JOURNAL OF COGNITIVE PSYCHOLOGY, 2011, 23 (2), 171–184

Psychology Press
Taylor & Francis Group

Comparison of RK and confidence judgement ROCs in recognition memory

Clara D. Martin[1,2], Jean-Yves Baudouin[3], Nicolas Franck[2], Fabrice Guillaume[4], François Guillem[5], Caroline Huron[6], and Guy Tiberghien[2]

[1]Universitat Pompeu Fabra, Barcelona, Spain
[2]Université de Lyon, Lyon, France
[3]Université de Bourgogne, Dijon, France
[4]Université de Provence, Marseille, France
[5]Hôpital Louis-H Lafontaine, Montréal, Canada
[6]Université de Strasbourg, France

Several indicators have been used to differentiate familiarity and recollection processes. One dualist theory stipulates that it is possible to decide whether memories come from a feeling of knowing or from a conscious retrieval of the encoding and storage conditions (remembering). Another dualist theory is based on an indirect estimation of familiarity and recollection via the subjective confidence associated with recognition responses, and from an analysis of the derived receiver operating characteristics (ROC). In the present study, participants were presented with target words or faces that they subsequently had to recognise among distractor words or faces. On the recognition phase, the old items were the same size or a different size. In two different conditions, participants had to report (1) their remember/know/guess judgement or (2) their confidence level for each of their recognition responses. The main goal of the experiment was to directly compare different indicators of familiarity and recollection. The results showed that it would be risky to consider the remember/know/guess method and the confidence-judgement method as strictly equivalent.

Keywords: Confidence judgement; Familiarity; Recollection; RK judgement; ROC.

INTRODUCTION

Two families of theories are currently competing to account for memory recognition. In the first type, recognition is the result of a decision based on an automatic and gradual process through which the familiarity of the information to be recognised emerges (Heathcote, 2003; Heathcote, Raymond, & Dunn, 2006; Slotnick & Dodson, 2005). In the second type, called "dual-process theories", there are two distinct states of conscious awareness of information about previously experienced events. The first reflects an automatic and graded decision based on the familiarity of the memory trace. The second, recollection, refers to the conscious and relatively controlled reactivation of the encoding conditions, such as the encoding context (Atkinson & Juola, 1974;

Correspondence should be addressed to Clara D. Martin, Departament de Technologies i Comunicacions, Universitat Pompeu Fabra, C. Roc Boronat, 138, 08018 Barcelona, Spain. E-mail: claramartin3@gmail.com

This research was supported by grant NEUR-012-041 (Neurosciences, neurologie et psychiatrie) from the Agence Nationale de la Recherche (France). C.M. was supported by the Spanish government (Consolider Ingenio 2010 CE-CSD2007-00121). Thanks are extended to Anne Cheylus for her technical assistance and to the volunteer participants. We thank Teresa Bajo, Kenneth Malmberg, and three anonymous reviewers for helpful comments on an earlier version of this paper.

http://www.psypress.com/ecp DOI: 10.1080/20445911.2011.476722

Hockley & Consoli, 1999; Jacoby, 1994; Mandler, 1980, 1989, 1991; Yonelinas, 1994, 1997, 2002).

Many theoretical and experimental studies have been conducted over the past decade in an attempt to empirically validate one or the other of these families of models (for a review, see Heathcote, Raymond, & Dunn, 2006; Yonelinas, 2002; Yonelinas & Parks, 2007). The problem of deciding which kind of model is the best is a difficult one because both types fit the experimental data well. Validating them also remains a complicated matter from the methodological standpoint because a variety of indicators are used to measure the hypothetical processes involved, and the question of the equivalency of these indexes has yet to be answered. The present paper was aimed at comparing two ways of measuring memory-based recognition: confidence judgements and RK judgements. Given that these two methods have often been used in attempts to validate dual-process theories (Cohen, Rotello, & Macmillan, 2008; Donaldson, 1996; Inoue & Bellezza, 1998; Wixted & Stretch, 2004), it is critical to verify whether the judgements at stake do in fact rely on the same underlying cognitive processes of familiarity and recollection.

Two methods for dual-process theories: RK judgements and confidence judgements

In reference to the dual-process hypothesis, several indicators have been described to differentiate familiarity and recollection processes. One dualist theory stipulates that it is possible to decide whether memories come from a feeling of familiarity ("I know", K) or from a conscious retrieval of the encoding and storage conditions ("I remember", R). According to this theory, the two cognitive states K and R can be directly dissociated based on subjective reports such as "I know I've seen it before" or "I remember seeing it before" (Arnold & Lindsay, 2007; Donaldson, 1996; Dudukovic & Knowlton, 2006; Gardiner, 1988; Gardiner & Java, 1993a, 1993b; Hirshman & Master, 1997; Knowlton & Squire, 1995; Rajaram, 1993).

Other researchers have proposed another dualist theory based on an indirect estimation of the two cognitive states. Derived from Signal Detection Theory (Dual-Process Signal Detection or DPSD) and based on previous work by Atkinson and Juola (1974), Jacoby (1991), Jacoby, Toth, and Yonelinas (1993), and Mandler (1980, 1989, 1991), this theory was developed by Yonelinas and collaborators (Yonelinas, 1994, 1997, 2001a, 2001b, 2002; Yonelinas, Otten, Shaw, & Rugg, 2005; Yonelinas & Parks, 2007). It states that an estimation of familiarity and recollection can be derived from the subjective confidence associated with recognition responses, and from an analysis of the derived ROC (receiver operating characteristic). The ROC is characterised by a discriminability index (d'), considered to estimate the familiarity of old information as compared to new information. If familiarity is the only determinant of recognition, then the predicted ROC will be curvilinear and symmetrical with respect to the positive diagonal. If an additional recognition process takes place, i.e., a recollection process (R_{old} or R_o) analogous to the R process in the previous theory, for instance, the correct recognition probability for high confidence levels will necessarily increase because of the all-or-none characteristic of R_{old} decisions. The result is a curvilinear ROC that is asymmetrical with respect to the positive diagonal and whose y-intercept is a measure of recollection.

If we represent the ROC in reduced normal units (zROC), the classic SDT model predicts a linear relation between z(hit) variation and z(false alarm) variation. The slope of this line is equal to the ratio of the standard deviation of the distribution of new-information familiarity to the standard deviation of the distribution of old-information familiarity. The model predicts that this slope is equal to 1, i.e., that the two distributions have the same variance (Equal Variance Signal-Detection model or EVSD). If the slope is less than 1, the variance of old-information familiarity is greater than the variance of new-information familiarity. If it is greater than 1 then the opposite is true. In general, this ratio is less than 1; in typical conditions it is about .80 (Unequal Variance Signal-Detection model or UVSD). Whenever the slope is less than 1, the existence of a recollection process (R_o) in addition to the familiarity process (measured by d') will produce a curvilinear U-shaped zROC. If the quadratic component is negative, the zROC has an inverted U-shape and indicates the existence of a response bias in the estimation of familiarity and/or confidence (Heathcote, 2003; Heathcote et al., 2006; Rotello, Macmillan, & Reeder, 2004; Wixted, 2007).

