Revisiting the Remember-Know Task: Replications of Gardiner and Java (1990)

Julia M. Haaf[1], Stephen Rhodes[1], Tony Sun[1], Hope K. Snyder[1], Moshe Naveh-Benjamin[1], & Jeffrey N. Rouder[2]

[1] University of Missouri

[2] University of California, Irvine

Author Note

Julia M. Haaf, Stephen Rhodes, Tony Sun, Hope K. Snyder, and Moshe Naveh-Benjamin, Department of Psychological Sciences, University of Missouri, Columbia, MO, USA. Jeffrey N. Rouder, Department of Cognitive Sciences, University of California, Irvine, CA, USA. JH, SR, MNB, and JR planned the study; JH, SR, TS, HK and JR designed the experiments; JH and SR analyzed the data; all authors contributed to the manuscript. We thank Ashley M. Meierhofer and Carson Burke for assistance with data collection.

Correspondence concerning this article should be addressed to Julia M. Haaf, 210 McAlester Hall, Columbia, MO, USA, 65203. E-mail: jhaaf@mail.missouri.edu

Abstract

Perhaps the most evidential behavioral result for two memory processes comes from Gardiner and Java (1990). Participants provided more remember than know responses for old words but more know than remember responses for old nonwords. Moreover, there was no effect of word/nonword status for new items. The combination of a crossover interaction for old items with an invariance for new items provides strong evidence for two distinct processes while ruling out criteria or bias explanations. Here, we report a modern replication of this remarkable study. In two experiments with larger numbers of items and participants, we were unable to replicate the stunning crossover. Instead, our data are more consistent with a single-process account. In a third experiment, we were able to replicate Gardiner and Java's baseline results with a sure-unsure paradigm supporting a single-process explanation. It seems that Gardiner and Java's remarkable crossover result is not replicable.

Keywords: Recognition Memory, Implicit Memory, Replication

Word count: 1,902 words for Introduction (including introductions of experiments), Discussion of Experiment 1, and General Discussion. Excluding Statistical Models for Data Analysis, methods, and result sections.

Revisiting the Remember-Know Task: Replications of Gardiner and Java (1990)

One major feature in the modern study of memory is a healthy respect for the distinction between different mnemonic processes. Perhaps no distinction has had more impact than that between conscious recollection of memories and automatic activation of familiar traces (Atkinson & Juola, 1973; Jacoby, 1991; Mandler, 1980; Yonelinas, 2002). This distinction is influential in the neurobiology of memory (Squire, 1994; Vilberg & Rugg, 2008), in understanding cognitive aging (Jennings & Jacoby, 1993, 1997; Prull, Dawes, Martin, Rosenberg, & Light, 2006), and in memory pathology research (Yonelinas, Kroll, Dobbins, Lazzara, & Knight, 1998).

Perhaps the simplest way of experimentally probing this distinction comes from Tulving (1985), who proposed that experimentalists simply ask participants *how* they recalled memoranda. Participants are first given a study list and are subsequently presented with a test list consisting of previously studied (old) and new items. For each test item, participants are asked to indicate whether it is new or old, and if they indicate that it is old, they are further asked whether they *remembered* this item or *knew* it. The endorsement of a *remember* response indicates a conscious, recollective recall, and the endorsement of a *know* response indicates automatic activation based on familiarity. This task is commonly called a *remember-know task*.

The remember-know task may be combined with experimental manipulations to validate the claim that it measures distinct memory processes. Gardiner (1988), for example, used a levels-of processing manipulation with the notion that deeply processed items are more likely to be consciously recollected. Table 1 shows the results from Gardiner (1988, Experiment 1). Indeed, only *remember* responses are affected by the processing depth manipulation.

Even though the above results are impressive, there are critiques. The one we find most persuasive originated with Donaldson (1996) and was further developed by Dunn (2004). Accordingly, the remember-know judgement cannot be considered a direct measure of

Table 1
Response proportions from Gardiner (1988, Exp. 1)

	Response		
	Remember	Know	New
Condition			
Deep	0.65 (0.66)	0.17 (0.15)	0.18 (0.19)
Shallow	0.35 (0.34)	0.17 (0.19)	0.48 (0.47)
Lure	0.05 (0.05)	0.07 (0.07)	0.88 (0.88)

Note. Predictions from a single-process model are provided in parentheses.

memory processes. Instead, a proper interpretation must account for the influence of decision processes. Key is the role of bias or criterial settings. An alternative to dual-process accounts is a single-process, signal-detection account where *remember* and *know* responses reflect different criteria on latent mnemonic strength. A manipulation at study may affect the criteria perhaps even as much as they affect the underlying memory process. Indeed, Dunn (2004) shows how the large corpus of remember-know results like those in Table 1 may be accommodated by a single-process signal-detection model. The predictions from a six-parameter signal-detection model for the Gardiner (1988) data are given in parentheses in Table 1. As can be seen, these predictions are fairly accurate given the constraint of the model. The signal-detection analysis is provided in the Online Supplement.

There are, however, remember-know results that seem immune to the Donaldson-Dunn critique. Perhaps the best example comes from Gardiner and Java (1990), and the results are shown in Table 2. Gardiner and Java had participants study words and word-like nonwords. At test, they judged four types of items: old words, old nonwords, new words, and new nonwords. Gardiner and Java hypothesized that words would better elicit recollection than nonwords.

The adoption of this 2×2 design is crucial as including both new words *and* new nonwords provides a means of comparing single- and dual-process accounts. The

Table 2
Response proportions from Gardiner and Java (1990, Exp. 2)

	Response		
	Remember	Know	New
Condition			
Old Word	0.28 (0.26)	0.16 (0.2)	0.56 (0.55)
Old Nonword	0.19 (0.19)	0.3 (0.3)	0.51 (0.51)
New Word	0.04 (0.05)	0.11 (0.09)	0.85 (0.86)
New Nonword	0.03 (0.03)	0.12 (0.12)	0.85 (0.85)

Note. Predictions from a single-process model are provided in parentheses.

dual-process account states that old words should elicit a greater proportion of *remember* responses than old nonwords. Moreover, if this effect reflects the study of old items, there should be no effect of lexical status for new items. If observed effects for old items are due to criterial shifts, i.e. people may respond *remember* to words more so than nonwords, then effects should be observed for old and new items alike.

Gardiner and Java's (1990) results, seen in Table 2, are stunning. For the old items, there is a perfect crossover with a greater proportion of *remember* responses to old words and a greater proportion of *know* responses to old nonwords. For new items, there is seemingly no effect of lexical status. The lack of effect for new items implies that lexical status does not affect the decision criteria. In our view, these results are perhaps the strongest of all remember-know results that we know of because they implicate two mnemonic processes while ruling out the most plausible alternative. Importantly, a signal-detection model cannot account for the data pattern (the predictions from such a model are given in parentheses in Table 2). Even though the misses are small they are nevertheless problematic because the model is so flexible and so heavily parameterized.

