The Collective Memory Performance in a Recognition Memory Task

Mark Steyvers

University of California, Irvine

Address for correspondence:

Mark Steyvers

University of California, Irvine

Department of Cognitive Sciences

2316 Social & Behavioral Sciences Gateway Building

Irvine, CA 92697-5100

E-mail: mark.steyvers@uci.edu **Phone:** (949) 824-7642 **Fax:** (949) 824-2307

Word count: 4300 (excluding references)

Number of figures and tables: 4

Abstract

I investigate the collective memory performance in a recognition memory task in which each individual in a group independently retrieves memories related to the same study items. For each test item in the recognition memory task, the aggregated memory judgment is based on the average of confidence ratings across individuals. Using a Bayesian Signal Detection Theory (SDT) analysis of the confidence ratings, I show that the aggregated confidence rating is associated with a discrimination performance that is substantially better than the best performing individual in the group and discuss a number of potential sources of noise related to encoding, retrieval and decision processes that are reduced as a result of averaging.

The Collective Memory Performance in a Recognition Memory Task

The aggregation of judgments across individuals in a group has been shown to lead to a group estimate that is better than most of the individual estimates. Demonstrations of this effect have focused on tasks where individuals produce subjective probability or magnitude estimates (e.g., Ariely et al., 2000; Budescu & Yu, 2007; Steyvers, Wallsten, Merkle, & Turner, 2014; Turner, Steyvers, Merkle, Budescu, & Wallsten, 2014; Wallsten, Budescu, Erev, & Diederich, 1997). In a now classic study, Galton (1907) asked over eight hundred individuals to estimate the weight of an ox. The median weight estimate, which corresponds to a simple form of aggregation, came within a few pounds of the true answer. This group estimate was much closer to the truth than the vast majority of individual estimates, a phenomenon that has become known as the Wisdom of Crowds effect (WoC; reviewed in Surowiecki, 2004). The most basic explanation of this effect is that the averaging across individuals reduces the noise associated with each individual decision – some individuals overestimate and others underestimate the underlying quantity – and aggregating cancels out some of these errors in judgment. The benefits of aggregating across individuals have also been demonstrated in more complex tasks involving rank-ordering judgments (Lee, Steyvers, & Miller, 2014), and optimization problems (Yi, Steyvers, & Lee, 2012). Recently, it has been shown that the benefits of averaging also extend to judgments within an individual (Vul & Pashler, 2008; Hourihan & Benjamin, 2010).

Studying the collective memory performance of a group of individuals has some real-world applications. One specific example is eyewitness testimony cases in which there are a number of individuals who all have witnessed the same set of events. If a researcher now queries each individual eyewitness and collects a series of memory judgments, it is important to understand what performance might be expected by combining the individually retrieved memories into a single judgment. Much of the existing research on collective memory has focused on developing an understanding of the conditions in which social interaction between group members can help or hurt memory performance (e.g., Ditta & Steyvers, 2013; Gagnon & Dixon, 2008; Harris, Paterson, & Kemp, 2008; Hinsz, 1990; Roediger, Meade, and Bergman, 2001).

In this chapter, I will investigate the collective memory performance that can be obtained by pooling retrieved memories across a number of individuals when there is no social interaction or information sharing of any kind between individuals in the group. Each individual independently provides a series a memory judgments and the aggregation is performed by the researcher. I will demonstrate that aggregation can lead to a WoC effect where the aggregated memory judgment is substantially better than the majority of individuals and even better than the best individual in the group. The main contribution of this research is to show how a Bayesian approach to Signal Detection Theory (SDT) can be used to assess the performance of individuals and the aggregate. Overall, the findings suggest that there are multiple possible explanations for the performance improvements, related to different ways in which noise at the encoding, retrieval, and decision stage are averaged out.

