Running head: COLLECTIVE MEMORY PERFORMANCE

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: 4200 (excluding references)

Number of figures and tables: 4

Abstract

Much of the research in memory builds on an assumption that the experimenter knows which stimuli were presented at study. However, in many real-world situations, such as cases involving eyewitness testimony, this kind of ground truth might not be available. The only observable data might consist of the verbal reports in the form of recognition or recall judgments. We 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, we calculate an aggregated memory judgment by simply averaging the confidence ratings across individuals. Using a Bayesian Signal Detection Theory (SDT) analysis of the confidence ratings, we show that the aggregated confidence rating is associated with a discrimination performance that substantially better than the best performing individual in the group.

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 in press; Turner, Steyvers, Merkle, Budescu, & Wallsten, in press; 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 (Steyvers, Miller, Lee, & Hemmer, 2009), 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).

We will investigate the collective memory performance that can be obtained by pooling retrieved memories across a number of individuals. The main question is whether

aggregation can lead to a WoC effect where the aggregated memory judgment is better than the majority of individuals or better than even the best individual in the group. In addition, we use a Signal Detection Theory (SDT) approach to assess the performance of individuals and the aggregate and use the estimated model parameters to better understand where this advantage is coming from.

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 contrast, we will investigate situations where there is no social interaction or communication 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.

Our investigation focuses on a standard recognition memory paradigm. Each subject is given the same study list of items and is tested on the same set of test items (presented in different order). For each test item, the subject produces a rating expressing

the confidence that the item was part of the study list. Because the set of study and test items are equivalent, we can aggregate the recognition confidence judgments across individuals. We propose a very simple method to combine the recognition memory confidence ratings across individuals by taking a simple average of the confidence ratings for the same item. We will show that averaging item level confidence rating leads to a level of performance that is better than any of the individuals in the group. In fact, we will show that the average confidence rating *substantially* outperforms the best performing subject. The SDT analysis shows that this increase in performance is associated with changes in the means as well as the variances of the signal and noise distributions. Therefore, these findings show that there can be multiple sources for the performance improvements.

The plan for this paper is as follows. We will first describe the previously published data that will be used for our analysis. We 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. We will then apply this model on the recognition memory data and compare the performance of individual subjects to the aggregate. We further 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)

We will analyze the WoC effect using 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. Our 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, we will refer to these studies 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 necessarily in the same order) and each individual was tested on the same set of items (again, not necessarily in the same order). Because the set of study and test items are equivalent (within each data set), we can aggregate the confidence judgments across individuals, exploring the performance of the aggregate as well as the importance of the number of possible confidence ratings.

For each experiment, we construct the aggregate by taking 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"), we record a 4 for the average rating

(the rating for "dog"). In case the averaging leads to fractional ratings, the rating is rounded, such that the aggregate confidence is based on the same response scale available to subjects. We proceed with this averaging for all test items in the list. It might be useful to think of this 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 our analysis, we 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

We will be using Signal Detection Theory to assess the ability of individuals and the aggregate to discriminate between targets and lures. We will use the unequal variance SDT model as the basis for our analysis (e.g., Wixted, 2007). The prototypical unequal variance model 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]

In our analysis, we assume 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 $d_a = \sqrt{2}\mu / \sqrt{1+\sigma^2}$.

Parameter Estimation

To assess the performance of the aggregate subject relative to the individual subjects, we apply the model to each subject separately. Therefore, for each individual (including the aggregate), we estimate the model parameters 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 we will show to be the case for the aggregate). Furthermore, in the case for many criteria cutoff points, the hit rates can reach ceiling and the false alarm rates can be so infrequent, complicating the construction of the z-ROC curve.

To achieve accurate estimates of model parameters across a wide variety of experimental settings, we 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 we will compare parameter estimates of d_a of the aggregate to individuals.

For this paper, we pursue a simple Bayesian estimation approach that allows us to estimate the SDT model parameters separately for each individual subject, including the aggregate. Our 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. We can use these estimates 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, we not only have point estimates available but also samples from the full posterior, which allows us 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 this, we can calculate corresponding distributions over other variables such as discriminability (d_a). We will focus 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, we ordered the subjects 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. 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 that are 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.

[INSERT FIGURE 3 HERE]

To better understand the nature of the improved discriminability of the aggregate, we explored the relationship of μ and σ across individuals. Figure 3 shows the relationship between the two model parameters. 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, values that are consistent with the literature. 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 closer, 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.

It should be noted that this illustration is produced on the basis of the mean parameter estimates (the mean of the posterior). Figure 2 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 behavior of this subject is more consistent with an equal variance model, caution should be taken in interpreting the particular point estimate of σ found with our estimation procedure.