Because Gardiner and Java's (1990) findings have the potential to distinguish between single- and dual-process predictions, and because their results seem to perfectly fit the

dual-process model, we decided to perform a preregistered replication study across two different labs.[1] In our Experiments 1 and 2, we attempted close replications of Gardiner and Java's remember-know task with words and nonwords (Gardiner and Java's Experiment 2). In our Experiment 3, we replicated Gardiner and Java's Experiment 3, which replaced the remember-know instructions with sure-unsure instructions. The replication attempt spanned the labs of Rouder and Naveh-Benjamin, who have somewhat opposing views on the usefulness of the distinction between recollection and familiarity. In previous publications, Naveh-Benjamin and colleagues have leveraged the explanatory power of this distinction (e.g. Naveh-Benjamin et al., 2009; Old & Naveh-Benjamin, 2008) while Rouder and colleagues have been skeptical (Pratte & Rouder, 2011, 2012).

Statistical Models for Data Analysis

The striking elements of Gardiner and Java's (1990) Experiment 2 data are the perfect crossover for old items in conjunction with no effect for new items. While it is clear that this data pattern supports a dual-process interpretation, it is unclear in general which possible data patterns would support or, alternatively, contradict a dual-process interpretation. In our view, identifying these patterns *before* data collection is key for a replication study. The hypotheses and models presented subsequently were documented in our preregistration document, and to us, are the most valuable part of the preregistration.

In Gardiner and Java's data analysis, they conducted two separate ANOVAs, one for old items and one for new items treating lexical status and remember/know choice as factors. The result was a significant interaction for old items and a non-significant interaction for new items. This analysis, however, is an unprincipled test of perfect crossover for old items in conjunction with a lack of effects for new items. Additionally, one cannot use ANOVA to specify and test alternative, theoretically meaningful patterns in the data. We therefore decided to assess evidence for or against dual-process theory in a Bayesian model comparison

[1] Preregistration of Experiment 1 can be found here: https://osf.io/873sg/; Experiments 2 and 3 are preregistered at https://osf.io/k2ve3/.

framework, most similar to Klugkist, Laudy, and Hoijtink (2005).

The first step for the analysis is to identify Gardiner and Java's (1990) hypotheses. We identify two hypotheses used in conjunction. Their first hypothesis, denoted H_1, is that *remember* responses would be more prevalent for old words than old nonwords. Their second hypothesis, denoted H_2, is that *know* responses would be more prevalent for old nonwords than for old words. Though not explicitly hypothesized, Gardiner and Java tested the lack of effect of lexical status for new items. We call this null hypothesis H_3. These three hypotheses are not exclusive, and all three were observed in their data.

Unfortunately, H_1 and H_2 are not independent: Any endorsement of a *remember* response necessarily implies a lack of endorsement of a *know* response leading to a negative correlation between *remember* and *know* response rates. This negative correlation was not accounted for by Gardiner and Java. We account for this negative correlation in our subsequent analysis.

These hypotheses may be conveniently specified in statistical models. We start with a general data representation. Let r_{ij}, k_{ij}, and n_{ij} denote the number of *remember*, *know* and *new* responses, respectively, for the ith participant and jth item type, $i = 1, \ldots, I$ and $j = 1, 2, 3, 4$, where the 4 item types in order are: old words, old nonwords, new words, and new nonwords. Hence $r_{ij} + k_{ij} + n_{ij}$ sums to the number of tested items for the ith participant in the jth condition. To account for the negative correlation, we consider a single measure per condition. We call this single measure the *scaled difference score*, denote it as Y_{ij}, and define it as the scaled difference between *remember* and *know* responses:

$$Y_{ij} = (r_{ij} - k_{ij})/(r_{ij} + k_{ij} + n_{ij}). \tag{1}$$

The value of the scaled difference is negative when *know* responses are preferred and positive when *remember* responses are preferred. The following is a set of statistical models on this scaled difference, Y_{ij}.

The most general model, the *unconstrained model*, is

$$\mathcal{M}_u : \quad Y_{ij} \sim \text{Normal}(\mu_j, \sigma^2),$$

where μ_j is the true mean scaled difference for the jth condition, and σ^2 is the variance. We can now place constraints on the four μ_js. The first model instantiates H_1, H_2 and H_3 simultaneously with the following restrictions:

$$\begin{aligned} \mathcal{M}_* : \quad & \mu_1 > \mu_2, \\ & \mu_3 = \mu_4. \end{aligned}$$

The inequality constraint corresponds to the higher prevalence of *remember* responses for old words than to old nonwords. The equality corresponds to the lack of effect of lexical status for new items. This model may be compared to the unconstrained model that does not impose any ordering restrictions on the collection of μ_j. In addition to the unconstrained model, we propose the following alternatives to competitively test model $\mathcal{M}_*$ against. The first model $\mathcal{M}_1$ captures the case that the lexical status has no effect on the scaled difference for old or new items:

$$\mathcal{M}_1 : \quad \mu_1 = \mu_2,\ \mu_3 = \mu_4.$$

The second model $\mathcal{M}_2$ captures the case that words, regardless of being old or new, enhance *remember* responses over *know* responses:

$$\mathcal{M}_2 : \quad \mu_1 > \mu_2,\ \mu_3 > \mu_4.$$

The third model $\mathcal{M}_3$ captures the opposite case that nonwords, regardless of being old or new, enhance *remember* responses over *know* responses.

$$\mathcal{M}_3 : \quad \mu_1 < \mu_2,\ \mu_3 < \mu_4.$$

To implement these models in a Bayesian framework, prior distributions are needed on the four μ_j and σ^2. We follow Haaf, Klaassen, and Rouder (In preparation) and Rouder, Morey, Speckman, and Province (2012) for prior settings using a g-prior approach. The critical setting here is the scale on g, and we used a default setting of $r = \sqrt{2}/2$. With this setting, model comparison with Bayes factors is straight-forward using analytic solutions (Rouder et al., 2012) and the encompassing approach (Klugkist et al., 2005). For the analysis, we used the `BayesFactor` package in `R` (Morey & Rouder, 2015). The code is provided at github.com/PerceptionAndCognitionLab/rm-gardiner-java.

Experiment 1

The goal of Experiment 1 was to closely replicate Gardiner and Java's Experiment 2. Even so, we decided to improve the experimental methods in four ways outlined subsequently.

Methods

In their Experiment 2, Gardiner and Java (1990) showed 20 participants 15 words and 15 nonwords on handwritten cards, sequentially, for 2 seconds each. Then after a 24 hour delay, participants were given a recognition test. Sixty items, again handwritten, were presented on a single piece of paper. These 60 consisted of 15 old words; 15 old nonwords; 15 new words; and 15 new nonwords. Participants were instructed to circle old words and then write "R" or "K" next to the item to indicate whether the old-item response reflect recollecting or knowing.