The plan for this paper is as follows. I will first describe the previously published data that will be used for the analysis. I will then describe the Bayesian analysis of the

SDT model to estimate the underlying signal and noise distribution in the context of these data and measure the ability of individuals to discriminate between targets and lures. This model is applied on the recognition memory data and the performance of individual subjects is compared to the aggregate. Finally, I provide some explanations for the WoC effect and discuss the potential reasons for the improvement of the aggregate.

Recognition Memory Data from Mickes et al. (2007)

The analysis of the WoC effect in this chapter is based on previously published recognition memory data from Experiments 1 and 2 of the Mickes, Wixted, & Wais (2007) study. In these experiments, subjects studied a list of 150 words and were tested on all target words and 150 lure words. Study and test order were randomized across participants. The study and lure words were randomly selected from a pool of three-to-seven letter words. Each target word was presented for 2 sec during the study phase. In Experiment 1, there were 14 subjects who produced confidence ratings on a 20-point scale. In Experiment 2, there were 16 subjects who gave confidence ratings on a 99-point scale. The analysis also included an unpublished study in which 12 subjects produced confidence ratings on a 6-point scale. In this study, the same list of study and test words was used as in Experiment 1 and 2 of the Mickes et al. (2007) study. In the rest of the paper, these studies will be referred to by the number of unique confidence ratings available to subjects: 20, 99, and 6.

An important property of the experiment is that all subjects were given the same study list of items (although not in the same order) and each individual was tested on the same set of items (again, not in the same order). Because the set of study and test items are equivalent (within each data set), the confidence judgments can be aggregated across individuals, exploring the performance of the aggregate as well as the importance of the number of possible confidence ratings.

For each experiment, the aggregate is based on the average confidence rating for a particular item. For example, if three subjects give confidence ratings 2, 4, and 6 to a particular test item (e.g., the word "dog"), a 4 is recorded for the average rating (the rating for "dog"). If the averaging leads to a fractional rating, the rating is rounded, such that the aggregate confidence is based on the same response scale available to subjects. This averaging process continues for all test items in the list. It might be useful to think of the aggregate confidence rating as the response from another subject in the experiment, whose task it is to respond with the average of all confidence ratings across subjects. In the analysis, I will compare the performance of the aggregate against each individual subject. A WoC effect is achieved in cases where the aggregate performance is as good as or even better than the best subjects.

Assessing Performance with a Bayesian Signal Detection Theory Analysis

Signal Detection Theory is a useful approach to assess the ability of individuals and the aggregate to discriminate between targets and lures. The focus will be on the unequal variance SDT model (e.g., Wixted, 2007) which is illustrated in Figure 1. It is assumed that each item at test has a memory strength that can be represented by a uni-dimensional continuum. The strengths for targets and lures are sampled from two separate distributions. Typically, the target distribution is assumed to have a higher mean as well as a higher variance than the lure variance (leading to the unequal variance model). This unequal variance accommodates findings from a ROC analysis of recognition memory data (e.g., Glanzer, Kim, Hilford, & Adams, 1999; Mickes, et al., 2007; Ratcliff, McKoon, Tindall, 1994; Ratcliff, Sheu, & Gronlund, 1992).

[INSERT FIGURE 1 HERE]

The assumption in the analysis is that the lure or noise distribution has a zero mean and unit variance. The target distribution has mean μ and standard deviation σ. Confidence ratings are produced by sampling strengths from the distribution associated with the test items and comparing the signal strengths to a set of criteria, $\mathbf{c} = (c_1,...,c_{K-1})$, where K is the number of unique confidence ratings produced by an individual. Each sampled signal strength falls in a region defined by the fixed set of criteria and each region is associated with a particular confidence rating. For example, a sampled strength that falls between c_1 and c_2 leads to a rating of "2", as illustrated in Figure 1. In the model, the ability of individuals to separate between lure and target items is not only dependent on μ but also on σ. Better discrimination performance can be expected when the mean of the target strength distribution increases, but also when the variance of the target strength distribution decreases. One standard measure of discrimination ability that combines the mean and variance of the target distribution is d_a (e.g., Macmillan & Creelman, 1991; Wickens, 2001) where $d_a = \sqrt{2}\mu \big/ \sqrt{1+\sigma^2}$.

