Journal of Memory and Language 59 (2008) 169–182

Contents lists available at ScienceDirect

Journal of Memory and Language

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

Two dimensions are not better than one: STREAK and the univariate signal detection model of remember/know performance ☆

Jeffrey J. Starns *, Roger Ratcliff

Department of Psychology, The Ohio State University, 225 Psychology Building, 1835 Neil Avenue, Columbus, OH 43210, USA

ARTICLE INFO

Article history:
Received 13 February 2008
Revision received 22 April 2008
Available online 10 June 2008

Keywords:
Memory models
Recognition
Remember/know

ABSTRACT

We evaluated STREAK and the univariate signal detection model of Remember/Know (RK) judgments in terms of their ability to fit empirical data and produce psychologically meaningful parameter estimates. Participants studied pairs of words and completed item recognition tests with RK judgments as well as associative recognition tests. Fits to the RK data showed that the univariate model provided a better fit than STREAK for the majority of participants. Although associative recognition relies primarily on specific memory evidence for the association, both the global and specific memory strength estimates from STREAK strongly correlated with associative recognition performance, and the correlation was actually nominally stronger for global strength. Thus, STREAK did not fit the data as well as the univariate model and did not produce interpretable estimates of global and specific memory strength. The success of the univariate model suggests that RK judgments are based on a single conglomerate of all available memory evidence and do not reflect qualitatively different memory systems or processes.

All can attest to the diversity of the subjective experiences that accompany recognition memory. Sometimes, one experiences a vivid revival of the precise situation in which the recognized item was previously encountered. Other times, one experiences only a vague impression that the item was previously encountered in some unspecified context. The Remember/Know (RK) procedure is a commonly-used tool for exploring the subjective experience of recognition (Tulving, 1985). In this procedure, participants first study a list of items. On a subsequent test, studied items are intermixed with words not appearing on the study list, and participants are asked to identify the items that they studied. Responses can be made on a confidence scale; for example, participants may respond on a scale ranging from 1 to 4 with a 1 indicating certainty that the item was not studied and a 4 indicating certainty that the item was studied. For all words participants claim to have studied, they are asked to report whether they recognized the item based on a specific recovered detail of the item's prior presentation (i.e., a "remember" or "R" response) or a global sense of familiarity (i.e., a "know" or "K" response). Our goal is to explore how specific and global information are integrated in RK judgments, and we approached this goal by applying formal models of the RK paradigm. The following paragraphs will define the models in our focus and outline our strategy for comparing them.

Rotello, Macmillan, and Reeder (2004) recently developed a detection model for the RK paradigm that provides separate estimates of the influence of specific and global information. The Sum–difference Theory of REmembering And Knowing (STREAK) assumes orthogonal dimensions of specific and global strength, with both types of strength distributed normally across items. Although the distinction between these two types of memory evidence is not precisely defined, global strength describes vague evidence of recent occurrence and specific strength describes the recovery of "qualitative" details from the learning event such as contextual information and elaborations (Rotello et al., 2004).

☆ Preparation of this article was supported by NIMH grant R37-MH44640 and NIA grant RO1-AG17083.

* Corresponding author.
E-mail address: starns.4@osu.edu (J.J. Starns).

0749-596X/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jml.2008.04.003

Fig. 1 provides a graphical depiction of STREAK, with panel A displaying the model's parameters and panel B displaying the response regions defined by the model. The horizontal dimension represents the global strength of memory candidates, and the vertical dimension represents specific strength. The circles are equal-likelihood contours of bivariate normal distributions representing memory evidence values, with separate distributions for studied and non-studied items (targets and lures, respectively). All of the model's parameters are expressed in units of the target distribution's standard deviation, which is scaled to be 1 in both dimensions.

STREAK uses three parameters to define characteristics of memory evidence: d_X indexes the distance between the means of the target and lure distributions on the global dimension, d_Y indexes the distance between the means of the target and lure distributions on the specific dimension, and σ_L indexes the standard deviation of the lure distribution (assumed to be equal in both dimensions). The remaining parameters define response criteria used to map the global and specific strength values onto the response categories. The lines labeled λ_{C1-3} are used for the initial confidence scale response. These lines have a slope of $-d_Y/d_X$, and each λ_C parameter measures the distance of the criterion from the mean of the lure distribution along a vector that is perpendicular to the criterion. The negative slope of the criteria indicates that both specific and global information contribute to recognition confidence. More specifically, confidence scale responses are determined by the weighted sum of the specific and global evidence with a weight of 1 for the specific evidence and a weight of d_Y/d_X for the global evidence. Higher values of the weighted sum lead to higher confidence that the item was studied (see Fig. 1B).

The lines labeled λ_{RK3-4} determine RK responses for items called "studied" (ratings of 3 or 4 in this example). Fig. 1 shows that the RK criteria may differ for different levels of confidence that the item was studied; for example, participants may be less willing to provide an R response after responding 3 versus 4. The RK criteria are perpendicular to the confidence scale criteria with a slope of d_X/d_Y. The positive slope of the RK criteria indicates that relatively high levels of specific versus global evidence promote R responses, whereas relatively low levels of specific versus global evidence promote K responses (see Fig. 1B). More specifically, the RK decision is driven by the weighted difference of the specific and global strengths with a weight of 1 for the specific strength and a weight of d_X/d_Y for the global strength. The λ_{RK} parameters measure the distance of each RK criterion from the mean of the target distribution on a vector that is perpendicular to the criteria. λ_{RK} values are negative to the top left of the target mean and positive to the bottom right, so higher values correspond to more liberality in making R responses.

STREAK assumes that specific and global forms of evidence contribute in different ways to confidence versus RK decisions; namely, increasing specific strength promotes higher confidence ratings and more R responses whereas increasing global strength promotes higher confidence ratings and fewer R responses. This assumption is intuitive given that the participant's goal in an RK task is to report qualitatively distinct memory experiences. However, an alternative assumption is that participants simply combine specific and global evidence into a single composite value which determines both the old/new and RK decisions (Wixted & Stretch, 2004). The latter assumption is consistent with single dimension or univariate detection models of the RK task.

The univariate model is portrayed in Fig. 2. This model assumes Gaussian evidence distributions for targets and lures, where the strength value for each item reflects a combination of all available evidence (i.e., both specific

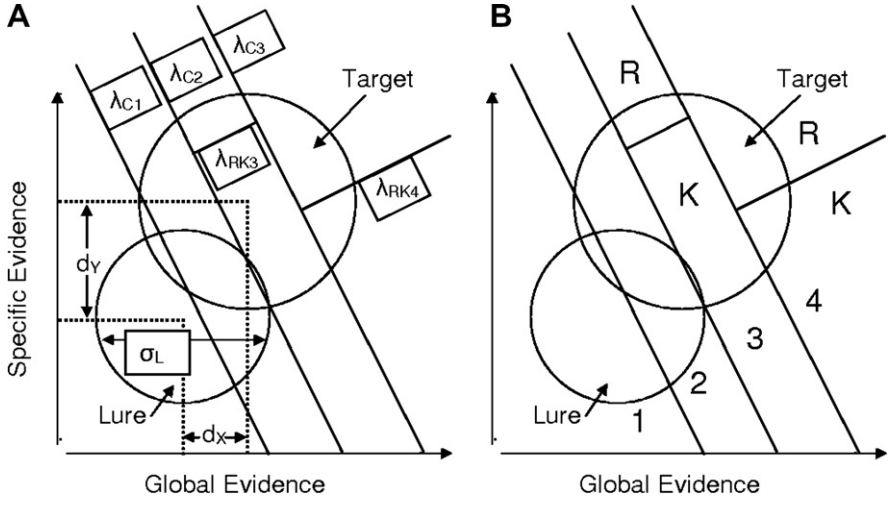

Fig. 1. Illustration of the STREAK model for recognition confidence and Remember–Know (RK) judgments. Circles represent bivariate normal distributions of memory evidence, and slanted lines represent response criteria. Panel A displays labels for each model parameter, and Panel B shows the response regions defined by the criteria. Regarding the parameters, d_X is the mean of the target distribution on the global dimension, d_Y is the mean of the target distribution on the specific dimension, σ_L is the standard deviation of the lure distribution, λ_{C1-3} are response criteria for the confidence rating, λ_{RK3} is the RK criterion following 3 ratings, and λ_{RK4} is the RK criterion following 4 ratings.