Here are the ways our experiment differed from Gardiner and Java. 1. To the best of our knowledge, the original materials are not available (we were not able to contact the authors). We therefore used different words and nonwords that were constructed following Gardiner and Java's generation rules. 2. Instead of a 24 hour retention interval, we used a 10 minute retention interval filled with a distractor task. The reason for this change is as follows: Gardiner and Java explicitly justify their 24-hour interval as a means of lowering performance to avoid ceiling effects (p. 24).

We decided that the better way to lower performance was to ask participants to remember more items. With more items, the statistical properties of the experiment increase and the experimenter has greater resolution to detect differences if they exist and greater confidence in null results otherwise. Moreover, asking participants to return is inconvenient and may result in the loss of some participants introducing a new bias into the sample. Hence, using a 10-minute delay with more items—in our case we doubled the number of to-be-remembered and to-be-judged items—is a preferred approach on all accounts to avoiding ceiling effects. 3. We increased the number of participants from 20 to 52 and doubled the number of items at study and at test. This increase of the number of observations results in a much better resolution of the data. 4. We used computer presented items rather than handwritten ones. Both study and test were performed in a sequential manner rather than simultaneously.

Participants and design. We initially planned to recruit 50 undergraduate students. In total, 53 undergraduates were recruited at the University of Missouri and participated for partial course credit. One participant was excluded from analysis due to overall performance below chance (accurate response in less than 50% of the trials). The study has a 2 (words vs nonwords) x 2 (old vs. new items) repeated measures factorial design, resulting in a total of $53 \times 2 \times 2 \times 30 = 6360$ collected observations.

Material. Criteria for material selection were taken from Gardiner and Java's Experiment 2. Sixty high familiarity concrete nouns with one syllable and four letters were taken from the MRC psycholinguistics database (Coltheart, 1981). Sixty pronounceable nonwords with four letters and two to four phonemes were selected from the ARC nonword database (Rastle, Harrington, & Coltheart, 2002). The words and nonwords used in this study are shown in Appendix B. In the original study, the authors formed two fixed study sets of 15 words and 15 nonwords and randomly selected one of the lists for each participant. In our study, 30 words and 30 nonwords were chosen at random to form the study set for each participant. Items were presented at the center of the screen in the Lucida Console font

with a height of 2° of visual angle at an approximate viewing distance of 50 cm. At test, 60 words and 60 nonwords were sequentially presented in a random order. Items were shown in the center of the screen together with two buttons (either labeled *OLD* and *NEW* or *R* and *K*, see below) that they could click on to respond. The buttons are circular with a radius of 2° and are presented 5° below and 5° to the left and right of the center of the screen.

Procedure. During the study phase, participants studied 60 items (30 words and 30 nonwords) in a randomly determined order. Each item appeared on the screen for 2 seconds (as in the original study) followed by a 0.5 second inter-stimulus-interval. The test phase followed after a ten minute retention interval. During the retention interval, participants were given a "spot-the-difference" task to complete before moving on to the recognition test. For this task, participants were asked to compare two pictures with small changes between them and circle these changes. Afterward, participants were given instructions for the recognition test phase. The instructions were presented on several screens and are provided in the online supplement. After the instructions were given on the screen, the experimenter gave a few examples of when *remember* and *know* responses may be appropriate. This approach was also used by Gardiner and Java (1990), but the exact examples they used could not be employed as they were not reported. During the recognition test participants were presented with items one at a time and characterized each item as *old* or *new* using the mouse to click on the corresponding button on the screen. Following an *old* response participants then made an additional remember-know judgment by using the mouse to click on buttons labeled *R* (for *remember*) or *K* (for *know*).

Results

Data were *born open* (Rouder, 2016), that is, they were uploaded to a public repository nightly during data collection, and are available at github.com/PerceptionCognitionLab/data1/tree/master/repGardinerJava. Average response proportions are shown in Table 3. Average accuracy on the old/new task was between 61%