Discussion and Conclusion

Across three separate studies, we have demonstrated that an aggregate of memory judgments is associated with a performance level that is superior to all of the individual subjects. The aggregate is based on a simple average of the confidence ratings of a particular test item across subjects. Generally, the result of averaging across subjects is that factors that contribute to subject variability are averaged out. One source of variability might be 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 diminished by averaging the decisions (for the same test item) across subjects.

It is important to note that although we used a particular SDT model to evaluate the effects of aggregation, the goal of this research is not to use the SDT model as a model that describes the underlying processes that give rise to a memory judgment. We merely explored the consequences of averaging confidence ratings across subjects and used the SDT model as a measurement tool. However, if we did expand the scope of the Signal Detection approach and considered it 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. We did in fact observe a decrease in the variability of target strength (σ) of the aggregate. However, we also observed a change in target strength mean (μ). Note that because of 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. It is possible that other sources of noise in recognition memory contributed as well. For example, note that in the particular recognition memory experiments we analyzed, both the test and study items were randomly ordered between individuals. Therefore, it is possible that some of our aggregation benefits are due to averaging out

recency and primacy effects as well as sequential effects among test items and overall drifts in attention due to fatigue.

Many models of recognition memory might make similar predictions regarding the effects of aggregation. For example, one popular alternative account to SDT is provided by dual process models (Yonelinas, 1994). In this account, performance is a mixture of two processes: familiarity and recollection. The familiarity process is modeled as an equal variance SDT model and recollection is modeled as high-threshold decision process -- an item is recollected if it occurred on the study list with some probability. If in this model, the recollection process is uncorrelated across subjects such that for the same item, some subjects are able to recollect the item whereas other subjects fail, we can expect a similar effect of averaging–the fluctuations in recollection should average out and could emerge as a decrease in target strength variability.

Further experiments and modeling will need to be done to better distinguish between the underlying causes of the WoC effect. One 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, 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.

Appendix

In this section, we provide 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. We assume that the subject produces K unique confidence ratings. 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, $\mathbf{c}$ = (c_1,...,c_{K-1}). For convenience, we also assume two additional 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 $\mathbf{c}$ are latent variables. We assume 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:

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

We set the hyperparameters of these priors 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), we opted for the uniform prior. We also found that the uniform prior gave good performance in parameter recovery studies.

Note that because we used a Normal prior on each individual criterion, the criteria are not ordered in any particular way. However, we assume that the criterion values are ordered at the time a confidence rating is produced from the model—this greatly simplifies the inference procedure especially when a large number of criterion values need to be inferred.

Parameter estimation was performed by an 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. In a Metropolis-Hastings procedure, a combination of single-variable proposals as well as block proposals was used. 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

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.

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.

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), XX-XX.

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.

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

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.

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.

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.

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., Lee, M. D., Miller, B., & Hemmer, P. (2009). The Wisdom of Crowds in the Recollection of Order Information. In J. Lafferty, C. Williams (Eds.) *Advances in Neural Information Processing Systems, 23*. MIT Press.

Steyvers, M., Wallsten, T.S., Merkle, E.C., and Turner, B.M. (in press). Evaluating Probabilistic Forecasts with Bayesian Signal Detection Models. *Risk Analysis*.

Turner, B.M., Steyvers, M., Merkle, E.C., Budescu, D.V., Wallsten, T.S. (in press). Forecast Aggregation via Recalibration. *Machine Learning*.

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.

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.