Moreover, the RKG[1] model could be interpreted in terms of Signal Detection Theory (SDT; Dunn, 2004; Wixted, 2007; Wixted & Stretch 2007), which authorises direct comparisons of these two procedures. One can argue here that R, K, and G correspond to different response thresholds (C_i criteria) on a graded familiarity dimension representing the level of response confidence, e.g., C_1 new, C_2 old G (uncertain), C_3 old K (moderately certain), and C_4 old R (highly certain). Based on this assumption, one can create ROCs and *z*ROCs from data obtained with the RKG paradigm. ROCs and zROCs obtained with the two procedures could then be directly compared.

If the methodology used to measure recognition made no difference, then we could predict substantial overlapping of the ROCs obtained for the two types of judgement. Moreover, it would be theoretically logical to find that the DPSD model fits well with the data obtained using the RKG-judgement method. Yonelinas and Parks (2007) presented data in favour of these interpretations but they were contested by Wixted (2007) and by Rotello and colleagues (Rotello, Macmillan, Hicks, & Hautus, 2006; Rotello et al., 2004; Rotello, Macmillan, Reeder, & Wong, 2005). It is important to notice that the interpretation of RKG data in terms of SDT is still a matter of debate and it is taken here as a hypothesis and not as a postulate. Another variable to consider is the type of discrimination: Studies validating the DPSD model and its neutrality with respect to the type of judgement required have mainly used discrimination tasks involving retrieval of an episodic detail, often associative in nature, or targets with similar foils. Data supporting an interpretation of the RKG paradigm in terms of SDT (uni- or multidimensional) based on familiarity alone have mainly been validated for discrimination tasks that do not involve retrieval of an associative episodic detail or targets with randomly similar foils. One of the characteristics of our experiment is its use of a discrimination task likely to be facilitated or perturbed by the retrieval of an episodic detail that is perceptual in nature, namely, its format (for a discussion of this issue, see Guillaume & Tiberghien, 2005; Malmberg, Holden, & Shiffrin, 2004; Murname, Phelps, & Malmberg, 1999).

From the methodological point of view, then, it is indispensable to compare RKG judgements to confidence judgements, since devising a unifying model of episodic recognition requires verifying how the behavioural indicators are related to the hypothetical cognitive processes elicited. Such a comparison is also necessary in order to clarify the hypotheses underlying the various models of recognition and, in particular, to answer the questions of the independence or overlapping of the two processes, and the continuous or discrete character of the processes at play.

Overview of the paper

In the present study, participants were presented with target words or faces (old items) that they subsequently had to recognise among distractor words or faces (new items). In the recognition phase, the old items shown were either the same size or a different size. In two different conditions, participants had to report (1) their RKG judgement or (2) their confidence level for each of their recognition responses. The experimental design was counterbalanced for all within-group factors (judgement, stimulus type, stimulus size, and testing order).

The main goal of the present experiment was to use two different methodologies (RKG vs. confidence judgement) to obtain various cognitive processing indicators assumed to be identical or at least analogous and to reflect the same processes of recollection and familiarity.

Using RKG judgements, recollection (R) was directly derived from judgements. Familiarity was derived from K values (Yonelinas & Jacoby, 1995, p. 630). Using confidence judgements, R and F were derived from the DPSD model.

To compare the two methods, two factors were manipulated: stimulus type (faces vs. words) and stimulus size (same vs. different), these variables being known to have different impacts on familiarity and recollection. Words and faces are thought to differentially affect the contributions of familiarity and conscious recollection to episodic recognition memory in normal subjects: Recognition is based mainly on conscious recollection and lexical or semantic associations when words are at stake, whereas it is based mainly on familiarity when unknown faces are being judged

[1]Certain studies using the remember/know procedure suggest that some know responses are not based on feelings of familiarity but rather on guessing: Participants guess that they previously studied an item without experiencing familiarity (knowing) and without recollecting any details from the learning phase (remembering). To distinguish between knowing and guessing, a third category of responses—guess responses—has been introduced.

(Farah, Wilson, Drain, & Tanaka, 1998). Furthermore, modification of stimulus size between the study and test phases can have a selective impact on familiarity but not on recollection when faces are presented briefly at study time (Nega, 2005). Thus, the effect of the method (RKG vs. confidence judgement) and its possible interaction with stimulus type and stimulus size were investigated on R (recollection) and F (familiarity) estimates. Systematic comparisons of the two estimation methods were applied to the recognition data gathered in the present experiment. The results were compared to the DPSD model's predictions and interpreted in terms of three main hypotheses:

Hypothesis 1: If the two methods are equivalent, the effects of independent factors (stimulus size and type) on R and F should be similar whatever the method (no Method × Size or Method × Type interactions on R and F).

Hypothesis 2: If the two methods are equivalent, the ROC points obtained with the RKG method should fall along the ROC that fits the recognition-confidence ratings.

Hypothesis 3: If the two methods are equivalent, the *z*ROC slopes derived from the two methods should be the same.

MATERIAL AND METHODS

Subjects

Twenty-four subjects participated in the experiment (7 women and 17 men; mean age 38.1 ± 9.8 years, range 24–57). They had normal or corrected-to-normal vision, no neurological illness, dyslexia, or symptoms of prosopagnosia. Participants with histories of traumatic brain injury, epilepsy, alcohol and/or substance abuse, other diagnosable neurological conditions, or a history of psychiatric illness were excluded from the study. After a complete description of the study, written informed consent was obtained from participants. The protocol was approved by a local ethics committee.

Stimuli

The stimuli were 256 front-view colour photographs of unfamiliar faces, and 256 French words (Figure 1). The photographed faces were of Caucasian adults (128 females and 128 males) without distinctive facial features that had been carefully edited to maintain uniform brightness and contrast. The words were 128 feminine-gender and 128 masculine-gender common French nouns (log[lexical frequency] 331.8 ± 48.7, range 200–400; imagery 4.0 ± 1.0, range 1–5; number of graphemes 6.0 ± .9, range 4–8; number of phonemes 4.5 ± .9, range 2–7; neutral affective valence). The faces and words were transposed onto a medium grey (50% black) background. The size of the pictures was 450 × 500 pixels (large size) or 225 × 250 pixels (small size). Four sets of stimuli were generated, each containing two series: a study series (64 stimuli) and a test series (64 old stimuli + 64 new stimuli). For each study series, 32 stimuli were small and 32 stimuli were large. For each test series, 32 old stimuli were presented in the same size, 32 old stimuli were presented in the other size, 32 new stimuli were presented in the small size, and 32 new stimuli were presented in the large size. For the study and test phases, stimuli were presented in random order. In the study phase, each stimulus was displayed for 2000 ms in the centre of a computer screen, following a fixation cross displayed for 1000 ms, and then a blank screen during which the response to the study task was recorded. In the test phase, a fixation cross was displayed for 1000 ms, followed by the stimulus, which remained in the centre of the screen until the participant responded.