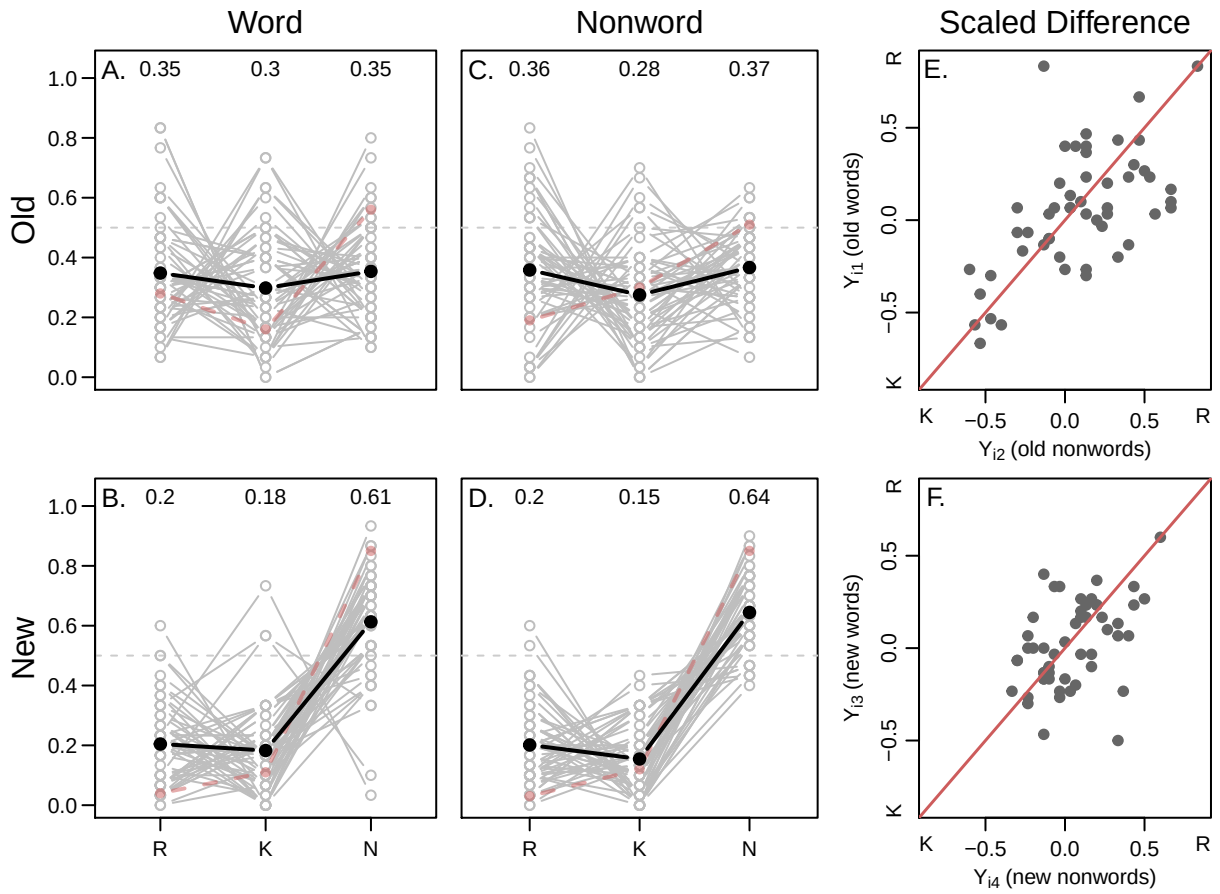

Figure 1. Results from Experiment 1. The dark lines shows average response rates for all participants; dashed lines show average response rates from Gardiner and Java's (1990) Experiment 1. Critically, there is no interaction between item type (i.e. word vs. nonword) and preferred response category (i.e. Remember vs. Know) for the replication data. The right two panels show the modeled scaled difference scores for nonwords relative to words. According to dual-process theory, the scaled difference scores should be above the diagonal for old items and on the diagonal for new items.

and 65% in all four conditions. This accuracy value is just a tad lower than the average accuracy, 66%, in Gardiner and Java's Experiment 2. All-in-all, our 10 minute retention period coupled with a doubling of items resulted in an overall performance level that was comparable to that from Gardiner and Java.

Descriptive Analysis. Participants in our study displayed far less bias than those in Gardiner and Java's. In our experiment, hit rates (0.63) and correct-rejection rates (0.64) are about the same in value indicating no particular bias to say old or new. This relative lack of bias contrasts to extreme bias in Gardiner and Java. In their experiments, hit rates

were low (0.47) while correct-rejection rates were high (0.85).

To assess the data pattern critical for the replication, we focus on proportions of *remember* and *know* responses as shown in Figure 1. The black lines in panels A.-D. show average response proportions. The two left panels show response proportions to old and new words, and the two right panels show response proportions to old and new nonwords. The original results by Gardiner and Java (1990) is shown by the dashed line. The critical comparison is between the left and right panels of each row. The expected data pattern for a successful replication of Gardiner and Java (1990) would show the following two signatures: 1. A marked difference between the left and right panels of the top row. In particular, recollection responses should be higher for old words than old nonwords and the reverse for know responses. 2. No differences between the bottom left and bottom right panels; that is, there should not be an effect of lexical status for new items. We did not observe the first signature. The top left-panel appears to be the same as the top-right panel. The invariance between the left and right panels indicates that there is no effect of lexical status on responses for old or new items. Nonwords seemingly act like words.

It may seem surprising that there is no effect of lexical status. However, note that Gardiner and Java (1990) also failed to find a main effect of lexical status (see Table 2). Instead, their analysis showed a perfect crossover interaction of lexical status and response category (*remember* vs. *know*). However, in the current study, there is no apparent interaction, let alone the stunning crossover.

On an average level, there is no differential preference for either *remember* or *know* responses across old and new items. Yet, individuals' response proportions vary drastically as shown by the grey lines in Figure 1. Some participants almost exclusively use *remember* responses to classify old items while others almost exclusively use *know* responses to classify old items This variability of preferences may have various explanations, one of them being that participants are not able to consistently classify their mnemonic experience as *remember* or *know*. We return to this issue when discussing Experiment 2, which aimed to better

instruct participants on the criteria for remember and know responses.

Model-based Analysis. To quantify the evidence for or against the replication we use the model-based approach explained previously. Figure 1 panels E-F show Y_{ij}, each individual's scaled difference between remember and know responses for the four item types (old words and old nonwords in panel E; new words and new nonwords in panel F). As a reminder, these scaled differences can be interpreted as the bias for *remember* responses compared to *know* responses. If an individual experienced differing processing for words and nonwords as proposed in Gardiner and Java (1990), we should observe positive scaled difference values for old words and negative values for nonwords in panel E. Yet, the scaled differences across conditions are highly correlated suggesting a more global bias to one of the two response options.

The data in Figure 1E-F are submitted to the model analysis, and the replication model, $\mathcal{M}_*$, is compared to alternative accounts using Bayes factor model comparison. The preferred model is Model $\mathcal{M}_1$, the model representing a straight-forward single-process criterion shift account. According to the model, proportions of *remember* and *know* responses are about the same for words and nonwords. Model $\mathcal{M}_1$ is preferred over the replication model $\mathcal{M}_*$ by 5.65-to-1. The second-best performing model is model $\mathcal{M}_3$ with a Bayes factor of 4.77-to-1 in favor of the winning model. The least preferred model is model $\mathcal{M}_2$ with a Bayes factor of 22.30-to-1 in favor of the winning model.

In summary, we were not able to replicate the data pattern in Gardiner and Java's Experiment 2 (1990). Instead, the Bayesian analysis yields evidence for the alternative model $\mathcal{M}_1$, capturing the case that the lexical status (nonword vs. word) has no effect on the scaled difference of *remember* and *know* responses for both old and new items.

Discussion

There are similarities and differences between our results and Gardiner and Java (1990). Although our participants have the same overall accuracy as Gardiner and Java, they

differ in bias. Our participants displayed no preference for old or new responses while Gardiner and Java's were heavily biased toward new responses. There are two possible differences in the procedure that may have contributed to this difference: 1. we used a sequential presentation at test that reduces dependencies among responses to different items; and 2. we used more items with a shortened retention interval to control overall accuracy. We think the lack of bias is an improvement from a psychometric point of view and have no desire to change our procedure to introduce such bias. We provide context for interpreting the procedural differences in the General Discussion.

There are two smaller concerns with Experiment 1. The first is that the overall accuracy is a bit low. From a statistical point-of-view, it would be more desirable to have accuracy closer to a target of .75.

In order to raise the level of accuracy, in Experiment 2 we slightly lowered the number of studied items from 60 to 50. Consequently, the number of to-be-judged items at test lowered from 120 to 100. The second concern we have is with our instructions for the test phase. Our on-screen instructions were standardized. Participants read these with an experimenter, and then the experimenter provided a few examples. This aspect of the procedure followed Gardiner and Java. However, we did not record the examples, and we cannot guarantee that different participants did not receive different examples. In Experiment 2, we standardized our examples as well as instructions.

Experiment 2

Methods

Participants. For the preregistration, we planned to at least collect 30 participants and up to 50 participants. We decided that Spring break 2018 would be our cutoff: If we collected more than 30 participants by then we would stop data collection; if not, we would continue until the end of the semester. Since all the confirmatory analyses are conducted in a Bayesian framework, optional stopping or data peaking was not considered problematic

(Rouder, 2014). In total, 51 undergraduates were recruited at the University of Missouri and participated for partial course credit. The experiment has the same design as Experiment 1, resulting in a total of $51 \times 2 \times 2 \times 25 = 5100$ collected observations.