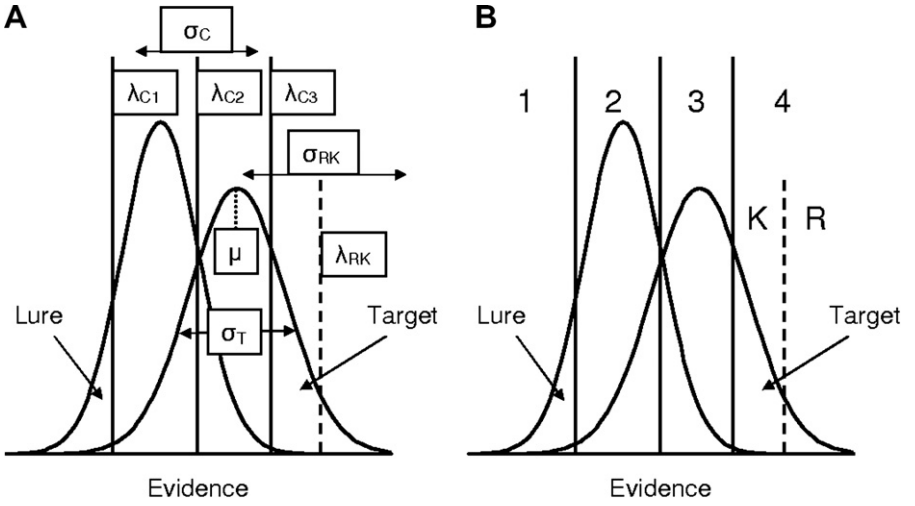

Fig. 2. Illustration of the univariate signal detection model of recognition confidence and Remember–Know (RK) judgments. Panel A shows labels for each model parameter, and Panel B shows the response regions defined by the criteria. Regarding the parameters, μ is the mean of the target evidence distribution, σ_T is the standard deviation of the target evidence distribution, $\lambda_{C1\text{-}3}$ are criteria for the confidence ratings, λ_{RK} is the criterion for the RK judgment, σ_C is the across-trial standard deviation in the positions of the confidence criteria, and σ_{RK} is the across-trial standard deviation in the position of the RK criterion. Although the RK criterion is shown in the region for a "4" response on the confidence scale, the trial-to-trial variability in criteria means that the RK criterion will sometimes fall in the region for a "3" response.

and global, Wixted, 2007; Wixted & Stretch, 2004). All model parameters are measured in units of the lure distribution's standard deviation, which is scaled to be one. The parameter μ represents the distance between the target and lure distributions, and σ_T measures the standard deviation of the target distribution. The criteria labeled λ_{C1-3} are used for the confidence rating response, with higher evidence values associated with higher confidence ratings as depicted in Fig. 2B. RK decisions are determined by an additional criterion, λ_{RK}. Evidence values above λ_{RK} are assigned R responses, and evidence values falling below λ_{RK} are assigned K responses.

When participants make both confidence and RK responses, remembering and knowing are associated with a range of confidence levels (Wixted & Stretch, 2004). In the basic univariate model, only one confidence category can be associated with both R and K responses. The RK criterion will fall in one of the regions demarcated by the confidence criteria, and only the confidence rating associated with this region will be split between R and K responses. For example, in Fig. 2 the RK criterion is located in the response region for a "4" on the confidence scale. Accordingly, only "4" responses are divided between "R" and "K" – "3" responses are always followed by a "K". Wixted and Stretch suggest that the distribution of R and K responses across confidence levels reflects between-trial variability in response criteria (see also Rotello, Macmillan, Hicks, & Hautus, 2006). We have incorporated this assumption in our univariate model fits by adding two parameters representing the standard deviation in criteria placement across trials: σ_C for the confidence criteria and σ_{RK} for the RK criterion. As a result of this variability, the RK criterion will fall in the evidence region for "3" responses on some trials and in the region for "4" responses on others; thus, both of these confidence categories will be associated with both R and K responses. We estimated variability in confidence and RK criteria separately to accommodate the possibility that the RK criterion is more variable because participants are less familiar with this sort of judgment than with assessments of confidence (Wixted & Stretch, 2004).

STREAK and the univariate model make dramatically different assumptions regarding how specific and global forms of evidence contribute to RK decisions. STREAK assumes that participants separately evaluate the different forms of evidence: specific evidence promotes "remember" responses while global evidence promotes "know" responses. The univariate model assumes that all available evidence, both specific and global, is integrated into a single evidence value. "Remember" responses simply indicate high levels of overall evidence. Thus, comparing these alternative models will provide insight into the mechanisms of evidence integration in memory judgments.

Distinguishing STREAK and the univariate model via fit

Previous comparisons of STREAK and the univariate model focus on each model's ability to fit empirical data (Dougal & Rotello, 2007; Rotello & Macmillan, 2006; Rotello et al., 2004, 2006). Evaluations of fit sometimes select more flexible but less appropriate models (Pitt, Kim, & Myung, 2003). Namely, if one of the alternative models is more flexible in accommodating a wide range of data patterns, it may outperform its competitor by accommodating error variability within a dataset. Thus, the better fitting model is not necessarily the most appropriate. We performed model recovery simulations to index overlap in the models' predictions. We generated a number of simulated datasets from each model. We fit each dataset with both models to determine if the model that generated the data clearly provided the better fit (see Wagenmakers, Ratcliff, Gomez, & Iverson, 2004). The cross-fits of a model

to the data generated by the other will gauge the relative flexibility of the models; that is, the more flexible model should provide close fits to data even when the processes that generated the data do not match the processes assumed by the model.

While preparing our manuscript, we learned that Cohen, Rotello, and Macmillan (in press) also performed model recovery simulations including STREAK and the univariate model. These researchers evaluated a variety of experimental paradigms, including one similar to our own in which an initial confidence rating was followed by a RK judgment (they call this the old-first paradigm). They found that the univariate model had comparable or slightly less flexibility in fitting data compared to STREAK. We are using different variants of the models that were evaluated by Cohen et al. Specifically, we fit a version of the univariate model with variability in both the RK and confidence criteria, whereas Cohen et al. evaluated a fixed-criterion univariate model for the old-first paradigm. Moreover, for our STREAK model we allowed the RK criterion to vary based on the initial confidence rating, whereas Cohen et al. use a single RK criterion. These differences may impact each model's flexibility. In addition to reflecting the exact models we used to fit data, our simulations are tailored to the details of our experiments. For instance, the simulated datasets have the same number of observations that we collected from each participant. Thus, our simulations are complementary to those of Cohen et al. but are specifically targeted for the empirical work to follow.

For our simulations, we focus on a paradigm in which participants first make 4-point confidence ratings then make RK decisions for all words called "old" (i.e., ratings of 3 or 4). Thus, we have 6 response frequencies for each item type; that is, the frequency of 1, 2, 3-K, 3-R, 4-K, and 4-R responses. Since these must sum to the total number of observations, there are 5 degrees of freedom for each item type and 10 total degrees of freedom for a dataset with target and lure responses. Each model uses 8 parameters to fit the 10 free response frequencies (see Figs. 1 and 2A), so the chi-squared tests of model fit have 2 degrees of freedom.

