To do so we fit an extension of Dobbin's et al.'s dual-process model to our accuracy, confidence and RK data. We also do the same with an extension of an alternative model of the confidence–accuracy dissociation proposed by Clark (1997). Clark's model is based on MINERVA 2, Hintzman's (1988) "single-process" theory, which assumes that episodic recognition decisions are based on a single type of memory representation and matching process. We then compare the accounts provided by the two alternative models. Before describing the extended models we first describe our experimental paradigm and discuss the differing ways in which RK data are interpreted by single and dual-process theories.

The face choice-similarity paradigm

In our paradigm participants briefly studied each of a sequence of pictures of male and female faces and were then tested by 2AFC decisions about pairs of new and old faces. The test pairs were either higher (same gender) or lower (different gender) in choice-similarity. In some cases the distracter or new faces were specifically chosen to be similar to a studied face. These new faces might appear with the studied face to which they are similar, in a high choice-similarity test pair. Alternately, they might be similar to an untested study face and appear in a low choice-similarity pair. On each test trial participants chose either the left or right member of the pair and rated their confidence in the choice. They then gave an RK response, indicating whether their decision was based on the recollection of details (remember) or familiarity (know).

This paradigm is based on the experiment reported by Dobbins et al. (1998), which in turn was based on several experiments reported by Tulving (1981). All of these experiments examined memory for scenic pictures rather than faces. Choice-similarity was manipulated by comparing performance for choices between studied and unstudied halves of the same scene versus different scenes. Participants were more confident in their decisions in the low choice-similarity (different scene) condition, but choice-similarity had the opposite effect on accuracy: accuracy in the high choice-similarity (same scene) condition was greater than in the low choice-similarity condition.

Note that the *choice-similarity effect* occurs when the *effect of memory-similarity* (i.e., the similarity between the new test item and memory traces)[1] is controlled. That is, the low choice-similarity pairs used in the comparison had a new item that was similar to an untested study item to the same degree that the new test item in the high

[1] We avoided the more commonly used term *target-lure similarity* because that could also apply to the similarity between choices in a two-alternative forced-choice test (i.e., what we call *choice-similarity*). The term *memory-similarity* emphasizes the relationship between a memory trace and a test lure, which is what our memory-similarity manipulation affects, while controlling for choice-similarity.

choice-similarity condition was similar to the old test item. Tulving (1981) found that the dissociation between confidence and accuracy only occurred when memory-similarity was sufficiently high. He manipulated memory-similarity based on the results of a calibration study in which his scene stimuli were rated on the similarity of their halves. The ratings were used to sort scenes into higher and lower memory-similarity sets, and each set was used to create higher and lower choice-similarity test pairs.

Heathcote et al.'s (2009) replication of the confidence–accuracy dissociation with face stimuli also used a rating method to create higher and lower similarity sets of face pairs in order to investigate the effect of memory-similarity. In contrast to Tulving (1981), they found that the dissociation occurred with both lower and higher memory-similarity sets. Likely this occurred because all of their face pairs were chosen to be similar to some degree, and because faces tend to be intrinsically more similar to each other than parts of scenes. Hence, even Heathcote et al.'s lower similarity faces pairs were sufficiently similar to support the confidence–accuracy dissociation.

Heathcote et al. (2009) used state-trace analysis (Bamber, 1979) to investigate the cause of the confidence–accuracy dissociation. In contrast to other methods, state-trace analysis can rigorously determine whether more than a single psychological dimension (i.e., a single latent variable, module, or process) is required to explain a dissociation (Dunn & Kirsner, 1988; Loftus, Oberg, & Dillon, 2004). They found that more than one dimension was required to explain the confidence–accuracy dissociation they observed.

Dobbins et al. (1998) took a different approach to investigating the cause of the confidence–accuracy dissociation, by having participants give an RK response following a simultaneous choice and confidence rating. As well as finding that remember responses were less common for higher choice-similarity pairs than lower choice-similarity pairs, they also found that the choice-similarity effect on accuracy occurred only in remember responses. In contrast, they found that the confidence effect occurred only in know responses. These results provide a potential psychological-process explanation for Heathcote et al.'s (2009) state-trace findings. That is, the confidence–accuracy dissociation is caused by different effects of choice-similarity on underlying recollection and familiarity processes. However, as we now discuss, other interpretations of RK data have been suggested.

The remember–know paradigm

The interpretation of RK findings is controversial. It has been suggested that confidence and RK judgments do not reflect separate recollection and familiarity processes. Rather, both depend on a single "memory strength" dimension, and both types of judgment are made using a decision process that compares memory strength to a set of criteria in the manner of signal detection theory (Donaldson, 1996). In this view, remember judgments are like high confidence judgments, in that they occur when memory strength is greater than a high criterion, with participants adopting different confidence and RK criteria depending on experimental demands. Wixted and Stretch (2004) elaborated this view, suggesting that RK criteria are more variable than confidence criteria because participants are highly practiced at making confidence judgments, whereas they only encounter the idea of a RK judgment just prior to the experiment.

This alternative interpretation is sometimes called the single-process or signal-detection RK model. For our purposes here the key difference between the two interpretations of RK data is that the dual-process approach assumes confidence and RK decisions are based on one type of evidence (i.e., recollection), or another (i.e., familiarity), but not both at once. In contrast, the single-process approach assumes that both types of decisions are based the same memory strength dimension, which combines all sources of available mnemonic evidence about the test item.

Determining whether the single- or dual-process RK interpretation is correct has not proved amenable to what is often thought to be the ideal test, strong inference (Platt, 1964). A strong inference test compares models by contrasting different predictions about the order of performance across experimental conditions. Such tests are only valid if the predictions are parameter free, that is, if the predictions hold for all reasonable values of the model's parameters. Dunn (2004) evaluated five tests that had been proposed by proponents of dual-process approach, and showed that none were valid. He also showed that quantitative fits of the single-process model were quite consistent with a large database of RK data.

Similarly, Gardiner, Ramponi, and Richardson-Klavehn's (2002) evidence favoring the dual-process model of RK data was found to be flawed by Macmillan, Rotello, and Verde (2005), due to Gardiner et al.'s use of the *A′* measure of accuracy (see also Benjamin, 2005). Macmillan et al. concluded that quantitative models are necessary to interpret RK data (see Dougal & Rotello, 2007; Kapucu, Rotello, Ready, & Seidel, 2008; Rotello, Macmillan, Hicks, & Hautus, 2006; Starns & Ratcliff, 2008, for applications of this approach to single item recognition).

However, the dual-process approach has been criticized by Wixted and Stretch (2004) because it has not been elaborated to provide a comprehensive quantitative account of RK data. In particular, they pointed out that most existing dual-process models do not account for false remember responses (i.e., incorrect recognition responses that are classified as being based on recollection). They suggested that a comprehensive dual-process account requires a false recollection process to account for remember errors, and that both correct and false recollection processes must provide graded evidence, as is assumed by some dual-process models (e.g., Cary & Reder, 2003). Rotello, Macmillan, Reeder, and Wong (2005) reported empirical results supporting the graded nature of the remember response in single item recognition.

Dunn (2008) made a potentially even more telling criticism of the dual-process account. He applied state-trace analysis to a database of 37 RK studies. His results provided little or no support for the claim that two processes are necessary to explain RK judgments. It is possible that one dimension was sufficient because the experimental manipulations in these studies caused strongly correlated changes in recollection and familiarity, despite the fact that was proposed not to be the case in the original interpretations of these results. If such a correlation is sufficiently high, state-trace analysis will indicate that only one dimension is necessary to explain the data.

However, given Heathcote et al.'s (2009) finding that more than one dimension is required to explain the confidence–accuracy dissociation they observe, the choice-similarity paradigm appears not to be subject to this problem. We note effects due to more than one dimension in this paradigm do not uniquely support the dual-process account. They could also be explained by a single-process model in which choice-similarity affects different parameters controlling a single memory strength dimension (e.g., its mean and variance). Hence, we examined both single and dual-process accounts of our data, rather than assuming one or the other to be correct.

Our examination of the alternative models is quantitative and comprehensive, in the sense that we fit each type of model to the full array of RK, confidence and 2AFC data (i.e., both correct and false response proportions broken down by both confidence and RK classifications). In order to do so we elaborated two existing accounts of the choice-similarity effect, Clark's (1997) single-process model and Dobbins et al.'s (1998) dual-process model. The elaborations were along the lines suggested by Wixted and Stretch (2004), adding a signal-detection account of RK responses to Clark's model and graded false and correct recollection to Dobbins et al.'s model. We first report the results of our experiment in terms of the confidence, accuracy and remember-response-proportion measures used in the analysis of previous choice-similarity experiments. We then provide details of the base models, and their elaborations, which we denote the single-process remember–know (SPRK) and dual-process remember–know (DPRK) models.

Experiment

Face stimuli (105 × 120 pixel black and white bitmaps) were taken from the FERET database (Phillips, Wechsler, Huang, & Rauss, 1998), which provides faces classified by gender and race (Black, Asian and White). Face images were sorted into 377 pairs and rated on the similarity between pair members by ten first year psychology students using a 5 point scale (1 = *very low* to 5 = *very high*). The face pairs were then rank ordered within gender and race categories using average similarity ratings. Higher and lower similarity sets, consisting of 150 pairs each, were created by selecting lowest and highest ranked pairs. The two sets were used to manipulate memory-similarity.

We crossed memory-similarity with a three level choice-similarity factor of the same type used by Tulving (1981), Experiment 1) and Dobbins et al. (1998). Following their nomenclature, studied test items are denoted A, unstudied (new) test items whose pair mates were studied are denoted by a prime (e.g., B' if B was studied), and new test items whose pair mates were never studied are denoted X'. Hence, the higher choice-similarity condition is denoted $A–A'$. The other two conditions, which have lower (and equal) choice-similarity, are denoted $A–B'$ and $A–X'$. $A–B'$ has the same average match to memory as $A–A'$, and $A–X'$ has a lower average match to memory compared to both $A–A'$ and $A–B'$, as the pair mate of X' was not studied.

In previous experiments with scenes (Dobbins et al., 1998; Tulving, 1981), random assignment was used to create lower choice-similarity test pairs. In contrast, our lower choice-similarity pairs were always made up of faces with different genders (higher choice-similarity pairs made up of faces of the same gender and study lists consisted of faces from only one race with equal numbers of each gender). This was done, in light of the structural similarity between randomly chosen pairs of faces, to insure clearly lower perceptual similarity between test pairs in the $A–B'$ and $A–X'$ conditions than in the $A–A'$ condition.

Note that Heathcote et al. (2009) performed experiments with a subset of the face stimuli used here and a similar design, except that they paired different race faces rather than different gender faces to create the lower choice-similarity condition, and they omitted the $A–X'$ condition (as in Tulving, 1981, Experiment 2). Analyses of accuracy and confidence produced largely the same pattern of results as here, indicating that gender and race related choice-similarity manipulations had similar effects.

Method

Participants

Participants were 64 introductory psychology students at the University of Newcastle, Australia, who received course credit for participation.

Apparatus and procedure

Table 1 provides the rating results and details of the gender and race classification for the 300 critical pairs. One member of each critical pair was randomly selected and used to create 15 study lists (2 Asian, 4 Black and 9 White), made up of 10 lower and 10 higher memory-similarity items, half male and half female. Assignment to study lists, and study order was random. One face appeared before, and one after, the critical faces as primacy and recency buffers which were not tested. The buffer

Table 1
Characteristics of the 300 critical experimental face pairs.

Gender	Race	Similarity	Mean rating (%)	Number
Female	Asian	Lower	36	10
		Higher	60	10
	Black	Lower	37	20
		Higher	66	20
	White	Lower	37	45
		Higher	60	45
Male	Asian	Lower	31	10
		Higher	71	10
	Black	Lower	35	20
		Higher	60	20
	White	Lower	37	45
		Higher	69	45

faces, and faces used for an initial practice study-test cycle, were drawn from the remaining 77 rated face pairs. Test pairs were presented side-by-side. Test lists had equal numbers of male and female faces. They consisted of 12 pairs, two from each of the six memory- and choice-similarity conditions. Note that the same critical pair was never used twice in a study list. For example, if the pair of items *A*–*B*′ was tested, a test pair containing *B* (e.g., *B*–*C*′) was never tested. Study faces were randomly allocated to choice-similarity conditions, and test pairs were presented in a random order.

