Why Does Higher Working Memory Capacity Help You Learn?

Kevin Lloyd (klloyd@gatsby.ucl.ac.uk)
Gatsby Computational Neuroscience Unit
25 Howland Street, London, W1T 4JG, UK

Adam Sanborn (A.N.Sanborn@warwick.ac.uk)
Department of Psychology, University of Warwick
University Road, Coventry, CV4 7AL, UK

David Leslie (d.leslie@lancaster.ac.uk)
Department of Mathematics and Statistics, Lancaster University
Lancaster, LA1 4YF, UK

Stephan Lewandowsky (stephan.lewandowsky@bristol.ac.uk)
School of Experimental Psychology and Cabot Institute, University of Bristol
Clifton, BS8 1TU, UK

Abstract

Algorithms for approximate Bayesian inference, such as Monte Carlo methods, provide one source of models of how people may deal with uncertainty in spite of limited cognitive resources. Here, we model learning as a process of sequential sampling, or 'particle filtering', and suggest that an individual's working memory capacity (WMC) may be usefully modelled in terms of the number of samples, or 'particles', that are available for inference. The model qualitatively captures two distinct effects reported recently, namely that individuals with higher WMC are better able to (i) learn novel categories, and (ii) flexibly switch between different categorization strategies.

Keywords: Bayesian inference; particle filter; working memory; category learning; knowledge restructuring

Introduction

Humans often behave in a manner consistent with Bayesian principles (Chater & Oaksford, 2008) yet how they achieve this is unclear. Though simple in principle, exact Bayesian calculations are frequently intractable in real-world settings, leading to a need for approximations. In statistics and computer science, this challenge has been met through the development of powerful, general-purpose techniques for approximate Bayesian inference, such as Monte Carlo methods, which allow practical application of Bayesian methods in complex domains. The practical success of these techniques has naturally prompted an interest in whether people deal with uncertainty in an analogous manner (Griffiths, Vul, & Sanborn, 2012). Importantly, such algorithms can approximate probabilistic inference arbitrarily well when sufficient time and memory are available, thereby providing a benchmark for ideal performance, but also display systematic deviations from the normative solution when resources are limited. These latter 'qualitative fingerprints' may be particularly illuminating when considering human cognition, where constraints on information-processing capacity are typically assumed. A salient example is provided by limits on working memory capacity (WMC; Cowan, 2001). While the exact nature of these limits remain the subject of debate, one prominent conception is that they reflect a limited resource which is shared across representations and processes in working memory (e.g., Just & Carpenter, 1992).

In the current work, we consider WMC limits within the context of Bayesian inference, asking whether WMC may be usefully modelled as a constraint on *inferential* resources. In particular, we model the learning process as one of *particle filtering*, in which a series of probability distributions is represented by a limited set of samples ('particles') which are sequentially updated over time (Griffiths et al., 2012). Higher WMC is then assumed to be implemented as a greater number of particles. This approach is applied to two recent experiments which indicate positive effects of higher WMC on two distinct aspects of categorization: (i) the facility with which novel categories are learned (Lewandowsky, 2011); and (ii) the ability to flexibly switch between different category representations or response strategies, referred to as *knowledge restructuring* (Sewell & Lewandowsky, 2012). We show that both of these effects are qualitatively captured by a single model in which WMC is equated with the number of particles available for inference — i.e., the *number of hypotheses* about category structure that an individual can concurrently entertain.

WMC and Category Learning

Lewandowsky (2011) measured participants' WMC before testing category learning performance on the six classical problem types of Shepard, Hovland, and Jenkins (1961) (henceforth 'SHJ'). Each involves learning to assign a set of stimuli to category *A* or *B* based on their values on binary dimensions, but the problem types vary in the number of stimulus dimensions required to correctly perform classification. Consistent with the classical results, participants generally learned the Type I problem fastest, Type VI the slowest, and Types II-V at an intermediate rate. Crucially, WMC score was found to be positively correlated with category learning performance: higher WMC individuals tended to make fewer errors across all problem types.

