Journal of Experimental Psychology: General
2011, Vol. 140, No. 2, 239–257

0096-3445/11/$12.00 DOI: 10.1037/a0023007

Strong Memories Are Hard to Scale

Laura Mickes, Vivian Hwe, Peter E. Wais, and John T. Wixted
University of California, San Diego

People are generally skilled at using a confidence scale to rate the strength of their memories over a wide range. Specifically, low-confidence recognition decisions are often associated with close-to-chance accuracy, whereas high-confidence recognition decisions can be associated with close-to-perfect accuracy. However, using a 20-point rating scale, the authors found that the ability to scale memory strength had its limitations in that a high proportion of list items received the highest rating of 20. Efforts to induce participants to differentiate between these strong memories using emphatic instructions and alternative scales were not successful. Remember/know judgments indicated that these strong and hard-to-scale memories were often based on familiarity (not just recollection). Providing error feedback on a plurals discrimination task finally produced a high-confidence criterion shift. The authors suggest that the ability to scale strong (and almost perfectly accurate) memories may be limited because of the absence of differential error feedback for very strong memories in the past (the kind of differential error feedback that may account for the memory-scaling expertise that participants otherwise exhibit).

Keywords: recognition memory, confidence and accuracy, signal-detection theory, feedback

Memories vary in strength, and people are generally quite adept at using a numerical confidence scale to indicate how strong different memories are. This ability is most apparent in studies of recognition memory, in which accuracy is typically strongly related to confidence (Ratcliff & Murdock, 1976). Mickes, Wixted, and Wais (2007) showed that the strong relationship between confidence and accuracy extends over a wider range than had been previously investigated. In that study, participants completed a standard recognition memory test during which targets and lures were presented one at a time, and participants were asked to rate each item using a 20-point scale. A rating of 1 indicated 100% certainty that the item was new, and a rating of 20 indicated 100% certainty that the item was old. Intermediate ratings indicated varying degrees of lesser certainty, with ratings of 10 and 11 indicating complete uncertainty that the item was new (10) or old (11). As shown in Figure 1, participants were very accurate when confidence was high (i.e., for ratings of 1 or 20), and accuracy declined in continuous fashion to chance levels as confidence decreased to complete uncertainty (toward the middle of the scale). Thus, as a general rule, participants are able to use a confidence scale to provide valid ratings of the strength of their memories. The fact that participants exhibit this expertise without being trained by the experimenter is a key consideration for the memory scaling issue that is the main focus of this article.

Figure 2a illustrates one interpretation of these results in terms of the standard unequal-variance signal-detection model. According to this account, items vary in memory strength, with the mean and variance of the target distribution (μ_{target} and σ^2_{target}, respectively) being greater than the mean and variance of the lure distribution (μ_{lure} and σ^2_{lure}, respectively). In this example, $\mu_{target} = \mu_{lure} + 1.5\sigma_{lure}$ (i.e., the mean of the target distribution is 1.5 standard deviations above the mean of the lure distribution), and $\sigma_{target} = 1.5\sigma_{lure}$ (i.e., the standard deviation of the target distribution is 1.5 times that of the lure distribution). The target and lure distributions in a signal-detection analysis are usually assumed to be Gaussian in form. According to the signal-detection account, confidence ratings are based on various decision criteria arrayed along the memory strength axis (i.e., confidence ratings index memory strength). For a 20-point rating scale, 19 confidence criteria are assumed. A test item with a memory strength that exceeds the highest criterion receives a rating of 20; a test item with a memory strength that exceeds the second highest criterion (but falls below the highest criterion) receives a rating of 19, and so on.

The predicted proportion correct for a confidence rating in the range of 11 to 20 is given by the area under the target distribution associated with that rating divided by the sum of the areas under the target and lure distributions associated with that rating. The predicted proportion correct for a confidence rating in the range of 1 to 10 is similar except that the numerator is the area under the lure distribution associated with a particular rating. Figure 2b shows the predicted relationship between confidence and accuracy for the model depicted in Figure 2a, and it is clear that the pattern is similar to the obtained pattern shown in Figure 1. The asymmetry in the confidence-accuracy function (higher accuracy for high-confidence "old" decisions than for high-confidence "new" decisions) reflects the unequal variance of the target and lure

This article was published Online First March 21, 2011.

Laura Mickes, Vivian Hwe, Peter E. Wais, and John T. Wixted, Department of Psychology, University of California, San Diego.

This work was supported by National Institute of Mental Health Grant R01MH082892 to John T. Wixted. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Mental Health or the National Institutes of Health. We thank Allen Lee, Drew Ellen Hoffman, Travis Carlisle, Sam Xie, and Caleb Gross for data collection.

Correspondence concerning this article should be addressed to John T. Wixted, Department of Psychology, 0109, University of California, San Diego, La Jolla, CA 92093-0109. E-mail: jwixted@ucsd.edu

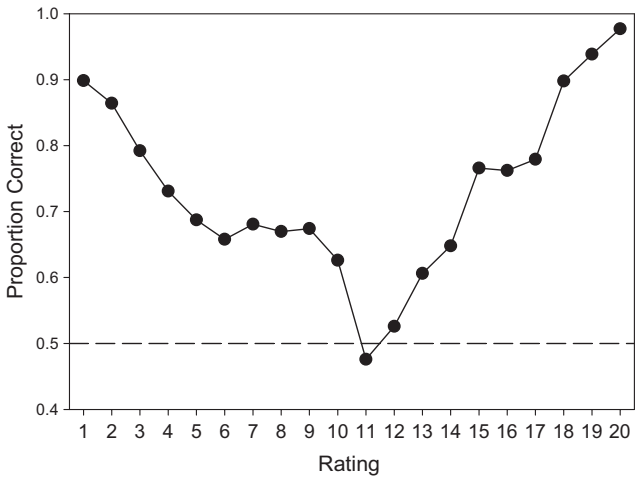

Figure 1. Accuracy (proportion correct) as a function of the confidence expressed in an old/new recognition decision (where 1 = *sure new* and 20 = *sure old*). The data are from Mickes et al. (2007).

distributions. The unequal-variance model illustrated in Figure 2a is consistent with many previous analyses of the receiver operating characteristic (ROC) for recognition memory (e.g., Egan, 1958; Ratcliff, Sheu, & Gronlund, 1992).

The confidence criteria in this example are equally spaced along the memory strength axis, but they need not be because participants are free to place the criteria anywhere. If participants did space them out at equal intervals, then the confidence ratings would provide an interval-scale measurement of memory strength over the range they cover. In that case, and if the range were wide enough, the confidence ratings could be used to compute directly the mean and variance of the target and lure distributions (instead of having to estimate them by fitting a Gaussian model to ROC data). Using this approach, Mickes et al. (2007) found that the standard deviation of the ratings to targets (s_{target}) was greater than that of the lures (s_{lure}). That is, in agreement with traditional ROC analyses, the direct ratings suggested an unequal-variance model. Moreover, at the level of the individual participant, the estimated ratio of the standard deviations based on Gaussian ROC analysis ($\sigma_{lure}/\sigma_{target}$), which is typically about 0.80, was significantly correlated with the estimated ratio computed directly from the ratings themselves (s_{lure}/s_{target}). Whereas ROC analysis specifically assumes a Gaussian model (but does not assume that confidence rating provide an interval-scale measure of memory strength), the direct rating method makes no distributional assumptions at all (but it does assume interval-scale measurement to a first approximation). The fact that the two methods are in such agreement provides further evidence that direct confidence ratings may provide valid scalar information about the strength of memory. The scaling expertise that participants exhibit in this regard is expertise that they bring with them to the laboratory.

If confidence ratings made using a 20-point scale provide valid scalar information, then the shape of the target and lure distributions can be directly examined by simply plotting the relevant frequency distributions. Mickes et al. (2007) reported that the frequency distributions for the targets and lures had the approximate form of two bell-shaped curves except that they were bunched on each end. The frequency distribution they reported is reproduced in Figure 3. The bunching effect at the extremes was especially apparent at the upper end of the scale. That is, a large number of targets received the highest rating of 20. Interpreted in terms of the signal-detection model, this means that the memory strength values associated with a substantial number of targets fell above the point at which the highest criterion was placed on the memory-strength axis (as illustrated in Figure 2a). Thus, according to this view, it is not the case that all of these target items had the same high memory strength, despite the fact that they were all given ratings of 20. Instead, the strong target items, like all other items on the recognition test, were presumably associated with varying degrees of strength, but the placement of the various confidence criteria was such that this presumed variability could not be identified. Criss (2009) also used a 20-point scale and reported the same bunching effect at the high end of the scale.

If the strongest memories do vary in strength and are distributed as part of a bell-shaped curve (just as memories associated with intermediate ratings appear to be), then it seems reasonable

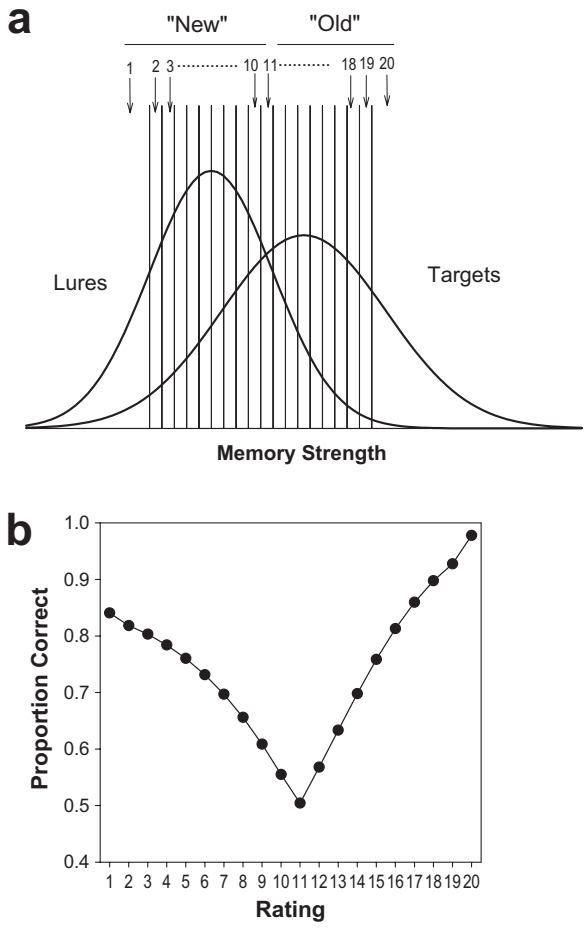

Figure 2. a. Hypothetical signal-detection model illustrating an equal-interval relationship between a 20-point rating scale and the memory strength scale. The vertical lines represent 19 confidence criteria. b. Predicted relationship between confidence and accuracy (proportion correct).

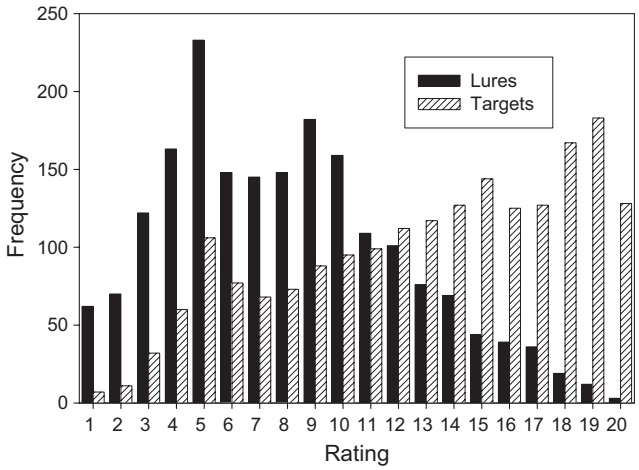

Figure 3. Frequency distribution showing the number of responses made to targets and lures for ratings of 1 through 20. The data are from Mickes et al. (2007).

to suppose that participants could be induced to use the rating scale in such a way that the right tail of the target distribution would become more evident in the distribution of ratings. That is, because participants are able to scale memories easily and accurately over such a wide range of strength without any special training, there is no obvious reason why they should not be able to do the same for stronger memories if they were so inclined. That was the issue we set out to address. The results of the first three experiments we report suggest that participants find it difficult to further scale their strongest memories in a meaningful way, despite their impressive ability to accurately scale their other memories. The results of the fourth experiment suggest that strong memories may be hard to scale whether they are based on recollection or on familiarity. The final experiment investigated the possibility that error feedback is what teaches people how to expertly scale memories in the first place. However, error feedback would not be able to teach people how to differentiate between the strengths of their strongest memories because those memories are essentially error free, and that may explain why strong memories are so hard to scale.