Material. Fifty words and nonwords were selected from Experiment 1, and the presentation parameters were identical. The selected words and nonwords are indicated in Appendix B.

Procedure. The general procedure was identical to that used in Experiment 1 with the following changes. Participants studied 50 items (25 words, 25 nonwords) in a random order and were tested on 100 items (50% old, 50% new). A major change was in the instructions presented prior to the recognition phase. Firstly, we felt, following interaction with participants in Experiment 1, that the phrasing of the written instructions reported by Gardiner and Java could be improved. These experiments were reported almost 30 years ago and were conducted on a UK sample. We attempted to make the remember/ know distinction clearer for our younger, US educated participants. The instructions are provided in the online appendix

Results

Data were made public after data collection and are available at github.com/PerceptionCognitionLab/data0/tree/master/rm-gardiner-java. Average response proportions are shown in Table 3. On average, participants performed better for new items with average accuracies of 69% and 78% for new word and new non-word, respectively. For old items, average accuracies remained similar to the levels in Experiment 1 with accuracies of 64% and 63%. Individuals' response proportions are shown in Figure 2.

Descriptive Analysis. Once again, the critical comparison is the comparison of panel A to panel C and panel B to panel D in Figure 2. This comparison yields almost no differences between the relative proportions of *remember* and *know* as a function of lexicality for either old items (top row) or new items (bottom row). Again, there is no sign of the

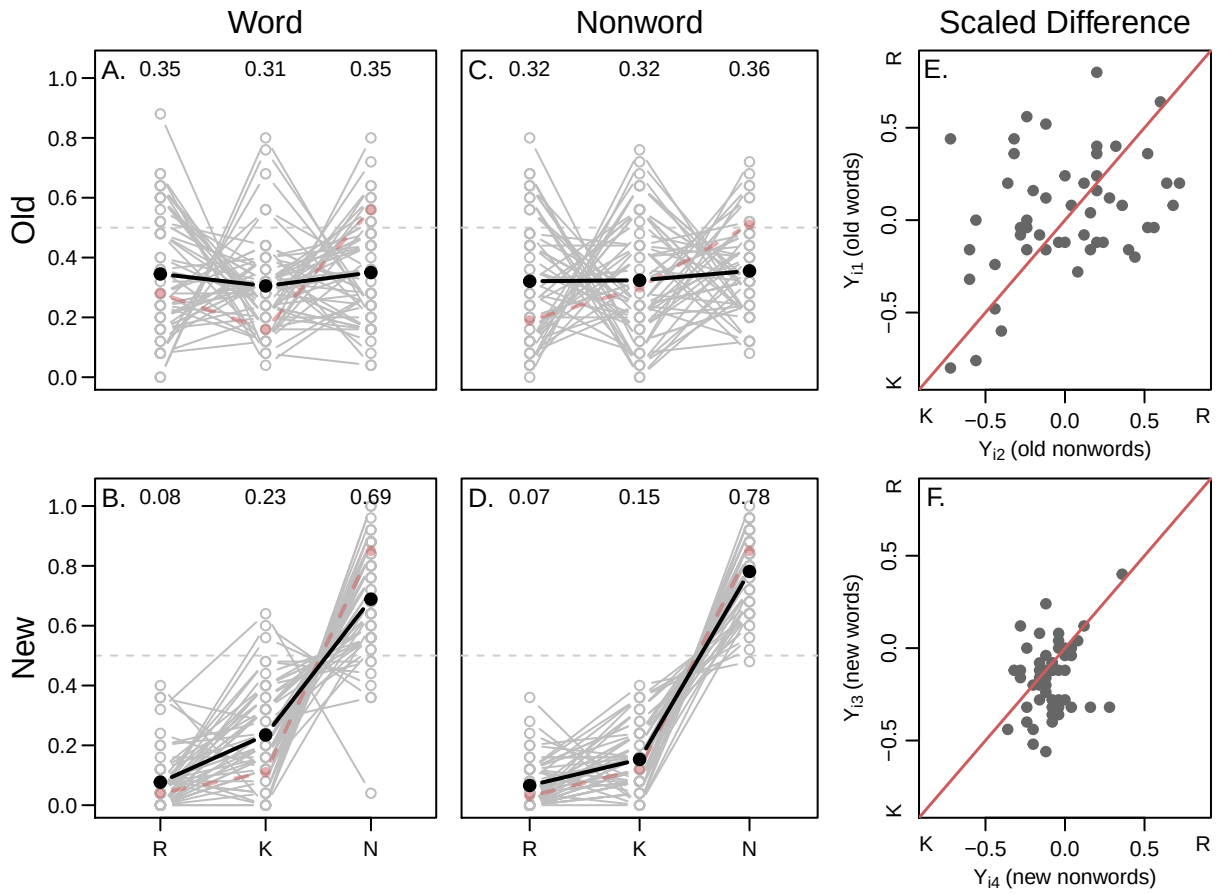

Figure 2. Results from Experiment 2. The dark lines shows average response rates for all participants; dashed lines show average response rates from Gardiner and Java's (1990) Experiment 1. Critically, there is no interaction between item type (i.e. word vs. nonword) and preferred response category (i.e. Remember vs. Know) for the replication data. The right two panels show the modeled scaled difference scores for nonwords relative to words. According to dual-process theory, the scaled difference scores should be above the diagonal for old items and on the diagonal for new items.

prominent crossover interaction of the original study. Additionally, we again find notable individual differences in the preference of either *remember* or *know* responses.

Model-based Analysis. Panels E and F in Figure 2 show the *remember* response bias for nonwords and words. As in Experiment 1, there is no sign that *remember* responses are preferred for old words while *know* responses are preferred for old nonwords. Instead, if anything, participants seem to have stable preferences across conditions.

Bayes factor model comparison again shows a preference model is Model $\mathcal{M}_1$, the model representing a straight-forward single-process criterion shift account. According to the

model, proportions of *remember* and *know* responses are about the same for words and nonwords. Model $\mathcal{M}_1$ is preferred over the replication model $\mathcal{M}_*$, which is the second-best performing model. The Bayes factor between $\mathcal{M}_1$ and $\mathcal{M}_*$ is 2.17-to-1 in favor of $\mathcal{M}_1$. The least preferred model is model $\mathcal{M}_2$ with a Bayes factor of 18.22-to-1 in favor of the winning model.

In summary, the main feature of Experiment 2 is a failure to replicate the stunning data pattern of Gardiner and Java's Experiment 2. In fact, we replicated our Experiment 1 finding in finding in that there is no effect of lexicality on recognition memory. We again found strong individual preferences to either *remember* or *know* responses. This finding may suggest that participants were not able to distinguish between these two distinct mnemonic experiences. To address this concern, we attempted to replicate Gardiner and Java's Experiment 3, where participants are instructed to state the certainty of their *old*-response instead of *remember/know*.