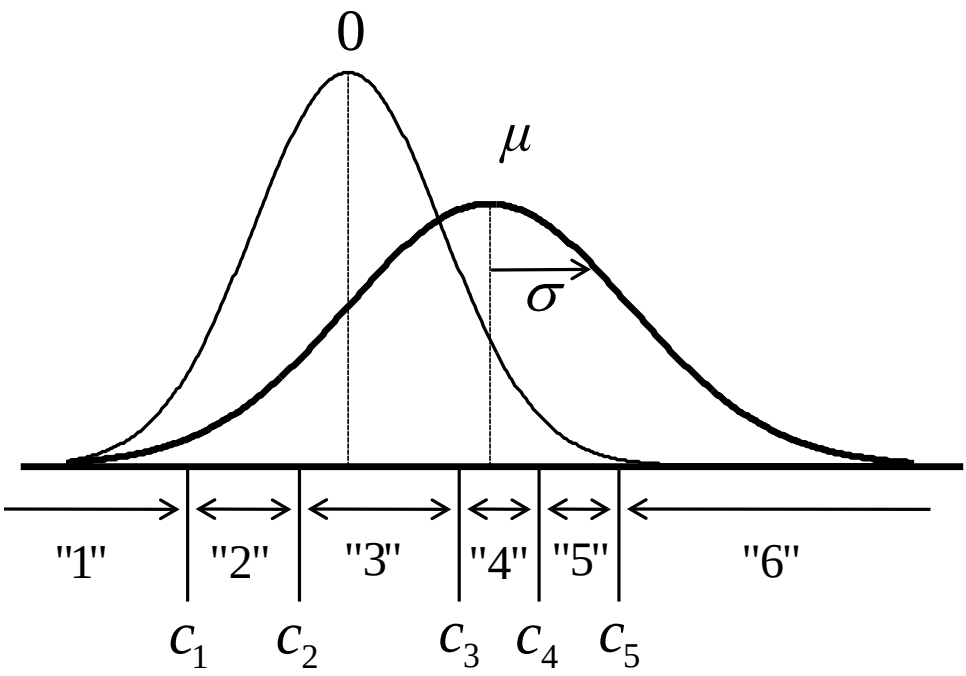
Figure Captions

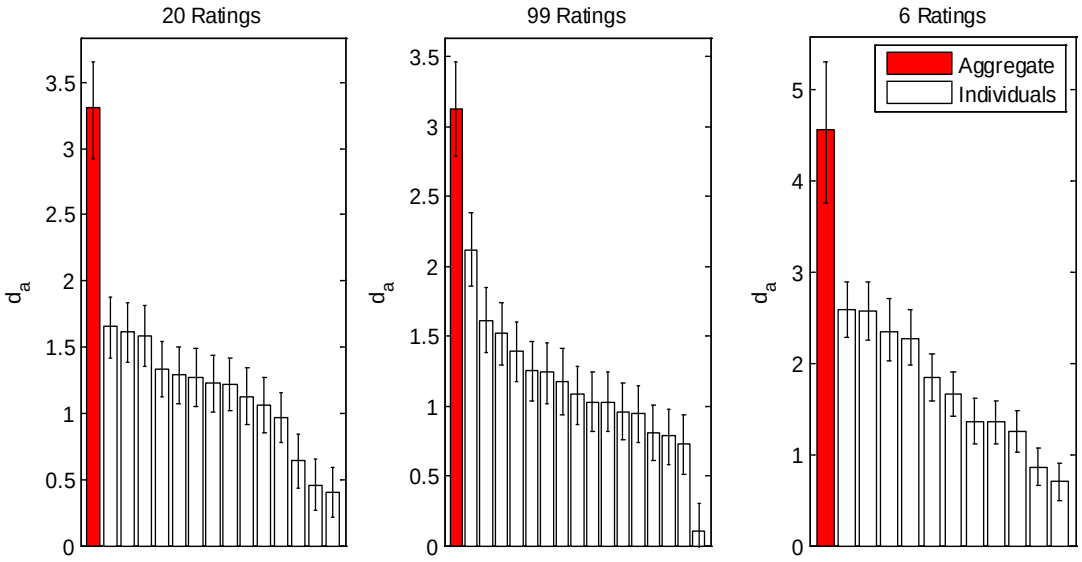
Figure 1. The signal detection theory model for ratings data.

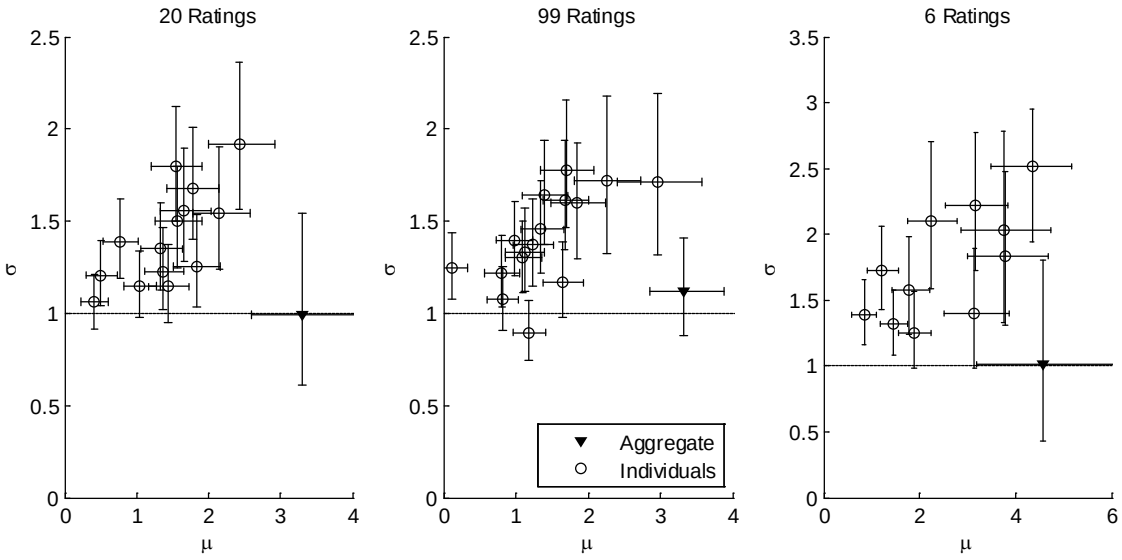
Figure 2. Estimated discrimination ability (d_a) of individuals and the aggregate. Error bars show the 90% confidence intervals for the parameter estimates and d_a values are ordered by magnitude.

