Supplementary Materials for

Does science have an inference crisis? Assessing theoretical conclusions with blinded inference

Jeffrey J. Starns, Andrea M. Cataldo, Caren M. Rotello, Jeffrey Annis, Andrew Aschenbrenner, Arndt Bröder, Gregory Cox, Amy Criss, Ryan A. Curl, Ian G. Dobbins, John Dunn, Tasnuva Enam, Nathan J. Evans, Simon Farrell, Scott H. Fraundorf, Scott D. Gronlund, Andrew Heathcote, Daniel W. Heck, Jason L. Hicks, Mark J. Huff, David Kellen, Kylie N. Key, Asli Kilic, Karl Christoph Klauer, Kyle R. Kraemer, Fábio P. Leite, Marianne E. Lloyd, Simone Malejka, Alice Mason, Ryan M. McAdoo, Ian M. McDonough, Robert B. Michael, Laura Mickes, Eda Mizrak, David P. Morgan, Shane T. Mueller, Adam Osth, Angus Reynolds, Travis M. Seale-Carlisle, Henrik Singmann, Jennifer F. Sloane, Andrew M. Smith, Gabriel Tillman, Don van Ravenzwaaij, Christoph T. Weidemann, Gary L. Wells, Corey N. White, Jack Wilson

Correspondence to: jstarns@psych.umass.edu

This PDF file includes:

Supplementary Text
Figs. S1 to S3
Table S1
Supplementary References

Supplementary Text

Contributors' Understanding of the Problem

Two concerns about our results center on whether our contributors were appropriately prepared for their task, and if they took the inferential task seriously. Regarding the first concern, consider the following excerpt from the emailed invitation to participate:

> The goal of this study is NOT to compare different memory models on model fit, statistical qualities, or psychological plausibility. Instead, **the aim is to compare the psychological inferences drawn when using different tools to understand recognition decisions**.
>
> To reach this goal, we are in the process of conducting several experiments in which various properties of recognition performance were explicitly manipulated. Specifically, we are using standard manipulations to influence the discriminability of memory evidence and the bias to respond "Studied." **You and other experts are asked to infer what was varied across conditions, using any tool, strategy, or model that you like.** That is, you will decide whether each 2-condition data set was generated by an experiment that manipulated discriminability [i.e., memory accuracy], bias, both, or neither across the conditions.

Thus, we feel confident that the contributors understood the task they would face. Supporting this claim, six individuals responded with substantive questions (about, e.g., what data would be shared; how many analyses they could submit), half of whom decided not to participate. Regarding the second concern, we note that of the 27 contributors who participated, 19 were willing to attach their names to their inferences, and 10 submitted more than one analysis of each

experiment. Thus, we believe their inferences reflected their true professional interpretations of the data.

Contributors' Methods

The variability in inferences described in the main text was matched by a wide range of variability in the analysis methods selected by our contributors. Submissions used a variety of techniques that are purportedly capable of distinguishing discriminability and response bias, including a) mathematical models of choice behavior like Signal Detection Theory (SDT; Macmillan & Creelman, 2005) and the two-high-threshold model (2HT; Riefer & Batchelder, 1988); b) mathematical models of choice and response time like the drift diffusion model (DDM, Ratcliff, 1978) and the Linear Ballistic Accumulator model (LBA, Brown & Heathcote, 2008); c) non-parametric measures like A' (Pollack & Norman, 1964) or area under the reaction-time-based ROC curve (Thomas & Myers, 1972); and d) miscellaneous techniques like visually interpreting plots of the data, reporting a 50% chance of an effect for every data set, using a general linear model (GLM), or using the Retrieving Effectively from Memory (REM) process model (Shiffrin & Steyvers, 1997). Many contributors used multiple methods (see OSF for details). Some contributors used traditional frequentist methods (e.g., maximum likelihood estimation or significance tests) and others used Bayesian methods (e.g., posterior distributions of parameters or model selection via Bayes Factors). None of the submitted analyses were exactly identical: even similar analysis strategies differed in terms of exclusion criteria, and the subset of contributors who did not exclude any data each used different analysis methods.

Simulation Method

Assessing the success of our contributors required us to determine the expected accuracy level for these inferences, given the variability inherent in random samples of data (like the

samples that we collected in Phase 1). Even a valid inference procedure will sometimes reach inaccurate conclusions due to sampling variability, so we needed to identify a benchmark accuracy level below which it would be reasonable to conclude that an invalid inference technique had been applied. To estimate this benchmark, we ran a large number of simulated experiments with the same methodological details as Phase 1 and the same number of (simulated) participants and trials per participant. We then applied a theoretically appropriate inferential tool to the results. That is, we ensured a valid measurement procedure by generating data from a signal detection theory (SDT) model and using inferential measures of discriminability and bias that were derived from this same model. In this section, we describe each step of our simulation procedure.