Experiment 1

The first experiment replicated that of Mickes et al. (2007), except that the instructions were designed to encourage participants to spread their confidence criteria out on the memory-strength axis such that ratings of 20 would be reserved for relatively few memories associated with very high memory strength. In terms of the model shown in Figure 2a, the instructions were designed primarily to induce a more conservative placement of the highest criterion. If participants adjusted their confidence criteria in accordance with these instructions, and if memory strengths are continuously distributed even when memory is strong, then the histogram for the target items should more clearly reveal the right tail of the target distribution.

Method

Participants. Nineteen undergraduates from the University of California, San Diego participated for lower division psychology course credit.

Materials and design. The word pool used consisted of three- to seven-letter words taken from the MRC Psycholinguistic Database (Coltheart, 1981), based on a concreteness rating range of 550–700. These criteria yielded 705 words from the database, of which 300 were selected randomly for testing (150 of which were selected randomly to be targets, whereas the remainder were lures). Instructions and stimuli were displayed for each participant on a computer monitor and were powered by a Dell computer. Stimuli were presented using an E-prime program (www.pstnet.com).

Procedure. Participants signed a consent form, were read instructions, studied 150 target words that were presented randomly for 2 s each, and completed a recognition test in which the 150 targets were randomly intermixed with the 150 lures. During self-paced testing, which followed immediately after presentation, participants indicated whether or not the word was on the presented list by pressing a number on the keypad ranging from 1 to 20 (with 1 meaning the word was definitely not on the list and 20 meaning the word was definitely on the list). The instructions were the same as those used by Mickes et al. (2007) with one exception. To prevent participants from using the highest (and lowest) rating so often, the instructions included the following admonition:

> Please be extremely cautious about using the end points of 1 and 20 and use them only if you are 100% positive about your answer. If you use 1 or 20, that means you CANNOT possibly make a mistake. That is, you are so confident in your answer that you would be willing to testify in a court of law, even in a life-or-death situation.

These instructions were emphasized during the practice session until participants indicated that they understood them.

Results

One participant appeared to have responded randomly (hit rate = .60, false alarm rate = .59) and was excluded from the analysis. Another participant appeared to have an extreme preference for ratings of 20, as evidenced by the fact that 18 of the lures received that rating. For the remaining participants, the average number of lures receiving a rating of 20 was 1.10, with a standard deviation of 1.52. Thus, this participant was judged to be an outlier and was excluded from the following analyses.

We first examined the degree to which distributional statistics computed directly from the ratings from the remaining 17 participants (an analysis that involves no assumptions about the form of the target and lure distributions) were consistent with distributional statistics estimated by fitting a Gaussian model to the confidence-based ROC data. To the extent that these distributional statistics agree, it would suggest that participants are capable of providing valid scalar information about memory strength (i.e., that they have expertise in rating memory strength). It would also provide evidence that the Gaussian assumption is valid because the agreement between the two approaches would otherwise have to be attributed to coincidence (see Rouder, Pratte, & Morey, 2010, and Wixted & Mickes, 2010b, for an exchange of views on this issue).

Table 1 shows the mean and standard deviations for the ratings made to the targets (m_{target} and s_{target}) and to the lures (m_{lure} and s_{lure}) for each participant. Across all participants, the mean rating for the targets was 12.11, and the mean rating for the lures was 7.43. The corresponding standard deviations were 5.36 and 3.77, respectively. Table 1 also shows, for each participant, the ratio of the standard deviation of the lure ratings to the standard deviation of the target ratings (s_{lure}/s_{target}). The mean ratio was .72, which is significantly less than 1.0, $t(16) = 8.70$ (unless otherwise indicated, an alpha level of .05 was used throughout). Thus, an analysis of the ratings (involving no distributional assumptions) suggests an unequal-variance model.

Similar conclusions are reached if the data are analyzed in terms of a Gaussian signal-detection model. More specifically, the ratio of $\sigma_{lure}/\sigma_{target}$ (obtained from fitting a straight line to each participant's *z*-ROC data) was, on average, .71 (shown in the rightmost column of Table 1). The estimates of s_{lure}/s_{target} (obtained from the ratings) and $\sigma_{lure}/\sigma_{target}$ (obtained from Gaussian ROC analysis) were positively and significantly correlated across participants, $r(15) = .49$.

The close agreement between methods can be illustrated by examining the group ROC data (see Figure 4). The smooth curve drawn through the probability ROC in Figure 4 does not represent a fitted function, as it typically would. Instead, it represents the Gaussian function that corresponds to the relative mean and standard deviation values computed directly from the pooled ratings. More specifically, in the pooled data, the mean rating for the targets was 12.11, and the mean rating for the lures was 7.43 (the same as when computed on an individual basis). The corresponding standard deviations were 5.95 and 4.25, respectively. These are both larger than when computed on an individual basis because, in the pooled data, variability across participants adds to the variability across items. On the basis of these values, the standardized mean of target distribution for the pooled data was estimated to be $(12.11 - 7.43)/4.25 = 1.10$ standard deviations above the mean of the lure distribution, and the standard deviation of the lure distribution was estimated to be $4.25/5.95 = 0.72$ times that of the target distribution. The smooth curve in Figure 4 corresponds to a Gaussian signal-detection model with those (directly computed, not fitted) parameters. The fit is not perfect in that some systematic deviation is apparent in the middle range, but it is remarkably good given that (a) no curve fitting was performed and (b) the function could deviate radically from the observed data if the measurement properties of the rating scale were systematically nonlinear with

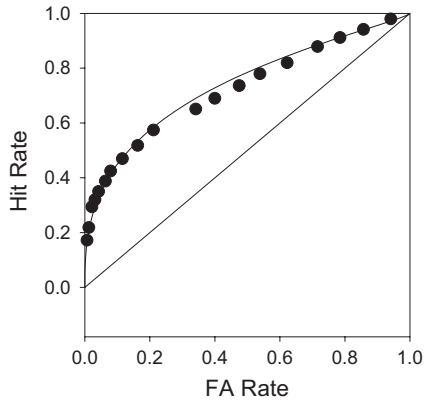

Figure 4. Receiver operating characteristic data from Experiment 1 pooled over participants. The smooth curve was not fit to the data but instead represents the Gaussian model specified by the means and standard deviations of the target and lure distributions computed directly from the ratings. FA = false alarm.

Table 1
Descriptive Statistics for the Rating Scale Data From Experiment 1

	Direct ratings method						
Subject no.	m_{target}	m_{lure}	s_{target}	s_{lure}	s_{lure}/s_{target}	d_r	ROC analysis slope
1	10.32	9.29	2.64	2.13	0.81	0.43	0.87
2	11.07	6.77	4.86	2.75	0.56	1.09	0.74
3	13.65	6.80	5.18	3.05	0.59	1.61	0.57
4	13.38	8.71	5.34	4.38	0.82	0.96	0.78
5	11.54	9.70	4.54	3.37	0.74	0.46	0.74
6	14.29	5.06	5.81	3.48	0.60	1.93	0.57
7	13.04	7.69	5.72	4.03	0.70	1.08	0.67
8	7.79	5.15	7.25	5.35	0.74	0.41	0.71
9	10.19	4.04	7.57	4.64	0.61	0.98	0.68
10	10.94	6.67	5.69	4.00	0.70	0.87	0.78
11	11.15	8.02	6.06	5.50	0.91	0.54	0.91
12	11.89	7.77	5.59	4.29	0.77	0.83	0.80
13	13.00	10.95	3.30	3.17	0.96	0.63	0.91
14	14.04	9.38	5.14	2.78	0.54	1.13	0.62
16	8.53	5.70	6.15	3.67	0.60	0.56	0.62
18	13.16	6.75	5.98	3.49	0.58	1.31	0.64
19	17.92	7.79	4.38	4.07	0.93	2.40	0.47
Mean	12.11	7.43	5.36	3.77	0.72	1.01	0.71

Note. m_{target} and m_{lure} represent the mean ratings to targets and lures, respectively; s_{target} and s_{lure} represent the corresponding standard deviations; d_r represents a discriminability measure equal to m_{target} minus m_{lure} divided by the root mean square of s_{target} and s_{lure}; ROC = receiver operating characteristic.

respect to memory strength (or if the true underling distributions deviated radically from the Gaussian model).

Figure 5a shows the accuracy for each level of confidence. One through 10 was scored as a correct response to lures (and as an incorrect response to targets), whereas the reverse was true for ratings in the range of 11 through 20. In accordance with the predictions of signal detection theory, and in agreement with the results reported by Mickes et al. (2007), accuracy varied continuously as the distance from the indifference point (10, 11) increased. Although the confidence–accuracy relationship is noticeably stronger for targets than for lures, confidence is a good predictor of accuracy in both cases. This result indicates in another way that participants can accurately scale the strength of their memories over a wide range.

The rating data and ROC data summarized above lend credibility to the Gaussian signal-detection model because the distributional statistics computed directly from the ratings agree with the distributional statistics estimated indirectly by fitting a Gaussian model (something that is not true when various non-Gaussian models are used; Wixted & Mickes, 2010b). But if the underlying distributions are Gaussian in form (even only approximately), then the distribution of memory strength values associated with the targets should exhibit a right tail. Figure 5b shows the obtained frequency distribution of ratings for the targets and the lures. Inspection of this figure immediately reveals that the emphatic instructional manipulation used in this experiment did little or nothing to reveal the right tail of what might be a continuously distributed target distribution. As in Mickes et al. (2007), the distributions appear to be bunched on both ends, with the effect being much more pronounced on the right end of the scale. Specifically, many of the target items (19% overall) received the highest rating of 20. More targets received that rating than any other rating by a factor of 2.5 or more. If strong memories vary in strength, and if participants are skilled in rating the strength of their memories (as much evidence suggests is the case), the bunching effect at the high end of the scale is puzzling.

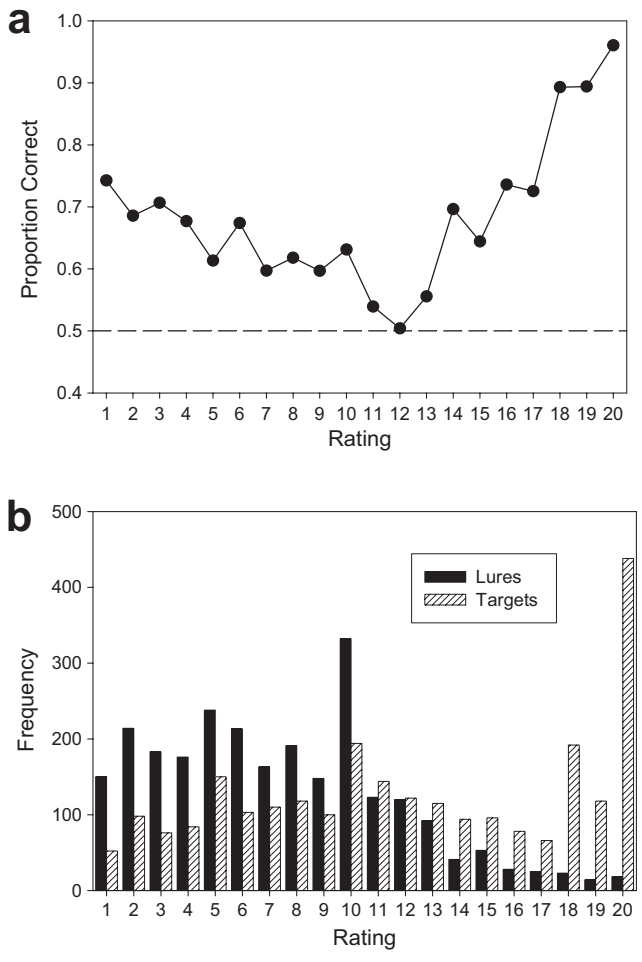

Figure 5. a. Accuracy (proportion correct) as a function of the confidence expressed in an old/new recognition decision in Experiment 1. b. Frequency distribution showing the number of responses made to targets and lures for ratings of 1 through 20 in Experiment 1.

