Supplementary Material

Systematic Differences in Visual Working Memory Performance are Not Caused by Differences in Working Memory Storage

Michael S. Pratte
Mississippi State University

Marshall L. Green
Mississippi State University

Hierarchical Bayesian Model of Iconic Memory Decay

Linking Data To Latent Memory Processes

The data are hits and false alarms, which follow binomial distributions with hit rates (h_{ijk}) and false alarm rates (f_{ijk}) for the ith person studying the jth display, in the kth retention interval condition. The first step is to specify a model of how latent memory processes map on to these change detection data. We use the simple Cowan high-threshold model (Cowan, 2001), which assumes that responses are determined by the probability that a probed items is in memory, and participant's propensity to guess "change". We allow the probability of guessing "change" (g_i) to vary across participants, and the probability of an item being in memory (p_{ijk}) to vary across the i participants, the j study displays, and the k retention interval conditions. These latent memory parameters map to hits and false alarm rates as

$$\begin{aligned} h_{ijk} &= p_{ijk} + (1 - p_{ijk}) \times g_i, \\ f_{ijk} &= (1 - p_{ijk}) \times g_i. \end{aligned}$$

Because our primary goal is to model latent iconic and working memory capacity, the next step is to specify how memory performance (p_{ijk}) varies across retention interval conditions. We use the typical exponential decay function (e.g. Lu et al., 2005) to describe how the probability of an item being in memory declines with lengthening retention intervals,

$$p_{ijk} = \Phi\left(W_{ij} + A_{ij} e^{-\lambda R_{ij}}\right)$$

where R_{ij} is the retention interval (in seconds) for the ith person viewing the jth display, and parameter λ determines the rate of this exponential decay. Parameters W_{ij} are performance in working memory conditions ith person viewing the jth display (i.e. at long retention intervals) and A_{ij} are the degree to which performance in iconic memory ($R_{ij} = 0$) is greater than working memory for the ith person viewing the jth display. The probability of an item being in memory is constrained to be between zero and one, and the CDF of the standard normal distribution, or "probit" function (Φ) is used to constrain the exponential decay function to this space. This transform makes the model similar to a signal detection model in which memory effects are placed on d', and as will be shown below it also greatly facilitates parameter estimation.

Additive & Hierarchical Structure on Memory Parameters

The goal is to estimate latent working memory performance for each person and display (W_{ij}) and iconic memory performance for each person and display ($I_{ij} = W_{ij} + A_{ij}$). However, because each person in the study only sees a display once, only in one retention interval condition, and only as a change or same trial, some constraint is necessary in order to make the model identifiable. First, additive structures are used such that performance is modeled as the sum of a grand mean (μ), participant effects (α_i) and display effects (β_j),

$$\begin{aligned} W_{ij} &= \mu^W + \alpha_i^W + \beta_j^W, \\ A_{ij} &= \mu^A + \alpha_i^A + \beta_j^A. \end{aligned}$$

Participant and item effects are constrained by placing hierarchical prior structures on the effects,

$$\begin{aligned} \alpha_i^W &\sim \text{Normal}(0, \sigma^2_{\alpha^W}), \\ \beta_j^W &\sim \text{Normal}(0, \sigma^2_{\beta^W}), \\ \alpha_i^A &\sim \text{Normal}(0, \sigma^2_{\alpha^A}), \\ \beta_j^A &\sim \text{Normal}(0, \sigma^2_{\beta^A}). \end{aligned}$$

The variances of these distributions are free parameters that reflect the degree to which effect parameters deviate from zero. In particular, for the effects on working memory the hierarchical distribution variances reflect the degree to which people (α_i^W) and displays (β_j^W) vary in their working memory abilities. Effects on parameters A_{ij} indicate the extent to which the increase in performance from working to iconic memory varies across people (α_i^A) and across displays (α_i^A). If the gain in performance from working to iconic memory is constant across people, the α_i^A will all be near zero, such that $\sigma^2_{\alpha^A}$ will tend toward zero as well. Estimates of iconic memory performance can therefore be computed by adding effects on W and A. For example, estimated performance in a pure iconic memory condition ($R_{ij} = 0$) for the ith participant and jth item is given by

$$I_{ij} = \mu^W + \mu^A + \alpha_i^W + \beta_j^W + \alpha_i^A + \beta_j^A.$$

Person effects on guessing biases are also treated hierarchically,

$$\begin{aligned} g_i &= \Phi(\mu^G + \alpha_i^G), \\ \alpha_i^G &\sim \text{Normal}(0, \sigma^2_{\alpha^G}), \end{aligned}$$

Where the probit function Φ acts to constrain guessing biases between zero and one.