Another standard measure in the context of recognition memory is the ratio of the standard deviation of the distractor and target distribution. Because we defined the distractor distribution to have unit variance, this ratio is simply the inverse of the target standard deviation: $s = 1/\sigma$. The inverse ratio is equivalent to the slope of z-ROC curve, in which the hits and false alarms are plotted against each other in z-space (e.g., Macmillan & Creelman, 1991; Wickens, 2001). An empirical regularity in most recognition memory experiments is that the standard deviation of the target distribution is larger than the distractor distribution (Glanzer et al., 1999; Mickes et al., 2007; Ratcliff et al.,1992, 1994). The standard deviation of the target distribution is typically around 1.25, which amount to an inverse ratio (z-ROC slope) of around 0.8.

Parameter Estimation

To assess the performance of the aggregate relative to the individual subjects, the SDT model is applied to each subject separately as well as the aggregate. Therefore, for each individual and the aggregate, the model parameter estimates consist of μ and σ for the target distribution, as well as the criteria values $\mathbf{c} = (c_1,...,c_{K-1})$. These model parameters can then be used to calculate discriminability d_a, at the level of individual subjects. A WoC effect corresponds to a much larger value of d_a for the aggregate relative to the individual subjects.

A common approach is to estimate these parameters through ROC analyses. The confidence ratings are converted to a hit and false alarm rate for a given criterion cutoff point. By varying the criteria cutoffs, the relationship between hit and false alarm rates can be plotted in an ROC plot. By z-transforming the hit and false alarm rates, a z-ROC plot is obtained which often reveals an approximate linear relationship. The slope of the regression line in the z-ROC plot can be used as an estimate for $1/\sigma$. Similarly, the regression parameters can be used to estimate measures of discriminability such as d_a. One drawback of this estimation procedure is that it is difficult to obtain stable estimates of the z-ROC regression line when subject performance levels are unusually high (as will be shown to be the case for the aggregate). Furthermore, in the case for many criteria cutoff points, the hit and false alarm counts are sparse, which complicates the construction of the z-ROC curve.

To achieve accurate estimates of model parameters across a wide variety of experimental settings, I use a Bayesian approach to estimate a SDT model for confidence judgments. This model is closely related to other Bayesian SDT models (Lee, 2008a, 2008b; Rouder & Lu, 2005; Rouder et al. 2007; Morey, Pratte, & Rouder, 2008) that have been applied to a number of tasks including recognition memory (Dennis, Lee & Kinnell, 2008). For example, Morey et al. (2008) developed a comprehensive hierarchical modeling framework that allows for the estimation of unequal variance models on the basis of confidence judgments. The hierarchical model is applied simultaneously to the memory judgments of all subjects and all items, allowing for the estimation of item and subject differences. One general advantage of the Bayesian approach is that it can give good estimates of SDT parameters even when the error rates are very low (Lee, 2008a, 2008b). Another advantage is that Bayesian estimation procedures for SDT models can produce confidence intervals on parameter estimates at the level of individual subjects. This is useful because the parameter estimates of d_a of the aggregate and the individuals need to be compared. Finally, the Bayesian estimation procedure is particularly helpful in the context of confidence scales with many possible responses (e.g., 99 confidence ratings) that can lead to sparse and incomplete data for many criteria – such data sparsity leads to major obstacles in computing d_a through ROC analysis.