An eight button response box, consisting of left and right hand clusters of three keys and a central pair of keys, was used to simultaneously record confidence and accuracy, with the left and right clusters labeled 3, 2, 1 and 1, 2, 3 from left to right. Larger numbers indicated greater confidence, and participants were instructed to make use of all confidence levels. Pressing a button in the left cluster indicated a left test choice, and pressing a button in the right cluster indicated a right choice. The central pair of keys was used to make RK judgments. Participants were instructed to press the left button, labeled *remember*, if they remembered seeing the face, or particular elements of the face, or the right button, labeled *familiar*, if the face was familiar but they did not remember the face or any particular elements of the face (these instructions were modeled on Dobbins et al., 1998). During study, buttons marked 1–3 were used to make face typicality ratings (ranging from 1 = *very typical* to 3 = *very unusual*). Study ratings were used to ensure attention to the faces and were not further analyzed.

Each participant was tested on a PC with a 1168 × 856 resolution monitor. The session began with participants reading instructions on the screen at their own pace. During study, faces were displayed one at a time in the middle of the screen for 2 s. After each face appeared, participants were prompted to make a typicality rating. The test phase began immediately after study. In the test phase, face pairs appeared one pair at a time. If no response was made after 6 s, the next face pair was displayed. The experiment lasted less than 1 h.

Results

Three participants were excluded from analysis, one because of a computer failure, one because they rarely used the lowest confidence rating and one because they rarely used the highest confidence rating. Linear mixed effects models (Bates. D. M., 2005) with random subject intercepts were used to obtain population mean estimates and standard errors, and to perform inference. We report the results of these analyses in terms of single degree of freedom tests, and test results where $p < .05$ are described as reliable.

Note that, as in Tulving's (1981) experiment, the crossing of choice and memory-similarity factors is not strictly orthogonal. For high choice-similarity (*A*–*A*′) pairs, for example, increasing memory-similarity also increases choice-similarity. In the same vein, *A*–*B*′ and *A*–*X*′ pairs are equated for choice-similarity but differ in memory-similarity. Hence, the single degree of freedom contrasts that we report are more appropriate than standard ANOVA main effect and interaction tests.

Fig. 1 presents population estimates of response proportion results; the percentages of correct and false remember responses and the percentages of accurate remember and know responses. We assumed a binomial-probit error model (McCullagh & Nelder, 1989) and used maximum-likelihood estimation to analyze this data. These analyses produced standard normal (*Z*) test statistics derived from the mixed model parameter estimate variance–covariance matrix. To avoid redundancy we report only the corresponding null-hypothesis probability values.

We also applied the same analysis to overall accuracy (i.e., without dividing responses into remember and know). As overall accuracy is not easily obtained from Fig. 1, we accompany reports of reliable effects in overall accuracy with the corresponding average percentages of correct responses.

Remember vs. know response proportions

As was found by Dobbins et al. (1998), remember responses were less common when choice-similarity was higher (*A*–*A*′) than when it was lower (*A*–*B*′ and *A*–*X*′). These differences were reliable for both correct ($p = .005$ and $p = .012$ respectively) and false ($p < .001$ and $p = .014$ respectively) remember responses. In contrast, the two lower choice-similarity conditions did not differ reliably, although false remember responses were marginally less common in the *A*–*X*′ than the *A*–*B*′ condition for lower memory-similarity pairs ($p = .051$). Memory-similarity had little effect on the frequency of remember responses, with the exception of a marginally higher overall percentage of correct remember responses for lower compared to higher memory-similarity pairs ($p = .078$).

Accuracy

Consistent with Tulving's (1981) results, overall accuracy in the *A*–*A*′ condition was reliably greater than in the *A*–*B*′ condition when memory-similarity was high (86% vs. 80%, $p < .001$), but not when it was low. In the *A*–*B*′ condition, overall accuracy was greater when memory-similarity was high than when it was low (86% vs. 80%, $p < .001$). No other effects of memory-similarity or choice-similarity on overall accuracy approached significance.

Results for our high memory-similarity condition replicated Dobbins et al. (1998) in that the accuracy of remember responses was less in the *A*–*B*′ condition than in the *A*–*A*′ and the *A*–*X*′ conditions ($p < .001$ and $p = .009$ respectively). These effects were not reliable when memory-similarity was lower ($p = .097$ and $p = .179$ respectively), although equivalent trends were evident. In contrast to Dobbins et al. (1998), the accuracy of know responses was reliably affected by choice-similarity, at least when memory-similarity was higher (*A*–*B*′ < *A*–*A*′, $p = .003$, and *A*–*B*′ < *A*–*X*′, $p = .012$). There was essentially no effect of choice-similarity on the accuracy of know responses when memory-similarity was lower ($p > .5$ in both cases).

The accuracy of remember responses was also reliably greater for lower than higher memory-similarity pairs in

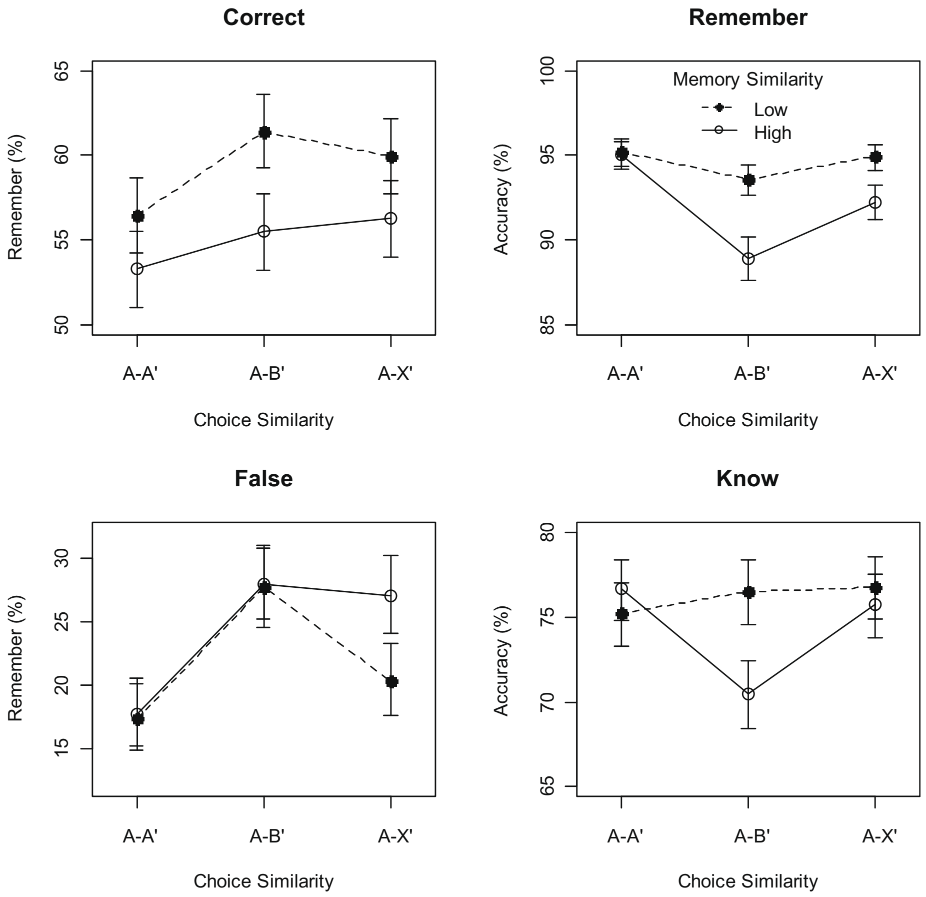


Fig. 1. Population mean estimates of the proportion of remember responses and accuracy, with 95% confidence intervals derived from the mixed model variance–covariance matrix.

the *A–B′* ($p < .001$) and *A–X′* ($p = .013$) conditions, but not in the *A–A′* condition. For know responses only the difference in the *A–B′* condition was reliable ($p = .031$).

Confidence

Fig. 2 shows confidence results broken down by accuracy and RK response. For ease of interpretation, confidence was transformed to 0–100 scale, which we refer to as a percentage, using $100 \times (r - 1)/2$, where r is the 1–3 integer confidence rating. We assumed a Gaussian error model and used restricted maximum-likelihood estimation in the analysis of confidence. Frequentist inferential results for the Gaussian error model (using *t*-statistics) were checked using Bayesian methods, as recommended by Baayen, Davidson, and Bates (2008). Both methods produced the same conclusions and we report null-hypothesis probability values obtained using the latter method.

We also applied the same analysis to overall confidence (i.e., without dividing responses into remember and know). As overall confidence is not easily obtained from Fig. 2, we accompany reports of reliable effects in overall confidence for correct and error responses with the corresponding average percentages of correct responses.

Choice-similarity affected overall confidence in error responses but had no reliable effect on overall confidence in correct responses. For error responses to low memory-similarity pairs, confidence was higher in the *A–B′* condition (35%) than in the *A–A′* condition (27%, $p = .002$) and the *A–X′* condition (30%, $p = .009$). When memory-similarity was high, error response confidence was also higher in *A–B′* condition (34%) than in the *A–A′* condition (25%, $p = .001$) but not the *A–X′* condition (33%). Confidence in correct responses was, however, reliably greater when memory-similarity was low than when it was high in both the *A–B′* (55% vs. 51%, $p = .011$) and *A–A′* (54% vs. 52%, $p = .04$) conditions, but not in the *A–X′* conditions. Memory-similarity did not have any reliable effects on error response confidence.

Replicating Dobbins et al. (1998), no choice-similarity effects were reliable for remember confidence. In contrast, there were reliable effects on know confidence, mainly for errors. Confidence effects for correct responses were restricted to lower memory-similarity pairs, with *A–B′* confidence reliably lower than *A–A′* ($p = .005$) and *A–X′* ($p = .03$) confidence. For error responses, confidence in the *A–B′* condition was reliably greater than confidence in the *A–A′* condition. This was true for both lower and higher

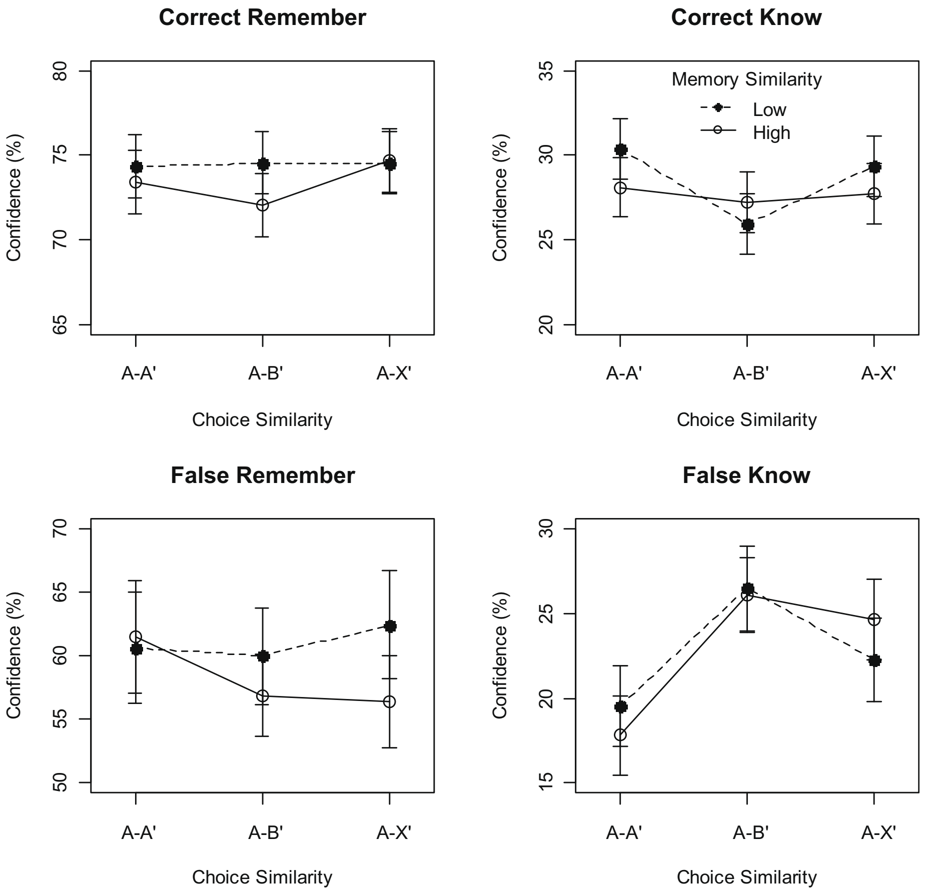


Fig. 2. Population mean estimates of confidence broken down by accuracy and remember–know, with 95% confidence intervals derived from the mixed model variance–covariance matrix.

memory-similarity pairs ($p = .013$ and $p = .001$, respectively). *A*–*B*′ confidence was reliably greater than *A*–*X*′ confidence for lower ($p = .03$) but not higher ($p = .73$) memory-similarity pairs. Memory-similarity had only one reliable effect, greater confidence for remember responses to low than high memory-similarity pairs in the *A*–*B*′ condition ($p = .02$).