WMC and Knowledge Restructuring

Sewell and Lewandowsky (2012) assessed the relationship between WMC and performance in a knowledge restructuring (KR) task. Participants were guided to use one particular categorization strategy in a binary classification task before being instructed to switch to an alternative, equally-effective strategy (Fig 1A). The stimuli, rectangles of varying height with a vertical bar located at different locations along their base, belonged to category *A* or *B* depending on their position in category space (Fig 1B). Crucially, training stimuli (filled circles) were clustered into two separate regions of category space (as indicated by different colours), with categories arranged so that partial category boundaries (solid lines) could not be integrated in a coherent manner; neither partial boundary could be extended so as to allow accurate classification of all stimuli in the other cluster. A third, binary 'context' dimension was systematically mapped onto the two training clusters so that stimuli belonging to distinct clusters appeared in different colours (see example stimuli, lower Fig 1B).

At the task outset, participants were given information designed to guide them towards using one of two different strategies for co-ordinating partial categorization rules: (1) a *knowledge partitioning* (KP) strategy was encouraged by imparting that the context variable (colour) could be used to determine which dimension to use (rectangle height or bar position) for categorization; (2) a *context-insensitive* (CI) strategy was instead encouraged by highlighting that bar position could be used to determine which partial boundary to apply (i.e., regardless of context). Both strategies could support perfect performance but predicted different patterns of generalization when applied to new stimuli (open squares, Fig 1B) in a transfer test, thereby revealing which strategy was in use (Fig 1C). A summary 'context sensitivity' (CS) measure was applied to participants' test patterns to quantify the degree to which they generalized in a manner consistent with the KP (high CS) or CI (low CS) strategy (Fig 1D).

Critically, Sewell and Lewandowsky found evidence that individuals with higher WMC were more adept at switching between these different categorization strategies when instructed to do so, as measured by how much their CS scores changed between tests. This was interpreted in terms of greater 'knowledge restructuring', i.e., ability to coordinate different category representations or response requirements.

Modelling Approach

Our model comprises three parts: 1) assumptions about how participants *represent* categories, specified in terms of an explicit generative process; 2) a procedure by which participants are assumed to *infer* categories in light of prior assumptions and experimental stimuli; and 3) a means for translating participants' beliefs into *choice* (i.e., a predicted category label). Our description focuses on how the modelling approach is applied to the KR task; the SHJ tasks are simpler and easily modelled with only minor modifications.

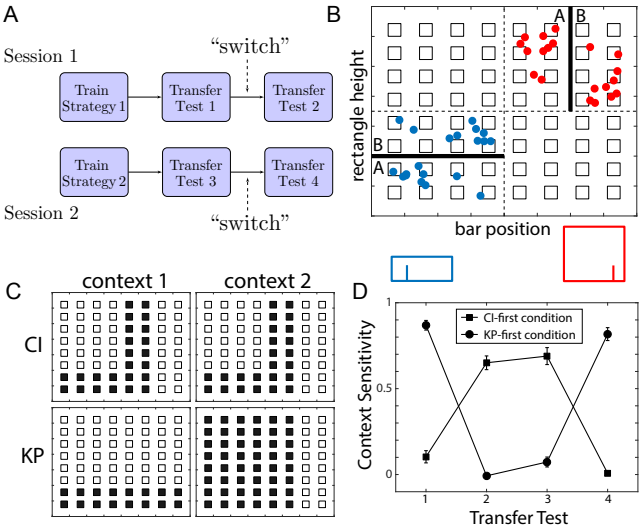

Figure 1: (A) Knowledge restructuring (KR) task design. (B) Experimental stimuli, depicted in category space: position of a vertically-oriented bar (*x*-axis) vs. height of rectangle (*y*-axis). Filled circles denote training stimuli; open squares denote test stimuli; solid lines indicate the partial rule boundaries. Two example stimuli are shown underneath. (C) 'Ideal' predicted response profiles given exclusive use of a context-insensitive (CI; top row) or knowledge-partitioning (KP; bottom row) strategy during test. Darker shading indicates a higher probability of classifying as category *A*. (D) Average context-sensitivity (CS) scores across participants during transfer tests, indicating use of CI (low) or KP (high) strategy. Figures B–D adapted from Sewell and Lewandowsky (2012).

Category Representation

A number of representational formats for categories have been discussed in the literature. Here, we opted to use *classification and regression tree* (CART) models (Breiman, Friedman, Olshen, & Stone, 1984). Firstly, these are well-suited to cases in which categories are readily described in terms of simple rules, particularly if an ordering on these rules is suggested (as in the KR task). Secondly, the classification boundaries generated by CART models lead naturally to 'axis-aligned' generalization patterns like those observed in the KR task (participants' response profiles were very similar to those shown in Fig 1C), whereas producing this behaviour is non-trivial for other category models.