Generating parameter values. We simulated data using a signal detection model (Fig. S1) with varying parameter values across simulated participants. To ensure that the distributions of parameter values across participants were consistent with the empirical data from Phase 1, we performed signal detection fits to the full empirical data set using hierarchical Bayesian methods. In the model, the strength distribution for unstudied items (lures) had a mean of zero and a standard deviation of one, satisfying the conventional scaling assumptions of SDT (Macmillan & Creelman, 2005). Each participant, p, had a unique μ parameter representing the mean of the strength distribution for studied items (i.e., targets) and a unique δ parameter representing the deviation of the response criterion from the midpoint between the lure and target strength distributions. For example, if $\mu = 1.6$ and $\delta = -0.2$, then the response criterion was placed at 0.6 ($= 0.5 \times 1.6 - 0.2$). Higher δ values indicate more conservative responding that results in fewer "old" decisions. The standard deviation of the target distribution was fixed at 1.25 for all participants in light of the substantial evidence for unequal variance in evidence strength for

recognition memory (Macmillan & Creelman, 2005). Thus, the likelihood of an "old" response on each trial was $1 - \Phi\left(\frac{\mu_p}{2} + \delta_p\right)$ for lures and $1 - \Phi\left(\frac{\left(\frac{\mu_p}{2}+\delta_p\right) - \mu_p}{1.25}\right)$ for targets, where Φ is the cumulative density function of a standard Gaussian and p indexes an individual participant. Each trial outcome, y_i, was modeled as a Bernoulli distribution parameterized with these response probabilities.

In the model, the μ and δ parameters followed Gaussian distributions across participants. As shown in Fig. S1, the parameters defining these across-participant distributions have a superscript to identify which individual-level parameter the distribution applies to and a subscript identifying the experimental condition. The distributions for the μ parameter had different means for conditions with words studied once, twice, or three times, denoted by μ_m^{μ}, where m could take values 1-3. The distributions for the δ parameter had different means for each of the nine conditions created by crossing number of study trials with the biasing conditions, denoted by μ_c^{δ}, where c could take values 1-9. The model also had two free parameters to define the across-participant standard deviation in the μ and δ parameters, σ^{μ} and σ^{δ}, respectively. We assumed that the across-participant variability in parameter values was equal across conditions. We used diffuse priors on the parameters defining the across-participant distributions; specifically, Gaussian distributions with a mean of zero and standard deviation of 10 for the across-participant distribution means and a uniform distribution from 0 to 10 for the across-participant standard deviations.

We used JAGS (Plummer & Plummer, 2003) to define the posterior distributions of parameter values. We were mostly interested in the parameters defining the across-participant distributions, and we generated point estimates of these values by taking the median of each relevant posterior distribution. Table S1 shows the estimates for the average μ and δ values

across participants in each condition (μ^{μ} and μ^{δ}, respectively). As expected, the μ^{μ} values increased with additional learning attempts, and the μ^{δ} values indicated that responding was more liberal (conservative) when participants were asked to specifically avoid misses (false alarms) compared to the condition that equally emphasized avoiding both types of errors. The standard deviation in parameter values across participants was .74 for μ and .33 for δ.

In the simulations, the parameter values for each simulated participant were randomly sampled from Gaussian distributions. Based on the condition assigned to the simulated participant, the means for these distributions were the appropriate values in Table S1. The standard deviation was .74 for μ and .33 for δ. Thus, the simulated participants matched the real participants both in terms of overall performance levels and variability in performance from one participant to the next.

Simulation procedure. We performed 5000 simulated replications of the study. Consistent with the empirical data, each replication comprised seven two-condition experiments with 24 simulated participants in each condition (i.e., the minimum sample size across the seven experiments) and 100 simulated trials for each participant (50 target and 50 lure trials). For each replication, the computer first selected across-participant distributions of parameter values to match the condition structure of the data analyzed by contributors. For example, in the first data set one condition had items studied three times with liberal bias instructions and the other had items studied three times with conservative bias instructions, and the means of the parameter-generating distributions were those reported in the corresponding cells in Table S1. Next, the computer randomly sampled parameter values for each simulated participant from these across-participant distributions and randomly sampled a data set for each simulated participant from an

SDT model with the participant-level parameters. We used the simulated data to calculate discriminability and bias measures for each simulated participant with the formulas

$$\mu' = 1.25\varphi(hr) - \varphi(fr)$$

$$\delta' = -\varphi(fr) - .5\mu'$$

where μ' is the estimated mean of the target distribution, δ' is the estimated deviation of the response criterion for the midpoint between the target and lure distributions, *hr* is the hit rate, *fr* is the false-alarm rate, and φ is the inverse of the cumulative distribution function for a standard normal distribution. These measures assume that memory strength follows Gaussian distributions and that the standard deviation of strength values is 25% higher for targets than for lures, which matches the model that generated the data.