We used a SIMPLEX parameter search routine (Nelder & Meade, 1965) to fit each model. Predictions from STREAK were derived using the equations reported by Rotello et al. (2004). Predictions from the univariate model were produced using Monte Carlo simulations. We simulated 100,000 trials for both targets and lures in each simulation run, and this large sample size assured that the predicted proportions were highly consistent from one run to the next. On each trial, an evidence value was sampled from the appropriate normal distribution (i.e., mean of 0 and standard deviation of 1 for lure trials; mean of μ and standard deviation of σ_T for target trials). An RK criterion for each trial was sampled from a normal distribution with a standard deviation of σ_{RK} and a mean of λ_{RK}. Additionally, a single value from a normal distribution with a mean of 0 and a standard deviation of σ_C was selected and added to each of the confidence criteria values defined by the λ_{C1-3} parameters. Thus, the confidence criteria varied in absolute value across trials but the distances between them were constant, which ensured that the criteria never crossed. Allowing independent variation for each confidence criteria would only increase model complexity, and this additional complexity was not needed to produce good fits to empirical data.

For each model, we generated 50 sets of parameter values to use in the model recovery simulations. The parameter values were randomly sampled from uniform distributions with ranges that were chosen to span the values commonly found in applications of the models and are reported in Table 1. Ten datasets were simulated using the parameter values in each of the 50 parameter sets, yielding 500 total datasets from each model. For our first set of simulations, each dataset had 1200 observations (600 for targets, 600 for lures), which is the same number of observations provided by each participant in Experiment 1. When the univariate model was used to generate datasets, this model produced a better fit for .87 of the datasets. The mean chi-squared was 2.34 for the univariate model and 7.98 for STREAK. For data simulated from the STREAK model, STREAK produced a better fit for .92 of the datasets. The mean chi-squared value was 2.09 for STREAK and 11.68 for the univariate model. We performed a second set of simulations using the sample size for each participant in Experiment 2 (320 observations for each item type). With this sample size, the univariate model provided the best fit to .76 of the datasets generated from this model, and STREAK provided the best fit to .87 of the datasets that it generated.

The most important result from the recovery simulations is that the correct model (i.e., the one that generated the data) provided the best fit for the majority of datasets. This indicates that the chi-squared fit statistic is appropriate for selecting the best fitting model, and we can expect that the more appropriate model will produce a lower chi-squared value than its competitor for the majority of participants. The models also appear to be roughly equated in flexibility for the ranges of parameter values we assessed. STREAK produced slightly lower chi-squared values when fitting univariate data than did the univariate model when fitting STREAK data, suggesting that this model is a little more flexible. However, the difference was not large enough to make STREAK look like the more appropriate model when the data were generated based on univariate evidence. Overall, our simulation results are quite consistent with the old-first paradigm simulations reported by Cohen et al. (in press), which suggests that our alterations to the models (i.e., adding variability in both confidence and RK criteria in the univariate model and allowing separate RK criteria for different confidence levels in STREAK) will not impair our ability to select the appropriate model based on fits to data.

To our knowledge, we report the first fits of the univariate model with variability in both the confidence and RK criteria. Rotello et al. (2006) and Dougal and Rotello (2007) fit both a fixed-criterion univariate model and a model with variability in the RK criterion only (i.e., with fixed confidence criteria). They found that these models provided roughly equivalent fits to empirical data. In initial explorations with our data, we found that criterion variability was critical to produce acceptable fits, and we saw

Table 1
Range of parameter values used to generate datasets for the recovery simulations

Univariate			STREAK		
Par.	Min.	Max.	Par.	Min.	Max.
μ	.50	2.00	d_X	.30	1.00
σ_T	1.15	1.50	d_Y	.30	1.00
λ_{C1}	$\mu/2 - 1.5$	$\mu/2 - 1.1$	σ_L	.70	.90
λ_{C2}	$\mu/2 - .2$	$\mu/2 + .2$	λ_{C1}	$(d_X + d_Y)/2 - 1.4$	$(d_X + d_Y)/2 - 1.0$
λ_{C3}	$\mu/2 + .65$	$\mu/2 + 1.05$	λ_{C2}	$(d_X + d_Y)/2 - .4$	$(d_X + d_Y)/2$
λ_{RK}	$\lambda_{C3} - .40$	$\lambda_{C3} + .40$	λ_{C3}	$(d_X + d_Y)/2 + .3$	$(d_X + d_Y)/2 + .7$
σ_C	.05	.30	λ_{RK3}	−.70	−.30
σ_{RK}	.50	1.00	λ_{RK4}	1.00	1.40

Note: Parameter values for the recovery simulations were randomly sampled from uniform distributions across the range defined by the minimum and maximum values in the table. The confidence rating criteria for the univariate model were determined based on the sampled value of μ using the equations reported in the table. The confidence rating criteria for STREAK were determined based on the sampled d_X and d_Y values using the equations reported in the table.

no reason to introduce variability in only one type of criteria. However, having separate estimates for the variability of confidence (σ_C) and RK (σ_{RK}) criteria could lead to problems if these two parameters were not separately identifiable. We evaluated the recovery simulations for the univariate model to ensure that both σ_C and σ_{RK} were accurately estimated in fits to data. The mean difference between the parameter values produced in fits and the parameter values used to generate datasets was .021 for σ_C and .001 for σ_{RK}. For example, if a dataset was generated with $\sigma_C = .1$ and $\sigma_{RK} = .4$, the prototypical values returned by the fitting routine would be $\sigma_C = .121$ and $\sigma_{RK} = .401$. Thus, both parameters were estimated accurately. The correlation between σ_C and σ_{RK} values produced by the fitting routine was a modest -.11, so we found no evidence of problematic trade-offs between these two parameters. These simulation results are important, because they suggest that we will be able to derive valid estimates of variability in both confidence and RK criteria in fits to data.

Previous fit comparisons have tended to support the univariate model over STREAK (Dougal & Rotello, 2007; Rotello & Macmillan, 2006; Rotello et al., 2006). However, no existing study has compared the fit of the full models currently under consideration; that is, the univariate model with variability in both the confidence and RK criteria compared to the STREAK model with separate RK criteria for each level of confidence. These model alterations could dramatically impact results. For example, we found that allowing multiple RK criteria in STREAK was critical in fitting empirical data, so prediction error for the STREAK model may be inflated in previous studies using a single RK criterion across all confidence levels (Dougal & Rotello, 2007; Rotello & Macmillan, 2006; Rotello et al., 2004, 2006).

An additional strategy for distinguishing STREAK and the univariate model

STREAK principally differs from the univairate model in assuming that global and specific evidence differ in their contributions to recognition and RK judgments, allowing STREAK to produce separate estimates for global and specific memory strength (d_X and d_Y, respectively). If STREAK truly separates the influence of different types of evidence, this would represent a marked advantage over the univariate model. However, if both old/new and RK decisions are based on a single conglomerate of all available information, then separating the influence of global and specific information would be impossible. The d_X and d_Y values produced by STREAK would be meaningless. Therefore, one can contrast the veracity of these two models by exploring the validity of specific and global strength estimates. As a validation strategy, we tested each model's ability to predict performance on a memory test that relies predominantly on specific memory evidence: associative recognition.

Participants studied lists of paired words, then completed both a single-item recognition test with RK judgments and an associative recognition test in which they were required to discriminate pairs in their studied configuration (intact pairs) from pairs that were recombined from study to test (rearranged pairs). Although STREAK is not an explicit model of associative recognition, model parameters should relate to associative recognition performance in principled ways. The single-item and pair tests followed identical encoding phases, so one might expect some degree of overlap in the evidence used on the two tests. Notably, instructions for the single-item test explicitly mentioned remembering the word paired with the test word at study as a basis for an R response. Moreover, performance levels on the two tests were highly correlated across participants (see results below), which supports the contention that certain forms of memory evidence contribute to performance on both tests. Critically, these overlapping forms of evidence would have to be specific in nature. Although both global and specific evidence support responding on the single-item test, a global sense of recency is not discriminative on the pair test. Only recently presented words appear on the pair test, so a general feeling of recency should be equally strong for both targets (intact pairs) and lures (rearranged pairs). Discrimination relies on memory for the specific pairing of words during encoding, and this associative information clearly falls under the rubric of specific evidence (i.e., it regards a qualitative detail of the encoding event). This logic forms the basis of the following predictions.