Experiment 2

Because even emphatic instructions failed to prevent participants from assigning the highest rating of 20 to a large number of target items, a different approach was used in the next experiment. A seemingly straightforward way to encourage participants to reveal the right tail of the target distribution would be to use a numerical rating scale with more than 20 levels of confidence. However, Mickes et al. (2007) also used a 99-point confidence scale, and this did not have the desired effect. Instead, many participants simply provided ratings on the fives (i.e., 5, 10, 15, 20, etc.), effectively transforming the scale back into a 20-point confidence scale. Moreover, a high percentage of targets again received the highest rating of 99.

In Experiment 2, we used another approach to encourage participants to scale strong memories. Specifically, after receiving general instructions in the use of the rating scale, participants were informed that people have a tendency to overuse the highest rating of 20 and that we were especially interested in their ability to scale strong memories for which they might be tempted to use that rating. As such, for each item that received a rating of 20, they were informed that a new scale would appear that could be used to rate the strength of these strong memories in a more fine-grained way. The new scale ranged from 20 to 30, and the verbal description associated with those values ranged from extremely strong memory (20) to, essentially, the strongest memory imaginable (30). Our assumption was that the distribution of ratings between 20 and 30 might take on the form of the right tail of a bell-shaped curve (with, presumably, only a very few items receiving the highest rating of 30).

Method

Participants. Fifteen undergraduates from University of California, San Diego participated for lower division psychology course credit.

Materials and design. These were the same as Experiment 1 in every respect.

Procedure. The procedure was the same as Experiment 1, except for the instructions and for the fact that a new scale was presented for each test item given a rating of 20. The instructions indicated that people have a tendency to use the end points much

more than they should and that we did not want them to use the extreme numbers often. In addition, the instructions indicated that if a rating of 20 must be used, then we would inquire into their ability to further scale those strong memories. To that end, a new scale ranging from 20 to 30 would appear for each test item that received a rating of 20. For this new scale, participants were urged to distribute their responses to fully cover this range and to reserve ratings of 30 for items with memories that are so strong that they would remember that the word was on the list for the rest of their lives. More specifically, the relevant part of the instructions stipulated the following about the 20-to-30 rating scale:

> Please remember that this is a scale we want you to use fully, and the highest response, a 30, indicates a memory so strong that you will remember that this specific word was on this list FOR THE REST OF YOUR LIFE. The higher twenties should mean that you will at least remember that word for years.

Results

Once again, we first analyzed the degree to which the rating data were compatible with a Gaussian signal-detection model (the model that suggests that a right-tailed distribution of target memory strengths ought to be observed in the confidence ratings). Table 2 shows the mean and standard deviations for the ratings made to the targets (m_{target} and s_{target}) and to the lures (m_{lure} and s_{lure}) for each of the 15 participants. Across all participants, the mean rating for the targets was 11.86, and the mean rating for the lures was 7.76. The corresponding standard deviations were 4.91 and 3.52, respectively. Table 2 also shows, for each participant, the ratio of the standard deviation of the lure ratings to the standard deviation of the target rating (s_{lure}/s_{target}). The mean ratio was .75, which is significantly less than 1.0, $t(14) = 6.07$. Thus, the results again suggest an unequal variance model, with the standard deviation of the targets being greater than that of the lures.

As in the first experiment, similar conclusions were reached when the data were analyzed in terms of a Gaussian signal-detection model. More specifically, the ratio of $\sigma_{lure}/\sigma_{target}$ (obtained from fitting a line to each participant's z-ROC data) was, on average, .77 (shown in the rightmost column of Table 2). As before, the estimates of s_{lure}/s_{target} (obtained from the ratings) and $\sigma_{lure}/\sigma_{target}$ (obtained from Gaussian ROC analysis) were positively and significantly correlated across participants, $r(13) = .62$. Thus, with regard to the ratio of the target and lure standard deviations, essentially the same information is obtained from direct ratings as from Gaussian ROC analysis.

Figure 6a shows the accuracy scores for each rating and, again, the strong relationship between confidence and accuracy is apparent. All of these results show that a 20-point rating scale provides information about the relative means and variances of the underlying distributions that corresponds closely to the information obtained from a Gaussian-based ROC analysis. If the Gaussian model is correct, this suggests that participants arrive at the laboratory with considerable expertise in rating memory strength. If so, one would expect them to be able to rate the strengths of strong memories.

With regard to the frequency distribution based on responses to the 20-point rating scale, the data again indicate that many targets received the highest rating of 20 (see Figure 6b). As shown in the figure, 220 targets received that rating, whereas only six lures did. Thus, it seems clear that this pattern (i.e., an apparently bunched target distribution) is a typical result. Also evident in this figure is an apparent bias to provide ratings of 10 (both targets and lures received that rating disproportionately) and a bias to provide ratings of 2, perhaps because, at the low end of the scale, participants heeded the instruction to avoid making extreme ratings if their certainty was less than 100%. However, no such caution was evident for ratings of 20 as many targets again received that rating (with accuracy being very high).

Table 2
Descriptive Statistics for the Rating Scale Data From Experiment 2

	Direct ratings method						
Subject no.	m_{target}	m_{lure}	s_{target}	s_{lure}	s_{lure}/s_{target}	d_r	ROC analysis slope
37	11.37	8.26	3.68	3.05	0.83	0.92	0.87
38	10.69	8.51	4.21	3.11	0.74	0.59	0.79
39	13.13	10.11	4.42	4.15	0.94	0.70	0.86
40	14.29	6.49	6.20	3.95	0.64	1.50	0.62
41	11.44	8.90	3.91	3.21	0.82	0.71	0.92
42	13.76	9.26	3.78	2.59	0.68	1.39	0.63
43	11.93	7.53	5.47	4.41	0.81	0.88	0.93
44	13.81	8.81	4.66	3.98	0.85	1.15	0.82
45	11.05	9.73	2.28	1.85	0.81	0.64	0.89
46	8.45	3.78	6.75	3.01	0.45	0.89	0.79
47	11.71	7.25	6.28	4.15	0.66	0.84	0.59
48	10.31	3.26	8.45	4.27	0.51	1.05	0.49
49	12.63	7.48	5.00	3.64	0.73	1.18	0.70
50	14.96	11.25	4.21	4.45	1.06	0.86	0.82
51	8.45	5.74	4.35	3.04	0.70	0.72	0.79
Mean	11.86	7.76	4.91	3.52	0.75	0.94	0.77

Note. m_{target} and s_{target} represent the mean ratings to targets and lures, respectively; s_{target} and s_{lure} represent the corresponding standard deviations; d_r represents a discriminability measure equal to m_{target} minus s_{target} divided by the root mean square of s_{target} and s_{lure}; ROC = receiver operating characteristic.

The question of primary interest in this experiment was whether or not a continuous distribution of memory strengths associated with the high-strength items would be observed when participants were given a second chance to rate the items using a scale that ranged from 20 to 30. Figure 7 shows the frequency distribution of the ratings that were provided when that scale was used. Obviously, the continuous pattern that might be expected if the targets follow a bell-shaped curve was not obtained. Instead, with relatively few exceptions, the items that were initially given a rating of 20 were approximately evenly distributed between the two extreme second-chance ratings (i.e., 20 and 30). Of the 220 targets that initially received a rating of 20, only 45 received subsequent ratings other than 20 or 30 (and these 45 responses were essentially evenly distributed between 21 and 29). This result could be taken to mean that strong memories are distributed in bicategorical fashion, in which case one could argue that strong memories are scalable after all. However, in our view, the simplest interpretation of this pattern of results is that participants were not able to scale meaningfully the strong memories that were initially given a rating of 20. Either way, there is no evidence that the target distribution has a right tail of continuously distributed (strong) memories.

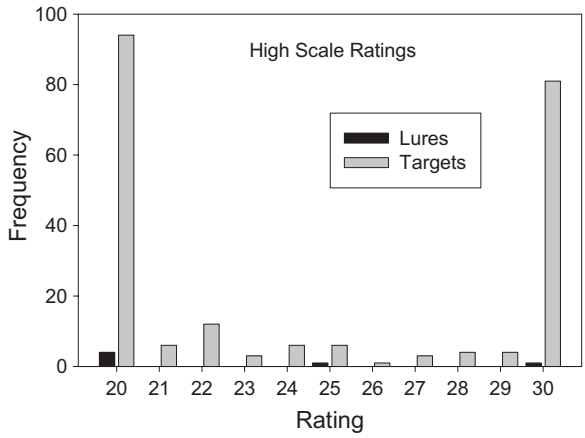

Figure 7. Frequency distribution of the number of responses made to targets and lures for ratings of 20 through 30 in Experiment 2.

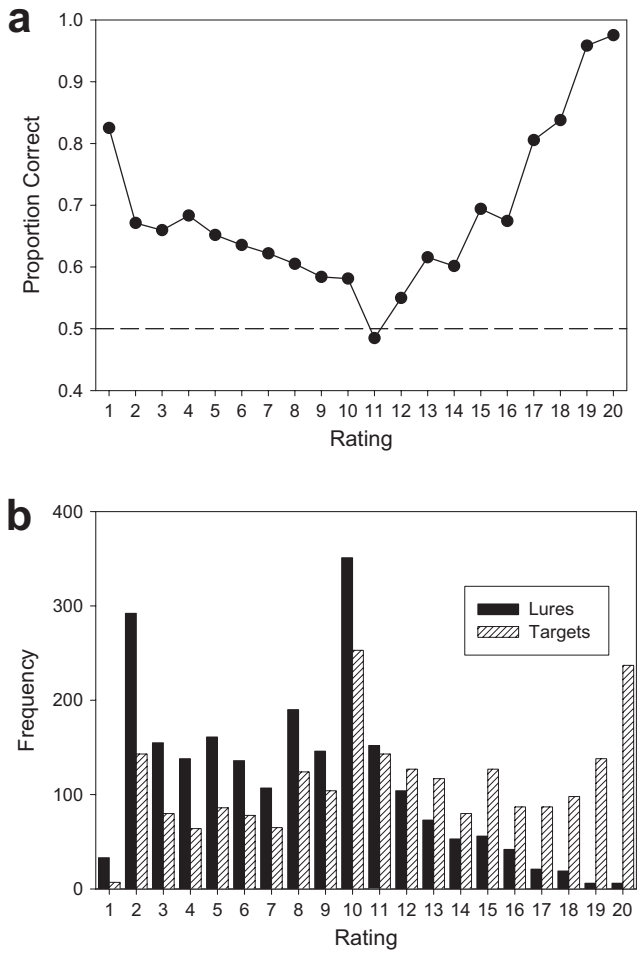

Figure 6. a. Accuracy (proportion correct) as a function of the confidence expressed in an old/new recognition decision in Experiment 2. b. Frequency distribution showing the number of responses made to targets and lures for ratings of 1 through 20 in Experiment 2.

Experiment 3

The use of a secondary scale for strong memories did not seem to provide meaningful ratings of strong memories. It is conceivable that participants would be better able to quantify differences between their strong memories if they were allowed to use a more unrestricted numerical rating scale instead of using the fixed rating scale provided by the experimenter. In the next experiment, participants were asked to use any numerical scale they wished to rate the memory strength of test items,[1] with the only restriction being that increasing certainty that the item was new should be indicated by increasingly negative numbers, and increasing certainty that the item is old should be indicated by increasingly positive numbers. Otherwise, they were free to use whatever numbers they wished to scale the strength of their memories.

Participants. Sixteen undergraduates from University of California, San Diego participated for lower division psychology course credit.

Materials and design. These were the same as Experiment 1 in every respect.

Procedure. The procedure was the same as Experiment 1 except for the method of supplying ratings to the test items. The instructions indicated that each test item should be numerically rated for confidence that it is a target or a lure. Any numbers could be used, except that increasing confidence that the item is a target should be expressed using increasingly positive numbers, and increasing confidence that the item is new should be indicated using increasingly negative numbers. A short practice session preceded the presentation of the 150 target words.