We submitted the μ' estimates to a Bayesian t-test with a Cauchy prior on standardized effect size (Rouder, Speckman, Sun, Morey, & Iverson, 2009) to define a Bayes Factor (*BF*) contrasting the hypothesis that each experiment involved a discriminability effect to the hypothesis that it did not. To translate the *BF* to a probability that there was a discriminability effect (i.e., the same judgment that out contributors were asked to submit), we used a prior odds of 1 (i.e., the effect and no-effect models were deemed equally likely *a priori*), so the posterior odds were equal to the *BF* for the test and the posterior probability of an effect equaled $\frac{BF}{BF+1}$. For each of the 5000 replications, we then calculated the same performance metrics that were used to evaluate our contributors' responses; that is, the number of experiments with correct inferences and the adjusted Brier score across the 7 experiments. Inferences were scored correct if the

outcome deemed more likely was the actual outcome (e.g., over 50% chance of a discriminability effect for experiments from two different levels of the study repetition variable, or under a 50% chance for experiments from the same level). Brier scores for each replication were calculated by averaging the squared deviations between the inferred probability of a discriminability effect and the actual outcome of the experiment for each of the seven experiments within the replication. Actual outcomes were scored as a 0 (no discriminability manipulation) or 1 (discriminability manipulation). For these raw Brier scores, lower values indicate better performance and possible values range from 0 to 1 with chance performance (guessing a 50/50 chance of a manipulation for every data set) at .25. To aid in interpretation, we re-scaled these Brier scores in the same way as we did for contributors such that higher values represent better performance and the possible scores range from -1 to 1, with 0 indicating chance performance. Specifically, scores at chance performance (.25) were adjusted to 0; scores better than chance (below .25) were assigned a value from 0 to 1 depending on their proportional position between chance performance and the best possible performance (e.g., a raw Brier score of .05 would get an adjusted Brier score of .8 because it is 80% of the way from chance to the best possible score: $\frac{.25-.05}{.25-0} = .8$); and below-chance scores (above .25) were assigned a value from -1 to 0 depending on their proportional position between chance and the worst possible performance (e.g., a raw Brier score of .5 would get an adjusted score of -.33 because it is 33% of the way from chance to the worst possible score: $\frac{.25-.5}{.25-1} = .33$).

Fig. S2 shows a histogram for the number of correct inferences across the seven experiments in each replication. The results show that inferential accuracy is generally high, but certainly not perfect, when data sets similar to those sent to our contributors are analyzed with a valid measurement technique (i.e., one that matches the true data-generating process). The

majority of replications had either five or six correct inferences out of the seven experiments, with less than 10% involving four or fewer. Fig. S3 shows a histogram of adjusted Brier scores across the simulated replications. The median value was 0.44, indicating that performance was typically about halfway between chance (guessing 50-50 for every data set) and perfect performance (indicating the correct answer every time with 100% confidence). A small proportion of the replications had values below zero, indicating that the sampled data sets were misleading as to the true effect status.

Our primary goal for the simulations was to establish benchmark values that we can use to define problematic inference procedures. We identified these cutoffs as the 10th percentile of performance across the simulated replications. For evaluating the number of correct inferences, this policy suggests that contributors at or below a value of four applied invalid inference procedures. For evaluating the Brier scores, the benchmark value was 0.13.

Of course, it is possible for a valid inference procedure to fall below our benchmarks if unlucky sampling produces misleading data (as in 10% of the simulated experiments). To evaluate whether our empirical data could be one of these unlucky samples, we applied the inference technique used for the simulations to the empirical data (i.e., Bayesian *t*-tests on μ' estimates). This produced correct inferences for 6 of the 7 experiments (all but Experiment 2) and an adjusted Brier score of 0.38. These values are comfortably above our cutoff for problematic inferences, suggesting that the empirical data are not a particularly unlucky sample compared to the simulation results. This also shows that a fairly simple inference procedure can beat our performance cutoffs, reinforcing the conclusion that contributors below this cutoff likely applied problematic inference procedures.