If participants separately consider global and specific information in the RK task, then STREAK should produce d_Y

estimates that are predictive of associative recognition performance. That is, participants who encode high levels of specific information should show both high d_Y values and acute ability to discriminate intact from rearranged pairs. Moreover, d_X estimates should bear little or no relation to associative performance, as global evidence is non-diagnostic on the associative test. The univariate model conflates specific and global evidence, so its strength estimate (μ) should be moderately predictive of associative performance.

A contrasting pattern of correlations is predicted if both confidence and RK responses are based on a single evidence value. This evidence dimension should not be thought of as a particular type of evidence (e.g., familiarity); rather, it is a conglomerate of all available evidence (Wixted, 2007). Some of the available evidence may be very specific, such as memory for associations formed at study, whereas some may be vague, such as fluent processing of the test stimulus based on the item's recent presentation in the study list. Notably, the specific forms of evidence contributing to associative performance should also contribute to the conglomerate of evidence used for the RK test, so the strength estimate from the univariate model should be correlated with associative performance.

STREAK cannot produce valid estimates of global and specific strength if they are conflated in RK judgments, so what predictions can be made for d_X and d_Y? Our parameter recovery simulations reveal how STREAK's parameters relate to memory strength when fit to RK data from the univariate model. Fig. 3 shows the relationship between strength estimates from STREAK and μ from the univariate model. Clearly, both estimates from STREAK are a direct linear function of the level of univariate evidence. As a result, both of these estimates should have about as strong a relationship with associative performance as the strength estimate from the univariate model.

Experiments 1 and 2

In the current experiments, participants completed both RK tests for single items and associative recognition tests for pairs of items after studying lists of word pairs. In the first experiment, participants completed both an RK and an associative recognition test following each study list. In Experiment 2, a single RK or associative test followed separate lists. The experiments are otherwise highly similar, so we report them together. We contrasted STREAK and the univariate model both in terms of their ability to fit the data and their predictions for the pattern of correlations with associative performance. If RK judgments involve separate assessments of specific and global information as assumed by STREAK, associative performance should be related to the level of specific strength on the RK test but have little or no relationship to global strength. If specific and global strength are integrated into a single signal as assumed by the univariate model, then both the global and specific strength estimates from STREAK should correlate with associative performance.

Method

Participants

Sixteen Ohio State University undergraduates participated in Experiment 1 to fulfill a course requirement, and 21 separate undergraduates participated in Experiment 2. Participants completed two sessions on subsequent weeks. Two participants in Experiment 1 were not available for a second session, so we analyzed only their first session data. Each session lasted for approximately 45 min.

Materials

Experimental stimuli were drawn from a pool of 812 words with frequencies ranging from 78 to 10,595 occurrences per million (Kucera & Francis, 1967). Each study list was constructed by randomly selecting 48 words from the pool and randomly grouping them into 24 pairs. Four pairs were buffers that appeared in the first two and last two positions of the list. The remaining 20 pairs appeared in a random order in the central list positions. For Experiment 1, an RK test and an associative test were constructed for each study list. Words from 10 of the studied pairs appeared on the associative test. Five pairs had the same configuration as in the study list (intact targets), and five pairs were recombined from study to test (rearranged lures). For the rearranged pairs, the right word from one studied pair appeared with the left word from another studied pair. The RK test consisted of 30 studied targets and 30 lures. Ten of the targets were words from the 10 pairs selected to appear on the associative test. Only one word from each pair appeared on the RK test, the right word from half of the

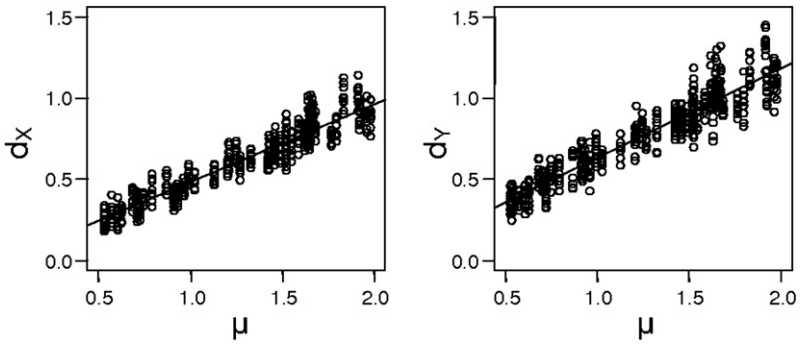

Fig. 3. Plots of d_X and d_Y parameters produced by STREAK in fits of data produced by the univariate model. μ denotes the mean of the target evidence used to generate the data from the univariate model.

pairs and the left from the other half. The 20 remaining targets were taken from the 10 studied pairs that did not appear on the associative test, and both the right and left members of each pair appeared as targets. For Experiment 2, RK and associative tests were constructed for separate study lists. The RK test consisted of 20 targets and 20 lures. Targets were words from each of the critical studied pairs, with half taken from the right position and half taken from the left position. For the associative tests, 10 studied pairs served as intact targets, and the words from the remaining ten pairs were recombined to create rearranged lures.

Procedure

The instructions informed participants that they would be studying lists of word pairs followed by a memory test for each list. Participants were also told that they would complete two different types of memory tests. Each study list began with a signal prompting participants to press the spacebar to begin studying words, and each test list began with a signal identifying the type of test and prompting participants to press the spacebar to begin. Participants in Experiment 1 completed 10 study/test cycles and participants in Experiment 2 completed 12. In Experiment 1, the RK test immediately followed each study list, and the associative test followed the RK test. In Experiment 2, an RK test followed eight study lists and an associative test followed the remaining four. Test types appeared in a random order in Experiment 2.

Instructions for the RK test closely corresponded to those reported by Rajaram, 1993). Participants were asked to respond R for "remember" when they recognized a word because they retrieved some specific detail surrounding the event of seeing the word in the study list. Participants were provided with multiple examples of specific details that they might remember, including remembering associations or images brought to mind during the study phase or remembering the word that was paired with the test word when it appeared at study. Participants were asked to respond K for "know" when no specific details of seeing the word on the study list came to mind, but they knew the word was studied nevertheless. Participants were told that K responses could be made with high confidence, and they were given the example of a word that felt so familiar that it must have been seen on the previous list. Instructions stressed that both the R and K responses could be made regardless of the confidence level reported for the initial judgment. For the RK test, a single word appeared on each trial, and participants first selected from a confidence scale with the options "Sure New", "Probably New", "Probably Old", and "Sure Old". For this response, stickers labeled "SN", "PN", "PO", and "SO" were placed on the "w", "r", "y", and "i" keys, respectively. If the participant responded with either of the "new" ratings, the next test candidate would appear on the screen. If the participant responded with either of the "old" ratings, the test word remained on the screen and a "Remember versus Know?" message appeared directly under it. Participants made RK decisions by pressing either an "R" sticker placed on the "d" key or a "K" sticker placed on the "j" key, after which the next test item appeared. Instructions for the associative test informed participants that they would see two words that either were or were not seen together in the study list. For each test pair, they pressed either a "T" sticker (on the "/" key) to respond "together" (intact) or an "S" sticker (on the "z" key) to respond separate (rearranged). Instructions noted how the rearranged test pairs were constructed, and participants were given the example that they may see the left word from the 2nd pair in the study list with the right word from the 15th pair in the study list.

Results and discussion

Model fits

The models were fit to the frequency of responses in each category defined by the combination of the confidence scale and RK responses. The critical chi-squared value for all fits is 5.99 at the .05 level of significance ($df = 2$). We fit the data for each subject as well as the overall data. The average chi-squared and parameter values from the individual subject fits appear in Table 2.