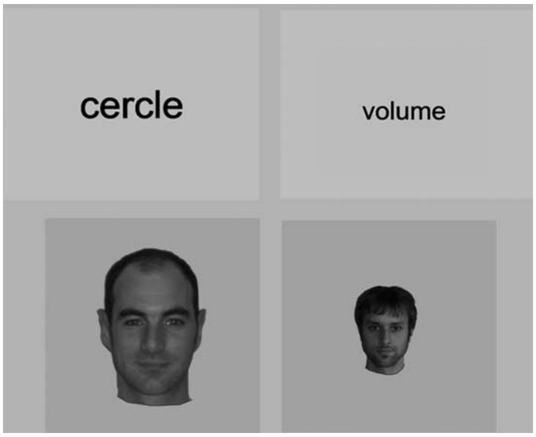

Figure 1. Examples of size, for faces and words used in the present experiment.

Experimental procedure

Each participant performed four tasks, two remember/know/guess tasks (RKG), and two confidence-judgement tasks, on faces and on words. Each task had a study phase and a recognition test. In the study phase, participants were asked to make a gender decision by pressing the "M" or "F" button on the response pad for each word or face presented. The recognition test occurred 10 min after the study phase. In each series, half of the words (faces) were of feminine gender (females) and the other half were of masculine gender (males).

For the RKG recognition test, participants were asked to respond "yes" (press the "Y" button on the response pad) to old items (regardless of their size) and to reject new items by responding "no" (press the "N" button on the response pad). Every time they gave a "yes" response, participants were asked to report their subjective state of awareness (i.e., remember, know, or guess) at the time they recognised the item. They were to respond "remember" if recognition was accompanied by the conscious recollection of what had happened or what they had experienced when the item was presented in the study phase. They were to respond "know" if recognition was associated with feelings of familiarity but no conscious recollection of any specific details about its occurrence on the study list. They could also give a "guess" response to items that elicited neither the experience of remembering nor the experience of knowing but that they thought might have appeared during the study phase. The procedure was the same for the second RKG task except that faces were replaced by words, or vice versa if the participants had begun the experiment with words.

The general procedure was the same for the confidence-judgement task. Participants had to decide as quickly as possible whether or not the face or word had been studied earlier (yes/no type of recognition). Once the response had been given, the subjects had to use a 3-point scale (1 = "not sure at all", 2 = "moderately sure", 3 = "very sure") to state how confident they were in the answer given.

Each participant was tested in two study–test blocks. In the first block, participants performed the RKG task on words and faces (two sessions); in the second, they performed the confidence-judgement task on words and faces (two sessions). The block and session orders and the study–test lists were counterbalanced across participants. The blocks were separated by a period of 1–2 weeks.

The validity of the counterbalancing in the experimental procedure was checked using an ANOVA with block (1 and 2), session (1 and 2), and list (1 to 4) as factors. The ANOVA showed that the counterbalanced factors had no significant effects on the behavioural measure H'*c* (H'*c* = %Hits −%FA): block ($F = 2.08$, $p = .17$), session ($F = 0.80$, $p = .38$), list ($F = 0.39$, $p = .54$). No one-way or two-way interactions between these factors reached the significance level, $0.06 < F\text{s} < 1.46$, and $.24 < p < .81$.

Confidence judgements

The hit and false alarm rates were computed for each confidence level and each experimental condition. The cumulative hit rate and the cumulative false alarm rate for each confidence level were determined and plotted to obtain 5-point aggregate ROC curves ([3 confidence levels × 2 response categories] – 1 *df*). The DPSD equation[2] was used with a method that minimised the sum square error (SSE) between the predicted and observed ROC data, along with the algorithm described by Yonelinas (1999). A fit was computed using two free parameters: d' (familiarity) and R_o (recollection of old items). For the UVSD model, a fit was computed using the maximum likelihood estimation method (Metz, 1998) with two parameters: d' (familiarity) and σ_{old} (standard deviation of the old-stimulus distribution). Corresponding *z*ROCs were plotted and their linearity was analysed based on a binormal model.

Familiarity (F) and recollection (R) were estimated for each subject using the DPSD model:

$$R = R_o$$
with $R_n = 0$

where R_o = recollection for old items; R_n = recol-recollection for new items,

$$F = F_o - F_n$$

with $F_o = P(\text{hit}) - R_o/(1 - Ro)$ and $F_n = P(FA)$

[2]In its standard form, this equation (in the Gaussian model) gives the proportion of hits as a function of the proportion of false alarms (FA), a target recollection parameter (Rt), a familiarity parameter (d'), and the decision criterion (c):

$$P(\text{"yes"/Target}) = P(\text{"yes"/Distractor}) + Rt + (1 - Rt) \times ((d'/2 - c) - (-d'/2) - c)$$

where F_o = familiarity for old items; F_n = familiarity for new items; P(Hit) = probability of hits; P(FA) = probability of false alarms.

Remember, know, and guess judgements

The absolute proportions of "yes", "remember", "know", and "guess" responses were calculated for old and new items by dividing the number of responses given by the number of items presented during the study phase in each condition.

Recollection was estimated for each subject:

$$R = R_o - R_n$$

Familiarity was estimated for each subject, based on the stochastic independence of the two processes (IRK procedure: Parks & Yonelinas, 2007; Wixted, 2007; Yonelinas, Kroll, Dobbins, Lazzara, & Knight, 1998):

$$F = F_o - F_n$$

with $F_o = K_o/(1 - R_o)$ and $F_n = K_n/(1 - R_n)$

where R_o = probability of "remember" response to old items; R_n = probability of "remember" response to new items; K_o = probability of "know" response to old items; K_n = probability of "know" response to new items.

The cumulative hit rate and the cumulative false alarm rate for each criterion C (C1 = R, C2 = R + K, C3 = R + K + G, C4 = R + K + G + New) were calculated and plotted on the empirical ROC obtained via the confidence-judgement method (Rotello et al., 2004; Wixted & Stretch, 2004). A three-point *z*ROC (4 criteria – 1 *df*) was plotted for each condition. As previously described, a corresponding linearity analysis was conducted.