For this chapter, I pursue a simple Bayesian estimation approach that makes it possible to estimate the SDT model parameters separately for each individual subject, including the aggregate. The estimation procedure results in estimates of the target strength mean μ and variability σ, as well as the criteria $\mathbf{c} = (c_1,...,c_{K-1})$, at the level of individual subjects. The model estimates can be used to calculate discriminability d_a. The estimation procedure allows for the presence of a large number of unique ratings used by an individual subject which necessitates the estimation of a large number of criteria values. For each model parameter, not only are point estimates available but also samples from the full posterior, which makes it possible to calculate confidence intervals (and potentially other measures of interest, such as correlations between model parameters) on all parameters. The Appendix provides more detail on how the Bayesian SDT model is defined and how estimation is performed.

Results

The Bayesian SDT model was fit separately to each subject (including the aggregate) in each of the three data sets. The model fits consist of the posterior distribution over the three variables in the model: the mean target-strength (μ), standard deviation (σ), and the set of criterion values $\mathbf{c} = (c_1,...,c_{K-1})$, from which variables such as discriminability (d_a) can be calculated. The focus in the analysis is on two measures extracted from the posterior distributions, the mean of the distribution as well as the 5% and 95% percentile estimates of the distribution. The latter estimates give us a 90% Bayesian credible interval.

[INSERT FIGURE 2 HERE]

Figure 2 illustrates the key finding of this paper. The Figure shows the estimated means and confidence intervals of the discriminability parameters (d_a) for each subject, including the aggregate, across the three data sets. For each data set, the subjects are ordered by their discriminability. As can be observed, the aggregate is associated with the highest level of discriminability in all three data sets. This pattern is consistent with the stronger version of WoC effect in which the aggregate outperforms the best individual even though the individuals are used to construct the aggregate. Note also that the performance levels of the aggregate are substantially higher than the next best subject. For the three data sets, the difference in d_a between the aggregate subject and the best subject was 1.66, 1.02, and 1.97 respectively. To put this in perspective, many experimental manipulations in recognition memory that result in reliable differences in discriminability are often associated with differences in d_a much smaller than 1. The fact that this WoC effect occurs across three separate data sets suggests that this effect is reliable and can be replicated with other recognition memory data sets. Note also that there appears to be no relationship between the magnitude of the WoC effect and the number of options on the confidence scale.

[INSERT FIGURE 3 HERE]

To better understand the nature of the improved discriminability of the aggregate, Figure 3 shows the relationship of μ and σ across individuals. It can be observed that the aggregate is different from the individual subjects in two respects: the mean (μ) of the target strength distribution is larger and the standard deviation (σ) of the target strength distribution is, in comparison to most subjects, smaller. Both of these effects contribute to the superior discriminability for the aggregate as shown in Figure 2. Note that for the individual subjects (excluding the aggregate), a range of standard deviations (σ) is observed, with most values above 1. The average value for individual subjects is 1.37, 1.35, and 1.67 respectively for the three experiments. This is consistent with the inverse ratio's (*s*) smaller than one that are found in the literature (Glanzer et al., 1999; Mickes et al., 2007; Ratcliff et al.,1992, 1994). In contrast, the standard deviations for the aggregate are some of the lowest values observed relative to the individual subjects and are close to 1 (the values are 1.00, 1.11, and, 1.01 respectively).

[INSERT FIGURE 4 HERE]

To explore this further, Figure 4 illustrates the estimated SDT models for a subset of subjects, including the aggregate subject (top row), best individual subject according to the discriminability (d_a) (middle row), and a typical subject with a performance level that was closest to the median discriminability. The figure also shows the inferred set of criteria values. The displayed values are the posterior means for each individual criterion value. It can be observed that individual subjects are characterized by unequal variance SDT models, a typical result in the literature. For the aggregate, the inferred SDT model is much closer to an equal variance model[1]. In addition, the criterion values for the aggregate are remarkably evenly spaced relative to the individual subjects. In the Discussion, we provide some reasons for this even spacing effect.