For the confidence data, model predictions are effectively displayed in the Receiver Operating Characteristic (ROC), which is a plot of the hit rate (positive responses to targets) on the false-alarm rate (positive responses to lures) across a range of bias levels (Egan, 1958; Ratcliff, McKoon, & Tindall, 1994; Ratcliff, Sheu, & Gronlund, 1992). For example, a conservative response policy would define only those items given a rating of 4 as recognized, with the proportion of targets given a 4 as the hit rate and the proportion of lures given a 4 as the false-alarm rate. A more neutral policy would define any items given a 3 or higher as recognized, and a liberal policy would define any items given a 2 or higher as recognized. Each level of bias contributes a point to the ROC plot. Hit and false-alarm rates increase as the policy becomes increasingly liberal, and the pattern across policies defines the empirical ROC function.

Figs. 4 and 5 show the overall fits from both experiments. Fig. 4 shows the confidence scale data in ROC form. Replicating the standard finding in recognition, the ROCs are clearly curved and asymmetric (Yonelinas & Parks, 2007). The solid line shows the predicted ROC from the univariate model, and the dotted line shows the predicted ROC from STREAK. Both models fit the confidence data very well, and they produce nearly identical predictions. As Table 2 shows, both models accommodate the ROC asymmetry by proposing that target evidence is more variable than lure evidence. The ratio of the lure evidence standard deviation to the target evidence standard deviation was always below one. For the univariate model, this ratio was .78 (1/1.28) in Experiment 1 and .74 (1/1.35) in Experiment 2. For STREAK, the ratio was .84 in Experiment 1 and .83 in Experiment 2.

Fig. 5 shows the proportion of R versus K responses made following 3 or 4 confidence ratings for targets and lures. Circles show the predictions of STREAK and squares show the predictions of the univariate model. Participants predominantly responded K following 3 ratings and predominantly responded R following 4 ratings. This pattern held for both targets and lures. The basic STREAK model predicts that the proportion of R responses is independent of confidence level (Rotello et al., 2004), so this model

Table 2
Parameter values and chi-squared statistics for both experiments

Univariate			STREAK		
Statistic	*E*1 value	*E*2 value	Statistic	*E*1 value	*E*2 value
μ	1.07 (.35)	1.30 (.53)	d_X	.54 (.17)	.60 (.21)
σ_T	1.28 (.12)	1.35 (.22)	d_Y	.73 (.28)	.80 (.29)
λ_{C1}	−.75 (.69)	−.67 (.61)	σ_L	.84 (.07)	.83 (.11)
λ_{C2}	.53 (.25)	.73 (.45)	λ_{C1}	−.60 (.56)	−.54 (.53)
λ_{C3}	1.35 (.58)	1.60 (.51)	λ_{C2}	.43 (.20)	.52 (.30)
λ_{RK}	1.15 (.85)	1.71 (.88)	λ_{C3}	1.08 (.47)	1.18 (.33)
σ_C	.13 (.13)	.27 (.27)	λ_{RK3}	−.55 (.68)	−.54 (.33)
σ_{RK}	.64 (.84)	.92 (1.22)	λ_{RK4}	1.32 (.58)	1.07 (.68)
χ^2	5.47 (5.62)	2.67 (2.29)	χ^2	7.75 (5.36)	5.65 (4.81)

Note: Values in parentheses are standard deviations across participants.

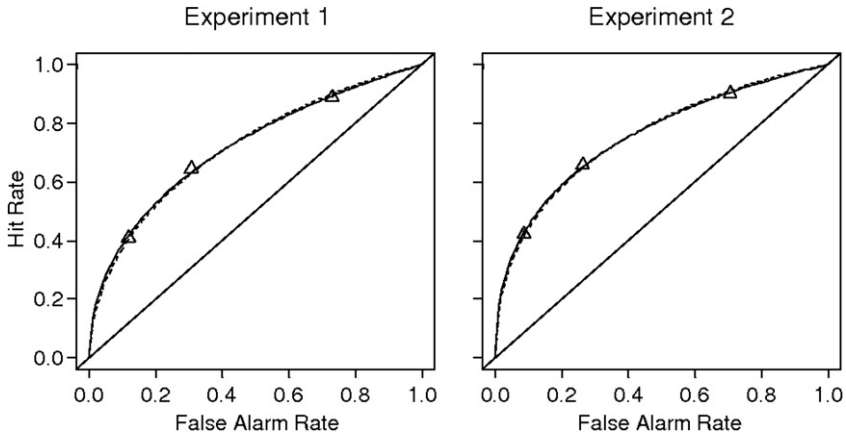

Fig. 4. ROC plots from both experiments. Triangles show the observed hit and false-alarm rates, the solid line shows the fit of the univariate model, and the dotted line shows the fit of STREAK.

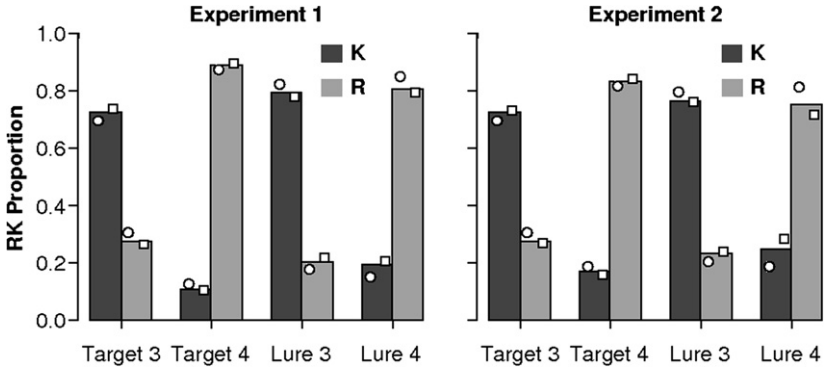

Fig. 5. Proportion of items receiving a "remember" (R) or "know" (K) response for targets and lures initially rated either a 3 or a 4 on the confidence scale. Circles show the fit of the STREAK model and squares show the fit of the univariate model.

clearly mispredicts the data. However, a version of STREAK with separate R criteria following 3 and 4 ratings can accommodate the observed pattern by assuming that participants require more evidence to say R following 3 ratings than following 4 ratings. Table 2 shows this very pattern in the λ_{RK} parameters (lower values indicate more conservative responding), and the circles in Fig. 5 show that STREAK roughly accommodates the differences across confidence. However, the univariate model predicts the RK data more closely than does STREAK. Overall chi-squared values for the univariate model were 20.64 in Experiment 1 and 12.52 in Experiment 2. The corresponding values for STREAK were 63.06 and 44.86.

Fits to individual participants also revealed a superior fit for the univariate model. In Experiment 1, the univariate model provided the best fit for 13 of 16 participants (.81).

In the recovery simulations, the univariate model provided a better fit for .87 of the datasets when it was the true model compared to .08 (1 – .92) of the datasets when STREAK was the true model. Obviously, the observed proportion is much closer to what would be expected if the univariate model generated the data. The average chi-squared value for the univariate model was 5.47 compared to 7.75 for STREAK. In Experiment 2, 14 of 21 participants (.67) were better fit by the univariate model. Cross referencing the recovery simulations with Experiment 2's sample size, the observed proportion better fit by the univariate model is again much closer to the proportion expected if the univariate model produced the data (.76) than the proportion expected if STREAK produced the data (.13). The average chi-squared value was 2.67 for the univariate model and 5.65 for STREAK. Along with the overall fits displayed in Fig. 4 and Fig. 5, the individual participant fits demonstrate that both models provide good fits to the data. However, directly comparing the models on the individual participant datasets reveals unequivocal evidence in favor of the univariate model. Our parameter recovery simulations revealed that the true model can be expected to provide the best fit for the majority of participants, and the univariate model clearly meets this standard.