Although including the normal hierarchical structures adds to the total number of model parameters, they effectively simplify the model by constraining effect parameters to follow the normal parent distribution. The amount of constraint depends on the hierarchical variances, which are themselves estimated from the data. The result is typically a slight "shrinkage" of the parameter estimates, pulling toward zero what might otherwise be estimates that are relatively large compared with the other estimates. Such hierarchical modeling typically provides for more efficient parameter estimation as the influence of extreme values, such as from outliers or estimates based on small amounts of data, are lessened by being pulled toward zero. However, because hierarchical structures paradoxically increase the number of model parameters yet add parsimony to the model, typical model comparison approaches such as AIC or BIC are inappropriate, as they quantify model complexity by counting parameters. Instead, DIC (Spiegelhalter et al., 2002) is typically used to compare hierarchical models, as it correctly takes constraint imposed by the hierarchical structures into account when determining and penalizing for the complexity of a fitted model.

Priors on grand means (μ^W, μ^A, μ^G) are standard normal distributions, which are non-informative on the probit-transformed space. Priors on the hierarchical variances (e.g. $\sigma^2_{\alpha^W}$) are inverse gamma distributions (a=b=.01). The prior on the rate parameter (λ) is an exponential distribution with a rate of 0.1.

Model Estimation: Augmented Data Approach

Even with the constraint gained from additive and hierarchical structures, there are still far too many parameters for this model to be estimated using conventional methods (e.g. maximum likelihood). For example, in our design with 173 participants and 420 displays the model includes 1,367 parameters. Fortunately, modern Bayesian Markov chain Monte Carlo (MCMC) approaches make estimation feasible (see Gelman et al., 2004; Rouder & Lu, 2005). Although there are general-purpose and simple-to-use approaches to fitting models via MCMC such as JAGS (Plummer, 2003), we have found that large non-linear models such as are used here are too complex for such generic approaches, and instead require custom-designed algorithms for stable and efficient estimation.

The goal is to estimate the joint posterior distribution of all model parameters conditioned on the data. The MCMC approach makes doing so possible because individual parameters (or groups of parameters) can instead be sampled from their *conditional* posterior distributions, that is, the distribution of some parameter conditioned on both the data and samples of all other model parameters. The Markov Theorem shows that, under certain assumptions and with enough samples, sampling from the conditional distributions iteritively (e.g. sample parameter a given a previous sample of b, then sample a new b given that sample of a, and so on) provides samples from the marginal posterior distributions of parameters, which is the goal in analysis. Doing so requires only that the conditional posteriors are known, and that random samples can be drawn from them. In complex models with many parameters it is also important that these samples can be drawn from their conditional distributions efficiently, such as by sampling from the joint distribution of many parameters simultaneously. Here, however, the data are binomially distributed and the complex relationship between any parameter and the likelihood function of the data makes it difficult to find such simple conditional distributions. However, a powerful technique in cases with binomial data is to posit "augmented data", such that parameters can be conditioned on the augmented data rather than the binomial data to make sampling easier. This approach is common when data are binomial (Albert & Chib, 1993), and in multinomial processing tree models like the high-threshold memory model used here (Klauer, 2010).