Discussion

We largely replicated effects related to the proportion of remember responses reported by Dobbins et al. (1998). Remember responses were less common when choice-similarity was higher (*A*–*A*′) than when it was lower (*A*–*B*′ and *A*–*X*′). We also found that memory-similarity did not have a reliable effect on the proportion of remember responses (note that Dobbins et al. did not manipulate memory-similarity). Our accuracy results differed from those of Dobbins et al. in that the *A*–*B*′ condition had lower accuracy than the *A*–*A*′ and *A*–*X*′ conditions for both remember and know responses, at least when memory-similarity was higher. In contrast, they found a reliable effect on accuracy only for remember responses. They also found that accuracy was greater in the *A*–*X*′ condition than the other two conditions, as did Tulving (1981), whereas accuracy was approximately equivalent in the *A*–*A*′ and *A*–*X*′ conditions in our experiment.

Our confidence results replicated some aspects of Tulving's (1981) and Dobbins et al.'s (1998) results, but not others. Like Dobbins et al. we found that choice-similarity effects on confidence were only evident for know responses. In our case the confidence effect was most evident for false know responses (i.e., incorrect recognition responses that are classified as being based on familiarity), with lower confidence when choice-similarity was higher (*A*–*A*′) than when it was lower (*A*–*B*′ and *A*–*X*′). For correct know responses our confidence effects were much weaker, and unexpectedly there was a small but reliable reversal of the usual finding for the low memory-similarity condition, with lower confidence in the *A*–*B*′ condition than the *A*–*A*′ and *A*–*X*′ conditions. Like Dobbins et al. we also found weaker overall effects of choice-similarity on error than correct responses, although they did report significantly less confidence in the *A*–*A*′ condition than the *A*–*B*′ condition for correct responses, whereas we did not find this difference to be reliable. Similarly, our confidence effects were weaker than those reported by Tulving (1981).

In summary, we replicated with faces many of the previous findings with scene stimuli (Dobbins et al., 1998;

Tulving, 1981), but there was also some divergence of results. This divergence brings into question the generality of the conclusions which we draw from the model analyses of our data reported in the next section. Fortunately, Dobbins et al. provided their results aggregated over participants (Table 3, p. 1311) in sufficient detail for us to fit our models. In order to address the question of generality, following the model analysis of our data we report the results of fitting versions of these models to their aggregate data.

Modeling

Clark (1997) showed his single-process model produces correct predictions for Tulving's (1981) accuracy and confidence data in an ordinal sense. Dobbins et al.'s (1998) dual-process model was in the main specified verbally. Neither theory has been tested, nor compared, in a comprehensive and quantitative manner. We first describe these models, and then describe elaborations enabling them to be applied to our paradigm. We then illustrate the elaborations by reporting the results of fitting a version of each type of model to our data. Next, we describe how we selected amongst different versions of each model type, and how we compared between types. Finally, we show that the models favored in our data also apply to Dobbins et al.'s data on memory for scenes. We conclude that the selected models provide a coherent overall account that requires only changes in the values of the same set of parameters to accommodate differences in stimulus type and experimental methodology.

Clark's (1997) model

In Clark's (1997) model, 2AFC decisions are based on evidence obtained by subtracting the old test item's match to memory from the new test item's match. Choices are determined by the sign of the difference, and confidence by how far the difference is from zero, relative to a set of symmetric confidence criteria arrayed around zero. The MINERVA 2 memory model (Hintzman, 1988) and arguably many other quantitative memory models (see Clark & Gronlund, 1996) predict that higher choice-similarity increases the correlation between the two matches. The variance of the difference between two random variables, Var($X - Y$), decreases as their correlation, and hence their covariance, Cov(X, Y), increases: Var($X - Y$) = Var(X) + Var(Y) − 2 × Cov(X, Y). Therefore, higher choice-similarity decreases the variance of the evidence (match difference) distribution. The decrease in variance increases accuracy, because it is less likely that the evidence falls on the wrong side of zero, but decreases confidence, because it is less likely that the evidence will be extreme enough to exceed the higher confidence criteria. Fig. 3 provides an example of the evidence distributions predicted by Clark's model.

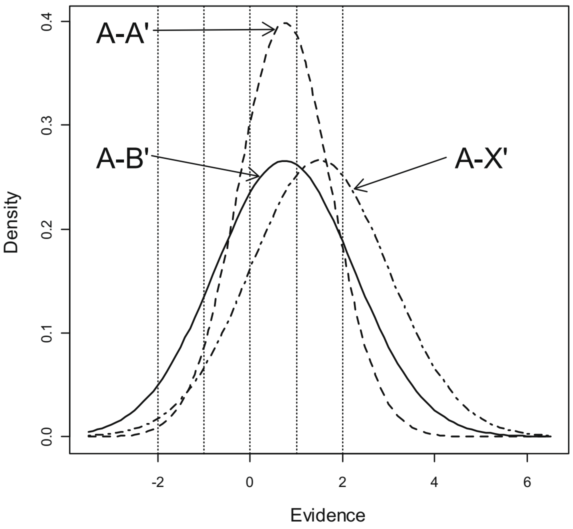


Fig. 3. Evidence distributions (old test item memory strength minus new test item memory strength) for Clark's (1997) model. Criteria for low/medium/high confidence responses are indicated by vertical dotted lines. The criteria are symmetric around zero. Evidence values less than zero result in an incorrect response and evidence values greater than zero result in a correct response.

Clark's (1997) model predicts a number of equalities and inequalities among the standard deviations (s) and means (d') of the evidence distributions in Fig. 3. Given X' and B' items should be equally uncorrelated with A items, A–B' and A–X' conditions will have equal standard deviations, which are greater than the standard deviation in the A–A' condition.

$$s(A\text{—}A') < s(A\text{—}B') = s(A\text{—}X') \quad (1)$$

The mean match to memory depends on the similarity between the test item and memory traces. All old test items have the same mean match, which is greater than the mean match for any new item, because of a strong match to the memory trace established when the old item was studied. Amongst new items, A' and B' items have an equal mean match, due to their similarity to the memory trace of their pair mate (A and B respectively), whereas X' items have a lower match, because they are not specifically similar to the memory traces of any particular test item.

$$d'(A\text{—}A') = d'(A\text{—}B') < d'(A\text{—}X') \quad (2)$$

Dobbins et al.'s (1998) model

Dobbins et al. (1998) proposed an explanation of their RK findings based on the dual-process theories of Jacoby (1991) and Yonelinas (1994). Know responses are assumed to be based on a continuous familiarity dimension. Remember responses are the result of a discrete state of awareness, although Dobbins et al. suggest, following Donaldson (1996), that this discrete state might be the result of a threshold placed on an underlying continuous recollection process. Jacoby's emphasis on the importance of recollection in strategic responding motivated the idea that participants make a remember response if *only one* test item causes recollection (i.e., when the result of recollection is unambiguous). A know response and confidence rating based on familiarity are made if *neither* or *both* test items cause recollection. That is, recollection is disregarded altogether if it is conflicting (i.e., if both test items produce a recollection when the participant knows only one was studied).

Dobbins et al. (1998) suggested that choice-similarity causes an increase in the probability of recollection, both for target (old) test items (p_T) and unstudied (new) test items (p_L), which in turn results in a decrease of the probability of a remember response, due to an increased incidence of conflicting recollections. They also suggested that false recollection does not occur in the *A*–*X*′ condition because *X*′ items are not specifically similar to any studied item. To account for the small but non-zero incidence of false recollection in the *A*–*X*′ condition, they introduced a guessing parameter into their model. Guessing was assumed to occur equally often in each experimental condition and to result in responses that are equally likely to be correct or incorrect and equally likely to have any level of confidence.

Dobbins et al. (1998) considered the idea that the probability of a decision based on either correct recollection (p_C), or false recollection (p_F), can be derived from the statistically independent probabilities of recollection for each test item (i.e., p_T and p_L). However, they rejected this idea because they did not think independence between target and new test items was plausible when choice-similarity is high. In light of this conclusion, we chose to directly estimate the probabilities that recognition decisions were based on either correct recollection (p_C) or false recollection (p_F), rather than make any assumptions about how these probabilities are related to individual target and new item recollection probabilities. The cost of this strategy is a proliferation in the number of estimated parameters, raising the possibility that the DPRK model might "over-fit" the data. Over-fitting occurs when a model provides a good fit by accommodating not only the systematic structure in the data but also the effects of measurement error. Over-fitting has the important practical consequence of poor prediction of new data, as measurement error will be different in new data.

One method of addressing over-fitting is to incorporate parameter constraints based on a model's psychological interpretation. Although Dobbins et al.'s (1998) model lacks the precise constraints discussed previously for Clark's (1997) model, several possibilities are arguable. First, as already discussed, decisions based on false recollection may be unlikely in the *A*–*X*′ condition, and so p_F can be fixed at zero in this condition. Second, if high choice-similarity causes false recollection for the new test item to always be accompanied by a correct recollection for the old test item, no recognition decisions will be based on false recollection in the *A*–*A*′ condition ($p_F = 0$). Together, these constraints suggest that only the *A*–*B*′ condition should have non-zero estimates of p_F:

$$0 = p_F(A\text{—}X') = p_F(A\text{—}A') < p_F(A\text{—}B') \tag{3}$$

The second set of constraint relates to the distribution of difference in familiarity between old and new items. The familiarity difference distribution is assumed to be the basis for know decisions, in much the same way that match difference distributions are the basis of responding in Clark's (1997) model. Dobbins et al. (1998) pointed out that their finding of equal accuracy for know judgments may be explained if "targets and distracters are assumed to have equal strength" (p. 1312). If the familiarity difference distributions for the *A*–*A*′ and *A*–*B*′ conditions differ in variance, as predicted by Clark's (1997) model, this explanation would not hold. However, it does hold for an equal variance model, which is also consistent with the assumption made about familiarity in Yonelinas's (1994) dual-process model. Hence, familiarity differences in the *A*–*A*′ and *A*–*B*′ conditions might be modeled by Gaussian distributions with the same mean and variance.

Given these assumptions it is also plausible that the familiarity difference distribution in the *A*–*X*′ condition has the same variance as the other conditions. However, it will have a higher mean, as the new (*X*′) items in this condition should be less familiar than the new items in the other conditions because they are not specifically similar to any studied item. Stated formally in terms of the mean (d') and standard deviation (s) of the familiarity difference distribution, these constraints are:

$$d'(A\text{—}A') = d'(A\text{—}B') < d'(A\text{—}X') \tag{4}$$

$$s(A\text{—}A') = s(A\text{—}B') = s(A\text{—}X') \tag{5}$$

In order to fix the scale for familiarity the usual convention (without loss of generality) is to fix $s = 1$ in one condition. Given (5) this implies $s = 1$ in all conditions.

The DPRK model

Fig. 4 displays the participant averages of the 72 response proportions (i.e., 12 response types in each of 6 experimental conditions) to which models were fit. If we adopt Yonelinas's (1994) assumption that recollection-based responding always results in the highest confidence rating, lower confidence remember responses will occur only due to guessing. The result is a constant remember response probability for lower confidence responses. As is evident in Fig. 4, this model's fit will be very poor as the frequency of correct medium (i.e., rating 2) confidence recollection responses is quite high.

To avoid these poor fits, we elaborated Dobbins et al.'s (1998) model by assuming that, when it occurs, recollection has a continuously varying strength. Note that, in this view (originally suggested as a possibility by Yonelinas, 1994, p. 1352), recollection is still a threshold process in that it can sometimes completely fail to provide any information (see Parks & Yonelinas, 2009, for recent evidence on this issue). Hence, as pointed out by Dobbins et al. (their Footnote 2), even if the strength of successful recollection is continuous, the success or failure of recollection still results in distinguishable cognitive states, which can be used determine what information contributes to a recognition decision. Hence, our elaboration remains consistent with Dobbin's et al.'s model of the choice-similarity effect, and differs from other proposals that recollection is not a threshold process but rather always provides continuously varying mnemonic information (e.g., Wixted, 2007).