Associative recognition

We performed a series of analyses to evaluate the relationship between the memory strength parameters produced by the models and associative recognition performance. We computed d' from each participant's associative recognition data by subtracting the *z*-score of the false-alarm rate from the *z*-score of the hit rate, where the hit rate was the proportion of "together" responses to intact test pairs and the false-alarm rate was the proportion of "together" responses for rearranged test pairs. Fig. 6 shows scatter plots of associative d' (y axis) and strength parameters. Each point represents a single participant, and we combined participants from both experiments because the results were quite consistent (see Fig. 4 and Fig. 5). The lines are the best fitting least squares regression functions, and the correlation coefficients and associated *p*-values are reported on the bottom right of each plot. Fig. 6 shows that all strength estimates were strongly related to associative recognition performance. The correlation coefficients were significant for all three strength estimates. In contrast to the predictions of STREAK, we found no evidence that d_Y was more strongly related associative recognition performance than d_X; in fact, the correlation coefficient for d_Y was nominally lower. This result suggests that d_X and d_Y cannot be interpreted as global and specific strength estimates, as both are strongly related to a test that requires specific forms of evidence. The results conform to the predictions of the univariate model; that is, participants consult a single strength value on the RK test that is related to the evidence used in associative recognition. When STREAK is applied to the data, both d_X and d_Y track the level of overall evidence (see Fig. 3) and should be comparably predictive of associative performance.

Before proceeding, we consider an assumption that may reconcile our pattern of correlations with the STREAK model. Specifically, what if one assumes that participants with high levels of global strength also tend to have high levels of specific strength? In this case, d_X may predict associative recognition performance not because global evidence is useful on the associative test, but because individual difference factors that influence global evidence also tend to influence specific evidence. STREAK assumes independence between global and specific evidence values at the item level, but this independence may not extend to the participant level. The assumption that global and specific strength covary across participants allows STREAK to predict a significant relationship between d_X and associative recognition performance, but can this assumption explain our full pattern of data?

We performed simulations to address the impact of participant-level correlations between global and specific strength. We randomly generated two sets of values from normal distributions to represent between-participant variability in factors influencing global memory evidence (g) and factors influencing specific evidence (s). For example, a high s value represents a participant whose characteristics promote effective encoding and retrieval of specific evidence. Our goal was to explore the effects of adjusting the correlation between g and s. A low correlation represents a situation in which there is little overlap in the individual differences affecting global and specific strength; a high correlation represents a situation in which most indi-

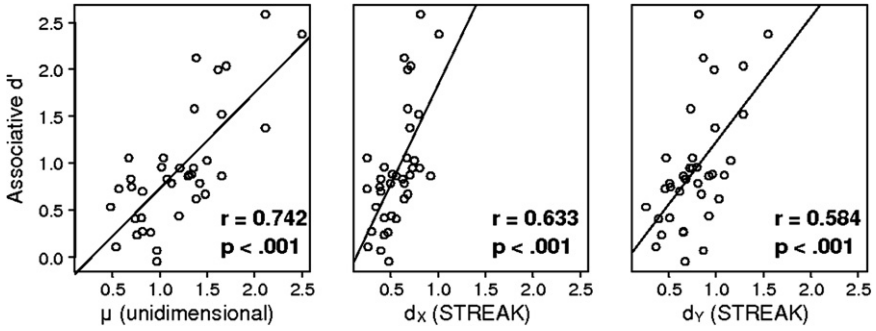

Fig. 6. Relationship between the memory strength parameters and associative recognition performance. The first plot shows results for the mean of the target evidence in the univariate model (μ), the second shows results for the global strength estimate from STREAK (d_X), and the third shows results for the specific strength estimate from STREAK (d_Y). Correlation coefficients and associated *p*-values appear in the bottom right corner of each plot.

vidual differences have similar effects on global and specific strength. After selecting g and s values for a given simulated participant, d_Y and associative recognition d' values were generated based on s with an additional normal error term to represent the influence of sources of error below the participant level, such as item-to-item variability in memorability. The additional error in the d_Y and d' estimates was set such that these variables had a correlation of .6, near the empirically observed correlation. A d_X estimate was generated for each simulated participant based on the g value and a normal error term with the same standard deviation as the error in the d_Y estimates. Finally, we evaluated the ability of d_X values to predict associative recognition d' across simulated participants.

When the correlation between factors influencing global (g) and specific (s) memory evidence was set to zero, d_X bore no relationship to associative recognition d', as expected. Even when g and s were strongly related ($r = .5$), the correlation between d_X and d' was only .3, half of the value of the correlation between d_Y and d'. In the empirical data, the relationship between d_X and d' was as strong as the relationship between d_Y and d'. To produce this result, we had to set the correlation between factors influencing global and specific memory evidence to 1. Thus, participant-level correlations between global and specific memory evidence can explain the data only if all or nearly all factors affecting global evidence have concomitant effects of specific evidence, and vice versa. Such an explanation seems antithetical to dual-process approaches like STREAK. If global and specific evidence truly made separate contributions to memory decisions, one would expect that some individual differences would selectively impact one form of evidence and have little impact on the other. To provide a few examples, differences in the effectiveness of the encoding strategy chosen by the participant may be expected to impact specific evidence to a greater extent than global evidence (see the level-of-processing and generation sections in Yonelinas, 2002). Differences in susceptibility to perceptual priming across participants should impact the fluency with which targets are re-processed at test, which should affect global evidence but not specific evidence (Johnston, Dark, & Jacoby, 1985). Thus, global and specific strength should be at least somewhat independent across participants. STREAK only captures our data with an implausible level of dependence between participants' levels of global and specific strength; therefore, we conclude that this model is inappropriate even when one allows for participant-level correlations between global and specific evidence.

The d_X and d_Y parameters from STREAK were not differentially predictive of associative recognition performance, which calls into question the claim that this model estimates separate global and specific strengths. In response, we evaluated two alternative models that make contrasting assumptions about the nature of specific and global evidence and the manner in which these forms of evidence contribute to memory decisions. If these models provide more accurate characterizations of how qualitatively different types of evidence are used in memory tasks, then their strength estimates may more reasonably relate to the associative test than those of STREAK.

TODAM

Murdock (2006) developed an extension of the TODAM model to accommodate RK data. TODAM (Murdock, 1982) represents items in a memory task as vectors of feature values. Information for all items is stored in a common memory vector. To store item information, the feature values for the item are added to the memory vector. To store an association between two items, the vectors for the items are convolved and the resulting convolution is added to the memory vector (see Murdock, 1982, for details of the convolution process). To perform item recognition, the model assesses the match of a vector representing the test item to the memory vector by taking the dot product. The model claims to recognize the test candidate if the match passes a criterion value. To perform associative recognition, the model convolves the vectors for the two test words and matches the convolution to the memory vector. The model claims that the items were studied together if the match passes a criterion. To perform cued recall, the model correlates the vector of the cue item with the memory vector, which produces a vector that approximates the feature values for the item that was associated with the cue item during learning. Thus, the model provides a unified framework for the retrieval of item and associative information in both recognition and recall tasks.

Murdock (2006) applies TODAM to RK judgments by assuming that participants separately evaluate item and associative information. Associative strength is evaluated by correlating the test item with the memory vector to produce a vector representing information associated with the test item. This retrieved vector is matched to memory, and the model responds with an R if the resultant match exceeds a criterion. When the associative match fails to support an R response, item information is evaluated by matching the vector for the test item to memory. The model responds K if the item match passes a criterion or responds "new" if the item match falls below the criterion. Item information is only assessed when the associative match fails, creating a two-step decision process. The item and associative match values are independent, and distributions of match values for both targets and lures are assumed to be normal with a standard deviation of 1.

We used Murdock's (2006) equations to compute the means of the associative match distribution (denoted μ_A) and the item match distribution (denoted μ_I) based on each of our participant's RK data. If the TODAM model accurately estimates item and associative strength from RK data, μ_A should more strongly predict associative performance than μ_I. That is, both R judgments and associative recognition judgments should be driven by the same associative information created by convolving items during learning.