The first step in the MCMC chain is to sample latent augmented normal data, conditioned on the observed binomal data and previously sampled parameters. Two sets of latent data are sampled, one corresponding to memory performance and one to guessing biases. For memory performance, on a particular trial the probed item may or may not have been in memory, with probability determined by p_{ijk} for that trial. For each trial, augmented data ($z_{ij}^{(m)}$) are sampled from a normal distribution with a mean of $\Phi^{-1}(p_{ijk})$ and variance of 1.0. If the binomial data and augmented guessing data for a particular trial imply that the item was not in memory on that trial, the $z_{ij}^{(m)}$ is sampled from the normal distribution truncated below zero (so forced to be negative). Alternatively, if the binomial and guessing data imply that the item was in memory for that trial, $z_{ij}^{(m)}$ is sampled from the normal truncated above zero (so forced to be positive). For example, consider a *change* trial in which the participant responded "change" (i.e. a hit), and the augmented guessing data for this trial and this sample in the MCMC chain implies that any guess response on this trial would have been "same". This pattern implies that the change response must have resulted from the item being in memory, so the augmented memory data for this trial ($z_{ij}^{(m)}$) is sampled from a normal with mean of $\Phi^{-1}(p_{ijk})$ that is truncated to be positive. Similarly, if on a *same* trial the participate responded "change" (a false alarm), the item could not have been in memory, and so the augmented memory data is truncated to be negative. In some cases the binomial data and latent guessing data do not indicate whether memory was successful or not, such as when a hit occurred and the guess for that trial was "change", such that the hit could have arisen from successful memory or from failed memory and a guess. In these cases the latent memory data are sampled from a normal without truncation.

Critically, by sampling the augmented memory data in this way across all trials within the MCMC chain, the resulting marginal distribution of $z_{ij}^{(m)}$ across all trials follows a normal distribution with mean equal to $\Phi^{-1}(p_{ij})$. The model parameters can be conditioned on the augmented normal data,

$$z_{ij}^{(m)} \sim \text{Normal}(W_{ij} + A_{ij}e^{-\lambda R_{ij}}, 1). \quad (1)$$

Because the $z^{(m)}$ are normally distributed, this model is now somewhat similar to a hierarchical linear model with Gaussian residuals, which makes sampling the parameters far more efficient than having to condition on the original binomial data. However, the binomial rates of the original data can be obtained from the augmented data, as the probit transform of the augmented data, $\Phi(z_{ij}^{(m)})$, are equal to the probabilities of an item being in memory p_{ij}.

Sampling the latent memory data requires also sampling augmented guessing data. That is, for each trial latent data ($z_{ij}^{(g)}$) are sampled that indicate whether a guess on that trial, should it have occurred, was a "change" or "same" guess. These data are sampled from truncated normal distributions in a similar fashion as the memory augmented data, but are conditioned on the binomial data (hits and false alarms) and the previously sampled augmented memory data ($z_{ij}^{(m)}$). For example, if a hit occurred on a *change* trial, but the sampled augmented memory data indicate that memory failed on that trial ($z^{(m)} < 0$), then the "change" response must have been a guess. This logic is used for each trial to sample $z_{ij}^{(g)}$ from a normal distribution with mean $\Phi^{-1}(g_{ij})$, and truncated to be positive when the guess was "change", negative when the guess was "same", and not truncated if the data and latent memory data do not imply one or the other. The resulting marginal distribution of $z_{ij}^{(g)}$ across all trials follows a normal distribution

$$z_{ij}^{(g)} \sim \text{Normal}(\mu^G + \alpha_i^G, 1).$$

The augmented data follow a linear model of guessing parameters μ^G and α_i^G with gaussian residuals, and so these parameters are easily sampled by conditioning on the augmented normal data. The probit transform of the guessing parameters $\Phi(\mu^G + \alpha_i^G)$ provides the guessing probabilities g_i. These guessing probabilities, along with the probabilities of items being in memory p_{ijk}, completely specify the binomial likelihood for the memory model.

Model Estimation: MCMC Parameter Sampling

The augmented data approach yields straightforward sampling of posterior distributions for most parameters. First, note that in Equation 1 the memory parameters W_{ij} and A_{ij} follow a simple linear model with gaussian residuals when conditioned on λ and the latent data $z_{ij}^{(m)}$, where W_{ij} are an intercept and A_{ij} are slope terms as a product of $e^{-\lambda R_{ij}}$. There are many powerful techniques for MCMC sampling in such situations, and we utilize *blocked sampling* whereby the intercept and slope terms are sampled together from their joint posterior distribution as a block (Roberts & Sahu, 1997). Because the joint posterior distribution of these parameters follows a multivariate normal distribution, they can be sampled simultaneously and independently from previous samples of these parameters, which leads to remarkably efficient sampling by minimizing sampling correlations between parameters and within parameters across MCMC samples. Because the number of parameters is so large across both W_{ij} and A_{ij}, they are sampled separately as a block of grand mean and participant effects ($\mu^W, \mu^A, \alpha_i^W, \alpha_i^A$) and a block of item effects (β_j^W, β_j^A). Guessing parameters (μ^G, α_i^G) are also sampled as a block, conditioned on the augmented data $z_{ij}^{(g)}$. Sampling the hierarchical variances ($\sigma^2_{\alpha^W}, \sigma^2_{\beta^W}, \sigma^2_{\alpha^A}, \sigma^2_{\beta^A}, \sigma^2_{\alpha^G}$) is straightforward given the sampled participant and item effect parameters: effects are normally distributed with inverse gamma prior, yielding an inverse gamma posterior distribution for each variance term (see Rouder & Lu, 2005, for an introduction to Bayesian modeling using similar approaches).