In particular, the DPRK model assumes recollection, when it occurs, is characterized by separate, but equal

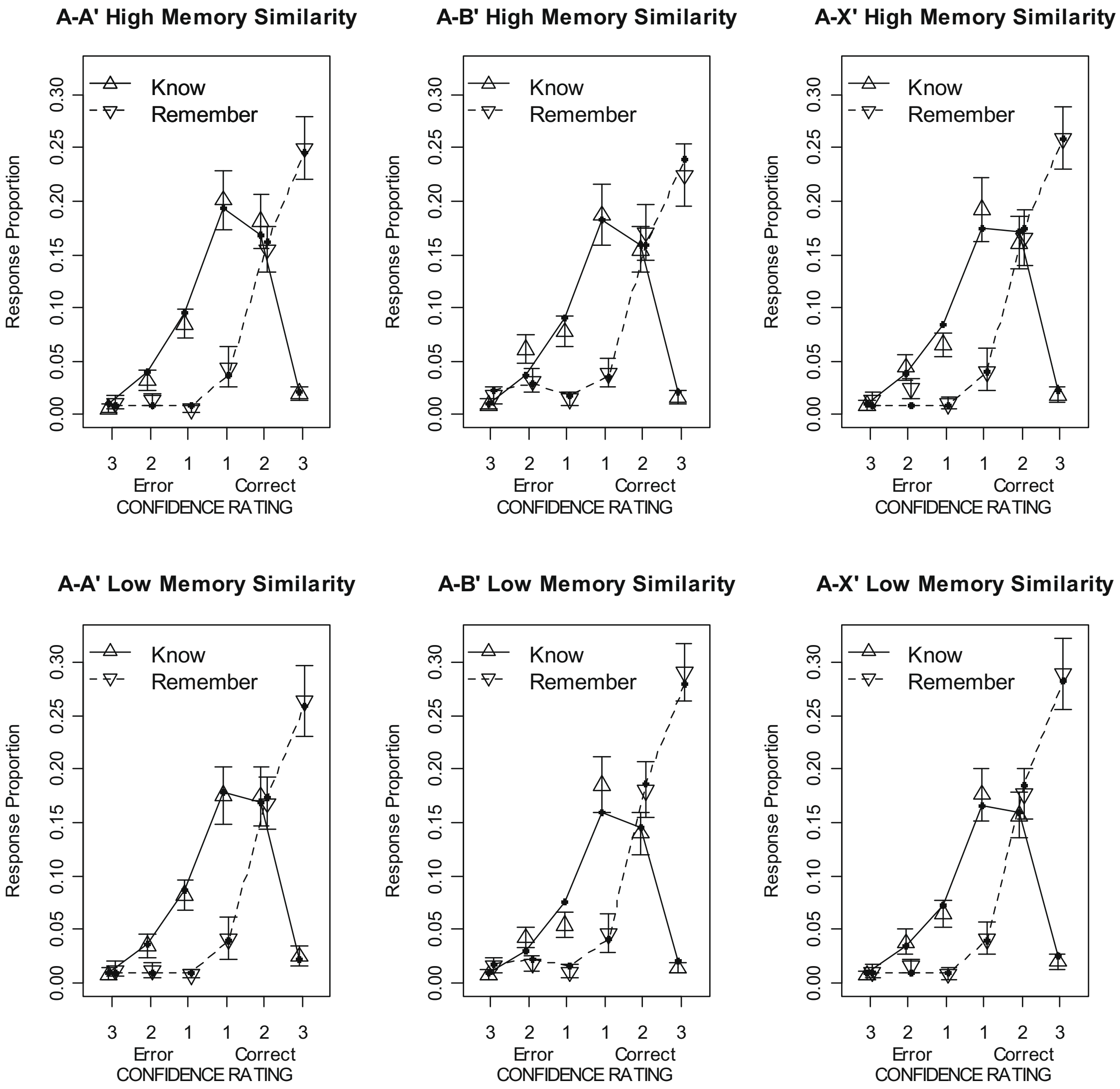


Fig. 4. Average proportions of remember and know responses (triangle symbols) with 95% confidence intervals and average fits (lines with solid points) of the 17 parameter DPRK model (see Table 2). Each panel gives results for one of the six experimental conditions. The x-axes specify rating of confidence, ranging from high (3) to low (1) for correct and error responses, with know results displaced slightly to the left and remember to the right in order to avoid overlap. Confidence intervals were calculated using 100,000 bootstrap replications (Efron & Tibshirani, 1993). Each replication was the average over participants of samples from binomial distributions for each participant, $B(p, n)$, where p and n are the observed proportion and number of trials for each participant.

variance, Gaussian distributions of true and false recollection strength. Recollection strength determines recollection confidence in the manner of signal detection theory, using the same criteria as for familiarity based confidence judgments. We note that, formally, Yonelinas's (1994) model is a special case of the DPRK model, occurring when the recollection means are large, in which case the entire recollection distribution falls above the upper confidence criterion, and so all recollection based decisions are always made with high confidence.

Fig. 5 provides a schematic illustration of the DPRK model. On each trial, participants attempt recollection for both test items. Recollection is either unsuccessful or successful, and if successful it results in recollection strength values with a unit variance Gaussian distribution with mean m_C for correct recollection (Fig. 5a) and mean m_F for false recollection (Fig. 5b). Fig. 5 illustrates the case, which we found to apply to our data, where correct recollection strength is on average greater than false recollection strength (i.e., $m_C > m_F$). If both or neither test item causes recollection the recognition decision and confidence are determined by the difference in their familiarities, and a know response is made. The DPRK model assumes that the familiarity difference distribution has a unit variance Gaussian distribution with mean d'_F (Fig. 5c). As in Clark's (1997) model, positive values indicate a correct response and confidence criteria are symmetric around zero.

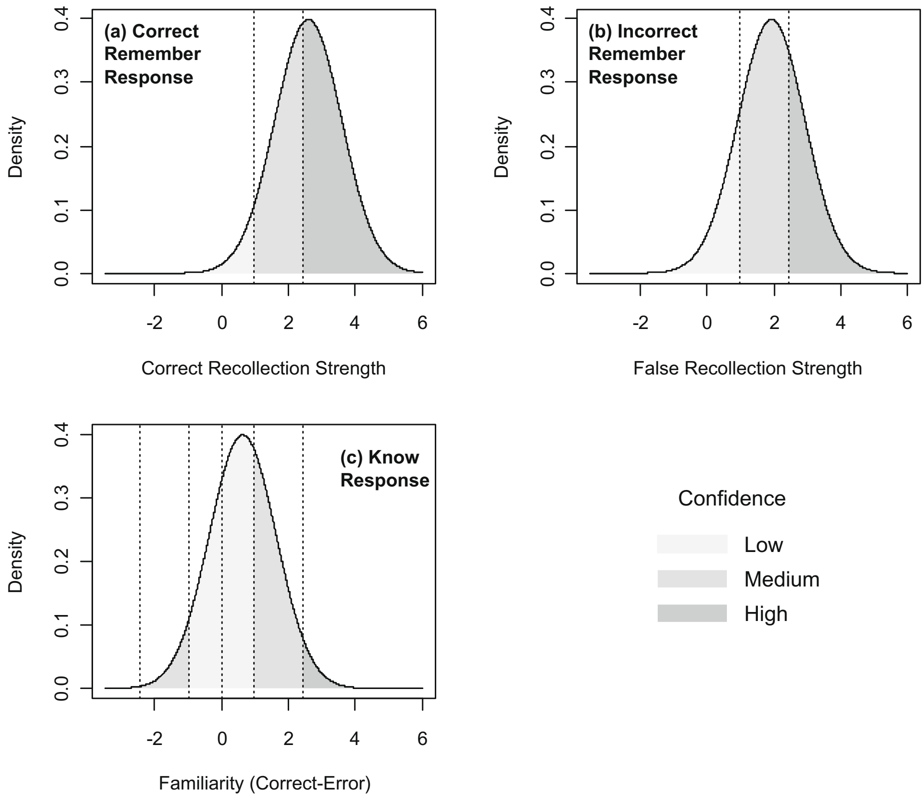


Fig. 5. The three strength distributions constituting the DPRK model, for (a) correct and (b) false recollection and (c) correct and false know responses. The distribution in (a) has mean m_C and the distribution in (b) has mean m_F. Proportions of each type of confidence response are shown by the shaded areas under each distribution.

Fig. 5 shows the proportions of confidence ratings of each type by shaded areas under the distribution functions, with heavier shading indicating greater confidence. The areas of the six shaded regions in Fig. 5c determine the proportions of correct and false know responses at each level of confidence. The same positive criteria used to determine familiarity based confidence are also used to determine recollection confidence. The areas of the three shaded regions in Fig. 5a determine the proportions of different confidence responses for correct recollections, and the areas of the three shaded regions in Fig. 5b determine the proportions of different confidence responses for false recollections.

We fit the DPRK model, subject to the equality constraints in (3)–(5), to each participant's data separately. We used the degree to which the resulting model's parameters respected the inequality constraints in (3)–(5) as a test of the plausibility of the model. The fit of this 17 parameter model, given the data have 66 degrees of freedom, was excellent, and it captures all of the systematic trends evident in Figs. 1 and 2. Table 2 gives the median parameter estimates over participants, and also the parameter estimates for a fit of this model to the data aggregated over participants. The model has a guessing parameter and six correct remember probability parameters (p_C), one for each experimental condition. It has two false recollection probability parameters (p_F), for *A*–*B*′ choices in the low and high similarity conditions. As predicted by the inequality in (3), these estimates, although small, are greater than zero.

Four parameters specify mean familiarity differences, one for the *A*–*A*′ and *A*–*B*′ conditions, and one for the *A*–*X*′ condition, with separate estimates for the low and high memory-similarity conditions. As predicted by the inequality in (4), the median estimates in Table 2 were greater in the *A*–*X*′ condition than in the other conditions, although this was true for only a small majority (55%) of individual participants. A further two parameters specify mean false and correct recollection strengths, and the remaining two parameters are the criteria dividing low and medium confidence responses and medium and high confidence responses. The lower mean for the false recollection distribution allows the model to accommodate the slightly elevated rates of medium (i.e., rating 2) confidence false recollection responses in the *A*–*B*′ condition evident in Fig. 4.

Note that the means of the recollection strength distributions, although much larger than the means of the familiarity distributions, are inconsistent with Yonelinas's (1994) model in which recollection always results in a high confidence response. In particular, the estimates in Table 2 indicate that high confidence responses are given for only around 25% of false recollections and 65% of correct recollections.

Table 2
Median parameter estimates from fits of the 17 parameter DPRK to participant data, with parameter estimates from fits to aggregated data in brackets. Note that the zero false remember probability values in the table are fixed and not estimated.

	Higher memory-similarity			Lower memory-similarity		
	A–*A*′	*A*–*B*′	*A*–*X*′	*A*–*A*′	*A*–*B*′	*A*–*X*′
Correct remember probability	.46 (.47)	.47 (.46)	.51 (.50)	.50 (.50)	.53 (.55)	.54 (.54)
False remember probability	0	.031 (.039)	0	0	.015 (.016)	0
Mean familiarity difference	.64 (.63)		.71 (.67)	.75 (.69)		.83 (.71)
Mean recollection strength	False = 2.1(1.9) correct = 3.3(2.9)			Guessing probability = .04 (.04)		
Confidence criteria	Low/medium = 1.02 (.97)			Medium/high = 2.97 (2.44)		

The SPRK model

We extended Clark's (1997) theory to address our RK data by adopting both Donaldson's (1996) idea that RK judgments are based on a criterion placed on the same dimension used to make confidence judgments (i.e., the match difference distribution in Clark's model), and Wixted and Stretch's (2004) idea that this RK criterion varies from trial to trail. The resulting SPRK model assumes that RK criterion has a normal distribution, $N(c, \sigma^2)$, with mean c and standard deviation σ. Hence, the SPRK model differs from Clark's model depicted in Fig. 3 through the addition of a single noisy RK criterion. We describe the technical details of how response probability predictions were derived for the SPRK model in Appendix to this paper.