Fig. 7 shows scatter plots of the relationship between item and associative strength estimates and associative recognition performance. Both estimates are significantly correlated with associative performance. The correlation coefficient for the associative strength estimate is higher than the coefficient for the item strength estimate, but the difference is slight. Thus, the results for this model recapitulate our findings with STREAK: parameters repre-

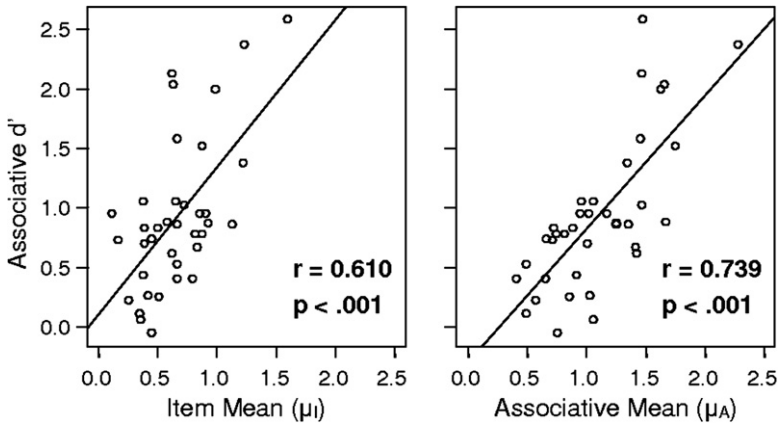

Fig. 7. Relationship between the item (μ_I) and associative (μ_A) strength estimates in the TODAM RK model and associative recognition performance. Correlation coefficients and associated *p*-values appear in the bottom right corner of each plot.

senting different types of memory evidence are not differentially predictive of associative recognition performance, which casts doubt on the claim that the model separately estimates item and associative information from RK data.

The dual-process model

We have focused on models assuming continuous forms of memory evidence, whether the evidence is univariate or bivariate. In contrast, the dual process model proposed by Yonelinas (1994) characterizes recollection as a threshold process that either succeeds or fails for a given item, where successful recollection supports complete certainty that the item was studied. On trials where recollection fails, participants evaluate a continuous familiarity value drawn from a normal distribution. When fit to confidence rating data, the model yields separate estimates of the probability of recollection (Rec) and the familiarity gained by targets as a result of their presentation in the study list (d'). If the threshold assumption better characterizes the recollection process than the continuous specific evidence in STREAK, then the dual process model may produce parameter values that are more consistent with the associative recognition data than the strength estimates from STREAK. Recollection should play a larger role than familiarity on the associative test (Yonelinas, 1997), so recollection estimates should be more strongly related to associative performance than familiarity estimates.

We fit the dual process model to the confidence rating data from each subject. Fig. 8 shows the relationship between d' and Rec estimates and associative recognition. Both model parameters are significantly correlated with associative performance, and the correlation coefficient for the Rec parameter (.41) is actually nominally lower than the correlation for the d' parameter (.52). As with STREAK and the TODAM model, the dual process model does not relate to associative performance in the manner expected given the constructs purportedly measured by its parameters. Ratcliff, Van Zandt, and McKoon (1995) note that confidence rating ROCs are inconsistent with the dual process model's characterization of recollection and familiarity processes. Concordantly, our results show that the dual process model fails to produce meaningful estimates of recollection and familiarity when applied to confidence rating data.

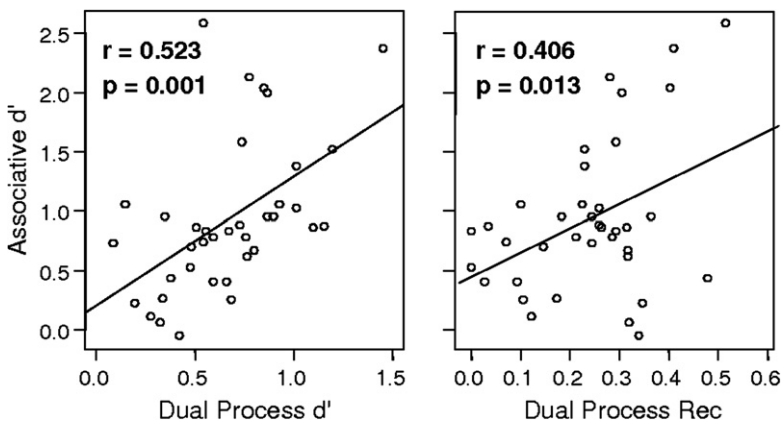

Fig. 8. Relationship between familiarity (d') and recollection (Rec) estimates from the dual process model and associative recognition performance. Correlation coefficients and associated *p*-values appear in the top left corner of each plot.

The dual process model was originally proposed to accommodate rating scale data (Yonelinas, 1994), but Rotello et al. (2006) extended the model to simultaneously predict RK and confidence scale judgments. We evaluated the fit of this extended model to more thoroughly explore the dual process perspective. The model has seven parameters: the probability of veridical recollection (Rec), the probability of false recollection (FRec), the proportion of recollected trials given a rating of 4 (R_4), familiarity based discriminability (d'), and three response criteria. The false recollection parameter is needed to accommodate R responses to lures. False recollection affects responding for all item types, so the predicted proportion of R's for lures equals FRec and the corresponding value for targets equals the sum of Rec and FRec. On all recollection trials (false or veridical), a 4 response is made with probability R_4 and a three response is made with probability $(1 - R_4)$. In the standard dual process model, the recollection process is always associated with the highest available confidence rating, corresponding to $R_4 = 1$. However, this assumption must be relaxed if R responses are distributed across confidence. Some of our participants made a high proportion of R responses following ratings of 3 on the confidence scale, so we fit a model with R_4 as a free parameter. When recollection fails, a rating response is made by comparing a familiarity value to the response criteria as in traditional SDT, and a K response is made following all "old" ratings.

We fit the extended dual-process model to each individual's data from both experiments. The mean chi-squared across participants was 16.37 compared to 3.88 for the univariate model and 6.56 for STREAK. Thus, the fit of the dual-process model was much worse than either signal detection approach. Only 6 of the 37 participants were better fit by the dual process model than by STREAK, and only 1 participant was better fit by the dual process model than by the univariate model. Moreover, the relationship between associative recognition performance and the Rec and d' estimates from the extended model was similar to the relationship seen with the standard dual-process model fit to the confidence rating data. Associative performance had a .55 correlation with Rec and a .58 correlation with d'. Thus, the extended dual-process model mirrors the problematic results of the traditional dual-process model.

General discussion

Our results strongly favor the univariate model over STREAK. The univariate model provided a superior fit at both the group and individual subject levels. The model recovery simulations suggest that STREAK and the univariate model have comparable flexibility in data fitting, so the superior fits for the univariate model indicate that this model accurately captures the processes that generated the data. The relationships between associative recognition performance and each of the memory strength parameters also support the univariate model. μ, d_X, and d_Y all bore a similar strong relationship to associative performance. This was the expected result if confidence and RK decisions were both based on a single conglomerate of evidence, some of which overlapped with the evidence driving associative recognition. In this situation, differences in the univariate strength would modulate the μ parameter of the univariate model, and d_X and d_Y from STREAK would both track the univariate strength as displayed in Fig. 3. This explains why all three parameters would be roughly equally related to associative recognition performance. STREAK's prediction that d_Y should more strongly predict associative performance than d_X received no support; in fact, the correlation coefficient for d_Y was nominally lower.

Critically, the success of the univariate model does not imply that qualitative differences in memory evidence do not exist or are not important. Wixted (2007) recently alleviated much theoretical confusion by clearly separating univariate models from single process memory theories. Regardless of the number of qualitatively distinct processes that influence memory decisions, a univariate model will be appropriate so long as all available evidence is integrated on each trial (see Ratcliff, 1978, p. 62 for a concordant claim). The strength estimates from the univariate model represent the conglomerate of evidence available from all of the memory processes contributing to the decision. Good fits of the univariate model to RK data suggest that participants integrate many different forms of evidence for both recognition and RK decisions. They do not begin to separate different forms of evidence simply because they are given RK instructions.