Simulated Inference with the Wrong Measurement Model

Our simulations above used the same process (i.e., an unequal variance signal detection model) to both generate and analyze simulated data sets. This provides a benchmark for performance when the only limiting factor on inference success is sampling variability in the data. Researchers analyzing empirical data can never be certain that they know the generating process in every detail, and must instead depend on having measurement models that are useful approximations of this process. Therefore, we used the simulated data sets from the previous section to explore the effect of applying a measurement model that does not match the data-generating process. Specifically, we performed the same inference procedure as the above simulation, except that we substituted a hits-minus-false-alarms measure (P_r, Snodgrass & Corwin, 1988) for the μ' measures. As one might assume, this measure is calculated by simply subtracting the observed false alarm rate from the observed hit rate. This performance measure is based on the two-high-threshold (2HT) model of recognition decisions, which essentially assumes that a participant either retrieves information that unambiguously identifies the study status of an item (i.e., target or lure) or fails to retrieve this information and makes a guess. The P_r measure is consistent with a particular version of the 2HT model in which there is an equal probability of experiencing retrieval states that identify targets as studied and lures as not studied. Notably, this model does not match the unequal variance signal detection model that generated the data sets, and applying the two models can produce different conclusions based on their different assumptions for the relative effect of bias on hit and false alarm rates (e.g., Dube, Rotello, & Heit, 2010; Rotello, Masson, & Verde, 2008). That said, the models can be fairly close approximations of one another when changes in bias are subtle (Dube, Rotello, & Heit, 2011; Rotello et al., 2008).

Applying the P_r measure to the simulated data sets produced relatively high accuracy levels, demonstrating that our contributors had the potential to make effective inferences even if they applied a measurement model that did not match the data-generating model in all respects. Specifically, using the P_r measure produced correct inferences for 76% of the simulated data sets, which is not far below the 79% accuracy achieved by the μ' measures that align with the data-generating model. Inferences with the P_r measure had a median adjusted Brier score of .38, which again was close to the .44 median for the μ' measures.

We found similar results when we reversed the process and generated simulated data sets from the 2HT model and measured discriminability with either 2HT or signal-detection measures. Again, the measures based on an incorrect measurement model (signal-detection) achieved a level of inference success that was just slightly below the measures based on the correct model (2HT). This reinforces the conclusion that inference success for our data sets did not depend on applying exactly the correct measurement model. Interested readers can confirm these claims using the code for the 2HT simulations (see OSF site; this includes fitting code to define posterior distributions of 2HT parameters so that the simulations could be based on parameter distributions that were consistent with our data).

Random Inference Simulations

Although a surprising number of contributors fell below our benchmarks for performance based on valid inference methods, some contributors seemed to be basically as accurate as they could be given the sampling variability in the data. Based on these results, we concluded that at least some of our contributors applied effective inference methods; however, we must consider the possibility that our high-performing contributors were simply lucky. In assessing the

potential role of luck, one should first note that all contributors received the same data sets, so variation in inference success could not be driven by sampling variability in itself. Instead, luck played a role only in terms of the interaction between sampling variability and the characteristics of the chosen inference method, meaning that luck played a (potentially much) smaller role than if contributors had received different random samples of data.

To more systematically assess the role of luck, we performed a third set of simulations to define expected performance levels based only on "guessing," by which we mean a scenario in which contributors had no information about whether or not each data set involved a discriminability manipulation. The valid-inference-technique simulations above demonstrate that the data sets that we sent to contributors do theoretically provide information about discriminability, but we are imagining a scenario in which contributors had no way to use this information, perhaps because they did not have available models that matched (or at least usefully approximated) the memory and decision processes that generated the data. If the results from our contributors as a whole are well outside of the distribution of results produced by guessing, then this will demonstrate that at least some of them were making effective inferences.

We began by comparing our contributors to chance-level accuracy on the task of determining whether or not each data set had a memory discriminability manipulation. The contributors included in this analysis made inferences for a total of 182 data sets, comprising 7 data sets each from 26 contributors[1]. The correct inference was selected on 124 (.68) of these attempts. Fig. S4 shows the expected binomial distribution for the number of correct inferences if each attempt was an uninformed guess (i.e., probability of success = 0.5), with the observed

[1] We excluded the contributor who reported a 50% probability of a memory effect for every data set, making it impossible to classify inferences as correct or incorrect.

number correct marked by the vertical line. Clearly, the observed accuracy of our contributors is well outside of the distribution of expected results based on pure guessing.