We fit a version of the SPRK model incorporating the equality constraints in (1) and (2) to each participant's data separately. The fit of this 12 parameter model was excellent, as shown in Fig. 6. Table 3 gives the median parameter estimates over participants, and also the parameter estimates for a fit of this model to the data aggregated over participants. Four parameters determined mean memory strength (d'), one for both the *A*–*A*′ and *A*–*B*′ conditions and one for the *A*–*X*′ condition for each level of memory-similarity. As predicted by the inequality in (2), the median estimates of d' in Table 3 were greater in the *A*–*X*′ condition than the *A*–*A*′ and *A*–*B*′ conditions, and this was also true for 63% of individual participants.

Four parameters determine the standard deviation of memory strength (s): one for both the *A*–*B*′ and *A*–*X*′ conditions and one for the *A*–*A*′ condition for each level of memory-similarity. Note that only three of these parameters are estimated to fit the data as (without loss of generality) the high memory-similarity *A*–*A*′ standard deviation was set to one to fix the scale on which memory strength is measured. As predicted by the inequality in (1), the median estimates of s in Table 3 were less in the *A*–*A*′ condition than in the *A*–*B*′ and *A*–*X*′ conditions, and this was also true for 62% of individual participants.

Two parameters determine confidence criteria, and one determines the RK criterion, with a further parameter for the standard deviation (*SD*) of the RK criterion. The latter parameter, in Table 3, indicates that the level of RK criterion variability is only 20% (i.e., $.45^2$) of the level of memory strength variability. Consequently the correlation between memory strength and the RK decision variable is quite high ($r \approx .9$). Finally, the SPRK model incorporates the same type of guessing parameter as the DPRK model, with both models estimating a similar low level of guessing.

Model selection

We tested the equality assumptions (1)–(5) underlying the versions of the DPRK and SPRK models just discussed by comparing these models with alternative DPRK and SPRK models that made different assumptions. In this section we describe the alternative models and the methods by which we selected the best model. Our model selection methodology is largely taken from Busemeyer and Wang (2000) and Wasserman (2000). The reader is referred to these papers for details, but to make our treatment self-contained we provide a summery here.

The first step in model selection is to determine the degree to which a model misfits the data. The deviance statistic is the appropriate misfit measure for our data, which consists of the number of responses (n) in each of the $i = 1\ldots6$ experimental conditions (*A*–*A*′, *A*–*B*′ and *A*–*X*′ for both low and high memory-similarity) and $j = 1\ldots12$ response types (correct/error × remember/know × low/medium/high confidence):

$$D = -2\sum_{i=1}^{6}\sum_{j=1}^{12}\log\left(\Pr_{ij}(n_{ij}|\boldsymbol{\theta})\right) \tag{6}$$

$\Pr_{ij}()$ is the probability of a response in condition i of type j given by a model with parameter vector $\boldsymbol{\theta}$ of length p, where $\boldsymbol{\theta}$ is chosen to minimize the deviance (i.e., maximum-likelihood estimation). Deviance has an asymptotically (i.e., in the large sample limit) χ^2 distribution, with larger values indicating greater misfit.

All models were fit by searching for parameter values that minimize the model's deviance. However, simply picking the model with the smallest deviance is misleading. For example, when the models being compared are nested (i.e., one model is a simpler version of the other because of restrictions on its parameter values) the more complex model always has a lesser or equal deviance compared to the simpler model. This occurs even when the simpler model is the true (i.e., data generating) model. One approach to these problems is to select a more complex model with p estimated parameters over a simpler nested model with q estimated parameters ($q < p$) if the decrease in deviance is significant relative to a $\chi^2(p - q)$ distribution. This approach is limited to the comparison of nested models, and, in any case, is known to pick overly simple models when sample size, and hence power, is low. Conversely, when sample size is large it picks overly complex models.

An alternative approach, which is not limited to nested models, is to pick the model which minimizes an information criterion. The most commonly used types are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). Both add to the deviance a penalty for complexity that is proportional to the number of estimated parameters (i.e., models with a larger number of parameters suffer a larger penalty). We focus here on BIC, as our conclusions were the same when we used AIC.[2] Given a model fit to a total of N responses ($N = 72$ for our data), BIC imposes the following penalty BIC:

$$P \times \ln(N) \tag{7}$$

BIC is designed to select the most probable model amongst a set of models. For a set of models $i = 1 \ldots K$ with BIC values BIC_i, the probability that a given model is the true model is:

$$p_i = \frac{e^{-\text{BIC}_i/2}}{\sum_{i=1}^{K} e^{-\text{BIC}/2}} \tag{8}$$

We describe this probability as p_{BIC} and use it to test observed differences in BIC.

DPRK models

Fig. 7 compares three versions of the DPRK model with the ordinate giving their deviance and BIC. Different models are indicated by labels on the abscissa corresponding to their number of estimated parameters. Each bar represents average participant results for one version of a model. The dark gray portion of the bar indicates deviance and the light gray portion of the bar represents the complexity penalty imposed by BIC. When comparing models the best fitting model has the lowest dark gray section and the best model according to BIC has the lowest bar overall.

Fig. 7 provides results for three versions of the DPRK model. The most complex version has 21 parameters, as it imposes only the two most basic restrictions on parameters as a function of choice-similarity: the equal variance assumption (5) and a zero false recollection probability in the *A*–*X*′ condition, as suggested by Dobbins et al. (1998). We examined even more complex models relaxing these constraints, but none reduced misfit much and all led to large increases in AIC and BIC. The 21 parameter model has 4 extra parameters relative to those shown in Table 2, two extra false recollection parameters for the high and low memory-similarity *A*–*A*′ conditions and two extra d′ parameters allowing the *A*–*A*′ and *A*–*B*′ conditions to have unequal mean familiarity differences for the low and high memory-similarity conditions.

The second version of the DPRK model, with results depicted in Fig. 7, has 19 parameters because it enforces the equality on d' between *A*–*A*′ and *A*–*B*′ conditions specified in (4). Finally, by adding the equality in (3), which sets false recollection in the *A*–*A*′ condition to zero, we arrive at the 17 parameter model considered in detail earlier. Fig. 7 shows that the extra constraints increase misfit but to a much lesser degree than the accompanying reductions in complexity penalties, so BIC chose the 17 parameter model. In particular, BIC for the 19 parameter model relative to the 21 parameter model improved for almost every participant (58/61). The overall probability of the 19 parameter model relative to the 21 parameter model was $p_{\text{BIC}} > .99$.

These results support the equality constraint specified in (4), and hence the assumption that the difference between new and old test item familiarity is the same in the *A*–*A*′ and *A*–*B*′ conditions. BIC for the 17 parameter model relative to the 19 parameter model also improves for almost every participant (60/61). The overall probability of the 17 parameter model relative to the 19 parameter model was $p_{\text{BIC}} > .99$. These results support the equality in (3), and hence the assumption that decisions in the *A*–*A*′ condition are never based on false recollection.

SPRK models

Fig. 7 also compares three versions of the SPRK model. Relative to the 12 parameter SPRK model discussed previously, the 17 parameter version relaxes the equality constraints (1) and (2) predicted by Clark's (1997) model, allowing different mean evidence parameters in the *A*–*A*′ and *A*–*B*′ conditions and different evidence variance in the *A*–*B*′ and *A*–*X*′ conditions. It also relaxes the assumption that only the RK criterion is variable by allowing a parameter[3] for variability in the criteria used to make the recognition and confidence decisions (see Benjamin, Diaz, & Wee, 2009, for further discussion of variability in criteria in item recognition tasks). The 13 parameter version imposes the equalities in (1) and (2) and the 12 parameter version also imposes the assumption that only the RK criterion is variable (i.e., it is the model with parameter estimates shown in Table 3).

Fig. 7 provides strong support for the equality assumptions predicted by Clark's (1997) model. BIC for the 13 parameter model, relative to the 17 parameter model, decreased for every one of the 61 participants, with an overall probability of $p_{\text{BIC}} > .99$. Fig. 7 also provides support for the assumption that only the RK criterion is variable. BIC for the 12 parameter model relative to the 13 parameter model, decreased for most participants (57/61), with an overall probability of $p_{\text{BIC}} > .99$.

Alternative models

Although we chose the simplest of the DPRK models, it still has five more parameters than the best SPRK model. Is this extra complexity in the dual-process model required? We tried reducing the complexity of the DPRK model by equating parameters which are similar in Table 2. The best

[2] AIC generally prefers more complex models than BIC. It might be thought that our conclusions, which generally preferred simpler models (e.g., SPRK) were due to this property of BIC, so that fact that results based on AIC (or more correctly, the small-sample adjusted version of AIC, see Wagenmakers & Farrell, 2004) were consistent with those based on BIC is reassuring. We also obtained consistent results when we made different assumptions about the value of N in the equation for BIC (e.g., that it equalled the total number of binary responses in a data set rather than the number of response proportions, which aggregates over binary responses in each condition).

[3] Only a single parameter is required on the assumption that all confidence criteria are perturbed equally by the sample of criterion noise (see Mueller and Weidemann (2008), for discussion of models in which criterion variability differs for different confidence criteria).

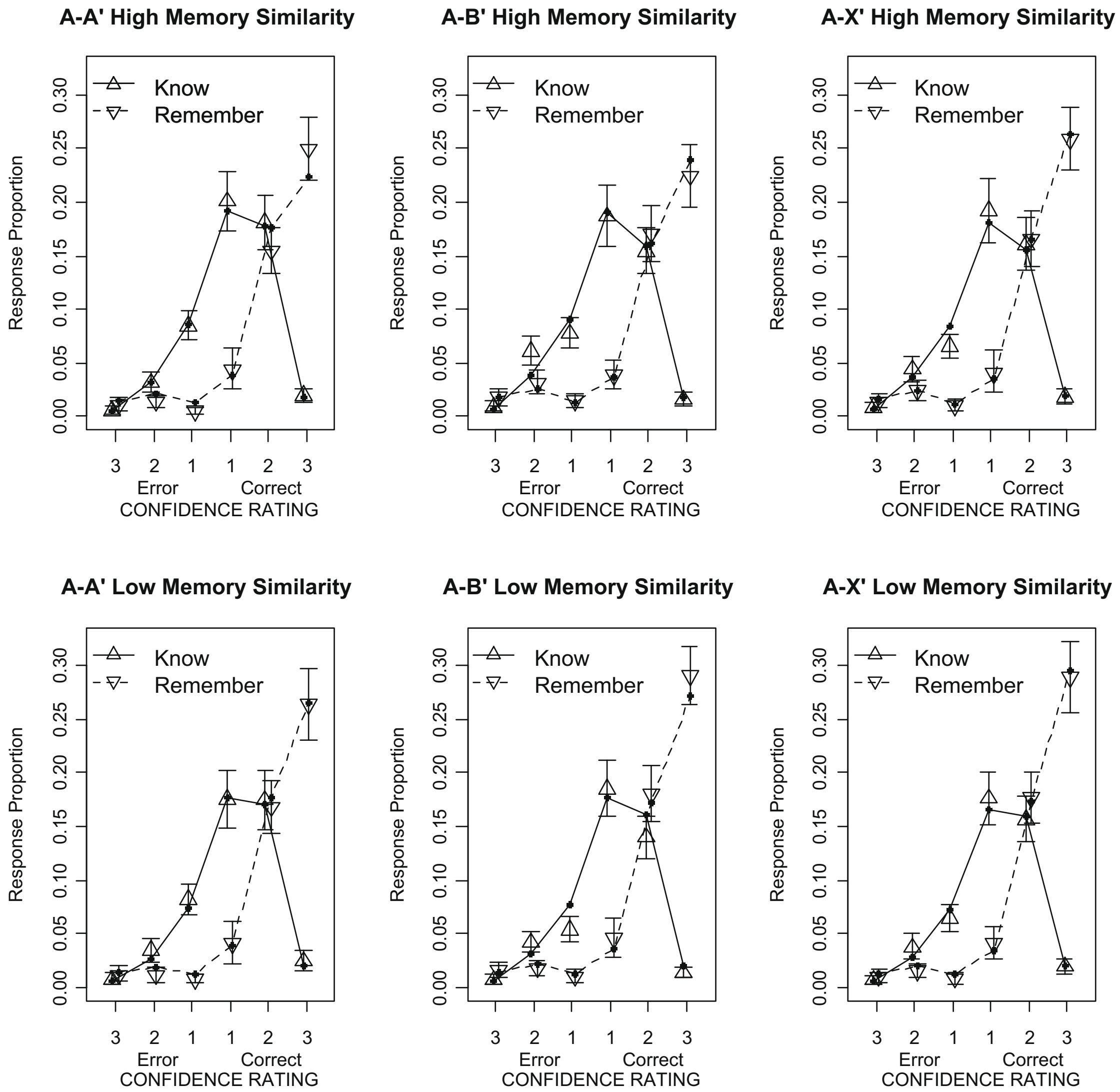


Fig. 6. Average proportions of remember and know responses (triangle symbols) with 95% confidence intervals and average fits (lines with solid points) of the 12 parameter SPRK model (see Table 3). Each panel gives results for one of the six experimental conditions (MS = memory-similarity). The legends apply to all panels. The *x*-axes specify rating of confidence, ranging from high (3) to low (1) for correct and error responses, with know results displaced slightly to the left and remember to the right in order to avoid overlap. Confidence intervals were calculated using 100,000 bootstrap replications (Efron & Tibshirani, 1993). Each replication was the average over participants of samples from binomial distributions for each participant, *B*(*p*, *n*), where *p* and *n* are the observed proportion and number of trials for each participant.