Results

Two participants were excluded because they appeared to respond randomly (i.e., their responses did not discriminate targets

[1] We thank Vic Ferreira for suggesting this experiment.

from lures). Before considering the frequency distributions, a preliminary analysis was conducted on the means and variances of the ratings. For this preliminary analysis, the ratings were normalized on the basis of the mean and standard deviation of each participant's ratings to the lures (denoted m_{lure} and s_{lure}, respectively). That is, the individual ratings to both targets and lures were normalized as follows:

$$z_i = (r_i - m_{lure})/s_{lure},$$

where r_i represents the observed rating and z_i represents the normalized rating. The normalized ratings for the targets are shown in Table 3. The mean and standard deviation of the normalized ratings to the lures were necessarily 0 and 1, respectively. Thus, by using the *z*-transformation shown above, the ratings were scaled with respect to the mean and standard deviation of the lure distribution, as in standard signal-detection analyses (in which the mean and standard deviation of the lure distribution are often set to 0 and 1, respectively). The mean normalized rating for the targets was 1.18, and the mean standard deviation was 0.99. Thus, unlike the previous experiments that used a fixed rating scale, these data are consistent with an equal-variance model. The average value of s_{lure}/s_{target} (i.e., the average value of $1/s_{target}$) was 1.07. By contrast, the Gaussian-based estimate of $\sigma_{lure}/\sigma_{target}$ obtained by fitting a line to the *z*-ROC data of each subject was .80 (see Table 3). This value is significantly less than 1, $t(13) = 3.59$, and, as usual, it suggests an unequal-variance model. Thus, in this case, the two estimates (i.e., s_{lure}/s_{target} and $\sigma_{lure}/\sigma_{target}$) do not agree. This could mean that the free rating scale is less likely to have interval-like properties than a fixed rating scale. In fact, Anderson (1961) argued long ago that for a scale to have interval-like properties, it needs to have well-defined end points, and that is exactly what the free rating scale lacks. Alternatively, the disagreement could instead indicate that the Gaussian signal-detection theory is wrong and that it is simply a coincidence that ROC data agree with the direct ratings data when a fixed scale is used (see Rouder et al., 2010, and Wixted & Mickes, 2010b, for a more detailed discussion of these issues).

The question of primary interest was the distribution of the ratings to the targets when participants were free to choose their own rating scale. Somewhat to our surprise, most participants spontaneously chose to use a 20-point rating scale. As shown in Figure 8, 10 of the 14 participants used a scale that ranged from approximately –10 to 10. One of these participants (S17) also made four apparently stray ratings that are not shown in the frequency distribution for that participant. Specifically, for the lures, all ratings except for three fell in the range of –10 to 10, with the three exceptions being –67, –19, and 67. For the targets, all ratings except for one fell in the range of –10 to 10, with the exception being 89. Of the four participants who used a range other than –10 to 10, two (S5 and S11) used scales that ranged from –100 to 100; another (S9) used a scale that ranged from –20 to 20, and another (S14) used a range from –5 to 5. All participants used whole numbers for every rating even though they were not instructed to do so. In many cases, participants avoided using the middle range of their freely chosen scales, which suggests that these ratings may not provide an approximately interval-scale measure (i.e., memories of intermediate strength are presumably abundant, but they did not receive intermediate ratings).

Not all of the ratings in a given range were used by every participant. For example, although S5 used a range of –100 to 100, only 14 separate ratings in that range were used (and most of those fell on the fives). The average number of separate ratings across participants was 17.1, which is not far from the 20-point scale used in Experiments 1 and 2. Thus, these findings are consistent with results reported by Mickes et al. (2007), who showed that the use of a rating scale that was more fine-grained than a 20-point scale yielded no apparent gain (as many participants effectively converted a fixed 99-point scale back into a 20-point scale by responding on the fives). When allowed to choose their own rating scale, most participants chose to use something close to a 20-point scale.

Table 3
Descriptive Statistics for the Rating Scale Data From Experiment 3

	Direct ratings method				
Subject no.	m_{target}	s_{target}	s_{lure}/s_{target}	d_r	ROC analysis slope
1	2.42	1.33	0.75	2.056	0.58
3	1.74	0.78	1.29	1.942	1.34
4	1.22	0.56	1.78	1.501	0.73
5	0.85	0.91	1.09	0.888	0.89
6	1.08	0.93	1.08	1.114	0.75
7	1.27	0.80	1.25	1.402	0.82
8	1.06	1.44	0.70	0.846	0.57
9	1.20	1.15	0.87	1.105	0.66
10	0.57	0.567	0.84		
11	0.72	1.03	0.97	0.711	0.93
13	1.27	0.69	1.46	1.485	0.93
14	0.73	1.11	0.90	0.686	0.86
15	1.85	1.35	0.74	1.557	0.49
16	0.52	0.86	1.16	0.560	0.81
Mean	1.18	0.99	1.07	1.17	0.80

Note. m_{target} and s_{target} represent the standardized mean and standard deviation, respectively, of ratings to targets, with $m_{lure} = 0$ and $s_{lure} = 1$; d_r represents a discriminability measure equal to m_{target} minus m_{lure} divided by the root mean square of s_{target} and s_{lure}; ROC = receiver operating characteristic. The slope estimate was obtained from fitting a straight line to the *z*-ROC of each participant.

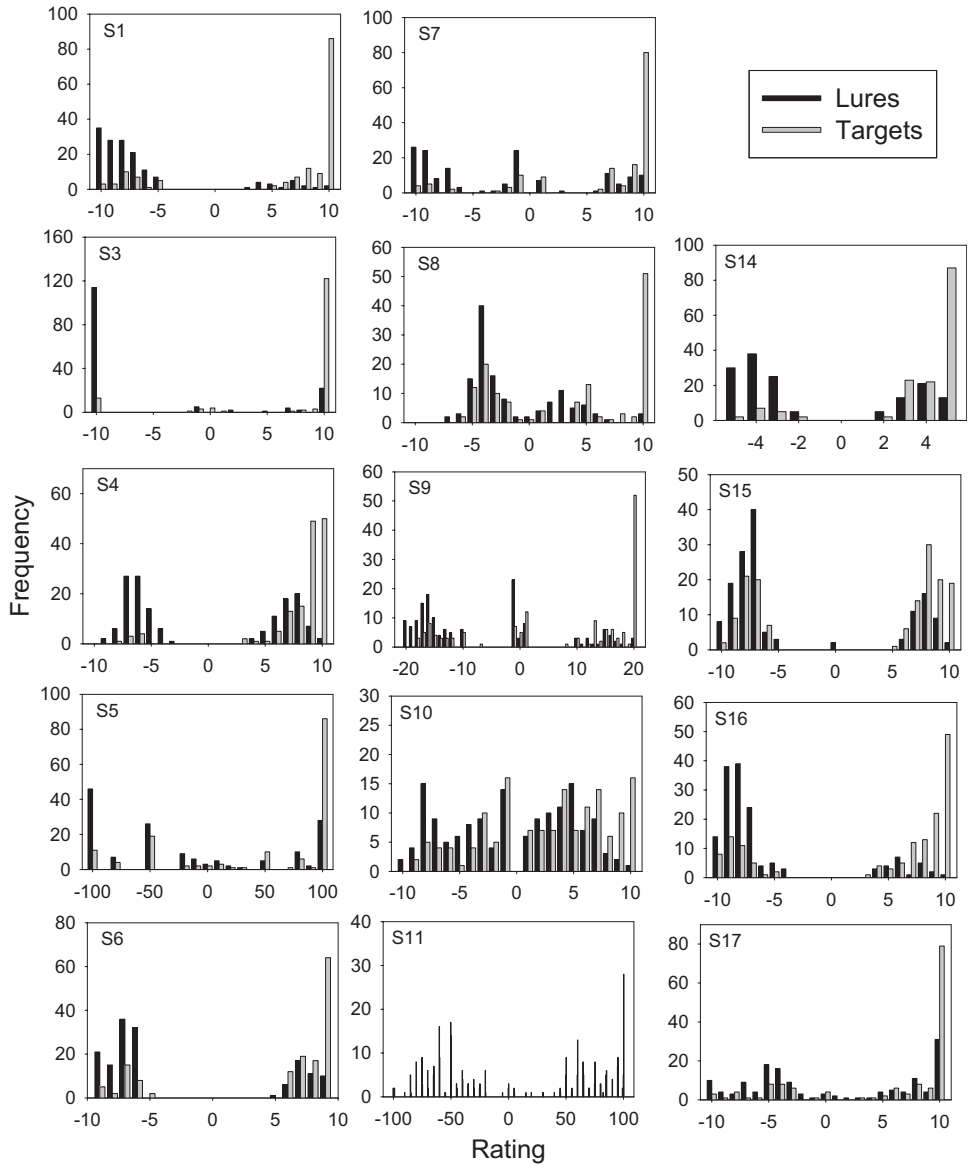

Figure 8. Frequency distributions for individual participants showing the number of responses made to targets and lures for the free-scale ratings in Experiment 3.

In light of the fact that participants freely adopted a rating scale that was not much different from the one used in Experiments 1 and 2, it is not surprising that many target items again received the highest rating (often a rating of 10 in this case) and that the form of the frequency distribution was that of an apparently bunched (but otherwise roughly bell-shaped) curve. As shown in Figure 8, every participant exhibited this phenomenon, which provides further evidence that strong memories are not easily scaled. For the most part, the effect is not as apparent on the other end of the scale, with S1, S3, and S5 being possible exceptions.

Experiment 4

The results of the first three experiments suggest that, at a minimum, participants cannot easily scale strong memories despite being adept at scaling all but their strongest ones. One possibility is that these strong and apparently unscalable memories reflect the "all" state of an all-or-none recollection process (Yonelinas, 1994). If recollection leads to the complete recovery of the original encoding experience, then it would not be possible to further scale these memories.

Recollection-based decisions are often thought to be identifiable using the remember/know procedure (e.g., Eldridge, Engel, Zeineh, Bookheimer, & Knowlton, 2005; Uncapher & Rugg, 2005; Yonelinas et al., 2002), although its utility in that regard is a matter of debate (e.g., Dunn, 2004; Wixted & Stretch, 2004). In its typical use, remember judgments are thought to reflect recollection-based decisions, whereas know judgments are thought to reflect familiarity-based decisions. In the next experiment, we combined

the remember/know procedure with the use of the 20-point rating scale. The question of interest was whether an appreciable number of items identified as old with a confidence rating of 20 would also be associated with know judgments. This should not be the case if those high confidence ratings primarily reflect complete recollection.

It has been argued that know judgments can be contaminated by random guessing, leading to lower accuracy than would otherwise be observed (e.g., Gardiner, Richardson-Klavehn, & Ramponi, 1997). The recommended solution is to include a guess option (in addition to the remember and know options), and we included that option in Experiment 4. During study, each list item was also presented with one of two source details (a question about the size of the word's referent or a question about its animacy), and participants attempted to recall that source question after making a remember, know, or guess judgment for each item. The ability to recollect this source information provided a validity test for high-confidence remember and know judgments.

Method

Participants. Fifteen undergraduates from University of California, San Diego participated for lower division psychology course credit.

Materials and design. These were the same as Experiment 1 in every respect, except the words were all randomly selected per subject from a large pool of words.

Procedure. The procedure was the same as Experiment 1 with a few exceptions. First, during presentation, words were presented one at a time at the center of the screen along with one of two questions ("Is this animate or inanimate?" or "Would this fit inside a shoebox?"). The question appeared below the word. Participants were instructed to mentally answer the question and to remember which question appeared with each word. Half the words were associated with one question and half with the other. The words were balanced in terms of these attributes. During the recognition test, participants provided a rating for each test item using the 20-point confidence scale. In addition, after each rating, they were asked to make a remember/know/guess judgment for any item that received a rating of 11 or higher. For these same items, participants were also asked to indicate which question was presented with the word at study (i.e., they were given a source recollection test). The targets and lures were displayed individually in the center of the screen, and the remember/know/guess and source questions appeared below the word. Subjects pressed the "R," "K," or "G" keys to record a remember, know, or guess judgment, respectively, and they pressed "1" or "2" to indicate which question they believed was presented with the word ("Is this animate or inanimate?" or "Would this fit inside a shoebox?" respectively).

The instructions for the remember/know judgments were based on Gardiner (1998) because these instructions are widely used (e.g., Yonelinas, 2001). The critical section of the remember/know instructions read as follows:

> If you indicate that the word was on the list that you studied, you will next be asked whether you "remember," "know," or are guessing that the word was one you saw on the presentation list. Only respond remember if you can remember some qualitative information about the items, such as recollecting what you thought about when the word was presented, what the word looked like, or sounded like (when you read it to yourself), or the question that was presented with it. Only respond "remember" if you can tell me what you recollect about the study event. You'll be asked about the specific question, but if you recollect other things about the word, press "remember." On the other hand, "know" means that you recognize a word on the test from the presentation list, but it does not bring back to mind anything specific. That is, it seems familiar, so you feel confident is was one you saw, even though you don't recollect anything you experienced when you saw it.