Data analysis

The experiment was conducted using a three within-participant factor ANOVA with repeated measures (2 × 2 × 2). The experimental factors were method (RKG judgement vs. confidence judgement), type of stimulus (word vs. face), and size of stimulus (same vs. different). The ANOVAs were applied to the measures of familiarity (F) and Recollection (R).

RESULTS

The mean hit rate and the mean false alarm rate in each experimental condition for confidence judgements are presented in Table 1. The mean hit and false alarm rates for RKG judgements are presented in Table 2.

Familiarity and recollection in RKG and confidence judgements

The mean probabilities of familiarity (F) and recollection (R) in each experimental condition for confidence and RKG judgements are presented in Table 3. The ANOVA for recollection (R) revealed a significant effect of stimulus type. The mean R proportion was significantly higher for words (R = 0.35, *SD* = 0.21) than for faces (R = 0.25, *SD* = 0.18), $F(1, 23) = 10.89$, $p = .0031$. The main effect of size was also significant, with the mean proportion of R being higher when the size did not change (R = 0.33, *SD* = 0.21) than when it did (R = 0.27, *SD* = 0.20), $F(1, 23) = 13.61$, $p = .0012$. The main effect of method was also significant, mean R proportion being higher for RKG judgement (R = 0.34, *SD* = 0.18) than confidence judgement (R = 0.26, *SD* = 0.22), $F(1, 23) = 6.61$, $p = .0171$. No interaction reached significance (all *p*s > .17).

TABLE 1
Hit rate and false-alarm rate, by size (same vs. different) in word and face recognition using the confidence-judgement method

	Word				*Face*			
	Same size		*Different size*		*Same size*		*Different size*	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Hit rate	0.768	0.118	0.764	0.158	0.747	0.124	0.650	0.144
False-alarm rate	0.193	0.121	0.193	0.121	0.249	0.130	0.249	0.130

M = mean, *SD* = standard deviation.

TABLE 2
Percentage of remember, know, and guess responses and overall hits and false alarms for targets and distractors, by size (same vs. different) in word and face recognition using the RKG-judgement method

	Word				*Face*			
	Same size		*Different size*		*Same size*		*Different size*	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Target								
Remember	0.431	0.173	0.426	0.200	0.379	0.197	0.293	0.168
Know	0.280	0.145	0.252	0.144	0.265	0.146	0.215	0.098
Guess	0.076	0.049	0.085	0.074	0.104	0.093	0.115	0.124
Total hits	0.787	0.150	0.763	0.168	0.748	0.111	0.623	0.159
Distractor								
Remember	0.040	0.063			0.045	0.081		
Know	0.096	0.120			0.103	0.073		
Guess	0.052	0.045			0.070	0.062		
Total false alarms	0.188	0.164			0.218	0.119		

M = mean, *SD* = standard deviation.

The ANOVA for familiarity (F) also indicated a significant effect of stimulus type. The mean F proportion was significantly higher for words (F = 0.40, *SD* = 0.18) than for faces (F = 0.30, *SD* = 0.17), $F(1, 23) = 16.78$, $p = .0004$. The main effect of size was significant, mean F proportion being higher when the size did not change (F = 0.37, *SD* = 0.18) than when it did (F = 0.33, *SD* = 0.19), $F(1, 23) = 4.45$, $p = .0459$. The main effect of method was also significant, mean F proportion being lower for RKG judgement (F = 0.30, *SD* = 0.17) than confidence judgement (F = 0.40, *SD* = 0.19), $F(1, 23) = 15.32$, $p < .0007$. The Stimulus type × Stimulus size interaction was significant, $F(1, 23) = 4.55$, $p = .0438$. The probability to recognise a face based on familiarity was significantly larger with no size change (F = 0.34, *SD* = 0.17) than when the size of the face differs between the encoding and recognition phase (F = 0.26, *SD* = 0.17), $t(23) = 4.15$, $p < .0004$. The probability to recognise a word based on familiarity was not influenced by size modification (same size: F = 0.40, *SD* = 0.19; different size: F = 0.40, *SD* = 0.19), $t(23) = 0.02$, $p = .982$. Interestingly, the Method × Stimulus size interaction was also significant, $F(1, 23) = 5.88$, $p = .0235$. In RKG judgement, the probability to recognise an item based on familiarity was significantly larger with no size change (F = 0.34, *SD* = 0.16) than when the size of the item differs between the encoding and recognition phase (F = 0.27, *SD* = 0.16), $t(23) = 3.33$, $p = .003$. On the contrary, in confidence judgement, the probability to recognise an item based on familiarity was not influenced by size modification (same size: F = 0.41, *SD* = 0.19; different size: F = 0.40, *SD* = 0.20), $t(23) = 0.99$, $p = .332$.

In summary, recognition based on recollection was independently affected by the three factors investigated in the experiment, the method,

TABLE 3
Mean probability of familiarity (F) and recollection (R), by size (same vs. different) in word and face recognition using the RKG and confidence-judgement methods

	Word		*Face*	
	Same size	*Different size*	*Same size*	*Different size*
Familiarity (F)				
RKG judgement	0.379 (0.183)	0.332 (0.172)	0.297 (0.130)	0.206 (0.127)
Confidence judgement	0.420 (0.173)	0.467 (0.184)	0.390 (0.199)	0.322 (0.187)
Recollection (R)				
RKG judgement	0.391 (0.165)	0.387 (0.197)	0.334 (0.175)	0.248 (0.144)
Confidence judgement	0.336 (0.245)	0.298 (0.231)	0.248 (0.217)	0.159 (0.165)

Standard deviations are reported in parentheses.

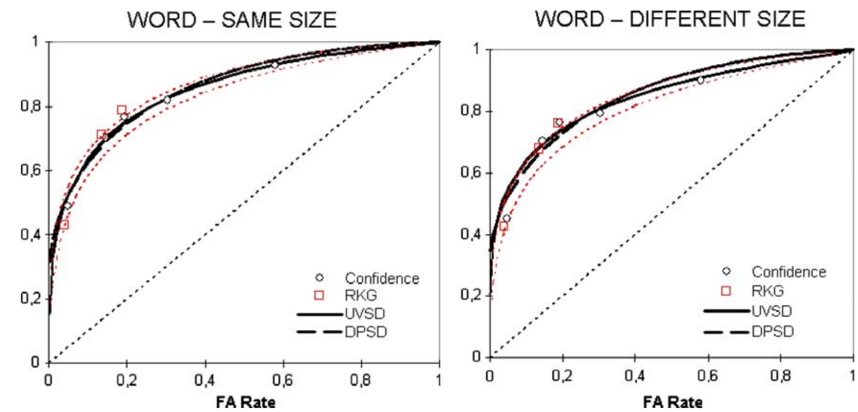

Figure 2. Confidence Receiver Operating Characteristics (ROC) for confidence judgement and remember/know/guess (RKG) judgement, by size (same vs. different) in word recognition. The figure presents the fit of the Dual-Process Signal Detection model (DPSD; solid line) and of the Unequal Variance Signal Detection model (UVSD; dotted line) for confidence judgements with the lower and upper bounds of the 95% confidence interval (red dotted line).

stimulus type, and stimulus size. Recognition based on familiarity was also affected by the three factors. Interestingly for our purpose, recognition based on familiarity was affected by stimulus size in the RKG judgement and was not affected by stimulus size in the confidence judgement.