Discussion and Conclusion

Across three separate studies, I have demonstrated that aggregating memory judgments across individuals leads to a performance level that is superior to all of the individual subjects. Note that there was no information sharing of any kind between individuals and this lack of communication might have contributed to the large aggregation benefits. Research on collaborative memory has shown that groups of people that remember together can remember fewer items than when the same people recall separately and their recall is subsequently combined (Thorley & Dehurst, 2007), a phenomenon called collaborative inhibition (Weldon & Bellinger, 1997). This could be because, when people remember in groups, they are subject to social variables that may influence their performance (e.g. social loafing; Latané, Williams, & Harkins, 1979), or

[1] It should be noted that this illustration is produced on the basis of the mean parameter estimates (the mean of the posterior). Figure 3 shows that there is considerable uncertainty associated with the estimates for the standard deviations. For the aggregate subject, the 5% and 95% percentile estimates of the confidence interval range from much smaller as well as much higher values of σ than 1. Therefore, even though the mean parameter estimates of σ for the aggregate (used to produce Figure 3) suggest that the aggregate is more consistent with an equal variance model, caution should be taken in interpreting the particular point estimate of σ found with the estimation procedure.

because of cognitive factors such as retrieval interference (Weldon, Blair, & Huebsch, 2000).

Generally, the result of averaging across subjects is that factors that contribute to subject variability are averaged out. In the context of recognition memory, one source of variability involves encoding factors that add to the variability in the internal memory representations for studied items (e.g., Wixted, 2007; DeCarlo, 2002). Attentional fluctuations might be one component of encoding variability -- during the study phase, subjects' attention might wander such that some items are better encoded than others. If attentional fluctuations are uncorrelated across subjects and therefore subjects pay differential amounts of attention to different study trials, this source of variability can be reduced by averaging the decisions (for the same test item) across subjects.

Another source of variability might be interference effects at retrieval where the retrieval cue interacts with traces in memory. Interference effects can arise when there are interfering episodic traces of the same item in different contexts or other items in the same context that might make it difficult to make accurate recognition memory decisions (e.g. Anderson & Neely, 1996; Norman & Waugh, 1968). Interference effects can also occur when test items leave episodic traces in memory that interfere with the retrieval of the target episodic trace (Criss et al. 2011; Malmberg, Criss, Gangwani, & Shiffrin, 2012). In this research, these output interference effects are a likely to have played a role in producing the WoC effects because the test items were randomly ordered between individuals in the particular recognition memory data that was used for the analysis. Therefore, a particular test item that comes early in the sequence for one individual provides less interference than the same item presented later in the test sequence for another individual, and aggregating across individuals reduces the noise associated with output interference.

Finally, one potential source of noise involves the decision process in which the internal evidence is translated to a confidence rating. Maintaining multiple criteria might lead to different response strategies (Malmberg, 2002) and also pose a burden to memory such that each point on the rating scale introduces some amount of variability to the decision process (Benjamin, Diaz, & Wee, 2009; Benjamin, Tullis, & Lee, 2013). Consistent with this view is the finding in Figure 4 that the aggregate is associated with estimated criterion values that are much more evenly spaced than individual subjects, suggesting that the effects of criterion noise and idiosyncratic uses of the response scale are partially reduced by averaging. However, while it is possible that the even criterion spacing of the aggregate is based on a reduction of decision noise, it seems unlikely that this explains the size of the WoC effect because effects of decision noise are typically small.

Further experiments and modeling will need to be done to better distinguish between the underlying causes of the WoC effect. It is important to note that although a particular SDT model was used to evaluate the effects of aggregation, the goal of this research was not to use the SDT model as a model that describes the underlying processes that give rise to a memory judgment. I merely explored the consequences of averaging confidence ratings across subjects and used the SDT model as a measurement tool[2]. An important direction for research is to investigate the WoC effect with process models such as SAM (Raaijmakers & Shiffrin, 1980, 1981) and REM (Shiffrin & Steyvers, 1997). Typically, these models do not distinguish between encoding and retrieval effects and are not specific about the nature of subject differences, but the models could be extended to incorporate such differences. The models could be designed to explain the size of the WoC effect as a function of experimental factors such as the variability of the study list, ordering of test items, number of subjects, the distribution of memory performance across subjects, etc. Overall, an important direction for future research is to use the WoC effect as an additional empirical finding to constrain memory models.