Briefly, CART models provide a flexible method for specifying the conditional distribution of a binary category label y given a p-dimensional stimulus feature vector $\mathbf{x} = (x_1, x_2, \ldots, x_p)$. In the KR task, for a given stimulus on trial t, we have $y_t \in \{A, B\}$ and a 3-dimensional input $\mathbf{x}_t = (x_{t,1} = \text{bar position}_t \in \mathbb{R},\ x_{t,2} = \text{height}_t \in \mathbb{R},\ x_{t,3} = \text{context}_t \in \{0, 1\})$. The models work by recursively partitioning the input space into axis-aligned cuboids (similar to the partial boundaries in Fig 1B) and applying a simple conditional model to each region (e.g., probability that category label =

A). The sequence of partitions can be represented as a binary tree (Fig 2).

Formally, a binary tree structure T consists of a hierarchy of nodes $\eta \in \mathsf{T}$. Nodes with children are *internal* nodes, while nodes without children are *leaf* nodes (Fig 2A). Each node is associated with a block $B(\eta) \subseteq \mathbb{R}^p$ of the input space as follows: the root node is associated with the entire input space, while each further internal node splits its block into two halves by selecting a single dimension $\kappa(\eta) = \{1,\ldots,p\}$ and location $\tau(\eta)$ on which to split (Fig 2B). The block of input space associated with a node η is determined by the ranges on each dimension j which it covers, and we denote the corresponding range $R_j^\eta = [R_j^{\eta,-}, R_j^{\eta,+}]$. We call the tuple $\mathcal{T} = (\mathsf{T}, \kappa, \tau)$ the *decision tree*.

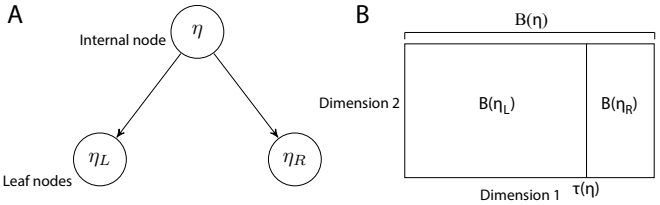

Figure 2: (A) Simple binary tree with (internal) root node η which splits into two 'leaf' nodes, η_L and η_R. (B) Corresponding split of a two-dimensional input space. The root node η is associated with the full input space, $B(\eta)$. Here, node η is split on dimension 1, $\kappa(\eta) = 1$, at a location $\tau(\eta)$. This splits the input space into two blocks, $B(\eta_L)$ and $B(\eta_R)$, associated with the leaf nodes η_L and η_R.

In addition to a decision tree $\mathcal{T}$ with K leaf nodes, a parameter $\Theta = (\theta_1, \theta_2, \ldots, \theta_K)$ associates parameter value θ_k with the kth leaf node. If a stimulus $\mathbf{x}$ lies in the region of the kth leaf node, then $y|\mathbf{x}$ has distribution $f(y|\theta_k)$ for some parametric family f. It is typically assumed that, conditional on $(\Theta, \mathcal{T})$, y values within a leaf node are i.i.d. and that y values across leaf nodes are independent. Thus, letting n_k denote the number of observations assigned to the kth leaf node and letting $y_{k,i}$ denote the ith observation of y assigned to leaf k,

$$p(y_{1:n}|\mathbf{x}_{1:n}, \Theta, \mathcal{T}) = \prod_{k=1}^{K} \prod_{i=1}^{n_k} f(y_{k,i}|\theta_k), \tag{1}$$

where $n = \sum_{k=1}^{K} n_k$ is the total number of observations.

Prior beliefs about category structure can be formalized as a prior distribution on decision trees, specified via a stochastic generative process. Following Chipman, George, and McCulloch (1998), we set the prior probability of a node η in tree structure T being split into children nodes to

$$p_{\text{SPLIT}}(\eta, \mathsf{T}) = \frac{\alpha}{(1+d_\eta)^\beta}, \tag{2}$$

where d_η denotes the depth of the node, and $\alpha < 1$ and $\beta \geq 0$ are parameters controlling expected tree size. Under this specification, the probability p_{SPLIT} is a decreasing function of node depth, and decreases more steeply for large β.