We next considered adjusted Brier scores. As discussed in the main text, a value of zero corresponds to chance-level performance for the adjusted Brier scores. Thus, by one reckoning, the distribution of expected Brier scores in the no-information scenario is a spike at zero, but that outcome is only predicted if all contributors acknowledge their lack of information by always reporting a 50% chance of a discriminability manipulation. Only one of our contributors chose to adopt that strategy, and those 7 responses were excluded from this analysis (see Footnote 1). The other contributors responded as if they had some useful information about discriminability, and their average adjusted Brier score is .13. To assess whether this level of performance could be achieved without any information about which experiments had discriminability manipulations, we simulated 20,000 studies in which we took the reported probabilities of a memory manipulation from each contributor, randomly assigned these probabilities to experiments, and calculated the average adjusted Brier score across contributors. Fig. S5 shows the results. The average Brier score from these scrambled probabilities was almost always below zero. This result occurs because the random responses are equally likely to ascribe high confidence to correct and incorrect inferences, but Brier scores penalize high-confidence errors to a greater extent than they reward high-confidence correct responses (Brier, 1950). Although the average adjusted Brier score sometimes exceeded zero just by chance, the observed average across our contributors (the vertical line in Fig. S5) is well outside of range of what one might expect from random responding, suggesting that our high-performing responders were not simply lucky.

References

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*(3), 153–178. https://doi.org/10.1016/j.cogpsych.2007.12.002

Dube, C., Rotello, C. M., & Heit, E. (2010). Assessing the belief bias effect with ROCs: It's a response bias effect. *Psychological Review*, *117*(3), 831–863. https://doi.org/10.1037/a0019634

Dube, C., Rotello, C. M., & Heit, E. (2011). The belief bias effect is aptly named: A reply to Klauer and Kellen (2011). *Psychological Review*, *118*(1), 155–163. https://doi.org/10.1037/a0021774

Macmillan, N. A., & Creelman, C. D. (2005). *Detection Theory: A User's Guide*. Lawrence Erlbaum Associates.

Plummer, M., & Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.13.3406

Pollack, I., & Norman, D. A. (1964). A non-parametric analysis of recognition experiments. *Psychonomic Science*, *1*(1–12), 125–126. https://doi.org/10.3758/BF03342823

Ratcliff, R. (1978). Theory of Memory Retrieval. *Psychological Review*, *85*(2), 59–108. https://doi.org/10.1037//0033-295X.85.2.59

Riefer, D. M., & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review*, *95*(3), 318–339. https://doi.org/10.1037//0033-295X.95.3.318

Rotello, C. M., Masson, M. E. J., & Verde, M. F. (2008). Type I error rates and power analyses for single-point sensitivity measures. *Perception & Psychophysics*, *70*(2), 389–401. https://doi.org/10.3758/PP.70.2.389

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. Retrieved from http://silk.library.umass.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2009-19040-001&site=ehost-live&scope=site

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*(2), 145–166. https://doi.org/10.3758/BF03209391

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: applications to dementia and amnesia. *Journal of Experimental Psychology. General*, *117*(1), 34–50. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2966230

Thomas, E. A. ., & Myers, J. L. (1972). Implications of latency data for threshold and nonthreshold models of signal detection. *Journal of Mathematical Psychology*, *9*(3), 253–285. https://doi.org/10.1016/0022-2496(72)90018-1