Experiment 3 – Sure vs. Unsure Instructions

Although our focus has been on Gardiner and Java's Experiment 2, these authors ran an additional experiment, their Experiment 3, to show that the crossover interaction was unique to the remember-know instructions, and, by extension, that remember and know can be interpreted as processes distinct from levels of confidence. In our Experiment 3, we aimed at replicating Gardiner and Java's (1990) Experiment 3.

Methods

In their Experiment 3, Gardiner and Java (1990) simply replaced *remember* with *sure* and *know* with *unsure* response options. In line with their expectation they found that, for both words and nonwords, participants responded *sure* more than *unsure* to old items, whereas for new words and nonwords *unsure* was selected more than *sure*. There were no effects of lexicality.

In our Experiment 3, we attempt to replicate Gardiner and Java's Experiment 3 as a demonstration of calibration. If we replicate Experiment 3 of Gardiner and Java (1990) using similar experimental procedures to those in our Experiments 1 and 2, then we have higher confidence that our failure to replicate the more theoretically contentious findings of Gardiner and Java's Experiment 2 is not due to procedural differences. We preregistered and conducted Experiment 3 at the same time as Experiment 2 and without knowing the results of Experiment 2.

Participants. For the preregistration, we stated the same decision rule as for Experiment 2. In total, 51 undergraduates were recruited at the University of Missouri and participated for partial course credit. The experiment has the same design as the previous experiments, resulting in a total of $51 \times 2 \times 2 \times 25 = 5100$ collected observations.

Material and Procedure. The same material as in Experiment 2 was used. The procedure was identical to Experiment 2 with two exceptions. First, participants received different instructions for the test phase guiding them on how to navigate sure/unsure responses. The instructions are provided in the online supplement. After the instructions, participants entered the test phase similar to Experiment 1 and 2. Participants were again presented with items one at a time and characterized each item as *old* or *new* using the mouse to click on the corresponding button on the screen. Following an *old* response participants then made a sure-unsure judgment instead of a remember-know judgment by clicking on buttons labeled S (for *sure*) or U (for *unsure*).

Results

Data were *born open* and are available at github. Average response proportions are shown in Table 3. On average, participants performed similarly for new and old items with average accuracies between 65% and 68%. On an individual level, accuracy varied between 24% and 96% when evaluated per condition. Individuals' response proportions are shown in Figure 3.

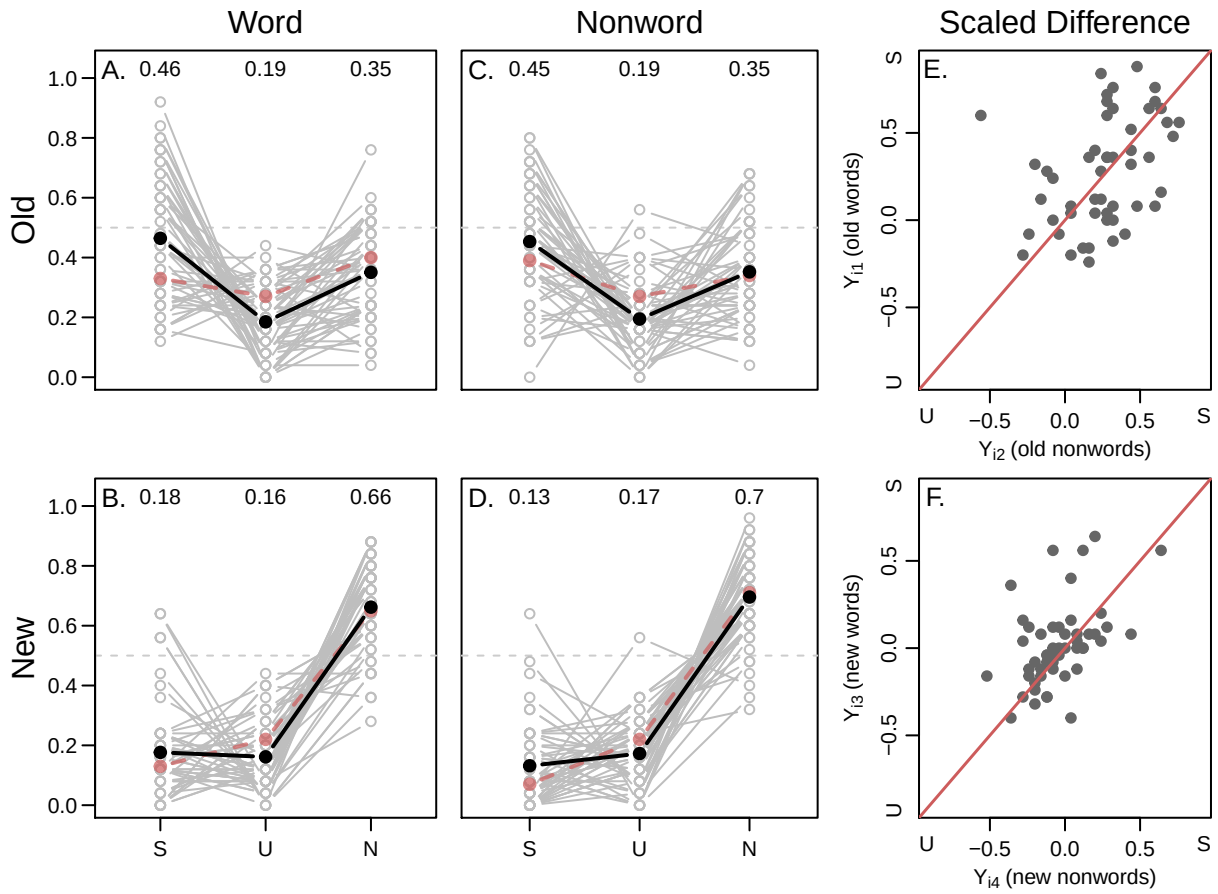

Figure 3. Results from Experiment 3. The dark lines shows average response rates for all participants; dashed lines show average response rates from Gardiner and Java's (1990) Experiment 1. The replication and original results are very similar.

Descriptive Analysis. The pattern of response proportions is fairly similar to the ones from Experiments 1 and 2 with the exception that there was a clear preference of *sure* responses over *unsure* responses for old items. In fact, the pattern of responses appears highly similar to Gardiner and Java's Experiment 3 as shown by the dashed lines in Figure 3. On an individual level, the majority of participants showed the response preferences for *sure* responses for old items, but there was no clear difference of preference between words and nonwords. For new items, *sure* and *unsure* responses were equally likely, again across words and nonwords.