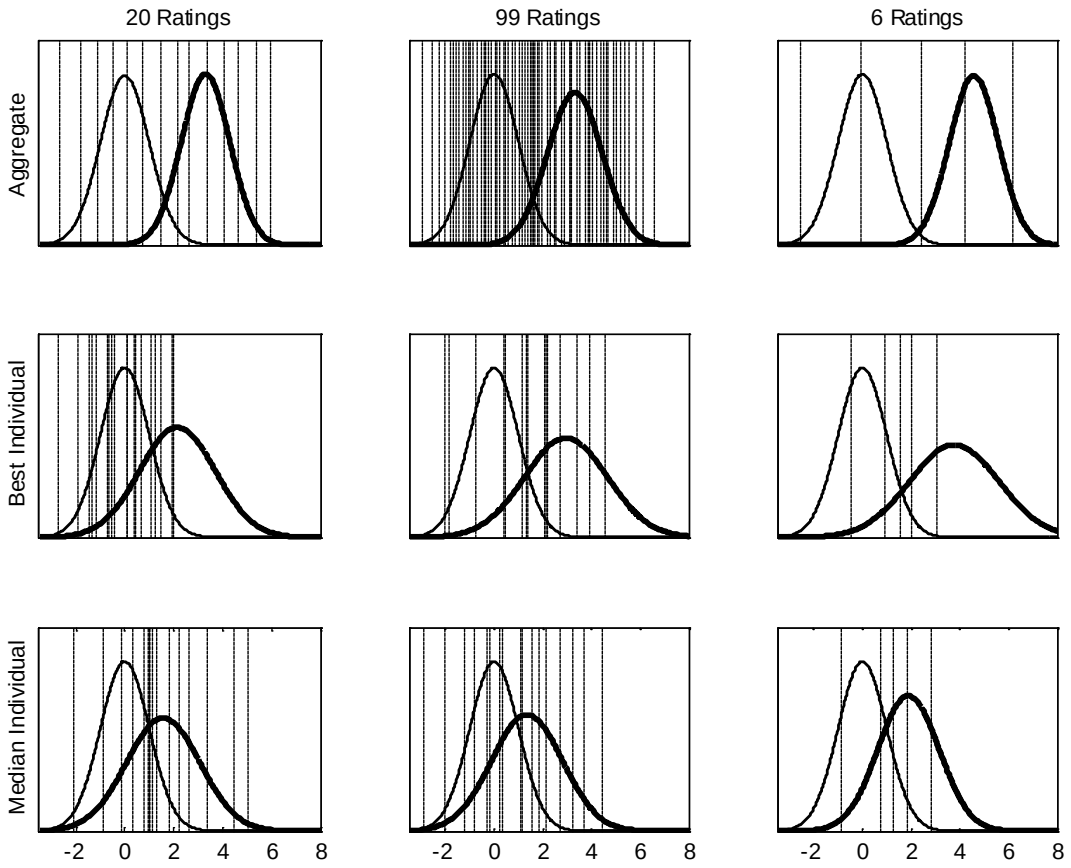
Figure 3. Estimated mean (μ_t) and standard deviations (σ_t) for the target distribution for each individual and the aggregate. Error bars show the 90% Bayesian credible interval of the parameter estimates.

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


0
μ
σ
"1"
"2"
"3"
"4"
"5"
"6"
c_1
c_2
c_3
c_4
c_5


20 Ratings
99 Ratings
6 Ratings
d_a
Aggregate
Individuals


20 Ratings
99 Ratings
6 Ratings
σ
μ
Aggregate
Individuals


20 Ratings
99 Ratings
6 Ratings
Aggregate
Best Individual
Median Individual
-2 0 2 4 6 8
-2 0 2 4 6 8
-2 0 2 4 6 8