During the brief practice session that preceded the experiment proper, any questions that participants had about the remember/know procedure were clarified. In addition, to minimize the likelihood that the remember/know judgments would be used merely as a proxy for confidence, participants were reminded during the practice session that know judgments mean that confidence is high (as indicated by high ratings on the 20-point scale) that the word was seen on the list even though nothing can be recollected about the word's prior occurrence. The idea that know judgments should be associated with high confidence has long been a standard component of remember/know instructions (e.g., Rajaram, 1993). Despite such instructions, participants often supply know judgments only when confidence is not high (e.g., Dunn, 2004), which is why we included this reminder during the practice sessions.

Results

We first analyzed the distributional statistics of the ratings made using the 20-point confidence scale. As shown in Table 4, the mean rating for the targets was 12.49, and the mean rating for the lures was 5.64. The corresponding standard deviations were 7.13 and 4.99, respectively, which means that the ratings again suggest an unequal-variance model. One participant, who had a very large standard deviation for the lures relative to the targets (s_{lure}/s_{target} = 2.10) was a clear outlier and was not included in any of the analyses for this experiment. The unusually large ratio occurred because, for this participant, nearly every target received a rating of 20 (142 out 150 targets received that rating). As a result, the standard deviation of the ratings to targets was unusually low. For the remaining participants, the average value of s_{lure}/s_{target} was .73, which suggests an unequal-variance model. By comparison, the average Gaussian-based estimate of $\sigma_{lure}/\sigma_{target}$ obtained from the slope of the best fitting line to each participant's z-ROC was .61, which also suggests an unequal-variance model. Although the mean ratio estimates differed to a greater degree than in Experiments 1 and 2, the correlation between the two estimates across participants was still significant, $r(12) = .64$. As in the previous experiments, the relationship between confidence and accuracy was strong (see Figure 9a), and many targets—but few lures—received the highest rating of 20 (see Figure 9b). However, unlike in the previous experiments, the distribution also reveals a bias at the lower end of the scale such that many lures and quite a few targets received a rating of 1. This phenomenon may be a side effect of asking for remember/know/guess judgments, and it may explain why the mean ratio estimates obtained from ratings and Gaussian ROC analysis did not agree as closely as they did in Experiments 1 and 2.

We next analyzed the remember, know, and guess judgments. The remember hit and false-alarm rates were 0.28 and 0.03,

Table 4
Descriptive Statistics for the Rating Scale Data From Experiment 4

	Direct ratings method						
Subject no.	m_{target}	m_{lure}	s_{target}	s_{lure}	s_{lure}/s_{target}	d_r	ROC analysis slope
1	16.83	4.94	5.60	5.43	0.97	2.16	0.61
2	14.41	5.42	6.82	4.86	0.71	1.52	0.54
3	12.23	2.99	8.51	4.06	0.48	1.39	0.39
4	13.49	5.13	6.45	4.20	0.65	1.54	0.51
5	11.73	1.89	9.07	3.64	0.40	1.42	0.43
6	9.67	1.73	9.09	3.23	0.36	1.16	0.48
8	9.84	2.45	8.76	4.40	0.50	1.07	0.71
9	15.13	9.40	6.84	6.75	0.99	0.84	0.63
10	12.53	9.39	5.23	4.75	0.91	0.63	0.86
11	10.87	7.21	5.72	4.13	0.72	0.74	0.70
12	8.50	6.60	5.07	3.32	0.65	0.44	0.64
13	11.58	7.52	8.13	6.72	0.83	0.54	0.57
14	15.61	8.17	6.57	7.21	1.10	1.08	0.72
15	12.40	6.16	8.03	7.20	0.90	0.82	0.80
Mean	12.49	5.64	7.13	4.99	0.73	1.10	0.61

Note. m_{target} and m_{lure} represent the mean ratings to targets and lures, respectively; s_{target} and s_{lure} represent the corresponding standard deviations; d_r represents a discriminability measure equal to m_{target} minus m_{lure} divided by the root mean square of s_{target} and s_{lure}; ROC = receiver operating characteristic.

respectively; the know hit and false alarm rates were 0.26 and 0.08, respectively; and the guess hit and false alarm rates were 0.09 and 0.08, respectively. The average confidence rating associated with all remember judgments was 19.1, and the corresponding values for know and guess judgments were 18.1 and 13.0, respectively. Figure 10a shows the frequency of remember judgments to targets and lures as a function of each item's confidence rating, and it is clear that most remember judgments were associated with ratings of 20. Figure 10b shows the frequency distribution for know judgments. As with remember judgments, more know judgments occurred for targets rated 20 than for any other individual rating. Figure 10c shows the distribution for guess judgments, and it is clear that most of those fall at the lower end of range.

The key finding is that ratings of 20 were not exclusively associated with remember judgments. Across all participants, 503 targets were rated 20 and judged to be remembered (R-20), 399 targets were rated 20 and judged to be known (K-20), and five targets were rated 20 and judged to be guesses (G-20). Thus, even memories that are subjectively based on familiarity can often achieve ratings of 20, yet such memories are not easily differentiated from each other in terms of strength (according to the results of Experiments 1, 2, and 3). This suggests that memories become hard to scale when they are very strong, not when they are based on recollection.

Do the K-20 and R-20 judgments differ on either old/new or source accuracy? Out of the 15 participants, 12 provided responses for both K-20 and R-20 decisions. Although old/new accuracy is similarly high for both R-20 and K-20 judgments (.96 and .93, respectively), source recollection accuracy is considerably (and significantly) higher for R-20 judgments than for K-20 judgments (.85 and .71, respectively), $t(11) = 2.74$. Findings like these suggest that know judgments do not indicate the absence of recollection but instead indicate that there was not *enough* recollection to warrant a remember response. This is consistent with prior work (Johnson, McDuff, Rugg, & Norman, 2009; Wais, Mickes, & Wixted, 2008) and with a signal-detection-based theory of remember/know judgments proposed by Wixted and Mickes (2010a). Although K-20 judgments are not associated with the absence of recollection (i.e., they do not reflect pure familiarity-based decisions), their high old/new strength is likely attributable mainly to familiarity, not to strong recollection. If K-20 judgments do in fact reflect decisions that are based largely on familiarity, then it would suggest that even strong memories based largely on familiarity are hard to scale.

This interpretation suggests that strong memories are differentiable in some respects. That is, participants can tell whether strong memories are based mainly on recollection or mainly on familiarity. However, based on the results of Experiments 1 through 3, those same strong memories appear difficult to distinguish in terms of memory strength.

Previous studies vary in how often high-confidence know judgments occur. In some prior studies that used a 6-point confidence scale, virtually no high-confidence know judgments were observed (Stretch & Wixted, 1998; Yonelinas, 2001), but in other studies, they have been observed more reliably (e.g., Rotello, Macmillan, Reeder, & Wong, 2005). It is not clear why this difference exists, but an absence of high-confidence know judgments would occur if participants used remember/know judgments as a proxy for confidence (where "remember" means high confidence and "know" means low confidence). It seems likely that remember/know judgments are often used in this way (e.g., Dunn, 2004). This is why we took steps to ensure that participants understood that remember/know judgments should be used to differentiate between recollection and familiarity (not to differentiate between high and low confidence).

Using a 20-point confidence scale, Wixted and Mickes (2010a) also reported the occurrence of many K-20 judgments, and this was true whether or not the procedure included a source memory question following each remember/know/guess judgment. This suggests that the frequent use of K-20 judgments in our Experiment 4 is not the result of probing memory for source information. Thus, we interpret the data to mean that the K-20 judgments reflect

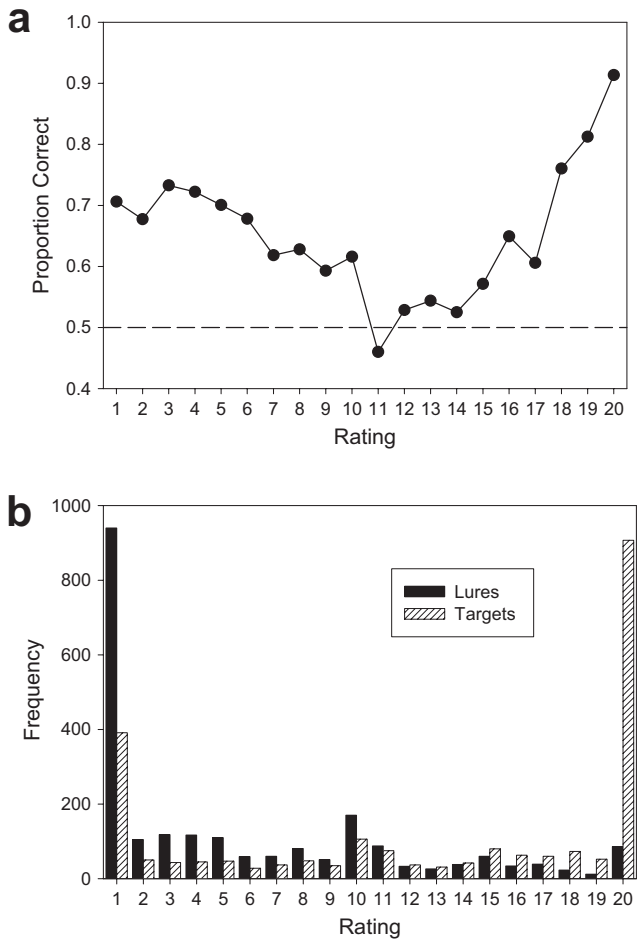

Figure 9. a. Accuracy (proportion correct) as a function of the confidence expressed in an old/new recognition decision in Experiment 4. b. Frequency distribution showing the number of responses made to targets and lures for ratings of 1 through 20 in Experiment 4.

strong memories based largely on familiarity (but strong memories appear to be unscalable anyway).

Experiment 5

Thus far, in all four experiments, participants bunched their responses at the ends of the scale, particularly at the high end of the scale. It therefore seems as though participants are unable to scale their strongest memories meaningfully, even when they are strongly urged to do so (as in Experiment 1), when they are provided with a second opportunity to scale the highest rated items (as in Experiment 2), and when they are not constrained to a specific scale (as in Experiment 3). Moreover, the results of Experiment 4 indicate that the strongest memories probably do not simply reflect the "all" state of an all-or-none recollection process (Yonelinas, 1994).

Why, then, are these strong memories unscalable? Perhaps memories, whether recollection based or familiarity based, simply have a ceiling level of strength (determined, for example, by neurophysiological constraints). In fact, Rouder et al. (2010) described a non-Gaussian model involving target and lure distributions that are bounded at the upper and lower ends of the strength scale instead of being unbounded (as the Gaussian model is). They produced this "probit model" by passing random Gaussian variables (X) through a $\Phi(2X/3)$ filter, where Φ is the standard normal cumulative distribution function, and they argued that it can account for data reported by Mickes et al. (2007) as well as the Gaussian model can. In effect, they suggested that the shape of the underlying memory strength distributions might correspond closely to the observed shape of the frequency distribution that Mickes et al. (2007) obtained using a 20-point rating scale.

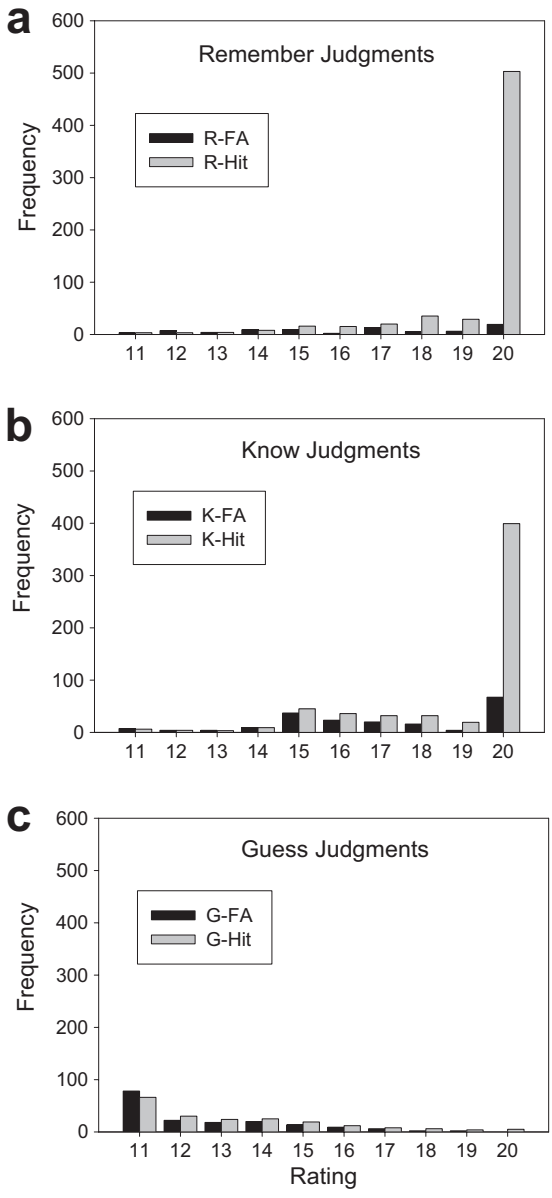

Figure 10. Frequency distribution showing the number of remember judgments (a), know judgments (b), and guess judgments (c) made to targets and lures for ratings of 11 through 20 in Experiment 4. FA = false alarm.