Model-based Analysis. Panels E and F in Figure 3 show the *remember* response bias for nonwords and words. As in the previous experiments, there is no sign for different

response biases for words and nonwords. For old items, however, almost all participants show a response bias in favor of *sure* responses resulting in positive values for Y_{ij}. The slight positive correlation in the two graphs shows that individuals who prefer *sure* responses for words tend to also prefer *sure* responses for nonwords; participants who prefer *unsure* responses for words tend to also prefer *unsure* responses for nonwords. This result on an individual level corresponds to Gardiner and Java's results across participants for Experiment 3.

Bayes factor model comparison again shows a preference for model $\mathcal{M}_1$, the model representing a straight-forward single-process criterion shift account. Model $\mathcal{M}_1$ is the replication model for Gardiner and Java's Experiment 3, and it is preferred over model $\mathcal{M}_*$, which is the second-best performing model. The Bayes factor between $\mathcal{M}_1$ and $\mathcal{M}_*$ is 4.70-to-1 in favor of $\mathcal{M}_1$. The least preferred model is model $\mathcal{M}_3$ with a Bayes factor of 114.11-to-1 in favor of the winning model.

Results across Experiments

After three replication attempts of remember-know and sure-unsure paradigms, we may revisit both the response patterns in the paradigms and the signal-detection modeling approach.

Remember-Know vs. Sure-Unsure

Our results provide for a speculation about the role of remember-know instructions compared to more conventional confidence-rating instructions. The confidence-rating experiment, Experiment 3, revealed a strong, consistent preference for the *sure* response relative to the *unsure* response with little individual differences. People are sure about what they know and they are clearly indicating so. As a result, the standard deviations for *unsure* response proportions are relatively low with 0.12 for nonwords and 0.12 for words. This preference can be contrasted with the response pattern for *remember* and *know* from Experiments 1 and 2. Here, we see a lack of preference as well as more variability across

Table 3
Response proportions for replication studies.

	Response		
	Remember	Know	New
Experiment 1			
Old Word	0.35 (0.37)	0.3 (0.26)	0.35 (0.37)
Old Nonword	0.36 (0.38)	0.28 (0.23)	0.37 (0.39)
New Word	0.2 (0.18)	0.18 (0.22)	0.61 (0.6)
New Nonword	0.2 (0.18)	0.15 (0.19)	0.64 (0.63)
Experiment 2			
Old Word	0.35 (0.33)	0.31 (0.33)	0.35 (0.34)
Old Nonword	0.32 (0.33)	0.32 (0.31)	0.36 (0.36)
New Word	0.08 (0.09)	0.23 (0.22)	0.69 (0.7)
New Nonword	0.07 (0.06)	0.15 (0.16)	0.78 (0.78)
Experiment 3			
Old Word	0.46 (0.46)	0.19 (0.19)	0.35 (0.35)
Old Nonword	0.45 (0.45)	0.19 (0.21)	0.35 (0.34)
New Word	0.18 (0.18)	0.16 (0.16)	0.66 (0.66)
New Nonword	0.13 (0.14)	0.17 (0.16)	0.7 (0.7)

Note. Predictions from a single-process model are provided in parentheses.

individuals. The pattern of individual response proportions is extreme: Some individuals almost exclusively respond *remember* to old items while others almost exclusively respond *know* to old items. The standard deviations for *know* responses are therefore somewhat higher with 0.21 for nonwords and 0.16. We speculate that participants have a vague idea at best what remember and know mean, and the vagueness leads to arbitrary, subjective decisions about their memory that are not indicative of underlying processes (see Naveh-Benjamin & Kilb, 2012). This vagueness can be contrasted with the treatment of the sure/unsure distinction where participants are more consistent and more sure of their memory.

Signal-Detection Revisited

In the beginning of this paper, we showed how the original Gardiner and Java results could not be well-accounted by a standard signal detection model (see also Dunn, 2004). How about our results? We fit the same signal detection model presented in Appendix A to our new data. Table 3 shows the predictions of a signal-detection model, and the values are presented in the parentheses. As can be seen, for both Experiments 2 and 3, the model fits fairly well, providing estimates close to the observed data. For Experiment 1, the estimates deviate more from the observed data, but the degree of misfit is not severe.

General Discussion

In this paper we sought to replicate Gardiner and Java (1990), Experiment 2. We consider this experiment to be the strongest direct behavioral evidence for two distinct memory processes of conscious recollection and automatic activation of familiar traces, and as such, the replication is timely and topical. Across two labs, the critical data patterns—a crossover interaction for old items with an invariance for new items—could not be found. Instead, there is seemingly no effect of lexical status. Moreover, the data pattern for all three experiments support a single mnemonic process model over the more complicated dual-process alternative.

Procedural and Analytic Differences

There are several procedural and analytic differences between our experiments and those from Gardiner and Java. For each, we think our choices are an improvement that rectifies a limitation in the original design. Here is a review of the major differences:

1. Increased sample sizes: Gardiner and Java ran experiments with 20 participants observing 60 test items for a total of 1200 observations. We ran experiments with a minimum of 50 participants observing a minimum of 100 test items for a total of 5000

observations. Hence, our experiments afford greater resolution to see effects and invariances.

2. Decreased retention interval: Our retention interval was 10 minutes rather than 24 hours. During this retention interval, all participants performed the same task. This shorter retention period allowed us to increase the number of items at study and test while maintaining a reasonable level of overall performance. Moreover, we could ensure that participants were having the same experience in the retention interval. Importantly, Gardiner and Java did not consider the long retention interval essential, and they note it was used only to avoid ceiling effects.

3. Computerized, sequential presentation: Gardiner and Java used hand-written items on cards and paper. We computerized the task. In doing so, we used a sequential presentation at test. This contrasts favorably with Gardiner and Java's simultaneous presentation at test, in which all test items were presented on a single piece of paper. Our approach is much more in line with the procedure employed nowadays by almost all recognition memory researchers, and the sequential nature reduces response dependencies across items. Additionally, the paper-method appears to have introduced a response bias in Gardiner and Java's procedure where participants preferred *new* responses (i.e. *not* circling an item) over *old* responses (i.e. circling an item). We eliminate this bias.

4. Analysis through model comparison: Gardiner and Java used ANOVA to analyze their data, and they analyzed response proportion as a function of response option (remember vs. know) and lexicality. Unfortunately, ANOVA is grossly inappropriate in this application. This negative correlation is not accounted for by ANOVA, and as a result, there is a marked tendency to overstate the significance of interactions. We take a more appropriate and sophisticated approach by instantiating different theoretical positions as formal statistical models and then use Bayesian model

comparison to draw inferences. This approach of using custom-tailored, theoretically specific linear models to answer critical questions should be attractive across cognitive psychology, and we refer interested readers to Rouder, Morey, and Wagenmakers (2016) and Rouder, Haaf, and Aust (2018).