STREAK failed to produce interpretable estimates of global and specific evidence, but this model is only one of a large class of possible multidimensional RK models. Other instantiations of the multidimensional approach may be more successful; that is, they may both fit the data and yield parameters that meaningfully relate to other memory tests, such as associative recognition. Thus, one should not reject the notion that multiple forms of evidence separately contribute to RK decisions based on the failure of a single model. Notably, we also evaluated two other approaches that posit separate sources of information: Murdock's (2006) TODAM model of RK data and the dual process model for recognition ROCs (Yonelinas, 1994). Much like STREAK, both of these models failed to produce parameter values that were consistent with associative recognition performance. We suggest that all of these models fail because they attempt to estimate two sources of evidence from decisions that are actually based on univariate evidence.

The close fits produced by the univariate model can be interpreted as evidence against the multivariate approach in general. If RK data were truly generated by a multivariate process, the univariate model should fail to fit certain aspects of the data. For example, our recovery simulations show that the univariate model often fails to provide a close fit to data generated from STREAK: 61% of the fits exceeded the chi-squared cutoff for model rejection at the .05 level. In contrast, when real data from individual participants were evaluated, the univariate model consistently provided acceptable fits. Concordant results are reported by Cohen et al. (in press), who also compared the univariate approach with approaches assuming multiple forms of evidence, such as STREAK and the dual-process model. They performed extensive simulations to define the fitting results expected if each alternative model truly character-

ized the processes involved in RK judgments. Given the simulation results, they were able to show that past studies clearly supported the univariate model over models proposing multiple forms of evidence (Dougal & Rotello, 2007; Rotello & Macmillan, 2006; Rotello et al., 2006). Finally, Dunn (2008) used state-trace analysis to garner evidence that RK judgments reflect unidimensional evidence values. Notably, this technique compares the entire class of unidimensional models to the entire class of multidimensional models; thus, Dunn reports direct evidence that the univariate approach best characterizes RK data.

One major challenge to the univariate approach was highlighted by Rotello et al. (2004), who argue that zROC slopes from confidence versus RK data are inconsistent with predictions from the univariate model. When ROCs are converted to *z*-scores (i.e., the *z*-score of the hit rate is plotted on the *z*-score of the false-alarm rate), univariate signal detection theory predicts a linear function with a slope equal to the ratio of the standard deviations of the lure and target evidence distributions. In an extensive review of RK studies, Rotello et al. found that zROC slopes from RK data were considerably higher than would be expected based on confidence rating slopes. Specifically, confidence rating slopes consistently have values near or less than .8 (Ratcliff et al., 1992, 1994), but the RK slopes in the studies evaluated by Rotello et al. had a mean value of 1.01. Thus, the variability in target and lure evidence appears to be roughly equal when RK data are considered, but target evidence appears to be more variable than lure evidence when confidence data are considered. Rotello et al. suggest that this result is inconsistent with the claim that RK and confidence responses are based on the same evidence dimension, as assumed by the univariate model. However, Wixted and Stretch (2004) showed that slopes can differ based on variation in response criteria from trial to trial. zROC slope approaches 1 as criteria variability increases, so slopes may be higher based on RK data because participants are less able to maintain a stable RK criterion than they are to maintain stable confidence criteria.

We used a version of the univariate model with separate free parameters for the standard deviation of confidence and RK criteria. To best fit the empirical data, the model produced higher variability estimates for the RK criterion than for the confidence criteria (e.g., .64 versus .13 in Experiment 1, see Table 2). That is, the data are consistent with a process in which the amount of evidence needed to make an R response changes substantially from trial to trial, whereas the standards used for the confidence responses are more stable. The difference in variability between the confidence and RK criteria was highly significant, $t(36) = 3.62$, $p < .001$, and the average chi-squared from a model in which the two types of criteria were constrained to be equally variable (11.06) was nearly three times higher than the model with separate variability parameters (3.88). Thus, the pattern that Wixted and Stretch (2004) proposed to explain the higher zROC slopes in the Rotello et al. (2004) review is precisely the pattern that emerges in fits to data. This suggests that the slope results do not provide strong evidence against the univariate model.

A recent model that predicts both response proportions and response times for the recognition confidence procedure suggests that researchers should mistrust direct interpretation of zROC slopes (Ratcliff & Starns, submitted for publication), which provides another reason not to regard the Rotello et al. (2004) results as problematic for the univariate model. Ratcliff and Starns showed that, once mechanisms for predicting response times are incorporated in a model of confidence judgments, zROC slope no longer directly estimates variability in memory evidence distributions. zROC slope is also sensitive to the placement of decision criteria in the response time model, which means that slope can be affected by variables imparting response biases (for concordant empirical results, see Van Zandt, 2000). Thus, from the standpoint of this model, confidence and RK slopes could differ for a number of reasons that have nothing to do with memory evidence. Notably, the Ratcliff and Starns model assumes univariate distributions of memory evidence but can produce accuracy patterns that have been interpreted as reflecting multiple sources of memory evidence, such as non-linear zROCs (Yonelinas, 1994, 1997). Thus, the model demonstrates the power of the univariate approach and shows that this approach succeeds even when faced with the additional constraints imposed by response time data. The ability of models proposing multiple forms of memory evidence to meet these standards remains in question.

Memory researchers commonly take RK judgments in an attempt to isolate memory processes that are conflated on a standard recognition test. Together with other findings supporting the univariate model (Donaldson, 1996; Dougal & Rotello, 2007; Dunn, 2004; Hirshman & Master, 1997; Rotello & Macmillan, 2006; Rotello et al., 2006; Wixted & Stretch, 2004), our results suggest that RK judgments are produced in the same way as recognition decisions; that is, a univariate evidence variable is compared to a set of criteria demarcating response regions. Thus, RK judgments provide no more of a theoretical foothold than recognition confidence judgments. Hirshman, Lanning, and Master (2002) suggest that RK judgments should be evaluated as reflecting separate memory processes only when a univariate model can be rejected. Our results suggest that not only can the univariate model successfully accommodate RK data, but also that it outperforms comparable models that posit separate processes for recognition and RK decisions.

References

Cohen, A. L., Rotello, C. M., & Macmillan, N. A. (in press). Evaluating models of remember–know judgments: Complexity, mimicry, and discriminability. *Psychonomic Bulletin & Review*.

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.

Egan, J. P. (1958). *Recognition memory and the operating characteristic* (Tech. Note AFCRC-TN-58-51). Hearing and Communication Laboratory: Indiana University.

Hirshman, E., Lanning, K., & Master, S. (2002). Signal-detection models as tools for interpreting judgments of recollections. *Applied Cognitive Psychology, 16*, 151–156.

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

Johnston, W. A., Dark, V. J., & Jacoby, L. L. (1985). Perceptual fluency and recognition judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*, 3–11.

Kucera, H., & Francis, W. (1967). *Computational analysis of present-day American English*. Brown University Press: Providence, RI.

Murdock, B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review, 89*, 609–626.

Murdock, B. (2006). Decision-making models of remember–know judgments: Comment on Rotello, Macmillan, and Reeder (2004). *Psychological Review, 113*, 648–655.

Nelder, J. A., & Meade, R. (1965). A simplex method for function minimization. *Computer Journal, 7*, 308–313.

Pitt, M. A., Kim, W., & Myung, I. J. (2003). Flexibility versus generalizability in model selection. *Psychonomic Bulletin & Review, 10*, 29–43.

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

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*, 59–108.

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

Ratcliff, R., Sheu, C-F., & Gronlund, S. (1992). Testing global memory models using ROC curves. *Psychological Review, 99*, 518–535.

Ratcliff, R., & Starns, J. J. (submitted for publication). Modeling confidence and response time in recognition memory.

Ratcliff, R., Van Zandt, T., & McKoon, G. (1995). Process dissociation, single-process theories, and recognition memory. *Journal of Experimental Psychology: General, 124*, 352–374.

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.

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

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

Wagenmakers, E. J., Ratcliff, R., Gomez, P., & Iverson, G. J. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology, 48*, 28–50.

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

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