Wixted and Mickes (2010b) showed that the probit model, despite being derived in view of the frequency data it was designed to explain, did not fit the ROC data as well as a Gaussian model (a model that was not developed in view of the to-be-explained data). Thus, as yet, no specific quantitative model involving bounded memory strengths offers an interpretation of the data that is as compelling as the account offered by the unbounded Gaussian model. In addition, with the possible exception of all-or-none recollection, there appear to be no a priori theoretical considerations that would anticipate a memory-strength distribution with a ceiling level of strength. Thus, although our data could be interpreted to simply mean that memory strength has a heretofore unnoticed (and possibly physiologically based) upper limit, we suggest that a more theoretically interesting explanation may account for the bunching effect.

On the basis of the data reported by Mickes et al. (2007) and the additional data reported here, it seems that the bunching effect tends to occur at the point on the memory strength scale where the lure distribution begins to play a negligible role (i.e., at the point on the scale where errors drop close to zero). Although it could be a mere coincidence, this intriguing fact suggests another possible explanation for our results. This explanation begins by asking a simple question: What accounts for the impressive scaling expertise that participants bring with them into the laboratory? That is, how did they become experts at scaling memory over such a wide range of strength in the first place? One possible explanation is that the ability to scale an internal dimension (such as memory strength) is not innate and that people learn to do so by means of error feedback. That is, based on experience, people may learn to give confidence ratings that reflect learned probabilities of making an error for a given level of memory strength. Indeed, a confidence rating could be construed as a number that is directly related to the learned probability of making an error given the strength of memory associated with a test item (rather than as a number that reflects memory strength, per se).

If error feedback were the mechanism that teaches people to expertly scale memory in the first place, there would be no mechanism to teach people how to differentiate between their strongest, error-free memories. According to this idea, people learn that when memory is weak, the probability of making an error is high (warranting a low confidence rating). When the strength of memory is moderately high, the probability of making an error is lower (warranting a higher confidence rating). But when memory strength is high, the probability of making an error is essentially zero, just as it is when memory strength is even higher than that. Thus, the mechanism that may teach people to scale memory strength would not be available to teach them how to differentiate between their strongest memories. Moreover, it is hard to imagine why people would care to differentiate between various strong memories that are error free.

This possible explanation is not easy to test because we cannot use error feedback to teach participants to discriminate between their strongest memories for the same reason that life experience may not teach them to form that discrimination. The problem is that when items are rated 20, virtually no errors are made (so no error-based feedback can be provided). Thus, to test the feasibility of this idea, we devised an experiment in which participants were set up to make errors associated with their confidence ratings of 20. We did this by using a plurals-discrimination procedure in which lures differed from targets presented on the list only by the presence or absence of the letter *s* (without warning participants of this subtlety).[2] Thus, if the word "shoe" appeared on the list, we expected participants to make a considerable number of high-confidence errors to the lure "shoes." Halfway through the recognition test, we provided error-based feedback (the kind of feedback that we hypothesize participants have received throughout their lives) to see what effect that would have on their subsequent ratings. Would such feedback induce them to become more conservative (and, therefore, more accurate) in their use of the highest rating on the confidence scale? That is, would feedback accomplish what emphatic instructions and alternative rating scales did not accomplish?

Method

Participants. Sixty undergraduates from University of California, San Diego who participated for lower division psychology course credit were randomly assigned to one of two groups, the feedback group ($n = 30$) or the control group ($n = 30$).

Materials and design. To create the lists, we extracted 346 nouns from the MRC database (Coltheart, 1981). Words that were selected were four to seven letters in length with a concreteness rating between 600 and 700. From that pool, we removed words with identical spellings for singular and plural versions (e.g., moose and moose), ablaut plurals (e.g., foot and feet), and irregular plurals (e.g., ox and oxen). Of the remaining words, 132 were randomly selected for presentation during List 1 and 132 for List 2. Sixty-six words were singular, and 66 were plural for the study phase of both lists. There were no "new" words introduced during the testing phase. Instead, 66 words were reversed for plurality or singularity, which we refer to as "lures" (whereas we refer to the 66 words that maintained their plurality as "targets"). Thirty-three targets and 33 lures were presented on the first half of the recognition test, and 33 of each were presented on the second half. The practice portion contained four words during the study phase and eight during the test phase. Participants were not informed that they needed to attend to the plurality of the words, nor did they ask. For both the feedback and control groups, there were two sets of words (Sets A and B). The only difference between the two sets was that the words that were plural in Set A were singular in Set B (and vice versa). List 2 contained the same words the participants saw in the first list, only this time, they were reversed for plurality (e.g., if List 1 consisted of the words from Set A, List 2 consisted of the words from Set B). Each participant was randomly assigned to Set A or B for List 1 and to the feedback or control group.

Procedure. After the practice phase was completed, target words were randomly presented and appeared one at a time for three seconds each (with a 250-ms interstimulus interval fixation cross). During the test phase, targets and lures were presented individually for a confidence rating using a 20-point rating scale. After they answered half of the test items, participants in both groups were instructed (via a message presented on the computer screen) to retrieve the experimenter. For the participants in the feedback group, feedback for the items they rated was then dis-

[2] We thank Jeff Starns for suggesting this experiment.

played on the monitor. First, participants were shown the last 12 words that they had correctly identified as old or new along with their corresponding ratings. Next, they saw the last 12 words that they had incorrectly identified as old or new. For both correct and incorrect feedback slides, the originally presented words were in the first column, followed by the corresponding test word and the rating the test word received. Finally, they saw a summary slide that displayed the number of correct answers, number of incorrect answers, and overall accuracy (percentage correct). While viewing the feedback screens, the experimenter informed the participants that plurality mattered and that they should try not to make errors when using the high end of the confidence scale (in accordance with the instructions they had received at the beginning of the experiment). The issue of most interest was what effect this feedback would have on the second half of the test that followed the first list (before participants had an opportunity to focus on plurality during encoding—an opportunity they would have when the second list was presented).

Participants in the control group were not provided with feedback. Instead, they took a break that was equivalent in duration to the time required to provide feedback for those in the Experimental group (about 3 min). They were told, "This is a rest period. By the way, you may have noticed, but in case you have not, plurality matters. If the plurality of the test item matches the studied item, then it is old. If not, then it is new." The experimenter then exited the room. For both groups, List 2 was presented immediately after the testing phase of List 1. During the presentation of this list, all participants were aware that plurality was the important dimension. After answering half of the test questions following List 2, there was a 3-min break, and feedback information appeared on the screen for those in the experimental group again. List 2 was presented as a "new" list, but the items were the same as those on List 1 with plurality reversed. We used a study-test, study-test design to assess whether overall accuracy would improve when participants were alerted to the nature of the task halfway through the first list (after they had already encoded the words on List 1) or if awareness of the plurality aspect of the task was required at encoding (as was true of List 2) for performance to improve. This was not our central question, but it was an interesting secondary question because of its relevance to Tulving's principle of encoding specificity (Tulving, 1983). This principle predicts that postencoding feedback would not increase overall accuracy (whether or not accuracy increased for high-confidence ratings) and that overall accuracy would increase only if the list items were specifically encoded with respect to plurality (as would be the case for List 2).

Results

Before considering the effect of feedback, we again present a distributional analysis of ratings made using the 20-point confidence scale. The means and standard deviations of the ratings to the targets and lures are detailed in Table 5 for both the feedback and control groups. For each group, the statistics were computed separately for the first half and second half of each recognition test. The data from List 1 are of most interest because participants were unaware of the plurality manipulation when they attempted to memorize the list.

As shown in Table 5, the statistics computed directly from the ratings do not suggest the usual unequal-variance Gaussian model. Instead, the standard deviation of the lure distribution was slightly greater than that of the target distribution. The same was generally true of the ratio estimates obtained from Gaussian ROC analyses, although the latter were more variable (and were higher for the control group). Averaged across the feedback and control conditions (and across List 1 and List 2), the ratio estimate based on the ratings (s_{lure}/s_{target}) was 1.04, whereas the corresponding estimate based on Gaussian ROC analysis was 1.03. This suggests that ratings made using a 20-point scale yield distributional statistics are in reasonably good agreement with those obtained from Gaussian ROC analysis, even when the slope of the ROC does not suggest that the target distribution has greater variance than the lure distribution (as it usually does). In this case, the ratio estimates tended to slightly exceed 1.0 even when memory performance improved for List 2. Thus, this phenomenon does not seem to be attributable to the fact overall recognition memory performance was quite low following List 1.

The fact that the slopes were, if anything, greater than 1 is consistent with prior research by Jeneson, Kirwan, Hopkins, Wixted, and Squire (2010), which also used target and lures that differed in only one small detail. In that study, the slope of the ROC was greater than 1, even when overall memory performance

Table 5
Average Statistics for the Rating Scale Data From Experiment 5 for the Control Group and Feedback Group

Group	Test half	m_{target}	m_{lure}	s_{target}	s_{lure}	s_{lure}/s_{target}	d_r	ROC slope
				List 1				
Control	1st	14.19	12.19	6.73	7.43	1.17	0.29	1.19
Control	2nd	12.68	9.88	6.62	7.09	1.08	0.41	1.13
Feedback	1st	14.33	11.58	6.67	6.91	1.05	0.43	0.89
Feedback	2nd	12.05	9.58	5.42	5.32	0.99	0.43	0.94
				List 2				
Control	1st	14.29	6.85	6.41	6.42	1.03	1.25	1.15
Control	2nd	13.08	7.33	6.35	6.33	1.02	0.95	1.09
Feedback	1st	13.91	6.51	4.90	4.87	1.02	1.50	0.96
Feedback	2nd	13.88	7.26	4.70	4.71	1.01	1.36	0.93

Note. m_{target} and m_{lure} represent the mean ratings to targets and lures, respectively; s_{target} and s_{lure} represent the corresponding standard deviations; d_r represents a discriminability measure equal to m_{target} minus m_{lure} divided by the root mean square of s_{target} and s_{lure}; ROC = receiver operating characteristic.

was very good (e.g., $d' > 2$). Jeneson et al. (2010) suggested that the lure distribution may have equal or greater variance than the target distribution under these conditions because the lures are associated with both recall-to-reject (when the incorrect detail is recollected, adding to high-confidence correct rejections when recollection is strong) and recall-to-accept (when the incorrect detail is not noticed but other thoughts about the item are recollected, adding to high-confidence false alarms when recollection is strong). This would have the effect of increasing the variance of the lure distribution relative to the target distribution, for which only a recall-to-accept process would apply. Whether or not this explanation is correct, our findings suggest that direct ratings and Gaussian ROC analysis show good agreement even when the slope of the ROC is close to 1.

The question of primary interest in this study concerned the effect of feedback on the proportion and accuracy of items rated 20 during the recognition test that followed List 1. Figure 11 shows the frequency distributions for the feedback and control conditions. During the first half of the recognition test, many targets and lures received the highest rating of 20. Our assumption is that, being unaware of the nature of the recognition test, participants placed the criterion for making confidence ratings of 20 high on the memory strength axis (as in the previous experiments). The reason for the high false-alarm rate for items rated 20 is that the lures generated a strong memory signal that often exceeded the highest criterion (unlike in the previous experiments). During the second half of the recognition test, the frequency of those high-confidence responses decreased in both groups, but the decrease was noticeably larger in the feedback group.