RKG and confidence ROCs

Empirical ROC data for words and faces in the confidence-judgement situations are presented in Figures 2 and 3. These ROCs were computed from 768 responses to old items (32 old items per condition $\times$ 24 participants) and 1536 responses to new items (64 items per condition $\times$ 24 participants). The fit of models DPSD and UVSD to these empirical data is also presented in Figures 2 and 3. Table 4 gives the values of the UVSD fit parameters, d' (familiarity) and σ_o (standard deviation of the old-item familiarity distribution) and the values of the DPSD fit parameters, d' (familiarity) and R_o (recollection). We can see that both models fit the empirical data very well (based on confidence judgement), which makes them difficult to dissociate on the basis of ROC data. However, model UVSD's SSE is still slightly below model DPSD's. On these same figures, we have plotted the lower and upper bounds of the 95% confidence interval of model UVSD (the best fit). The three points of the empirical ROCs obtained with the RKG method are also shown in these figures.

As can be seen in Figures 2 and 3, for the same-size word, different-size word, and same-size face conditions, only the two RKG points

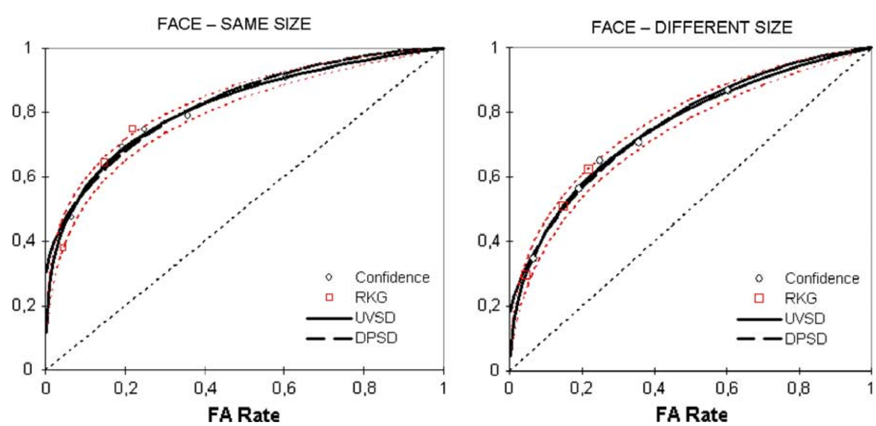

Figure 3. Confidence Receiver Operating Characteristic (ROC) for confidence judgement and remember/know/guess (RKG) judgement, by size (same vs. different) in face recognition. The figure presents the fit of the Dual-Process Signal Detection model (DPSD; solid line) and of the Unequal Variance Signal Detection model (UVSD; dotted line) with the lower and upper bounds of the 95% confidence interval (red dotted line).

TABLE 4
Recognition memory ROC parameters for UVSD and DPSD fits by size (same vs. different) for word and face recognition using the confidence-judgement method

	Best fit UVSD model			*Best fit DPSD model*		
	d'	σ_o	*SSE*	d'	R_o	*SSE*
Word-same size	1.75	1.27	.00038	1.22	0.29	.00089
Word-different size	1.84	1.33	.00138	1.10	0.33	.00306
Face-same size	1.51	1.28	.00040	0.97	0.29	.00093
Face-different size	1.08	1.18	.00040	0.78	0.18	.00068

d' = familiarity; σ_o = standard deviation for old items; R_o = recollection; SSE = sum square error.

corresponding to the more conservative decision criterion were within the symmetrical 95% confidence interval for the UVSD fit of the confidence data. For the different-size faces, all three RKG points were within the symmetrical 95% confidence interval for the UVSD fit.

Results of the *z*ROC linearity analysis

ROCs reduced to normal units (*z*ROCs), which correspond to the ROCs obtained with the confidence-judgement method in the different conditions, are plotted in Figure 4. The best fit indicated by linearity analysis is also shown. Figure 4 also shows the data for the *z*ROCs computed with the RKG-judgement method. Quantitative data for these analyses are given in Table 5.

It was possible to fit a line to the different *z*ROCs: The determination coefficient R^2 varied between .954 and .995 for the confidence-judgement paradigm, depending on the condition, and between .997 and 1.000 for the RKG-judgement paradigm. Thus, we can accept the normality postulate for the familiarity distributions. The *z*ROC slopes were significantly below 1 when derived from confidence judgements (the value 1 was always outside the symmetrical 95% confidence interval). They were close to .80, which is a classical observation for recognition tasks. The slope of the regression lines was equal to the ratio of the standard deviation of the new-item familiarity distribution to the standard deviation of the old-item familiarity distribution. One can therefore conclude that the variance of the familiarity distribution was larger for old items than for new items on confidence judgements.

Finally, it is justified to test for the quality of a polynomial fit (linear + quadratic) when the slope is less than 1. The fit of a polynomial function (linear + quadratic) to the empirical *z*ROCs obtained from confidence judgements always gave rise to a negative quadratic *c*-component (inverted-U function), which indicates the presence of response biases in all conditions. However, the quadratic *c*-component significantly differed from 0 only for word recognition: same size, $c = -0.09$, $SD = 0.06$, $t(4) = -3.35$, $p = .029$; different size, $c = -0.21, SD = 0.12, t(4) = 3.91, p = .017$. Figure 4 shows the inverted-U quadratic functions for these last two conditions. For the other two conditions (faces), the linear functions are plotted.

The RKG-judgement slopes never differed from 1 (the value 1 was always within the symmetrical 95% confidence interval), so the variances of the old-item and new-item familiarity distributions can be considered equal. The linear functions for the four conditions are plotted in Figure 4.