[2] However, if the Signal Detection approach was considered as a model for recognition memory judgments, there would be some challenges in explaining the experimental results. For example, the process of averaging out encoding noise would only affect target items in the SDT model. The results did in fact show a decrease in the variability of target strength (σ) of the aggregate. At the same time, the results also showed a change in target strength mean (μ). Note that because of the particular parametrization of the SDT model, the variance of the noise distribution was fixed at 1. Therefore, a decrease in the noise variance translates into an increase in the target strength mean (μ). At present, it is not obvious in a Signal Detection framework how to explain the increase in mean (μ) or equivalent reduction in variance for the lures.

Appendix

This section provides a more detailed description of the SDT model for ratings data. The distribution of memory strengths are modeled by Normal (Gaussian) distributions with the lure distribution centered at 0 and unit (1) variance. The target strengths have mean μ and standard deviation σ. For each trial j=(1,...,N) in the experiment, let the variable x_j encode whether the trial is a target (x_j=1) or lure (x_j =0). Let the variable y_j represent the confidence rating produced by the subject on trial j. Let K represent the number of unique confidence ratings that the subject produces. It is convenient to map these confidence ratings to consecutive integer values 1 to K. In the model, confidence ratings are generated by sampling strengths from the target or lure distribution and comparing the sampled value against a set of K-1 criterion values, **c** = (c_1,...,c_{K-1}). It is convenient to add two fixed criterion values c_0 =-∞ and c_K =+∞. The probability of a particular confidence rating is then given by:

$$p\left(y_j = k \mid x_j\right) = \begin{cases} \Theta\left(c_k, \mu, \sigma\right) - \Theta\left(c_{k-1}, \mu, \sigma\right) & x_j = 1 \\ \Theta\left(c_k, 0, 1\right) - \Theta\left(c_{k-1}, 0, 1\right) & x_j = 0 \end{cases}$$

Therefore, in this model, the variables y_j and x_j are observed outcomes and μ, σ, and **c** are latent variables. The model assumes a Normal prior on μ with precision τ, a uniform prior on σ with values between [0,..., σ_{max}] and a Normal prior with precision τ_c on each criterion value:

$$\mu \sim \text{Normal}\left(0, \tau\right),\ \sigma \sim \text{Uniform}\left(0, \sigma_{\max}\right),\ c_k \sim \text{Normal}\left(0, \tau_{\text{c}}\right)$$

The hyperparameters of these priors are set to the following: τ=0.05, σ_{max}=3, and τ_c=0.05. There are a number of alternative priors for the standard deviation that could be used such as the inverse gamma (e.g., see Jackman, 2009). Because it has been argued that one should be careful using the inverse gamma (e.g., Gelman, 2006), a uniform prior was chosen. The uniform prior gave good performance in parameter recovery studies.

Parameter estimation was performed by a Markov chain Monte Carlo (MCMC) procedure written in Matlab. The procedure results in samples from the posterior distribution over μ_t, σ_t, and **c**. From these samples, we can calculate the posterior mean and use this as a point estimate. We can also calculate credible intervals on these variables to assess the uncertainty associated with the parameter estimate.

In the MCMC procedure, each chain was initialized with μ =1, σ =1. The criteria were initialized by an equal spacing between -1 and +2, which spaces the criteria between one standard deviation below the lure mean and one standard deviation above the target mean. Note that because a Normal prior was placed on each individual criterion, the criteria are not ordered (a priori) in any particular way. However, after sampling the initial criterion values, the model reorders the criterion values such that $c_1 < c_2 < \ldots < c_{K-1}$. This ordering is maintained during inference. In a Metropolis-Hastings procedure, a combination of single-variable proposals as well as block proposals was used for the set of criterion values. In the simulations described in this research, each chain was run for 2500 iterations and samples were taken after a burnin of 1500 iterations. A total of 7 chains were used.