Figure 12 focuses specifically on the decisions of most interest, namely, those made with the highest level of confidence (20). Figure 12a shows the proportion of all responses (summed across targets and lures) that received the highest rating of 20 in the first and second halves of the test. A mixed 2 (Testing Block: first half vs. second half) $\times$ 2 (Group: feedback vs. control) analysis of variance conducted on the proportion data shown in Figure 12a revealed a significant main effect of testing block, $F(1, 51) = 107.91$, $p < .001$; a marginally significant main effect of group, $F(1, 51) = 3.37$, $p = .073$; and a significant interaction, $F(1, 51) = 8.21$, $p < .01$. The interaction indicates that the decrease in the proportion of items rated 20 was greater in the feedback group than in the control group. Figure 12b shows the accuracy of decisions made with a confidence rating of 20. Because the plurals discrimination procedure had the intended effect of inducing many high-confidence decisions to both targets and lures, accuracy for items rated 20 was low (approximately 62% correct averaged across both groups). This result is in stark contrast to the much more impressive high-confidence accuracy scores observed in Experiments 1–4. An analysis of variance conducted on the accuracy scores for the high-confidence ratings shown in Figure 12b revealed a significant main effect of testing block, $F(1, 51) = 13.53$, $p < .001$; a significant main effect of group, $F(1, 51) = 12.21$, $p < .01$; and a significant interaction, $F(1, 51) = 4.43$, $p < .05$. The interaction indicates that feedback provided halfway through the recognition test improved accuracy for high-confidence ratings, but simply learning about the nature of the experiment (i.e., that plurality matters), halfway through had a lesser effect. The slight improvement in accuracy for the control group (first half to second half) did not approach significance.

The information provided to the feedback group following List 1 did not enhance the overall ability to discriminate targets from lures. That is, as shown in Table 5, for both the feedback group and

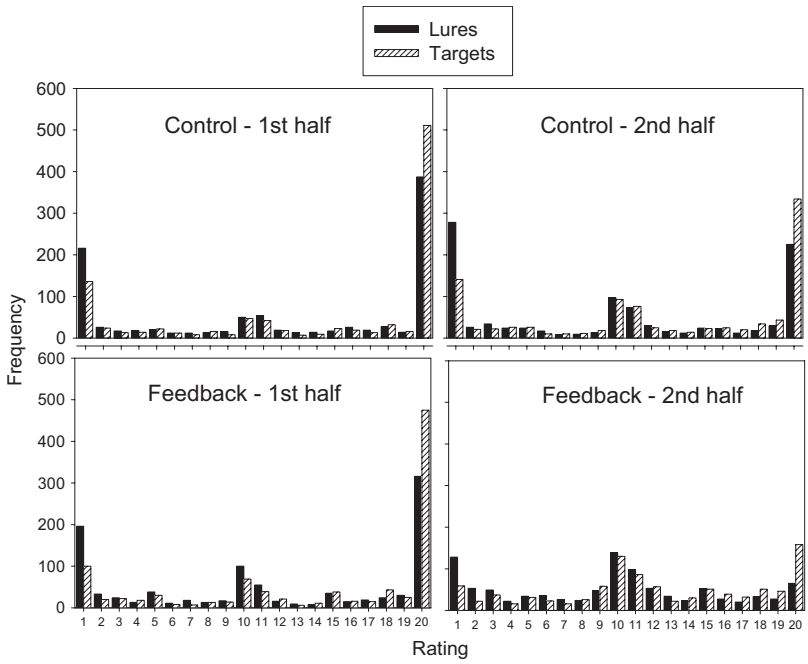

Figure 11. Frequency distributions showing the number of responses made to targets and lures for ratings of 1 through 20. The data are shown for the control group and the feedback group, separately for the first and second halves of the recognition test following List 1.

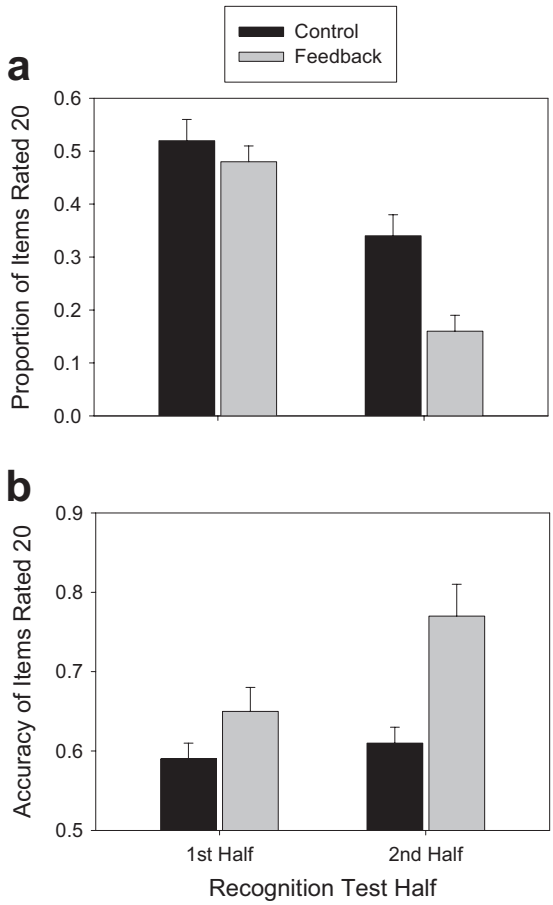

Figure 12. a. Mean proportion of responses given ratings of 20 for the control group and feedback group in the first and second halves of the recognition test in Experiment 5 (List 1). b. Mean proportion correct for responses given ratings of 20 for the control group and feedback group in the first and second halves of the recognition test in Experiment 5 (List 1).

the control group, the d_r discriminability measure computed from the ratings is similarly low during the first and second halves of the test (traditional d' scores are very similar to the d_r scores). The control group showed some increase in discriminability in the second half, but the difference was not significant ($p = .11$). It was only following List 2 that d_r scores noticeably increased (see Table 5). As indicated earlier, this pattern of results accords with Tulving's encoding specificity principle because information about the importance of plurality improved memory only when that information was taken into consideration at encoding.

Our results were consistent with other recognition memory experiments in which feedback was provided (e.g., Han & Dobbins, 2008; Kantner & Lindsay, 2010; Verde & Rotello, 2007), which resulted in a criterion shift without changing overall performance. Although overall accuracy remained unchanged following midlist feedback, the accuracy associated with ratings of 20 increased considerably (as shown in Figure 11b). This suggests that feedback caused participants to be more conservative about expressing high confidence (i.e., they adjusted the high-confidence decision criterion), an effect that we hypothesize occurs in everyday life in response to error feedback.

The conservative shift in the confidence criteria is clearly evident for the feedback group, postfeedback, and it is not restricted to the high-confidence criterion. This is shown in Figures 13A and 13B, where, in the first half of the test following List 1, the points along the ROC curve are clustered together for both groups. However, in the second half of the test, the points for the feedback group spread out along the ROC curve, as shown in Figure 13B. This reflects a general spreading of the 19 confidence criteria as the highest criterion is moved to a more conservative (i.e., higher) position on the memory strength scale. The control group shows some evidence for a criterion shift from the first half to the second half as well, but it is clearly smaller, and the points on the ROC appear to shift in lockstep. Even this small criterion shift may be due to indirect error feedback as participants deduce that they have been making many high-confidence errors (indeed, subjects sometimes spontaneously commented on their sudden awareness of the fact that they had been making high-confidence errors).

An alternative approach to inducing criterion shifts would be to train participants with feedback that does not draw attention to the critical dimension (namely, plurality). This could be done, for example, by using trial-by-trial error feedback in which each recognition decision is followed by nothing more than "correct" or

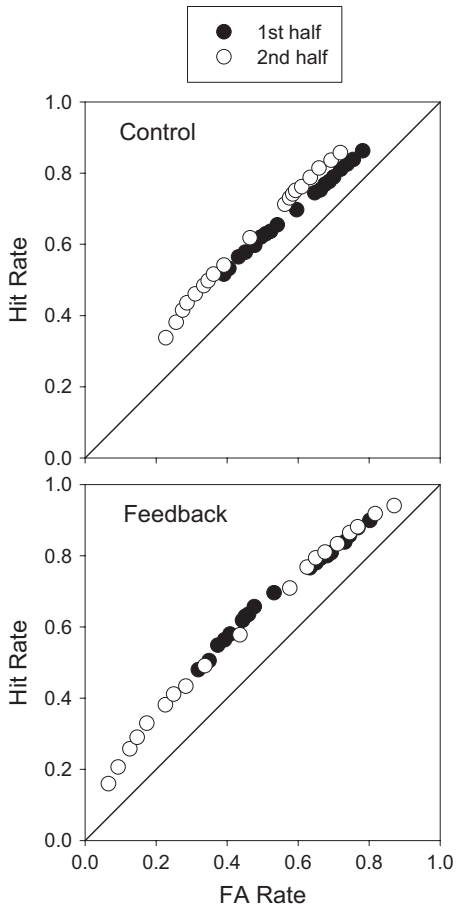

Figure 13. a. Receiver operating characteristic (ROC) for the control group on List 1 in Experiment 5. b. ROC for the feedback group on List 1 in Experiment 5. FA = false alarm.

"incorrect." Our results do not indicate whether or not participants would eventually hit upon the critical dimension with this kind of error feedback (though we assume they would). However, our assumption is that feedback in everyday life usually involves more information than simply indicating the correctness of the decision. For example, when children make high-confidence recognition errors, the feedback they receive may not only contain information that the decision is incorrect but may also contain information about the nature of the error (e.g., an explanation of how the incorrectly recognized individual looks a lot like, but not exactly like, Grandpa). It is this kind of more generalized feedback that our participants received in Experiment 5.

General Discussion

The novel finding from this set of experiments is that strong and accurate memories associated with the highest levels of confidence are difficult to differentially scale. That is to say, participants find it difficult to meaningfully assign numerical ratings to differentiate between a relatively large number of their strongest memories. This is true even though they are quite adept at using a numerical scale to differentiate between other memories, including moderately strong ones. This ability to accurately scale memory strength for most items is shown by the fact that as confidence ratings increase from low levels to high levels, accuracy increases from chance performance to, in some cases, nearly 100% correct (e.g., Figure 1). In addition, target and lure distributional statistics computed directly from the ratings that were made using a 20-point rating scale correspond closely to the distributional statistics estimated by fitting a Gaussian model to the ROC. Thus, the different levels of confidence used to gauge the strength of different memories provide valid information. But this ability to supply valid scalar information about memory strength apparently does not extend to a significant subset of memories that are very strong. As a result, confidence ratings for target items invariably take on the appearance of a distribution that is bunched on the right end. This is true even though, in all other respects, the data are consistent with a Gaussian signal-detection model. If the Gaussian model is correct, then strong memories should be continuously distributed, and a right-tailed (not a bunched) distribution of ratings to targets should have been observed.

Why are strong memories so resistant to further scaling? One possibility is that the strongest memories are hard to scale simply because they represent a memory-strength ceiling. If so, then the strongest memories would not differ from each other in memory strength (just as the ratings suggest), and the underlying distribution would have a shape similar to the observed distribution of ratings (e.g., Figure 3) instead of having a Gaussian or Gaussian-like distribution. Indeed, Rouder et al. (2010) argued that a bounded non-Gaussian model can accommodate the data as well as the traditional unbounded Gaussian model can. Although Wixted and Mickes (2010b) showed that the particular model described by Rouder et al. (2010) did not fit the data as well as a Gaussian model, it seems reasonable to suppose that some other bounded model could.

Why might memory strength be bounded? The only a priori theoretical consideration that might help to explain the bunching effect is all-or-none recollection (Yonelinas, 1994). It is conceivable that items receive a rating of 20 when recollection succeeds, whereas familiarity is a continuous process associated mainly with memories of lesser strength and lesser degrees of confidence. If successful recollection entails the complete retrieval of an encoding experience, it would seem to follow that recollection-based memories would receive the highest confidence rating and would not be distinguishable from each other. However, using the remember/know procedure, we found that a substantial number of old/new decisions associated with confidence ratings of 20 were associated with know judgments. That is, participants indicated that many of these decisions were based on familiarity (something they would be unlikely to do if they were in the "all" state of a recollection experience). Even so, strong memories are hard to scale. The fact that error feedback ultimately succeeded in getting participants to scale many of their strong memories (Experiment 5) also suggests that those memories do not reflect an "all" state.

If the apparent limitation on the ability to scale strong memories does not indicate all-or-none recollection, then perhaps it reflects some previously unrecognized constraint, such as a physiological limit on the magnitude of the neural activity that underlies memory strength (Rouder et al., 2010). This possibility also seems inconsistent with the results of Experiment 5, which showed that error feedback can make many strong memories scalable. That is, post-feedback, confidence criteria were adjusted such that ratings of 20 were reserved for especially strong memories (though whether or not *those* memories are scalable is not known).