In addition to this prior on tree structure T, we generally assume that the probability of splitting on each dimension is equal,

$$p(\kappa(\eta) = j) = 1/p, \quad j = 1,\ldots,p, \tag{3}$$

and that split location is then drawn uniformly from the node's range,

$$\tau(\eta)|\kappa(\eta) = j \sim \mathcal{U}(R_j^{\eta,-}, R_j^{\eta,+}). \tag{4}$$

However, in the KR task, participants were guided towards a particular strategy by being told in the first instance that stimulus colour (KP-first condition) or bar position (CI-first condition) reliably indicated whether height or bar position was diagnostic of stimulus category. To incorporate this additional information, we assume a bias term $b \leq 1$ which assigns higher probability to splitting the root node η_0 on the dimension j^* highlighted by instruction:

$$p(\kappa(\eta_0)) = \begin{cases} b & \text{if } \kappa(\eta_0) = j^*, \\ \frac{1-b}{2} & \text{otherwise.} \end{cases} \tag{5}$$

The generative model is completed by the conditional probabilities of stimulus labels given the tree structure, $p(y_{1:t}|\mathbf{x}_{1:t}, \mathcal{T})$. We assume that the kth leaf node has an associated probability θ_k of generating label A,

$$p(y_t|\theta_k, \mathbf{x}_t) = \theta_k^{y_t}(1-\theta_k)^{1-y_t}, \tag{6}$$

and that this probability is an i.i.d. draw from a Beta distribution, $\theta_k \overset{iid}{\sim} Beta(a_0, b_0)$. Standard analytical simplification then yields the marginal likelihood

$$p(y_{1:t}|\mathcal{T}, \mathbf{x}_{1:t}) = \left(\frac{\Gamma(a_0+b_0)}{\Gamma(a_0)\Gamma(b_0)}\right)^K \prod_{k=1}^{K} \frac{\Gamma(n_{kA}^t + a_0)\Gamma(n_{k\cdot}^t - n_{kA}^t + b_0)}{\Gamma(n_{k\cdot}^t + a_0 + b_0)}, \tag{7}$$

where n_{kA}^t and $n_{k\cdot}^t$ are respectively the number of instances of category A and the total number of data points in the partition of leaf k up to trial t. Note that for a given tree, this likelihood is higher for leaves assigned observations with homogenous labels, and these are exactly the partitions that constitute 'good' solutions to the categorization problem.

Inference

Participants are assumed to approximate the sequence of posterior distributions $\{p(\mathcal{T}|\mathbf{x}_{1:t}, y_{1:t})\}_{t=1}^T$ over trials. Given the implausibility of enumerating all possible trees, participants are assumed to represent a relatively small number of samples, i.e. hypotheses, from these posterior distributions which can be updated over time. In other words, we assume participants perform particle filtering.

Two aspects of the inference process which we now describe draw parallels with working memory. Firstly, similar to the idea of a limit on the number of items that can be held in working memory (Cowan, 2001), we assume there is a bounded number of hypotheses about category structure — in this case, the particles which correspond to specific tree structures — that can be entertained at a given time. Secondly, similar to the notion that working memory is *active* (Baddeley, 1992), involving manipulation rather than merely

passive storage of items, we assume that inference involves a continual process whereby local transformations to current hypotheses are proposed, and which may be accepted or rejected. The latter process promotes diversity in the hypothesis set and continuous exploration of the hypothesis space.

In detail, we assume that on trial t, a participant's beliefs are represented by a small set of L possible trees $\{\mathcal{T}^{(l)}\}_{l=1}^{L}$ with associated importance weights $\{w_t^{(l)}\}_{l=1}^{L}$. This set of trees constitutes the limited set of hypotheses putatively maintained in a working memory of capacity L. With the observation of the stimulus and category label on the next trial $t+1$, a proper reweighting of the lth tree is given by the following update (Chopin, 2002):

$$w_{t+1}^{(l)} \propto w_t^{(l)} p(y_{t+1}|\mathcal{T}^{(l)}, \mathbf{x}_{t+1}, y_{1:t}). \tag{8}$$

As standard within particle filtering methods, this reweighting process is alternated with a *resampling* stage in which very unlikely trees, i.e., those with very low weights, are discarded and replaced by replicates of more probable trees. A simple way of doing this is to sample L times with replacement from the set $\{T^{(l)}\}$ with probabilities proportional to the updated weights $\{w_{t+1}^{(l)}\}_{l=1}^{L}$ (Gordon, Salmond, & Smith, 1993). Following this resampling step, all particle weights are equalized to $1/L$.