References

Anderson, M. C., & Neely, J. H. (1996). Interference and inhibition in memory retrieval. In E. Bjork & R. A. Bjork (Eds.), *Memory* (pp. 237–313). San Diego, CA, USA: Academic Press.

Ariely, D., Au, W. T., Bender, R. H., Budescu, D. V., Dietz, C. B., Gu, H., Wallsten, T. S., & Zauberman, G. (2000). The effects of averaging subjective probability estimates between and within judges. *Journal of Experimental Psychology: Applied*, *6*(2), 130-147.

Benjamin, A. S., Tullis, J. G., & Lee, J. H. (2013). Criterion noise in ratings-based recognition: Evidence from the effects of response scale length on recognition accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 1601-1608.

Benjamin, A. S., Diaz, M. L., & Wee, S. (2009). Signal detection with criterion noise: Applications to recognition memory. *Psychological Review, 116,* 84–115.

Budescu, D. V. & Yu, H. (2007). Aggregation of opinions based on correlated cues and advisors. *Journal of Behavioral Decision Making, 20,* 153-177.

DeCarlo, L. T. (2002). Signal detection theory with finite mixture distributions: Theoretical developments with applications to recognition memory. *Psychological Review, 109*(4), 710-721.

Criss, A.H., Malmberg, K. J., & Shiffrin, R.M. (2011). Output Interference in Recognition Memory. *Journal of Memory and Language, 64*(4), 316-326.

Dennis, S., Lee, M. D., & Kinnell, A. (2008). Bayesian analysis of recognition memory: The case of the list-length effect. *Journal of Memory and Language, 59,* 361-376.

Ditta, A.S., & Steyvers, M. (2013). Collaborative Memory in a Serial Combination Procedure. Memory, 21(6), 21(6), 668-674.

Gagnon, L. M., & Dixon, R. A. (2008). Remembering and retelling stories in individual and collaborative contexts. *Applied Cognitive Psychology, 22*, 1275-1297.

Galton, F. (1907). Vox Populi. *Nature, 75*, 450-451.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis 1*, 515–533.

Glanzer, M., Kim, K., Hilford, A., & Adams, J. K. (1999). Slope of the receiver-operating characteristic in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 500-513.

Harris, C. B., Paterson, H. M., & Kemp, R. I. (2008). Collaborative recall and collective memory: What happens when we remember together? *Memory, 16*, 213-230.

Hinsz, V.B. (1990). Cognitive and consensus processes in group recognition memory performance. *Journal of Personality and Social Psychology, 59*(4), 705-718.

Hogarth, R. M. (1978). A note on aggregating opinions. *Organizational Behavior and Human Performance, 21*(1), 40-46.

Hourihan K. L., Benjamin, A. S. (2010). Smaller is better (when sampling from the crowd within): Low memory span individuals benefit more from multiple opportunities for estimation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 1068–1074.

Jackman, S. (2009). *Bayesian analysis for the social sciences.* Hoboken, NJ: Wiley.

Latané, B., Williams, K., & Harkins, S. (1979). Many hands make light the work: The causes and consequences of social loafing. *Journal of Personality and Social Psychology*, 37, 822–832.

Lee, M. D. (2008a). BayesSDT: Software for Bayesian inference with signal detection theory. *Behavior Research Methods, 40,* 450–456.

Lee, M. D. (2008b). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin & Review, 15*, 1–15.

Lee, M.D., Steyvers, M., & Miller, B.J. (2014). A cognitive model for aggregating people's rankings. *PLoS ONE, 9*.

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide*. New York: Cambridge University Press.