An alternative and perhaps more interesting possibility is that the ability to meaningfully assign different levels of confidence to memories that vary in strength (and that correspondingly vary in their likelihood of being correct) is derived from past experience. That is, experience may be what teaches a participant to express high confidence when memory is strong (and likely to be accurate) and to express low confidence when it is weak (and likely to be inaccurate). Perhaps because of that past experience, participants can immediately and effectively use a confidence scale in the laboratory without any special training or instructions (unlike in the case of remember/know judgments, which typically require detailed instructions that are often misunderstood anyway). However, memories at the strongest end of the scale, being almost perfectly accurate, would not be associated with differential feedback in everyday life. As a result, participants may never have learned—indeed, may never have had any reason to learn—to differentiate between them.

The closer the correspondence between a participant's confidence and their actual performance, the more calibrated the participant is said to be (e.g., Lichtenstein, Fischhoff, & Phillips, 1982). Previous research suggests that, on a variety of tasks, young children exhibit confidence judgments that are poorly calibrated to accuracy compared with the confidence judgments of older children. In fact, young children have been referred to as "eternal optimists" because of their tendency to express high confidence in all of their decisions (Newman & Wick, 1987). With error feedback provided in the laboratory, however, the calibration exhibited by both younger and older children improves. The same kind of training may occur throughout our daily lives, presumably beginning at a young age.

In a study of general knowledge questions in adults, Stock, Kulhavy, Pridemore, and Krug (1992) found that participants spent more time studying feedback (before moving on to the next question) for high-confidence errors than for low-confidence errors, a

result they interpreted in the following way: "Those results were explained by the proposition that people try to reduce discrepancies between what they think they know and what feedback indicates they know" (p. 654). Similarly, more recent studies by Butterfield and Metcalfe (2006) and Fazio and Marsh (2009) suggest that people pay particular attention to (and learn from) high-confidence errors on tests of general knowledge. The results of our Experiment 5 further suggest that participants not only attend to high-confidence errors, they also make adjustments to decrease the likelihood of making such errors in the future. It seems reasonable to suppose that the same kind of training happens outside the laboratory and that this, perhaps, accounts for why people acquire the ability to appropriately scale the strength of their memory (which they can then easily display in the laboratory without having to be trained at all in the use of a 20-point confidence scale).

The key consideration for purposes of understanding why participants are apparently unable to scale their strongest memories is that the learning process that may account for a participant's general ability to scale memory strength involves differential *error feedback*. Such training may be necessary for people to make effective use of their own internal sense of memory strength. In this regard, Skinner (1953) once made the following argument: "Strangely enough, it is the community which teaches the individual to 'know himself'" (p. 261). More specifically, Skinner argued that certain aspects of mental life remain undifferentiated in the absence of explicit discrimination training. To make this point, he used the example of color: "Anyone who as suddenly been required to make fine color discriminations will usually agree that he now 'sees' colors which he had not previously 'seen' " (p. 260). It is conceivable that it is the same way with the subjective sense of memory strength. Through discrimination training involving differential error feedback, people come to be able to accurately gauge to the strength of their own memories such that the relationship between confidence and accuracy becomes quite strong (as is evident in Figure 1).

According to this interpretation, strong memories do, in fact, vary in strength, as the standard Gaussian signal-detection model assumes they do. However, as illustrated in Figure 2a, these strong memories are distributed along the memory strength scale in a region where lures rarely reach unless special procedures are used (e.g., Roediger & McDermott, 1995). As such, differential error feedback would not provide any indication of differences in memory strength in that high region of the memory strength scale, so participants may never have learned to discriminate any differences in memory strength that may exist. Indeed, it is hard to imagine why they would ever care to do so.

Error feedback is, of course, known to play an important role in teaching fine discriminations in other contexts. Kornell and Bjork (2008), for example, taught participants to discriminate paintings by different artists. In this case, initial training was conducted by simply pairing each painting with the artist's name. However, during testing, further training occurred (and performance further improved) when participants attempted to identify the artist of unfamiliar paintings and received feedback about the accuracy of their decisions. Our suggestion is that learning to discriminate accurate from inaccurate memories is a bit like learning to discriminate paintings by different artists, except that only the error-feedback mechanism is available to teach the discrimination.

Our emphasis on the strong relationship between confidence and accuracy in recognition memory may seem diametrically opposed to other research suggesting that this relationship is weak and that confident eyewitnesses are sometimes badly mistaken when testifying in courts of law. In our own data, the relationship between confidence and accuracy is invariably strong, especially for targets (as shown in Figure 1). Juslin, Olsson, and Winman (1996) reported a similarly strong confidence–accuracy relationship in the context of a photo lineup study in which participants viewed a videotape of a staged crime scene (two men stealing a bicycle) and were later asked to identify the culprits in photo lineups in which the lures were selected by experienced police officers. Thus, the relationship between confidence and accuracy is strong even under more realistic conditions that are relevant to the legal setting. Juslin et al. (1996) suggested that the widespread impression that the confidence–accuracy relationship is weak stems partly from the use of the point–biserial correlation in prior studies to measure that relationship. This correlation coefficient can be (and often is) misleadingly low, even when the confidence–accuracy relationship (expressed as the probability of being correct for each level of confidence) is actually quite strong.

Despite the strong relationship between confidence and accuracy in our study and despite the very high accuracy associated with confidence ratings of 20, our results should not be taken to mean that, generally speaking, verbal expressions of high confidence are infallible. When a 20-point confidence scale is used, we find that most participants place the high-confidence criterion high enough on the memory strength scale such that no errors are made for targets (i.e., for ratings of 20). However, participants do make an appreciable number of high-confidence errors for lures (i.e., for ratings of 1). Thus, an expression of high confidence, per se, is not an indication that the recognition decision is necessarily correct. Moreover, had a 6-point confidence scale been used, errors may have been more common for expressions of high confidence, even for targets (i.e., for ratings of 6). This would occur if the criterion for making a rating of 6 were placed lower on the memory strength scale than the criterion for making a rating of 20, even though both ratings might be associated with the verbal label "high confidence."

These considerations suggest that there is no contradiction between the idea that the relationship between confidence and accuracy is strong in recognition memory and the idea that eyewitnesses (like participants in the laboratory) can be mistaken about their confident recognition memory decisions. Even if the confidence–accuracy relationship is strong, the mere verbal expression high-confidence cannot reasonably be equated with infallibility (especially not in a court of law). Indeed, expressions of confidence can be influenced by a variety of variables other than memory strength (e.g., Busey, Tunnicliff, Loftus, & Loftus, 2000), and to the extent that these variables play a role in a recognition memory decision, accuracy may be impaired. High-confidence errors can occur, and the relationship between confidence and accuracy in recognition memory is strong. Both are true. Our suggestion is that the relationship is strong because people have learned from the error feedback they have received during a lifetime of making recognition memory decisions. If so, then when memories become very strong—so strong that they are essentially error free—it should no longer be possible to differentiate between

them in terms of memory strength, and this may be why strong memories, unlike weaker memories, are so hard to scale.

References

Anderson, N. H. (1961). Scales and statistics: Parametric and nonparametric. *Psychological Bulletin, 58,* 305–316. doi:10.1037/h0042576

Busey, T. A., Tunnicliff, J., Loftus, G. R., & Loftus, E. F. (2000). Accounts of the confidence–accuracy relation in recognition memory. *Psychonomic Bulletin & Review, 7,* 26–48. doi:10.3758/BF03210724

Butterfield, B., & Metcalfe, J. (2006). The correction of errors committed with high confidence. *Metacognition Learning, 1,* 69–84. doi:10.1007/s11409-006-6894-z

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

Criss, A. H. (2009). The distribution of subjective memory strength: List strength and response bias. *Cognitive Psychology, 59,* 297–319.

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

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

Eldridge, L. L., Engel, S. A., Zeineh, M. M., Bookheimer, S. Y., & Knowlton, B. J. (2005). A dissociation of encoding and retrieval processes in the human hippocampus. *Journal of Neuroscience, 25,* 3280–3286.

Fazio, L. K., & Marsh, E. J. (2009). Surprising feedback improves later memory. *Psychonomic Bulletin & Review, 16,* 88–92.

Gardiner, J. M. (1998). Functional aspects of recollective experience. *Memory & Cognition, 16,* 309–313.

Gardiner, J., Richardson-Klavehn, A., & Ramponi, C. (1997). On reporting recollective experiences and "direct access to memory systems." *Psychological Science, 8,* 391–394.

Han, S., & Dobbins, I. G. (2008). Examining recognition criterion rigidity during testing using a biased-feedback technique: Evidence for adaptive criterion learning. *Memory & Cognition, 36,* 703–715.

Jeneson, A., Kirwan, C. B., Hopkins, R. O., Wixted, J. T., & Squire, L. R. (2010). Recognition memory and the hippocampus: A test of the hippocampal contribution to recollection and familiarity. *Learning and Memory, 17,* 63–70.

Johnson, J. D., McDuff, S. G. R., Rugg, M. D., & Norman, K. A. (2009). Recollection, familiarity, and cortical reinstatement: A multivoxel pattern analysis. *Neuron, 63,* 697–708.

Juslin, P., Olsson, N., & Winman, A. (1996). Calibration and diagnosticity of confidence in eyewitness identification: Comments on what can be inferred from the low confidence–accuracy correlation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 1304–1316.

Kantner, J., & Lindsay, D. S. (2010). Can corrective feedback improve recognition memory? *Memory & Cognition, 38,* 389–406.

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction?" *Psychological Science, 19,* 585–592.

Lichtenstein, S., Fischhoff, B., & Phillips, L. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases.* New York, NY: Cambridge University Press.

Mickes, L., Wixted, J. T., & Wais, P. E. (2007). A direct test of the unequal-variance signal-detection model of recognition memory. *Psychonomic Bulletin & Review, 14,* 858–865.

Newman, R. S., & Wick, P. L. (1987). Effect of age, skill, and performance feedback on children's judgments of confidence. *Journal of Educational Psychology, 79,* 115–119.

Rajaram, S. (1993). Remembering and knowing: Two means of access to the personal past. *Memory & Cognition, 2,* 89–102.

Ratcliff, R., & Murdock, B. B., Jr. (1976). Retrieval processes in recognition memory. *Psychological Review, 83,* 190–214.

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

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 803–814.

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

Rouder, J. N., Pratte, M. S., & Morey, R. D. (2010). Latent mnemonic strengths are latent: A comment on Mickes, Wixted, and Wais (2007). *Psychonomic Bulletin & Review, 17,* 427–435.

Skinner, B. F. (1953). *Science and human behavior.* New York, NY: Free Press.

Stock, W. A., Kulhavy, R. A., Pridemore, D. R., & Krug, D. (1992). Responding to feedback after multiple-choice answers: The influence of response confidence. *Quarterly Journal of Experimental Psychology, 45,* 649–667.

Stretch, V., & Wixted, J. T. (1998). On the difference between strength-based and frequency-based mirror effects in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 1379–1396.

Tulving, E. (1983). *Elements of episodic memory.* Oxford, England: Clarendon Press.

Uncapher, M., & Rugg, M. (2005). Encoding and the durability of episodic memory: A functional magnetic resonance imaging study. *Journal of Neuroscience, 25,* 7260–7267.

Verde, M. F., & Rotello, C. M. (2007). Memory strength and the decision process in recognition memory. *Memory & Cognition, 35,* 254–262.

Wais, P. E., Mickes, L., & Wixted, J. T. (2008). Remember/know judgments probe degrees of recollection. *Journal of Cognitive Neuroscience, 20,* 400–405.

Wixted, J. T., & Mickes, L. (2010a). A continuous dual-process model of remember/know judgments. *Psychological Review, 117,* 1025–1054.

Wixted, J. T., & Mickes, L. (2010b). Useful scientific theories are useful: A reply to Rouder, Pratte, and Morey (2010). *Psychonomic Bulletin & Review, 17,* 436–442.

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

Yonelinas, A. P., Kroll, N. E., Quamme, J. R., Lazzara, M. M., Sauve, M. J., Widaman, K. F., & Knight, R. T. (2002). Effects of extensive temporal lobe damage or mild hypoxia on recollection and familiarity. *Nature Neuroscience, 5,* 1236–1241.

Received April 28, 2009
Revision received November 23, 2010
Accepted November 29, 2010 ■