Additionally, this resampled particle set can then be *rejuvinated* (Chopin, 2002), reintroducing diversity and allowing continuous exploration of alternative solutions. This is the 'active' step which, we suggest, recalls conceptions of working memory as involving active manipulation of currently-stored items. Specifically, we may, without altering the targeted posterior distribution, propose transformations of trees from a Markov chain transition kernel $q_{t+1}(\cdot|\mathcal{T}^{(l)})$ with appropriate stationary distribution $p(\mathcal{T}|\mathbf{x}_{1:t+1}, y_{1:t+1})$. Closely following the transition kernel suggested by Chipman et al. (1998), we consider the scheme where for each tree $\{\mathcal{T}^{(l)}\}$, a new tree $\mathcal{T}^{(l)*}$ is proposed by randomly choosing among 3 possible transformations: (1) *grow*: randomly select a leaf node, then draw a splitting dimension and location from the prior; (2) *prune*: randomly select an internal node, then turn it into a leaf node by deleting all nodes below it; or (3) *change*: randomly select an internal node, then reassign it a splitting dimension and location by a draw from the prior. The proposed tree $\mathcal{T}^{(l)*}$ is then accepted with probability

$$\alpha(\mathcal{T}^{(l)}, \mathcal{T}^{(l)*}) = \min\left\{ \frac{p(\mathcal{T}^{(l)*}|\mathbf{x}_{1:t+1}, y_{1:t+1})/q_{t+1}(\mathcal{T}^{(l)*}|\mathcal{T}^{(l)})}{p(\mathcal{T}^{(l)}|\mathbf{x}_{1:t+1}, y_{1:t+1})/q_{t+1}(\mathcal{T}^{(l)}|\mathcal{T}^{(l)*})} \right\},$$

as per the standard Metropolis-Hastings algorithm.

We also need to model the effect of an instruction to switch categorization strategy. We assume that the effect is to *change the prior distribution* over trees, which is then combined with past observations to produce an updated posterior distribution. This update can be implemented via a simple *reweighting* operation on the set of trees.

To see how this works, consider the specific example where a participant has initially been guided to use the CI strategy and after t training sessions has in mind the set of weighted trees $\{\mathcal{T}^{(l)}, w_t^{(l)}\}_{l=1}^{L}$ approximating the target distribution under the prior appropriate to the CI strategy. We denote this target distribution $p_{CI}(\mathcal{T}|\mathbf{x}_{1:t}, y_{1:t})$. The experimenter then instructs the participant to change to using the KP strategy. Assuming that the set of trees remains fixed, the associated tree weights now need to be changed to reflect the new target distribution $p_{KP}(\mathcal{T}|\mathbf{x}_{1:t}, y_{1:t})$. This can be achieved by an *importance weighting* step, treating $p_{CI}(\mathcal{T}|\mathbf{x}_{1:t}, y_{1:t})$ as the importance distribution. In particular, denoting a particle's weight before and after the instruction to switch as $w_t^{(l)-}$ and $w_t^{(l)+}$, respectively, the relevant reweighting is

$$w_t^{(l)+} \propto w_t^{(l)-} \frac{p_{KP}(\mathcal{T}^{(l)}|\mathbf{x}_{1:t}, y_{1:t})}{p_{CI}(\mathcal{T}^{(l)}|\mathbf{x}_{1:t}, y_{1:t})}. \tag{9}$$

To switch in the reverse direction — from the KP to CI strategy — the appropriate reweighting instead uses the ratio $p_{CI}(\mathcal{T}^{(l)}|\mathbf{x}_{1:t}, y_{1:t})/p_{KP}(\mathcal{T}^{(l)}|\mathbf{x}_{1:t}, y_{1:t})$.

Choice

Participants are assumed to predict category labels based on their current hypotheses. Assuming a newly-resampled particle set with equal weights $1/L$, a sample-based approximation to the predictive probability that a stimulus $\mathbf{x}_{t+1}$ has label $y_{t+1} = A$ is given by

$$\begin{aligned} p(y_{t+1} = A|\mathbf{x}_{1:t+1}, y_{1:t}) &\approx \frac{1}{L}\sum_{l=1}^{L} p(y_{t+1} = A|\mathbf{x}_{1:t+1}, y_{1:t}, \mathcal{T}^{(l)}) \\ &= \frac{1}{L}\sum_{l=1}^{L} \mathbb{E}_{\theta_k|\mathbf{x}_{1:t+1}, y_{1:t}, \mathcal{T}^{(l)}}[\theta_k]. \end{aligned} \tag{10}$$