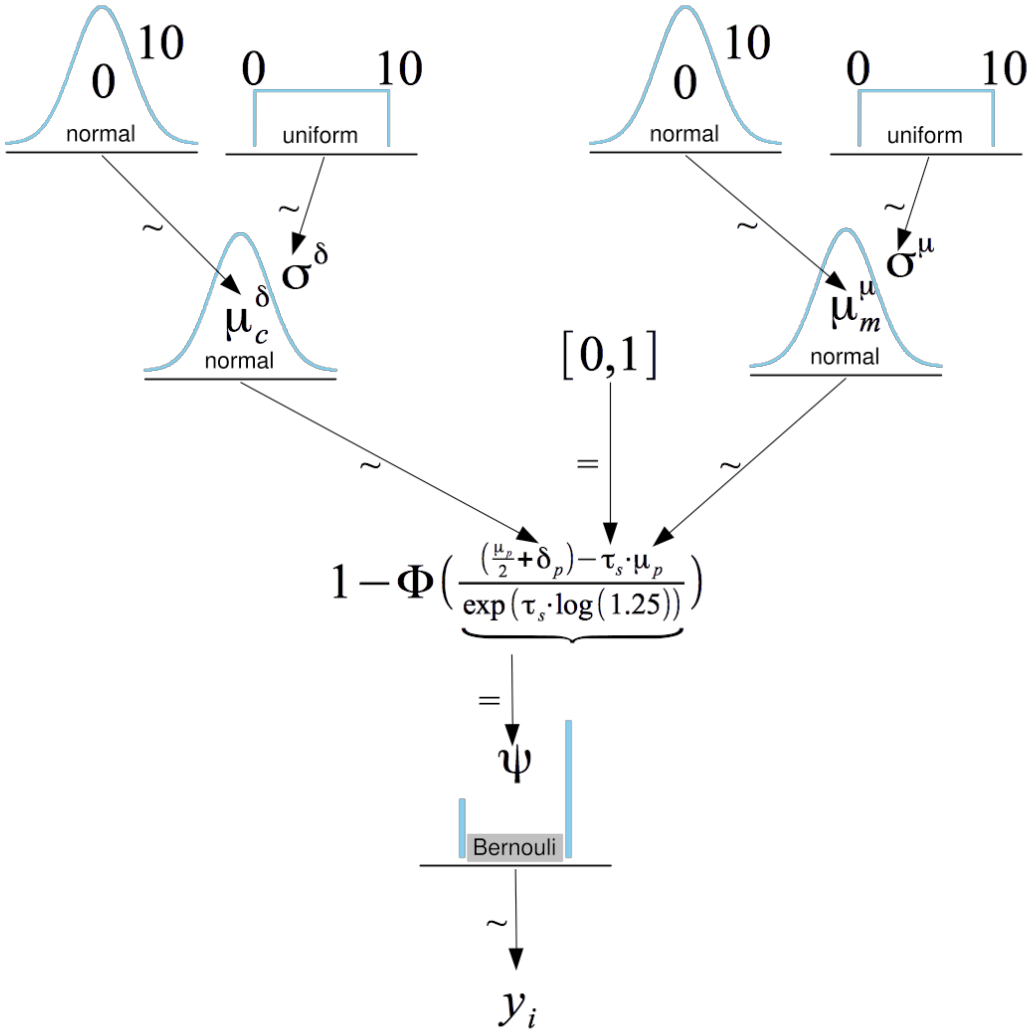

Fig. S1. Diagram of the Bayesian hierarchical unequal variance signal detection model.

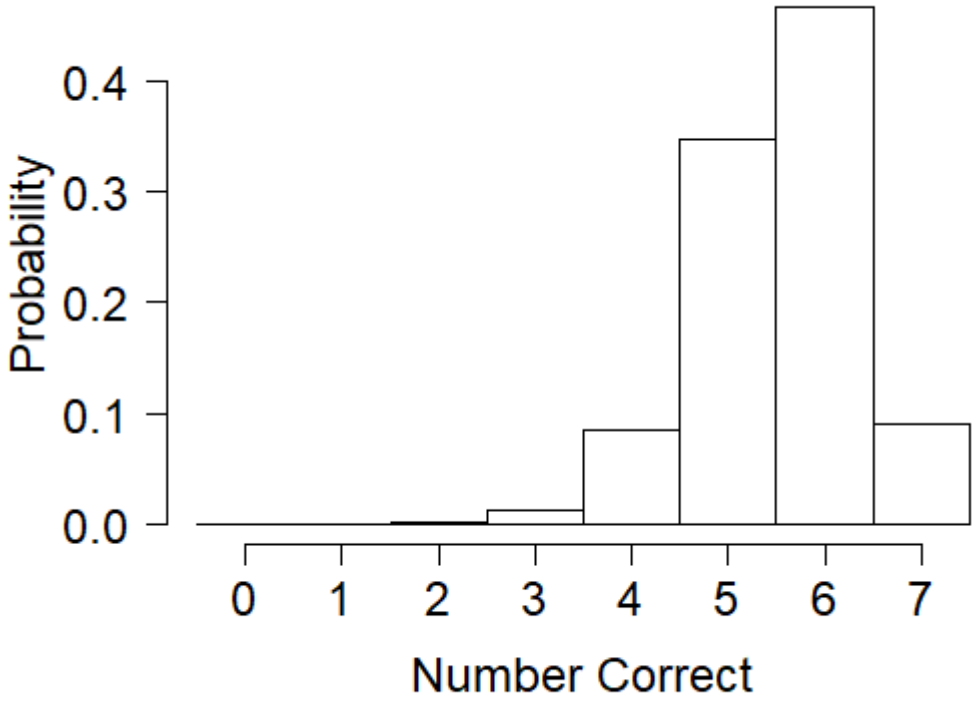

Fig. S2. Histogram of the number of correct inferences about a discriminability effect out of the seven experiments in each simulation.