Approaches such as blocked sampling in which the conditional posterior distribution is of known form and can be sampled directly (e.g. from a multivariate normal or inverse gamma distribution) is referred to as Gibbs sampling (Geman & Geman, 1984), which is typically the most efficient method of sampling in MCMC. Unfortunately, the rate parameter λ in the exponential memory decay function does not have such a known posterior distribution, however, the posterior can be derived up to a constant of proportionality and sampled from using other methods. The posterior is equal to the likelihood (conditioned here on the normally distributed augmented data) times the prior, which is an Exponential distribution on λ with rate τ. The resulting log posterior distribution of λ is proportional to

$$l(\lambda|z_{ij}^{(m)}, W_{ij}, A_{ij}, \tau) \propto \frac{\sum_i \sum_j \left(z_{ij}^{(m)} - \left(W_{ij} + A_{ij}e^{-\lambda R_{ij}}\right)\right)^2}{-2} - \tau\lambda.$$

Samples were drawn from this posterior using Metropolis Hastings sampling (Hastings, 1970). This approach yields some auto-correlation in the MCMC chains, particularly due to trade-off between λ and the baseline working memory performance μ^A: a particularly high sample of λ (fast decay) will result in high sample of μ^A, which will then make the next sample of λ low, and so on. This correlation is mitigated to some degree by using a de-correlating step between μ^A and λ within the MCMC chain, by perturbing the sample of μ^A and λ in opposite directions with a normally distributed value, and using a Metropolis-Hastings sampler to accept the new perturbed values on approximately 40% of the MCMC samples.

Simulations show that this model estimation approach provides for efficient and accurate parameter estimation, but with some remaining auto-correlation. For the main analyses the MCMC chains were therefore run for 50,000 iterations, with a burn-in period of 5000 discarded samples that are discarded. The chains were then thinned such that only every 10th sample was retained, which further reduces correlation within each chain. This procedure provided 4500 MCMC samples for each model, which were used for parameter estimation and computing model fit statistics (DIC).

DIC Results for Simulated Data

Our main goal in developing the hierarchical model was to obtain estimates of latent working memory ability (W_{ij}) and latent iconic memory ability ($I_{ij} = W_{ij} + A_{ij}$) for each person and item, and measure the extent to which they are related. To do so, the full model presented above was fit to the experimental data. In this full model iconic and working memory parameters are free to be completely unrelated, yet estimated parameters (posterior means) from this model were found to be highly correlated. A restricted model was also fit to the data in which the additive parameters A_{ij} were forced to be a linear function of working memory performance (see main manuscript),

$$\begin{aligned} W_{ij} &= \mu^W + \alpha_i^W + \beta_j^W, \\ A_{ij} &= \mu^A + \gamma\alpha_i^W + \phi\beta_j^W. \end{aligned}$$

Because iconic memory performance (I_{ij}) is the sum of working memory performance and A_{ij}, in this restricted model

$$I_{ij} = (\mu^W + \mu^A) + \alpha_i^W(1+\gamma) + \beta_j^W(1+\phi)$$

That is, iconic memory for a particular person or item is constrained to be equal to working memory plus a constant (μ^A), and possibly with person and item effects scaled by γ and ϕ for people and items, respectively. By forcing iconic memory and working memory performance to be linearly related across people and items, this model implies that they are perfectly correlated.