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

Malmberg, K. J. (2002). On the form of ROCs constructed from confidence ratings. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(2), 380-387.

Malmberg, K. J., Criss, A.H., Gangwani, T. & Shiffrin, R.M. (2012). Overcoming the Negative Consequences of Interference that Results from Recognition Memory Testing. *Psychological Science, 23*, 115-119

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.

Morey, R. D., Pratte, M. S., & Rouder, J. N. (2008). Problematic effects of aggregation in zROC analysis and a hierarchical modeling solution. *Journal of Mathematical Psychology, 52*, 376-388.

Norman, D. A., & Waugh, N. C. (1968). Stimulus and response interference in recognition memory experiments. *Journal of Experimental Psychology, 78*, 1–59.

Raaijmakers, J.G.W., & Shiffrin, R.M. (1980). SAM: a theory of probabilistic search of associative memory. In G.H. Bower (Ed.), *The psychology of learning and motivation* (Vol 14, pp. 207-262). New York: Academic Press.

Raaijmakers, J.G.W., & Shiffrin, R.M. (1981). Search of associative memory. *Psychological Review*, 88, 93-134.

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

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

Roediger, H. L., Meade, M. L. & Bergman, E. (2001). Social contagion of memory. *Psychonomic Bulletin & Review, 8*, 365-371.

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review, 12*, 573-604.

Rouder, J. N., Lu, J., Sun, D., Speckman, P., Morey, R., & Naveh-Benjamin, M. (2007). Signal detection models with random participant and item effects. *Psychometrika*, *72*, 621-642.

Shiffrin, R.M. & Steyvers, M. (1997). A model for recognition memory: REM: Retrieving Effectively from Memory. *Psychonomic Bulletin & Review, 4*(2), 145-166.

Steyvers, M., Wallsten, T.S., Merkle, E.C., and Turner, B.M. (2014). Evaluating Probabilistic Forecasts with Bayesian Signal Detection Models. *Risk Analysis*, 34(3), 435-452,

Thorley, C., & Dewhurst, S.A. (2007). Collaborative false recall in the DRM procedure: Effects of group size and group pressure. *European Journal of Cognitive Psychology, 19*, 867–881.

Turner, B.M., Steyvers, M., Merkle, E.C., Budescu, D.V., Wallsten, T.S. (2014). Forecast Aggregation via Recalibration. *Machine Learning*, 95(3), 261-289.

Surowiecki, J. (2004). *The Wisdom of Crowds*. New York, NY: W. W. Norton & Company, Inc.

Vul, E. & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science, 19*(7) 645-647.

Wallsten, T. S., Budescu, D. V., Erev, I. & Diederich, A. (1997). Evaluating and combining subjective probability estimates. *Journal of Behavioral Decision Making, 10,* 243-268.

Weldon, M.S., & Bellinger, K.D. (1997). Collective memory: Collaborative and individual processes in remembering. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 1160–1175.

Weldon, M. S., Blair, C., & Huebsch, P. D. (2000). Group remembering: Does social loafing underlie collaborative inhibition? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(6), 1568-1577.

Wickens, T. (2001). *Elementary Signal Detection Theory*. Oxford University Press, USA.

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

Yi, S.K.M., Steyvers, M., & Lee, M.D. (2012). The Wisdom of Crowds in Combinatorial Problems. *Cognitive Science, 36*(3), 452-470.

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.

Figure Captions

Figure 1. The signal detection theory model for ratings data.

Figure 2. Estimated discrimination ability (d_a) of individuals and the aggregate. Error bars show the 90% confidence intervals for the parameter estimates and d_a values are ordered by magnitude.

Figure 3. Estimated mean (μ_t) and standard deviations (σ_t) for the target distribution for each individual and the aggregate. Error bars show the 90% Bayesian credible interval of the parameter estimates.

Figure 4. The SDT models estimated for the aggregate, best and median individuals. Dashed lines indicate the estimated criteria settings.