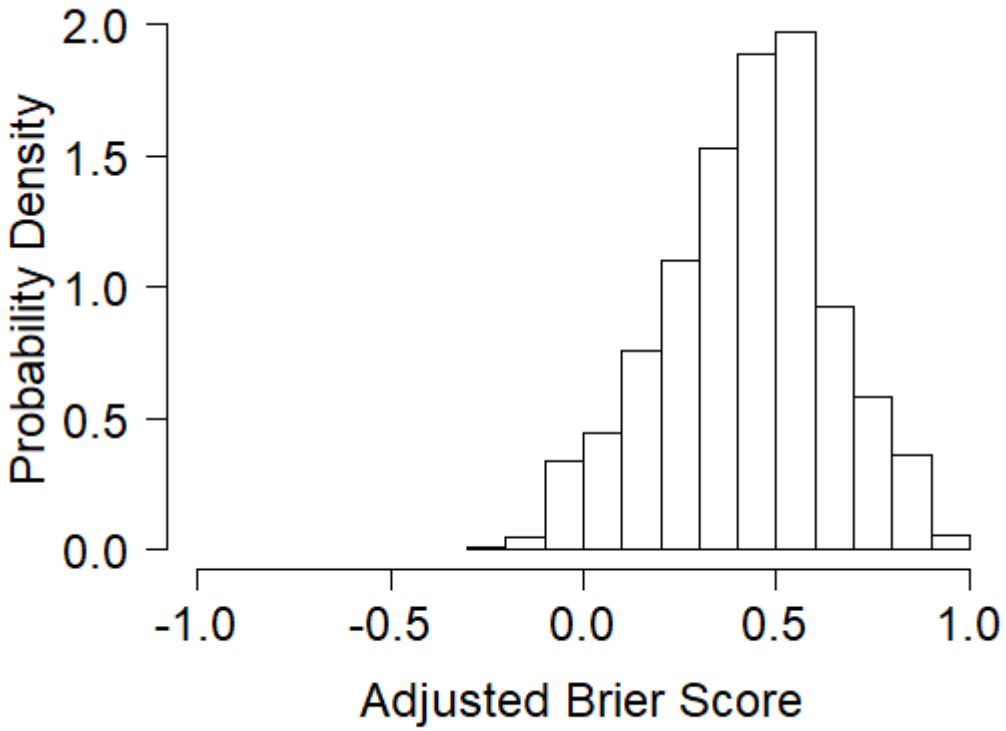

Fig. S3. Histogram of the 5000 adjusted Brier scores across simulated replications.

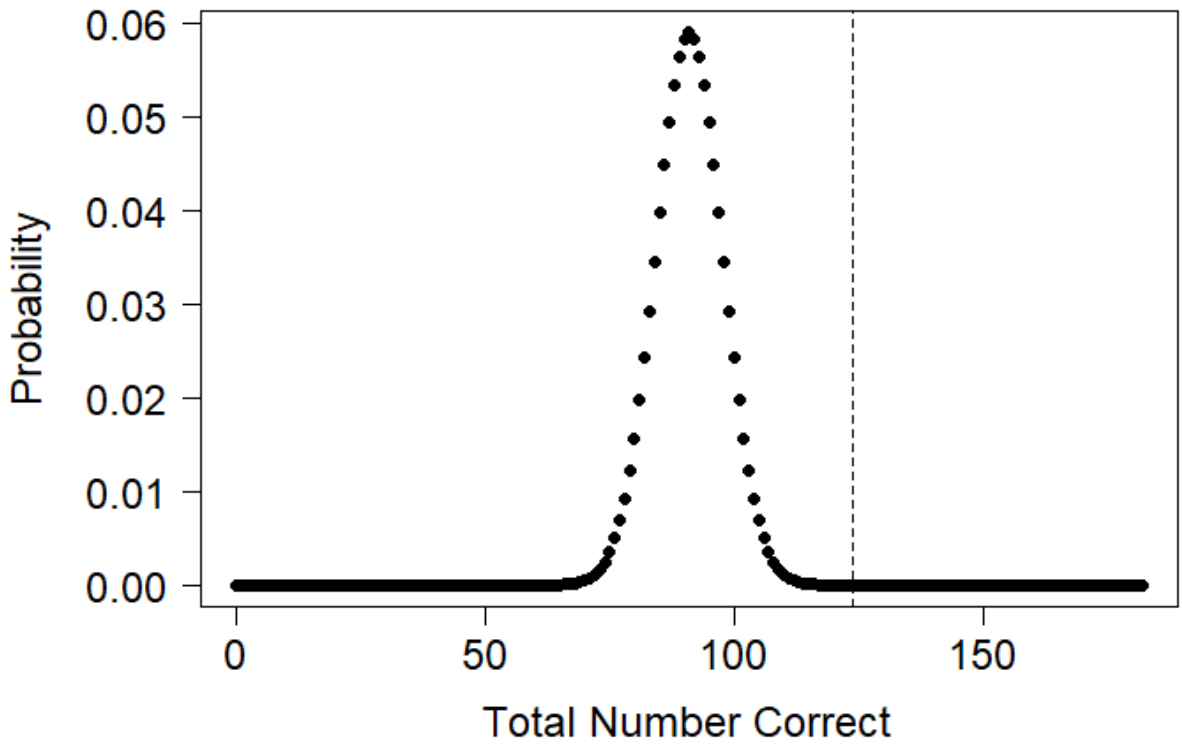

Fig. S4. The expected binomial distribution for the number of correct inferences if each attempt was an uninformed guess (i.e., probability of success = 0.5). The observed number correct is marked by the vertical line.