Table 3
Median parameter estimates from fits of SPRK to participant data, with parameter estimates from fits to aggregated data in brackets. *s*(*A*–*A*′) = 1 in the higher memory-similarity condition by assumption.

	d′		*s*	
	A–*A*′ and *A*–*B*′	*A*–*X*′	*A*–*A*′	*A*–*B*′ and *A*–*X*′
Higher memory-similarity	1.05 (1.03)	1.10 (1.11)	1	1.09 (1.13)
Lower memory-similarity	1.18 (1.18)	1.26 (1.22)	.99 (1.03)	1.09 (1.08)
Confidence criteria	Low/medium = .73 (.72)		Medium/high = 1.79 (1.70)	
RK criterion	Mean = 1.17 (1.16)	*SD* = .45 (.62)	Guessing probability = .03 (.09)	

resulting model, which had the same number of parameters as the SPRK model (12), assumed: (a) the same mean familiarity for all conditions, (b) the same mean for correct and false recollection strength, and (c) a zero remember probability for lower memory-similarity pairs in the *A*–*B*′ condition. As might be expected by the opportunistic

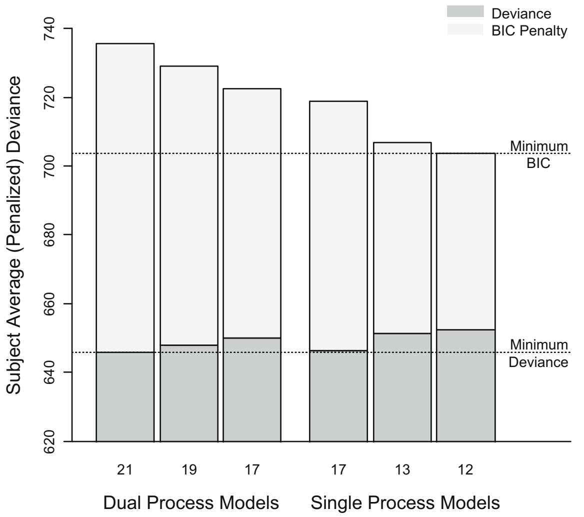


Fig. 7. Model selection results. Each bar represents average participant results for one version of a model with the type of the model (DPRK or SPRK) and its number of parameters indicated on the *x*-axis label. The dark gray portion of each bar indicates deviance and the light gray portion of the bar represents the complexity penalty imposed by BIC. Horizontal dotted lines indicate the minimum values of each statistic over all models.

way in which these reductions were made, they were supported by BIC ($p_{BIC} > .99$, 61/61).

Even with this simplified model the 12 parameter SPRK model was still preferred ($p_{BIC} > .99$, 40/61), indicating that it fit better than the 12 parameter DPRK model. In order to make a fairer comparison, we also applied the same approach to the SPRK model. As shown in Table 3, memory-similarity had virtually no effect on the *s* parameter, suggesting a simplification from 12 to 10 parameters. The resulting model was preferred according to BIC relative to both the 12 parameter SPRK model ($p_{BIC} > .99$, 59/61) and the 12 parameter DPRK model ($p_{BIC} > .99$, 61/61).

We also made a more wide ranging search for better DPRK models, including simplifications suggested by reviewers (e.g., equating correct remember probability across choice-similarity conditions), and models that allowed unequal familiarity and/or recollection strength variance while simplifying other aspects of the model, but were unsuccessful in finding any alternative model that reduced BIC. Although we do not claim that our search was exhaustive, we were convinced after many unsuccessful attempts that the results presented here are at least close to the best alternative within our DPRK framework. We also agree with reviewers of this paper who rightly cautioned us against the validity of this ad hoc approach to model selection, which is likely to capitalize on chance variation in our results. Hence, in the following comparison of DPRK and SPRK models we focus on the principled alternatives described in the last two sections.

DPRK vs. SPRK

Fig. 7 also enables a comparison of the DPRK and SPRK models. The 21 parameter DPRK model provides the best fit (i.e., has minimum deviance) overall, although it is only very marginally better than the 17 parameter SPRK model. In contrast, the 12 parameter SPRK model provides the best account according to BIC. The preference for the SPRK models is very clear; every SPRK model, even the most complex, is preferred to any DPRK model, even the least complex.

The preference for the SPRK model indicated by Fig. 7 relies on the appropriateness of the penalties applied for extra model parameters. However BIC does not take account of *functional form complexity*, that is, model flexibility differences that arise because of differences in the way that model equations combine parameters and data (the same is true for AIC). Pitt and Myung (2002) discuss minimum description length and Bayesian methods of addressing this limitation. As we found these methods difficult to apply to the models considered here, due to their relatively large number of parameters, we used a different approach to this issue proposed by Busemeyer and Wang (2000). This approach selects a model based on an intuitively plausible criterion, the generalizability of a model's data-fitting abilities.

Busemeyer and Wang's (2000) generalization criterion involves obtaining parameter estimates by fitting a model to one set of data (the calibration sample) and then measuring the misfit of predictions based on those parameters for data from a new experimental design (the validation sample). This approach differs from cross-validation, which as mentioned previously is asymptotically equivalent to selection by AIC, where the validation sample is drawn from the same experimental design. Busemeyer and Wang's validation sample came from an extrapolation design, which measures data for quantitative independent variable values outside the range examined in the validation sample (see also Wagenmakers, Grunwald, & Steyvers, 2006).

Here we use a validation sample drawn from an interpolated design, which used all but the *A–X′* condition of the present experiment and a subset of the same face images (Heathcote et al., 2009). Apart from the omission of the *A–X′* condition and measuring a different group of participants, this design was new in the sense that it used study lists with mixed race (but same gender) faces, and created a low choice-similarity condition by testing faces from different races rather than different genders. In particular, the validation data comes from the 38 of the 99 participants tested by Heathcote et al. who made RK decisions as well as making 2AFC recognition decisions with confidence ratings.

For each participant's data in the validation sample we calculated the average of the deviances obtained using the parameters estimated for each of the 61 participants in the present experiment, omitting parameters specific to the *A–X′* condition. Fig. 8 presents results averaged over the 38 participants in the validation sample for each of the model's examined in Fig. 7. Deviance results for the validation sample are in the opposite order to deviance for the calibration sample. In contrast, they are in agreement with the order of the BIC results, demonstrating the utility of this criterion in selecting models.

In all but one case, the differences between models in validation deviance are much larger than differences in

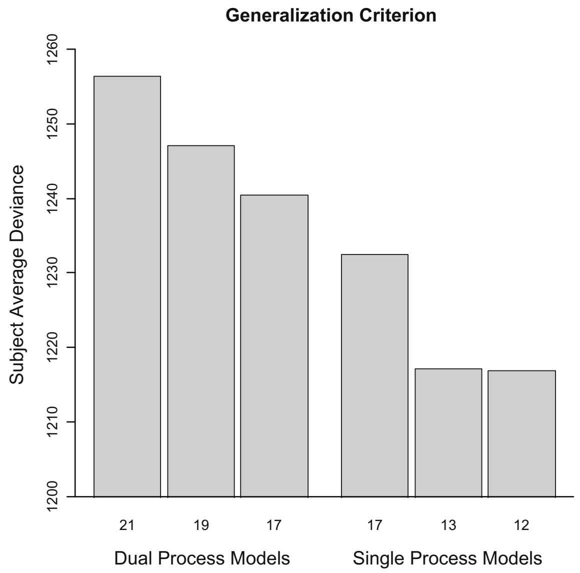


Fig. 8. Generalization-criterion results based on misfit to data from Heathcote et al.'s (2009) participants who gave RK responses for predictions made by the models and associated parameter estimates fit to the present data.

calibration deviance and similar to BIC differences. The one exception concerns the comparison between the two simplest SPRK models, where the 12 parameter model does not differ much from the 13 parameter model, although its generalization deviance is numerically smaller (1216.9 vs. 1217.6). Hence, the generalization-criterion results confirm the superiority of the SPRK models to DPRK models as a class, the constraints for the DPRK model, and the constraint derived from Clark (1997) for the SPRK model, but are equivocal on whether only the RK criterion is variable (i.e., the 12 parameter model) or whether confidence criteria are also variable (i.e., the 13 parameter model).

Up to now we have not addressed the model parameter differences that explain the memory-similarity effect. In part this was because the models are not bound to make strong predictions, as memory-similarity is a between item manipulation (i.e., different items make up high and low memory-similarity sets). As a result, outcomes may be affected by item related factors other than memory-similarity (e.g., items in the low similarity set may be more distinctive and so more likely to support correct recollection). However, one aspect deserves mention because, for both models, it converges with Heathcote et al.'s (2009) state-trace findings. Choice-similarity effects in the SPRK model are mediated by a parameter that is little affected by memory-similarity, that is, evidence variance (see Table 3), consistent with a dissociation between the underlying causes of these effects. For the DPRK model false recollection could fulfill a similar role to evidence variance, although the dissociation is less clear cut.

Memory for scenes

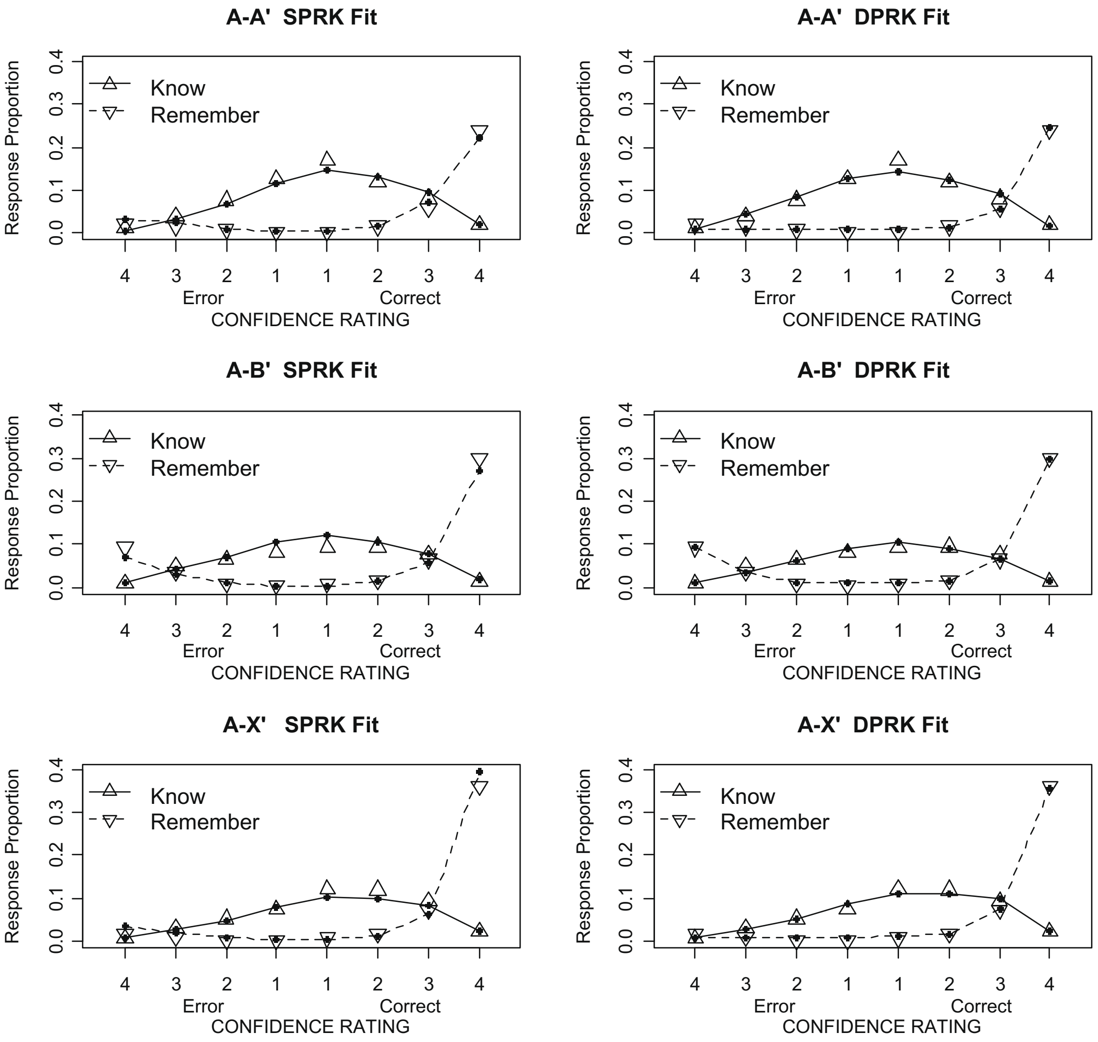
In order to investigate the generality of our findings with faces, we applied our model based analysis to Dobbins et al.'s (1998) data on memory for scenic stimuli. The focus of this analysis was to determine whether the best DPRK and SPRK models could provide an accurate account of the scene data, and whether the corresponding parameter estimates were consistent with our findings for faces. Neither could be taken for granted in light of some marked differences in our findings, such as stronger effects of choice-similarity on know confidence for the scene stimuli. The design of Dobbins et al.'s experiment differed from ours in that memory-similarity was not manipulated, and participants gave a four-level confidence judgment. Data aggregated over participants were taken from their Table 3 (p. 1311) and are plotted in Fig. 9.