TABLE 5
*z*ROC linearity analysis results, by size (same vs. different) and judgement (confidence vs. RKG judgement) for word and face recognition (R^2_{lin} = linear coefficient of determination)

	Confidence judgement		*RKG judgement*	
	R^2_{lin}	*Slope*	R^2_{lin}	*Slope*
Same-size word	.990	0.79 (0.71–0.88)	1.000	1.12 (0.94–1.30)
Different-size word	.954	0.75 (0.66–0.83)	.999	1.05 (0.88–1.22)
Same-size face	.992	0.78 (0.70–0.86)	.999	1.07 (0.91–1.23)
Different-size face	.995	0.85 (0.76–0.93)	.997	0.95 (0.80–1.09)

Lower and upper bounds of the 95% confidence interval are reported in parentheses.

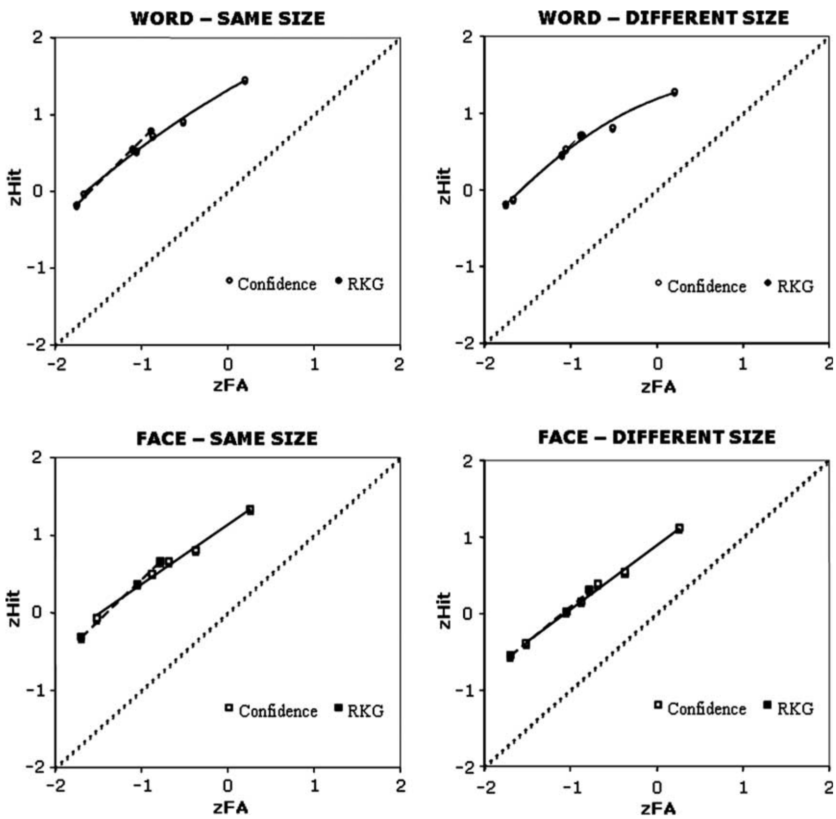

Figure 4. *z*ROC by size (same vs. different) and items (words vs. faces) using the confidence-judgement and RKG-judgement methods. The figure presents the best fit (linear or linear + quadratic).

When the *z*ROC slopes of RKG judgements were compared to those of confidence judgements, they were found to differ significantly except when the task was face recognition and the size was different at recognition time. In this case, the RKG-judgement slope fell within the symmetrical 95% confidence interval of the confidence-judgement slope.[3]

The *z*ROC linearity analysis thus allows us to conclude that a standard SDT model based solely on familiarity can account for the data obtained here, for both confidence and RKG judgements. Model UVSD nicely describes the *z*ROCs obtained in the confidence-judgement condition: linear *z*ROCs (with a slight bias for words) and slopes less than 1 (old-item familiarity distribution variance greater than new-item familiarity distribution). Model EVSD accounts for the *z*ROCs obtained in condition RKG: linear *z*ROCs and slopes equal to 1 (equal variances for the old- and new-item familiarity distributions).

DISCUSSION

The indicators used to measure same or similar processes should be affected in the same way by various experimental manipulations. The data obtained in this experiment only partially validated this hypothesis. The type of material studied (words vs. faces) had the same impact on the probability of recollecting (R) using both methods (RKG and confidence judgement), with higher R values for words than for faces. Other studies have also found that recollection plays a

[3]However, Malmberg and Xu (2006) showed that the *z*ROC slope is higher when calculated on the H and FA mean than on the mean slope of the individual *z*ROCs. They showed how averaging can distort the shape of the *z*ROCs such that differences in *z*ROC slopes are observed because they become nonlinear (see also Ratcliff, McKoon, & Tindall, 1994; Ratcliff & Starns, 2009). In the experiment described here, the *z*ROCs and their slopes were determined from aggregate *z*ROCs, so we cannot rule out the possibility of skewing due to the calculation method. If we compute the slopes from the mean of the individual *z*ROC slopes rather than from the slope of the aggregate *z*ROC, the differences between the two calculation methods (confidence and RKG) remain small. More importantly, the conclusions are the same no matter which way the slope is calculated. In fact, Macmillan and Kaplan (1985) showed that the two calculation methods were very similar whenever the interindividual variance was low.

greater role in word recognition than in face recognition (Farah et al., 1998; Umiltà & Moscovitch, 1994).[4] Modifying the stimulus size between the study phase and the retention test also acted in the same way on R values measured with RKG or confidence-judgement methods. This indicator was lowered by a change in size. Overall, the analysis of recollection estimate (R) revealed no significant effect of the method used to measure recollection.

The result patterns differed when we compared the effects of the independent variables on familiarity estimates (F) obtained with the two methods. Recognition based on familiarity was affected by stimulus size in the RKG judgement and was not affected by stimulus size in the confidence judgement. Changing stimulus size between the encoding and retrieval phase decreased significantly the probability of recognition based on familiarity (F) if F is estimated from RKG judgement. On the contrary, when F is estimated from confidence judgements, it is not affected by stimulus size congruency. These results are only partially consistent with Yonelinas and Jacoby's research (1995) on recognition, in which the old geometric shapes they used were either size congruent (same size at study and test) or size incongruent. In fact, their results showed that changing stimulus size between the encoding and retrieval phase led to a decrease in both recollection and familiarity.

Thus, strict equivalence between parameters F obtained using the confidence judgement or the RKG method could not be demonstrated here. Moreover, if there were no differences between the two assessment methods, then the three points of the RKG ROC would always land on the confidence-judgement ROC; this was not observed in only one of our four experimental conditions. These results are inconsistent with studies showing a satisfactory overlap between confidence-based ROCs and RKG-based ROCs (Malmberg & Xu, 2006; Stretch & Wixted, 1998; Wixted & Stretch, 2004; Yonelinas et al., 1998; for a review, see Yonelinas & Parks, 2007).