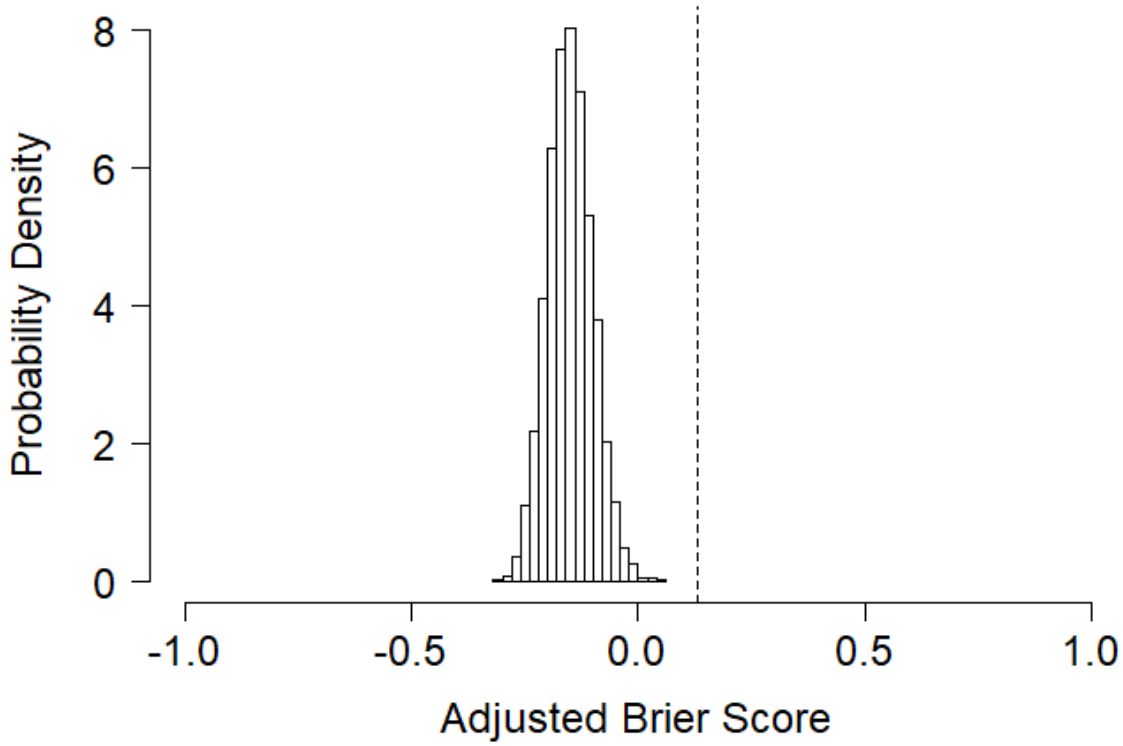

Fig. S5. Average adjusted Brier score for each of 20,000 simulated studies in which contributors' reported probabilities of a memory manipulation were randomly assigned to each of the 7 experiments.

Discriminability	Bias	μ^{μ}	μ^{δ}
1	Liberal	2.05 (.07)	-0.30 (.05)
	Neutral	2.05 (.07)	-0.25 (.06)
	Conservative	2.05 (.07)	0.02 (.05)
2	Liberal	2.63 (.07)	-0.37 (.06)
	Neutral	2.63 (.07)	-0.29 (.05)
	Conservative	2.63 (.07)	-0.14 (.06)
3	Liberal	2.86 (.07)	-0.46 (.06)
	Neutral	2.86 (.07)	-0.35 (.06)
	Conservative	2.86 (.07)	-0.24 (.06)

Notes. Discriminability is coded in terms of how many times (1, 2, or 3) the targets were studied. The μ^{μ} parameter represents the mean of the across-participant distribution of μ, the mean of the target strength distribution for each participant; μ^{δ} represents the mean of the across-participant distribution of δ, the deviation of the response criterion from halfway point between target and lure distributions for each participant. Values in parentheses are standard deviations of the posterior distribution for each parameter.

Table S1. Selected parameter estimates (medians of the posterior distribution) from the Bayesian hierarchical signal detection model.