Fig. 9 illustrates that, although a similar pattern to our data is evident, with less accurate know responses accompanied by low and medium confidence ratings and more accurate remember responses accompanied by higher confidence ratings, there are also clear differences. The proportion of remember responses at the highest confidence rating was greater, and a substantial proportion of false remember responses at the highest confidence are evident in the *A*–*B*′ condition. In the DPRK model the latter finding is consistent with false recollection unchecked by correct recollection, but it is also consistent with a less variable RK criterion in the SPRK model when combined with greater evidence variability and low accuracy in the *A*–*B*′ condition.

In the left-hand column, Fig. 9 shows the fit of the version of the SPRK model that had 12 parameters for our design. For Dobbins et al.'s (1998) design, this model has 9 parameters, with estimates shown in Table 4. Low accuracy in the *A*–*B*′ condition and higher accuracy in the *A*–*X*′ condition were attributed to larger differences between the *d*′ and *s* parameter estimates than in our data. Guessing was at approximately the same level, and, as expected, the RK criterion was less variable than in our data. The extra confidence rating was modeled by a very low criterion. Confidence criteria were generally lower relative to our data, with the second criterion approximately equal to our first, and the RK and highest criteria less than the estimates for our data.

The right-hand column of Fig. 9 shows the fit of the version of the DPRK model that had 17 parameters for our design. For Dobbins et al.'s (1998) design this version has 12 parameters with estimates shown in Table 5. As would be expected, the mean estimate was greater for correct than false recollection, recollection probability was greatest in the *A*–*X*′ condition, and least in the *A*–*A*′ condition. Both guessing and the probability of false remember responses were much higher than for our face data, as was the advantage for the *A*–*X*′ condition in the mean familiarity difference.

As might be expected, given its larger number of parameters, misfit for the DPRK model was less than for the SPRK model. However, the difference was only very slight, with a deviance per subject of 426.9 vs. 427.3 (i.e., a .1% difference). Because the data in this case are aggregated over participants, none of the model selection methods, which we used with our data to account for model complexity differences (i.e., BIC and the generalization criterion), were straightforwardly applicable. Hence, the main conclusion


A-A' SPRK Fit

A-A' DPRK Fit

A-B' SPRK Fit

A-B' DPRK Fit

A-X' SPRK Fit

A-X' DPRK Fit

Fig. 9. Average proportions of remember and know responses from Dobbins et al. (1998) and fits of the 9 parameter SPRK model and 12 parameter DPRK model (see Tables 4 and 5). The *x*-axes specify rating of confidence, ranging from high (4) to low (1) for correct and error responses.

Table 4
Parameter estimates for fits of SPRK to Dobbins et al.'s (1998) aggregated data. Note that *s*(*A*–*A*′) = 1 by assumption.

	d′		*s*	
	A–*A*′ and *A*–*B*′	*A*–*X*′	*A*–*A*′	*A*–*B*′ and *A*–*X*′
Evidence	.58	1.03	1	1.28
Confidence criteria	Low/medium = .41	Medium/high = .79	High/highest = 1.26	
RK criterion	Mean = 1.08	*SD* = .34	Guessing probability = .05	

Table 5
Parameter estimates for fits of DPRK to Dobbins et al.'s (1998) aggregated data. Note that false remember probability is fixed at zero in the *A*–*A*′ and *A*–*X*′ conditions.

	A–*A*′	*A*–*B*′	*A*–*X*′
Correct remember probability	.33	.41	.50
False remember probability	0	.13	0
Mean familiarity difference	.24		.50
Mean recollection strength	Correct = 3.28	False = 3.06	Guessing probability = .13
Confidence criteria	Low/medium = .60	Medium/high = 1.24	High/highest = 2.37

to be drawn from these analyses is that the SPRK and DPRK models both provide accurate and parametrically coherent accounts of Dobbins et al.'s (1998) data. Both models also provide a coherent and unified account of memory for faces and memory for scenes in the choice-similarity paradigm. Differences between these cases are accounted for by variations in the models' parameter values that are consistent with the procedural differences between the experiments.

General discussion

The results of our experiments answer empirical questions about recognition memory for faces and our model based analyses answer questions about the underlying psychological causes of these findings. On the empirical front we found that accuracy was better when a recognition choice was between similar (i.e., same gender) faces than when it was between dissimilar (i.e., different gender) faces, where in the latter case the incorrect choice was specifically similar to a studied face which was not tested. This effect was not as large as the corresponding effect that has been found with natural scenes (e.g., Tulving, 1981), but it was large in a relative sense. That is, accuracy for same gender choices was equal to accuracy for a different gender choice when the unstudied test face was not specifically similar to any studied face. In contrast, with natural scenes the latter type of choice was the most accurate.

Dual-process theory

Dual-process theory can account for these findings in terms of false recollection, which occurs when an unstudied face is specifically similar to a studied face, and a strategy to reduce the effects of false recollection. The strategy is applied when both choices cause recollection. When such conflicting recollections occur, participants can strategically use their knowledge that only one face was studied to disregard recollection altogether and base their choice on familiarity. This strategy is particularly effective with similar recognition choices, as for these choices false recollection is accompanied by correct recollection. Hence, accuracy is increased when choices are similar, because the strategy eliminates wrong responses based on false recollection. The same mechanism also explains a decrease in the proportion of responses classified as based on recollection when choices are similar, as measured by Tulving's (1985) remember–know (RK) procedure.

The dual-process model was able to account for confidence results, but only when it was assumed that the strength of recollection is graded. When recollection strength is graded, recollection does not always result in a highly confident recognition decision. Instead, confidence is less when the recollected details are vague. We found that for faces the average strength of false recollection was substantially less than that of correct recollection. Hence, responses based on correct recollection resulted in a greater proportion of high confidence ratings, whereas responses based on a false recollection resulted in a greater proportion of medium confidence ratings.

The same type of dual-process model that provided an accurate account of the confidence–accuracy dissociation with face stimuli also provided an accurate account of the same phenomena with scenic stimuli (Dobbins et al., 1998). Differences in performance for the two types of stimuli were accounted for by differences in model parameter estimates. One difference was that responses based on false recollection were more common for scenes. Another difference was that for scenes mean false recollection strength was not much less than mean correct recollection strength. As a result, high confidence responses based on false recollection were more common for scenes than faces. These results illustrate the utility of a model based analysis; it is able to show that empirically divergent results can be explained by a common underlying mechanism.

Although the way in which we added graded recollection to Dobbins et al.'s (1998) dual-process model provides an accurate, coherent and general account of the confidence–accuracy dissociation, we do not claim that alternative extensions might not also do so. A more parsimonious model might be obtained, for example, if the same graded strength that determines recollection confidence also determines whether recollection occurs (e.g., when strength exceeds a threshold). However, if this was the case, when conflicting recollections occur it would make sense for participants to use the difference in recollection strengths to inform their recognition decision rather than discounting recollection entirely, suggesting a more radical departure from Dobbins et al.'s account. We leave investigation of this and other possibilities, such as Wixted's (2007) suggestion that recollection always occurs to some degree, to future research.

Single-process theory

Single-process theory also provides an accurate account of the confidence–accuracy dissociation caused by choice-similarity. Findings related to recognition accuracy and confidence are explained by Clark's (1997) original model, in which decisions are based on the difference between the match of each test choice to memory. Findings related to RK responses are explained by comparing the match difference to a RK criterion. We found that the RK criterion typically falls between medium and high confidence criteria. Remember responses are less common for choices between faces of the same gender or parts of the same scene for the same reason that confidence is reduced in these cases; the match difference tends to be less extreme in these conditions, and so is less likely to exceed the RK criterion. A high remember criterion also explains why remember confidence does not vary with choice-similarity; most remember responses are made with the highest confidence, and so there is less chance for large confidence differences to occur. Conversely, know confidence is more able to vary, and so differences in know confidence can more easily occur across conditions.

These points all apply to Donaldson's (1996) original explanation of RK performance assuming a constant RK criterion. However, trial to trail variability in the RK criterion, as suggested by Wixted and Stretch (2004), was required to provide a good quantitative account of the joint frequency distribution of RK and confidence responses. In particular, a variable criterion allows some lower confidence responses to be classified as "remember" and some higher confidence responses as "know". Our results were also consistent with some variability in the recognition criteria (see Benjamin et al., 2009; Mueller & Weidemann, 2008, for evidence supporting this type of variability). Although we could not clearly adjudicate between these two possibilities with the methodology used here,[4] the key point is that, in both cases, recognition and remember decisions were highly, but not perfectly, correlated because they are both based on a decision variable that shares a large common component with the recognition decision variable, that is, the memory match difference.

Dobbins et al. (1998, p. 1306) claimed that Clark's (1997) model can not "mimic" remember performance. Clearly our results make it unlikely that the success of our elaborated version of Clark's model is due to mere mimicry by an overly flexible model. If anything, that possibility must be considered as applying to the dual-process model, because it has many more parameters than the single-process model, and is less able to predict new data. However, the idea that the single-process model succeeds through mimicry might have appeal given the common subjective experience that confidence is boosted by recollection in everyday recognition. Correspondingly, in the laboratory, why would participants disregard instructions and base their RK response on test item matches?

An alternative interpretation of our single-process model might provide an explanation. Suppose retrieval of details from memory is occurring, or at least being attempted, as proposed by Gillund and Shiffrin's (1984) SAM model of recognition and recall. In SAM, study creates associations among attended aspects of the internal and external study context, including self-strength, an association of an item to itself. Recognition and recall performance depend on the strength of this web of associations, which increases with study attention. For example, the probability of recognizing a test item depends on its association strength (match), and that of the test context, to memory. The probability of recollecting a memory trace depends in part on the same match, but this match is divided by the sum of matches for all memory traces, and recall only occurs when this ratio exceeds a threshold.

Because recognition and recollection both depend on the level of attention during study, they tend to be correlated; a test item that is easily recognized also tends to be a good cue for recollection. However, the correlation is not perfect, because recollection depends on factors that do not affect recognition, such as the strength of memory traces that compete for recollection. For example, in Mandler's (1980) "butcher-on-the-bus" scenario the butcher is recognized because many previous encounters promote self-strength, but recollection does not immediately occur because the supermarket detail is out-competed by other details that are associated with the bus context. In our experimental scenario, in contrast, there is no such context shift so recognition and recollection tend to be quite highly correlated. As a consequence, our model of remember responses, as being dependent on a decision variable that is highly correlated with the recognition decision variable, provides quite an accurate characterization of performance when participants are basing remember responses on the occurrence of recollection.