A glance at the slope of the *z*ROCs derived from these ROCs confirms this conclusion. The hypothesised normality of the familiarity distributions was validated no matter what method was used since it was possible to fit a line to the data obtained with either method (RKG or confidence judgements). However, the *z*ROC slope based on confidence judgements was always below 1, which validates the UVSD model, whereas the *z*ROC slope based on the RKG method did not differ significantly from 1, which is in line with the EVSD model.

Comparing the results of the ANOVAs on F and R and the linearity analysis (*z*ROC) thus creates a paradox. Indeed, on one side, the ANOVAs indicate a differential effect of stimulus size on F and R estimates. But on the other side, the *z*ROCs are always linear, which leads us to conclude that there is only one underlying process—familiarity—in both confidence judgement and in RKG judgement. In other words, no matter what method is used, the UVSD model fits the data better. This paradox could be resolved by assuming that R and K are not different cognitive states but two, more-or-less strict positions of the decision criterion along a single axis, familiarity (Donaldson, 1996). Moreover, Wixted and Stretch (2007) developed such a one-dimensional SDT model (equal variances) that fits our data, with a low condition-dependent SSE (ranging from .001 to .005).[5]

Another hypothesis would be that recognition is always based on familiarity but that familiarity can be separated into two orthogonal dimensions: a global familiarity dimension and a specific familiarity dimension. The old/new decision would be derived from the sum of the states observed on these two axes at a given time, whereas the RK decision would be derived from the weighted difference between these two states. This multidimensional model of recognition, based on familiarity alone, was formalised by Rotello et al. (2004) under the acronym STREAK ("Sum-difference Theory of REmembering And Knowing"). This model has given rise to a number of developments within the past few years (Cohen et al., 2008; Hautus, Macmillan, & Rotello, 2008). We applied the STREAK model to the data presented in this paper, but the SSE

[4]Although faces are usually recognised better than words (Bruce & Young, 1986), there are many exceptions. The face superiority effect—and, more generally, the superiority of pictorial material over verbal material—has mainly been observed with between-subject experimental designs. On the other hand, and although it is not clear why, the opposite effect or no effect has often been observed when within-subject designs are used, as we did here (Roediger III, 2008).

[5]Note, however, that the accuracy level attained for this model was not higher than for UVSD or DPSD, all the while requiring a greater number of parameters.

was still higher than that observed for the models tested previously (STREAK SSE between .006 and .016), and the number of STREAK parameters was larger.

In any case, even with a familiarity-based model, the confidence-judgement method and the RKG-judgement method still differed: In the former, the *z*ROC slope was less than 1, which means that the underlying old-item familiarity distribution had a higher variance than the new-item one. In the latter, the *z*ROC slope was equal to 1, which means that the variances of the two distributions were equal. In other words, model UVSD was validated by the confidence-judgement data, whereas EVSD was validated for RKG. A higher RKG *z*ROC slope than confidence *z*ROC slope was also found by Rotello et al. (2005) in a review of 375 experiments: The observed mean RKG slopes were close to or greater than 1 (see also Dunn, 2004; Rotello et al., 2006). The cause of this discrepancy has not yet been discovered. A partial explanation of these differences could be found by looking at various factors, such as the number of levels on the response scale (binary decision vs. continuous decision) or the use of different decision rules. But it could also be that the decision criterion was not the same in the RKG and confidence judgements.[6]

CONCLUSION

It would be risky to consider that the remember/know/guess method and the confidence judgement method are strictly equivalent. If we accept the dualist postulate, the data obtained in this experiment suggest that the measures of recollection (R) based on RKG or confidence judgements are highly similar indicators of a single memory-retrieval process. On the other hand, it is less certain that the measures of familiarity (F) based on RKG or confidence judgements are interchangeable measures of one and the same hypothetical familiarity process. Moreover, although the data obtained with the two methods can be described and interpreted within the framework of a Signal Detection Theory (SDT) while bringing to bear a single process (familiarity), substantial differences remain. First of all, the ROCs produced using the RKG judgement method could not be superimposed upon those produced by the confidence judgement method. Second, the slopes of the corresponding *z*ROCs were not the same for the two methods. The SDT model can be well adjusted to data derived from both the RKG and the confidence-judgement methods, but the first method validates the UVSD model and the second one the EVSD model. As Yonelinas and Parks (2007, p. 824) said, the origin of this difference is still a "mystery". Elucidating it would no doubt require using a judgement method to consider differences in the rules and decision criteria utilised, and, more generally, conducting a systematic study of individual ROCs to take inter-individual variability in model testing into account (Malmberg & Xu, 2006).

Original manuscript received June 2009
Revised manuscript received January 2010

[6]The overall decision criterion is always more strict in an RKG-judgement situation than in a confidence-judgement situation. However, the difference observed here was not statistically significant for words, but only for faces, $t(23) = 2.67$, $p = .014$ for unchanged size; $t(23) = 2.36$, $p = .027$ for changed size. The type of material and the stimulus size did not have a significant effect on the decision criterion.

REFERENCES

Arnold, M. M., & Lindsay, D. S. (2007). "I remember/know/guess that I knew it all along": Subjective experience versus objective measures of the knew-it-all-along effect. *Memory and Cognition*, *35*(8), 1854–1868.

Atkinson, R. C., & Juola, J. F. (1974). Search and decision processes in recognition memory. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology: Vol. 1. Learning, memory and thinking* (pp. 243–293). New York: Academic Press.

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology*, *77*, 305–327.

Cohen, A. L., Rotello, C. M., & Macmillan, N. A. (2008). Evaluating models of remember-know judgments: Complexity, mimicry, and discriminability. *Psychonomic Bulletin and Review*, *15*(5), 906–926.

Donaldson, W. (1996). The role of decision processes in remembering and knowing. *Memory and Cognition*, *24*(4), 523–533.

Dudukovic, N. M., & Knowlton, B. J. (2006). Remember-know judgments and retrieval of contextual details. *Acta Psychologica*, *122*(2), 160–173.

Dunn, J. C. (2004). Remember-know: A matter of confidence. *Psychological Review*, *111*(2), 524–542.

Farah, M. J., Wilson, K. D., Drain, M., & Tanaka, J. N. (1998). What is "special" about face perception? *Psychological Review*, *105*(3), 482–498.

Gardiner, J. M. (1988). Functional aspects of recollective experience. *Memory and Cognition*, *16*, 309–313.

Gardiner, J. M., & Java, R. I. (1993a). Recognising and remembering. In A. F. Collins, S. E. Gathercole, M. A. Conway, & P. E. Morris (Eds.), *Theories* *of memory* (pp. 163–188). Hove, UK: Lawrence Erlbaum Associates Ltd.