In summary, although our experiments differ in a few aspects from Gardiner and Java, we feel that our choices provide clear improvements. We thought carefully and deliberately about each, understood why we were making the change, and documented each in the preregistration documents (https://osf.io/873sg/ and https://osf.io/k2ve3/). Despite these procedural differences, we were able to replicate Gardiner and Java's Experiment 3. If any of the design changes are responsible for our failure to replicate Experiment 2, these changes did not affect the relative success in replicating Experiment 3.

Noise or Signal?

The remaining question is why our results are different from Gardiner and Java's? Some readers, especially those predisposed to the dual-process account, may remain unsure if our failure to replicate reflects these changes. We suspect most readers will not object to our computer presentation, appropriate analysis, or increased sample size. Some may wonder about the effect of the 10-minute vs. 24-hour retention period or the effect of sequential vs. simultaneous testing. We note that there is no theoretical reason to think that dual-process signatures would be observable only after a day or only with simultaneous tests. In fact, it stretches common sense that such a fundamental mnemonic signature, if it existed, would be observable in such an unanticipated, limited set of conditions. Moreover, if a 24-hour retention period or simultaneous testing are needed to observe the critical dual-process pattern, then the vast majority of remember-know experiments in the literature are fatally flawed.

It is highly plausible that Gardiner and Java have misinterpreted noise for signal. Their studies were relatively underpowered and their analysis is characterized by high true

Type I error rates in interaction contrasts from naturally occurring negative correlation across response options. When we correct these flaws, we see no signature of two processes, and our data is highly concordant with extant single process models.

References

Atkinson, R. C., & Juola, J. F. (1973). Factors influencing the speed and accuracy of word recognition. In S. Kornblum (Ed.), *Attention and performance iv* (pp. 583–612). New York: Academic Press.

Aust, F., & Barth, M. (2017). *papaja: Create APA manuscripts with R Markdown*. Retrieved from https://github.com/crsh/papaja

Bates, D., & Maechler, M. (2016). *Matrix: Sparse and dense matrix classes and methods*. Retrieved from https://CRAN.R-project.org/package=Matrix

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology*, *33A*, 497–505.

Donaldson, W. (1996). The role of decision processes in remembering and knowing. *Memory & Cognition*, *24*, 523–233.

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

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

Gardiner, J. M., & Java, R. I. (1990). Recollective experience in word and nonword recognition. *Memory & Cognition*, *18*, 23–30.

Haaf, J. M., Klaassen, F., & Rouder, J. N. (In preparation). *Using systems of orders to capture theoretical constraint in psychological science*.

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

Jennings, J. M., & Jacoby, L. L. (1993). Automatic versus intentional uses of memory:

Aging, attention, and control. *Psychology and Aging*, *8*(2), 283–293.

Jennings, J. M., & Jacoby, L. L. (1997). An opposition procedure for detecting age-related deficits in recollection: Telling effects of repetition. *Psychology and Aging*, *12*(2), 352–361.

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: A bayesian approach. *Psychological Methods*, *10*(4), 477.

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

Morey, R. D., & Rouder, J. N. (2015). *BayesFactor: Computation of bayes factors for common designs*. Retrieved from https://CRAN.R-project.org/package=BayesFactor

Naveh-Benjamin, M., & Kilb, A. (2012). How the measurement of memory processes can affect memory performance: The case of remember/know judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(1), 194–203.

Naveh-Benjamin, M., Shing, Y. L., Kilb, A., Werkle-Bergner, M., Lindenberger, U., & Li, S.-C. (2009). Adult age differences in memory for name–face associations: The effects of intentional and incidental learning. *Memory*, *17*(2), 220–232.

Old, S. R., & Naveh-Benjamin, M. (2008). Memory for people and their actions: Further evidence for an age-related associative deficit. *Psychology and Aging*, *23*(2), 467–472.

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence diagnosis and output analysis for mcmc. *R News*, *6*(1), 7–11. Retrieved from http://CRAN.R-project.org/doc/Rnews/

Pratte, M. S., & Rouder, J. N. (2011). Hierarchical single- and dual-process models of recognition memory. *Journal of Mathematical Psychology*, *55*, 36–46.

Pratte, M. S., & Rouder, J. N. (2012). Assessing the dissociability of recollection and familiarity in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Prull, M., Dawes, L., Martin, A., Rosenberg, H., & Light, L. (2006). Recollection and

familiarity in recognition memory: Adult age differences and neuropsychological test correlates. *Psychology & Aging*, *21*, 107–118.

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rastle, K., Harrington, J., & Coltheart, M. (2002). 358,534 nonwords: The arc nonword database. *The Quarterly Journal of Experimental Psychology: Section A*, *55*(4), 1339–1362.

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*, 301–308. Retrieved from http://dx.doi.org/10.3758/s13423-014-0595-4

Rouder, J. N. (2016). The what, why, and how of born-open data. *Behavioral Research Methods*, *48*, 1062–1069. Retrieved from 10.3758/s13428-015-0630-z

Rouder, J. N., Haaf, J. M., & Aust, F. (2018). From theories to models to predictions: A Bayesian model comparison approach. *Communication Monographs*, *85*, 41–56. Retrieved from https://doi.org/10.1080/03637751.2017.1394581

Rouder, J. N., Morey, R. D., & Wagenmakers, E.-J. (2016). The interplay between subjectivity, statistical practice, and psychological science. *Collabra*, *2*, 6. Retrieved from http://doi.org/10.1525/collabra.28

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*, 356–374. Retrieved from http://dx.doi.org/10.1016/j.jmp.2012.08.001

Squire, L. (1994). Declarative and nondeclarative memory: Multiple brain systems supporting learning and memory. In D. Schacter & E. Tulving (Eds.), *Memory systems 1994* (pp. 203–231). Cambridge, MA: MIT Press.

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

Vilberg, K. L., & Rugg, M. D. (2008). Memory retrieval and the parietal cortext: A review

of evidence from a dual-process perspective. *Neuropsychologia*, *46*, 1787–1799.

Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software*, *21*(12), 1–20. Retrieved from http://www.jstatsoft.org/v21/i12/

Wickham, H. (2011). The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, *40*(1), 1–29. Retrieved from http://www.jstatsoft.org/v40/i01/

Wickham, H. (2016). *Rvest: Easily harvest (scrape) web pages*. Retrieved from https://CRAN.R-project.org/package=rvest

Wickham, H. (2017). *Stringr: Simple, consistent wrappers for common string operations*. Retrieved from https://CRAN.R-project.org/package=stringr

Wickham, H., Hester, J., & Ooms, J. (2017). *Xml2: Parse xml*. Retrieved from https://CRAN.R-project.org/package=xml2

Xie, Y. (2015). *Dynamic documents with R and knitr* (2nd ed.). Boca Raton, Florida: Chapman; Hall/CRC. Retrieved from https://yihui.name/knitr/

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., 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*, 323–339.