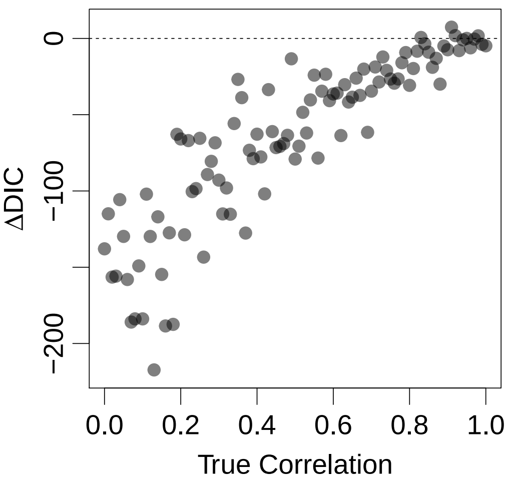

Figure 1. Simulation Results. For each simulation, differences in DIC across the models (y-axis) reflect evidence for the full model (negative values) or the restricted model (positive values). In each simulation the true correlation between effects on working and iconic memory were equal to a value between zero and one (x-axis).

Comparing the ability of the full and restricted models to fit the experimental data (using DIC) suggested that the restricted model was preferred. To examine the extent to which this result captures the true underlying relationship between working and iconic memory performance, in a simulation study data were generated from a model similar to the full model presented above, with true parameters and design similar to the experiment (200 participates and 420 items). However, effects on working memory and effects on iconic memory were sampled from bi-variate normal distributions. The correlation between participant effects on working and iconic memory, and the correlation between item effects on working and iconic memory, were varied across simulations from zero to one (in steps of .01). That is, from the case in which person and item effects are completely independent across working and iconic memory, to the case where they are precisely equal. The full and restricted (linear) models were fit to each sample, and the resulting DIC scores compared.

Figure 1 shows the resulting DIC differences as a function of the true correlation, where negative DIC values imply that the full model was preferred. For nearly all simulations in which the true correlation between working and iconic memory effects was less than 0.9, DIC correctly preferred the full model in which the effects are not forced to be linearly related. It is important to note that this full model can certainly be reduced to a restricted model, by allowing the variances of effects on A_{ij} (α_i^A, β_j^A) to be small, thus shrinking these effects toward zero resulting in equivalent person and display effects on iconic and working memory. Because the DIC statistic is designed to respect such constraint in computing model parsimony, the full model can essentially become a restricted model in which effects on iconic and working memory are equivalent, both in parameter estimates and model fit statistics. Nonetheless, in cases with moderate true correlation, the full model is providing for accurate parameter recovery, and even small deviations from a perfect linear relationship between effects on working and iconic memory is sufficient for DIC to prefer the full model. Critically, when the true correlations were extremely high (above about .9), DIC typically favored the restricted model, suggesting that when working and iconic memory are very highly related, the DIC approach correctly prefers the simpler model which forces them to be related. The DIC advantage of the restricted model is fairly small even for high correlation values, such that the preference for this model over the full model for the experimental data suggests a very strong relationship between effects on working and iconic memory.

References

Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, *88*(422), 669–679. https://doi.org/10.1080/01621459.1993.10476321

Cowan, N. (2001). The magical number 4 in short term memory. A reconsideration of storage capacity. *Behavioral and Brain Sciences*, *24*(4)arXiv 0140-525X, 87–186. https://doi.org/10.1017/S0140525X01003922

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian Data Analysis*. Boca Raton, CRC Press / Chapman; Hall. https://doi.org/10.1007/s13398-014-0173-7.2

Geman, S., & Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-6*(6), 721–741. https://doi.org/10.1109/TPAMI.1984.4767596

Hastings, W. K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, *57*(1), 97. https://doi.org/10.2307/2334940

Klauer, K. C. (2010). Hierarchical multinomial processing tree models: A latent-trait approach. *Psychometrika*, *75*(1), 70–98. https://doi.org/10.1007/s11336-009-9141-0

Lu, Z. L., Neuse, J., Madigan, S., & Dosher, B. A. (2005). Fast decay of iconic memory in observers with mild cognitive impairments. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(5), 1797–1802. https://doi.org/10.1073/pnas.0408402102

Plummer, M. (2003). JAGS: A program for analysis of Bayesian models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing; Vienna, Austria*, *124*(125.10), 1–10.

Roberts, G. O., & Sahu, S. K. (1997). Updating Schemes, Correlation Structure, Blocking and Parameterization for the Gibbs Sampler. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *59*(2), 291–317. https://doi.org/10.1111/1467-9868.00070

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*(4), 573–604. https://doi.org/10.3758/BF03196750

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583–639. https://doi.org/10.1111/1467-9868.00353