Gardiner, J. M., & Java, R. I. (1993b). Recognition memory and awareness: An experiential approach. *European Journal of Cognitive Psychology*, *5*(3), 337–346.

Guillaume, F., & Tiberghien, G. (2005). Electrophysiological study of contextual variations in a short-term face recognitition memory. *Cognitive Brain Research*, *22*, 471–487.

Hautus, M. J., Macmillan, N. A., & Rotello, C. M. (2008). Toward a complete decision model of item and source recognition. *Psychonomic Bulletin and Review*, *15*(5), 889–905.

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

Heathcote, A., Raymond, F., & Dunn, J. (2006). Recollection and familiarity in recognition memory: Evidence from ROC curves. *Journal of Memory and Language*, *55*, 495–514.

Hirshman, E., & Master, S. (1997). Modeling the conscious correlates of recognition memory: Reflections on the remember-know paradigm. *Memory and Cognition*, *25*(3), 345–351.

Hockley, W. E., & Consoli, A. (1999). Familiarity and recollection in item and associative recognition. *Memory and Cognition*, *27*(4), 657–664.

Inoue, C., & Bellezza, F. S. (1998). The detection model of recognition using know and remember judgments. *Memory and Cognition*, *26*(2), 299–308.

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

Jacoby, L. L. (1994). Measuring recollection: Strategic versus automatic influences of associative context. In C. Umiltà & M. Moscovitch (Eds.), *Attention and performance XV: Conscious and nonconscious information processing* (pp. 661–680). Cambridge, MA: MIT Press.

Jacoby, L. L., Toth, J. P., & Yonelinas, A. P. (1993). Separating conscious and unconscious influences of memory: Measuring recollection. *Journal of Experimental Psychology: General*, *122*, 139–154.

Knowlton, B. J., & Squire, L. R. (1995). Remembering and knowing: Two different expressions of declarative memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(3), 699–710.

Macmillan, N. A., & Kaplan, H. L. (1985). Detection theory analysis of group data: Estimating sensitivity from average hit and false-alarm rates. *Psychological Bulletin*, *98*(1), 185–199.

Malmberg, K. J., Holden, E. H., & Shiffrin, R. M. (2004). Modeling the effects of repetitions, similarity, and normative word frequency on old-new recognition and judgments of frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(2), 319–331.

Malmberg, K. J., & Xu, J. (2006). The influence of averaging and noisy decision strategies on the recognition memory ROC. *Psychonomic Bulletin and Review*, *13*(1), 99–105.

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

Mandler, G. (1989). Memory: Conscious and unconscious. In P. R. Solomon, G. R. Goethals, C. M. Kelley, & B. R. Stephens (Eds.), *Memory: Interdisciplinary approaches* (pp. 84–106). New York: Springer Verlag.

Mandler, G. (1991). Your face looks familiar but I can't remember your name: A review of dual-process theory. In W. E. Hockey & S. Lewandowsky (Eds.), *Relating theory and data: Essays on human memory in honor of Bennet B. Murdock* (pp. 207–225). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Metz, C. E. (1998). Rockit (Ver. 0.9.1 B, beta version). Chicago.

Murnane, K., Phelps, M. P., & Malmberg, K. (1999). Context-dependent recognition memory: The ICE theory. *Journal of Experimental Psychology: General*, *128*(4), 403–415.

Nega, C. (2005). Perceptual effects and recollective experience in face recognition. *Experimental Psychology*, *52*(3), 224–231.

Parks, C. M., & Yonelinas, A. P. (2007). Moving beyond pure signal-detection models: Comment on Wixted (2007). *Psychological Review*, *114*(1), 188–202.

Rajaram, S. (1993). Remembering and knowing: Two means of access to the personal past. *Memory and Cognition*, *21*(1), 89–102.

Ratcliff, R., McKoon, G., & Tindall, M. (1994). Empirical generality of data from recognition memory receiver-operating characteristic functions and implications for the global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 763–785.

Ratcliff, R., & Starns, J. J. (2009). Modeling confidence and response time in recognition memory. *Psychological Review*, *116*(1), 59–83.

Roediger, H. L., III. (2008). Relativity of remembering: Why the laws of memory vanished. *Annual Review of Psychology*, *59*, 225–254.

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 and Cognition*, *34*(8), 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*(3), 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 and Review*, *12*(5), 865–873.

Slotnick, S. D., & Dodson, C. S. (2005). Support for a continuous (single-process) model of recognition memory and source memory. *Memory and Cognition*, *33*(1), 151–170.

Stretch, V., & Wixted, J. T. (1998). Decision rules for recognition memory confidence judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 1397–1410.

Umiltà, C., & Moscovitch, M. (Eds.) (1994). *Attention and performance XV: Conscious and nonconscious information processing*. Cambridge, MA: MIT Press.

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review*, *114*(1), 152–176.

Wixted, T., & Stretch, V. (2004). In defense of the signal detection interpretation of remember/know judgments. *Psychonomic Bulletin and Review*, *11*(4), 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*(6), 1341–1354.

Yonelinas, A. P. (1997). Recognition memory ROCs for item and associative information: The contribution of recollection and familiarity. *Memory and Cognition*, *25*(6), 747–763.

Yonelinas, A. P. (1999). The contribution of recollection and familiarity to recognition and source memory judgments: A formal dual process model and an ROC analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 1415–1434.

Yonelinas, A. P. (2001a). Components of episodic memory: The contribution of recollection and familiarity. *Philosophical Transactions of the Royal Society London: Biological Sciences*, *356B*(1413), 1363–1374.

Yonelinas, A. P. (2001b). Consciousness, control, and confidence: The 3 Cs of recognition memory. *Journal of Experimental Psychology: General*, *130*(3), 361–379.

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.

Yonelinas, A. P., & Jacoby, L. L. (1995). The relation between remembering and knowing as bases for recognition: Effects of size congruency. *Journal of Memory and Language*, *34*, 622–643.

Yonelinas, A. P., Kroll, N. E. A., Dobbins, I., Lazzara, M., & Knight, R. T. (1998). Recollection and familiarity deficits in amnesia: Convergence of remember-know, process-dissociation, and receiver operating characteristic data. *Neuropsychology*, *12*(3), 323–339.

Yonelinas, A. P., Otten, L. J., Shaw, K. N., & Rugg, M. D. (2005). Separating the brain regions involved in recollection and familiarity in recognition memory. *Journal of Neuroscience*, *25*(11), 3002–3008.

Yonelinas, A. P., & Parks, C. M. (2007). Receiver operating characteristics (ROCs) in recognition memory: A review. *Psychological Bulletin*, *133*(5), 800–832.