In summary, our findings for 2AFC recognition join a growing body of evidence from single item recognition that both the objective and subjective aspects of recognition confidence, and the RK paradigm, are equally well explained, if not better explained, by single-process models than by dual-process models (Dunn, 2004, 2008; Rotello & Macmillan, 2006; Rotello et al., 2006). Given the large body of evidence taken as supporting single- and dual-process theories (e.g., Parks & Yonelinas, 2007; Wixted, 2007) our findings cannot decisively adjudicate between these very different alternatives. However, we have at least provided a clearly specified extension of these theories to the domain of 2AFC recognition, which might form the basis of future development and testing. We also hope that we have illustrated the potential and importance of a comprehensive quantitative modeling approach to these issues, both for providing objective evidence about the psychological processes proposed by each theory, and for providing an objective basis for comparing theories.

Acknowledgments

Thanks to Mark Steyvers for providing the initial set of face pairs taken from the FERET database and sorted into similar pairs by members of his laboratory and to Andy Yonelinas for clarifications about the proposed threshold nature of recollection. Thanks also to Neil Macmillan, Shane Mueller, Neil Dobbins and anonymous reviewers for their guidance. This research was supported by ARC Discovery Project DP0558407.

A. Appendix

In this appendix, we describe the method by which we obtained predicted response probabilities for SPRK models. Given a RK criterion with a normal distribution, $N(c, \sigma^2)$, the probability of a remember response is $\Pr[N(d', s^2) > N(c, \sigma^2)]$, given memory strength with a mean d' and standard deviation s. This remember probability can be expressed in terms of the probability that a random variable incorporating both memory strength and criterion variability is greater than a fixed criterion: $\Pr[N(d', s^2 + \sigma^2) > \mathrm{c}]$. The latter mathematical form assumes independence of memory strength and the remember criterion, but would not be changed in any important way for our purposes here if there was some dependence.

We use these results to obtain the SPRK model's predictions about the recognition decision, confidence, and RK decision using a formally equivalent model where both types of decisions are based on fixed criteria placed on separate normal distributions. The two normal distributions are correlated due to a memory strength component com-

[4] A reviewer suggested that the relative variability of RK and confidence criteria may be investigated by comparing the variability in reaction times for making RK and high vs. low confidence decisions.

mon to both. The correlation equals the total proportion of the RK decision variable's variance ($s^2 + \sigma^2$) that comes from memory strength (s^2): $r^2 = s^2/(s^2 + \sigma^2)$. Predictions for the equivalent model are obtained by integrating a bivariate normal distribution over the appropriate areas as depicted in Fig. 10.

In Fig. 10 the y axis represents memory strength and the x axis represents the random variable on which RK decisions are based, relative to a fixed RK criterion. Consequently the bivariate normal distribution, depicted by equal probability contours in Fig. 10, has a mean of d' for each axis, variances of $s^2 + \sigma^2$ and s^2 for the x and y axes respectively, and a covariance of s^2. The correlation depicted in Fig. 10 is $r \approx .9$, indicating much greater variability in memory strength than the RK criterion. Note that, although the values in Fig. 10 were taken from fits to the high memory-similarity A–A' condition, the figure could represent any condition with appropriate changes in d' and s.

Fig. 10 is divided by thick solid lines into four rectangular regions corresponding to know and remember responses that are either correct or false. Within each rectangle, the areas corresponding to each level of confidence are indicated by shading. As in Clark's (1997) model, the SPRK model assumes that recognition and confidence responses are determined relative to symmetric criteria around zero. Memory strength greater than zero (i.e., cases where the old test item has a greater match than the new test item) results in a correct recognition decision, whereas memory strength less than zero (i.e., cases where the new test item has a greater match than the old test item) results in a false recognition decision. The SPRK model also assumes that RK decisions are based on criteria that are symmetric around zero. Importantly, the RK criteria are the same for all choice and memory-similarity conditions, so differences between conditions can only be explained by Clark's (1997) underlying model of memory strength.

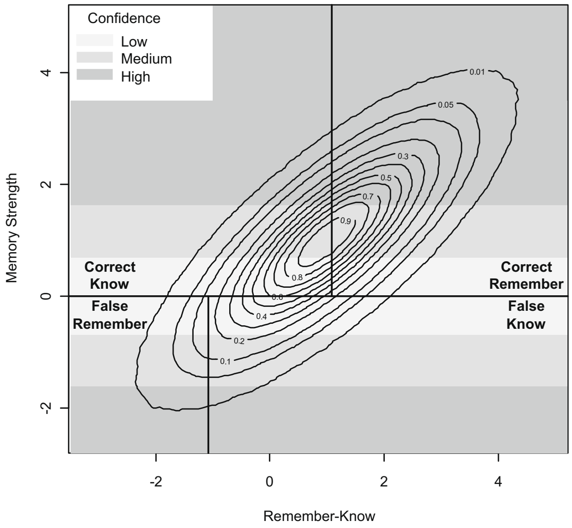


Fig. 10. The bivarite normal distribution that predicts response probabilities for the SPRK model. The distribution is represented by elliptical equal probability contours, with the probability of a sample falling outside the ellipse indicated by numbers on each contour. Rectangles demarcated by the thick solid lines correspond to correct and false remember and know decisions, and shaded areas within each rectangle correspond to different levels of confidence.

The criterion that is used to make the RK decision depends on the preceding recognition decision. When the recognition decision is correct it is based on a positive memory strength value, and values of the RK decision variable greater than the positive RK criterion (indicated by the thick vertical line to the upper right in Fig. 10) produce remember responses. Otherwise a know response is given. Conversely, when the recognition decision is incorrect it is based on a negative memory strength value, and values of the RK decision variable less than the negative RK criterion (indicated by the thick vertical line to the lower left of Fig. 10) produces remember responses. Otherwise a know response is given. Note that the side on which the remember and know regions occurs swaps for correct and incorrect responses because a remember classification is made if the value on the RK dimension is extreme in the same direction as the recognition decision (i.e., positive for a correct recognition decision and negative for an incorrect recognition decision).

It is important to note that Fig. 10 does not depict a two-dimensional signal detection model with different types of mnemonic information contributing to confidence and RK decisions (e.g., Rotello, Macmillan, & Reeder, 2004). Instead, both decisions are based on a common sample of mnemonic information (memory strength) that is used for both the confidence decision and the subsequent RK decision. The common mnemonic information causes a positive correlation between these decisions. However, the correlation is not perfect due to variability in the RK criterion. The bivariate distribution in Fig. 10 formally captures this relationship between the decisions, as well as the variation that affects each individual decision. Hence, integration of the bivariate normal distribution over each of the 12 shaded areas depicted in Fig. 10 produces the SPRK model's predictions for each of the 12 types of possible response.

References

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390–412.

Bamber, D. (1979). State-trace analysis: A method of testing simple theories of causation. *Journal of Mathematical Psychology, 19*, 137–181.

Bates, D. M. (2005). Fitting linear mixed models in R. *R News, 5*, 27–30.

Benjamin, A. (2005). Recognition memory and introspective remember/know judgments: Evidence for the influence of distractor plausibility on "remembering" and a caution about purportedly nonparametric measures. *Memory & Cognition, 33*, 261–269.

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

Busemeyer, J. R., & Wang, Y.-M. (2000). Model comparisons and model selection based on generalization criterion methodology. *Journal of Mathematical Psychology, 44*, 171–189.

Cary, M., & Reder, L. M. (2003). A dual-process account of the list length and strength-based mirror effects in recognition. *Journal of Memory and Language, 49*, 231–248.

Clark, S. E. (1997). A familiarity-based account of confidence–accuracy inversions in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 232–238.

Clark, S. E., & Gronlund, S. D. (1996). Global matching models of recognition memory: How the models match the data. *Psychonomic Bulletin & Review, 3*, 37–60.

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

Donaldson, W. (1996). The role of decision processes in remembering and knowing. *Memory & Cognition, 24*, 523–533.

Dougal, S., & Rotello, C. M. (2007). "Remembering" emotional words is based on response bias, not recollection. *Psychonomic Bulletin & Review, 14*, 423–429.

Dunn, J. C. (2004). Remember–know: A matter of confidence. *Psychological Review, 111*, 524–542.

Dunn, J. C. (2008). The dimensionality of the remember–know task: A state-trace analysis. *Psychological Review, 115*, 426–446.

Dunn, J. C., & Kirsner, K. (1988). Discovering functionally independent mental processes: The principle of reversed association. *Psychological Review, 95*, 91–101.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Gardiner, J. M., Ramponi, C., & Richardson-Klavehn, A. (2002). Recognition memory and decision processes: A meta-analysis of remember, know, and guess responses. *Memory, 10*, 83–98.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91*, 1–67.

Heathcote, A., Freeman, E., Etherington, J., Tonkin, J., & Bora, B. (2009). A dissociation between similarity effects in episodic face recognition. *Psychonomic Bulletin & Review, 16*, 824–831.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review, 95*, 528–551.

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

Kapucu, A., Rotello, C. M., Ready, R. E., & Seidel, K. N. (2008). Response bias in "remembering" emotional stimuli: A new perspective on age differences. *Journal of Experimental Psychology: Learning, Memory and Cognition, 34*, 703–711.

Loftus, G. R., Oberg, M. A., & Dillon, A. M. (2004). Linear theory, dimensional theory, and the face-inversion effect. *Psychological Review, 111*, 835–865.

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

Macmillan, N. A., Rotello, C. M., & Verde, M. F. (2005). On the importance of models in interpreting remember–know experiments. *Memory, 13*, 607–621.

Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review, 87*, 252–271.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. London: Chapman & Hall.

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

Parks, C. M., & Yonelinas, A. P. (2007). Moving beyond pure signal detection models: Comment on Wixted (2007). *Psychological Review, 114*, 188–202.

Parks, C. M., & Yonelinas, A. P. (2009). Evidence for a memory threshold in second-choice recognition memory responses. *Proceedings of the National Academy of Science USA, 106*, 11515–11519.

Phillips, P. J., Wechsler, H., Huang, J., & Rauss, P. (1998). The FERET database and evaluation procedure for face recognition algorithms. *Image and Vision Computing Journal, 16*, 295–306.

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences, 6*, 421–425.

Platt, J. R. (1964). Strong inference. *Science, 4*, 79–95.

Rotello, C. M., & Macmillan, N. A. (2006). Remember–know models as decision strategies in two experimental paradigms. *Journal of Memory and Language, 55*, 479–494.

Rotello, C. M., Macmillan, N. A., Hicks, J. L., & Hautus, M. J. (2006). Interpreting the effects of response bias on remember–know judgments using signal detection and threshold models. *Memory & Cognition, 34*, 1598–1614.

Rotello, C. M., Macmillan, N. A., & Reeder, J. A. (2004). Sum–difference theory of remembering and knowing: A two-dimensional signal detection model. *Psychological Review, 111*, 588–616.

Rotello, C. M., Macmillan, N. A., Reeder, J. A., & Wong, M. (2005). The remember response: Subject to bias, graded, and not a process-pure indicator of recollection. *Psychonomic Bulletin & Review, 12*, 865–873.

Starns, J. J., & Ratcliff, R. (2008). Two dimensions are not better than one: STREAK and the univariate signal detection model of remember/know performance. *Journal of Memory and Language, 59*, 169–182.

Tulving, E. (1981). Similarity relations in recognition. *Journal of Verbal Learning and Verbal Behaviour, 20*, 479–496.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology, 26*, 1–12.

Wagenmakers, E-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review, 11*, 192–196.

Wagenmakers, E.-J., Grunwald, P., & Steyvers, M. (2006). Accumulative prediction error and the selection of time series models. *Journal of Mathematical Psychology, 50*, 149–166.

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology, 44*, 92–107.

Wickelgren, W. A. (1977). *Learning and memory*. Englewood Cliffs, NJ: Prentice Hall.

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review, 114*, 152–176.

Wixted, J. T., & Stretch, V. (2004). In defense of the signal detection interpretation of remember/know judgments. *Psychonomic Bulletin & Review, 11*, 616–641.

Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: Evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 1341–1354.

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