Thus, an approximation to the predictive probability is given by an unweighted average of posterior means for θ_k, where k for the lth particle is the index of the leaf node relevant to the input $\mathbf{x}_{t+1}$ in $\mathcal{T}^{(l)}$. In our case, the posterior mean is

$$\mathbb{E}_{\theta_k|\mathbf{x}_{1:t+1}, y_{1:t}, \mathcal{T}^{(l)}}[\theta_k] = \frac{n_{kA}^t + a_0}{n_{k\cdot}^t + a_0 + b_0}. \tag{11}$$

Results

Rate of Learning

Lewandowsky (2011) found that WMC was positively correlated with category learning performance. We hypothesized that a greater number of particles, i.e. increasing L, would have a similar effect since, on average, one might expect the search for a 'good' (i.e., more probable) category structure to progress faster, and with less chance of getting stuck in local maxima, with a higher number of particles.

Figure 3 displays average simulated learning curves for the SHJ tasks when the number of particles is increased from 1 (Fig 3A) to 100 (Fig 3B). Though the effect is subtle, there is a general steepening of learning curves and a downward shift in initial error rate for problem Type I. A more systematic gauge of the effect is obtained by fitting exponential functions to such learning curves and comparing the size of the fitted coefficients as the number of particles is increased (a

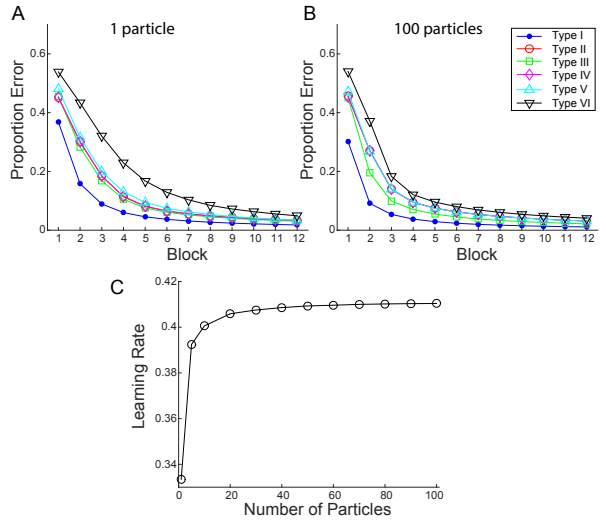

Figure 3: Increasing the number of particles leads to faster category learning. Simulated learning curves for (A) 1 particle, and (B) 100 particles. Learning curves are averages over 100 simulations with other model parameters fixed ($a_0 = b_0 = 1; \alpha = 0.95, \beta = 1$). (C) Learning rate as a function of number of particles. For each setting, the model is run 100 times and exponential curves fit to each individual learning curve. The resulting coefficients are averaged over both simulation runs and problem types to yield an aggregate 'learning rate'.

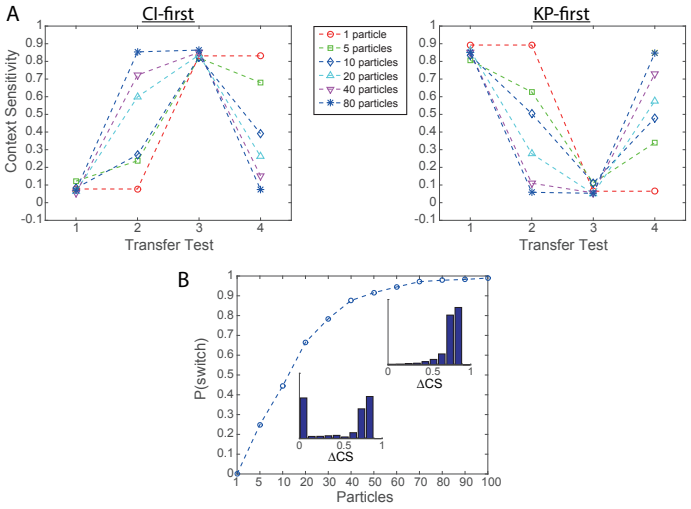

Figure 4: (A) In both the context-sensitive (CI)-first (left) and knowledge-partitioning (KP)-first (right) conditions, increasing the number of particles L leads to a greater change in context sensitivity (CS) score on average when prompted to change strategy (1500 simulation runs per condition). (B) This is due to an increased probability $P(\text{switch})$ of a successful switch ($\Delta CS > 0.5$). Lower inset: with fewer particles ($L = 20$), it will frequently occur that the model completely fails to switch ($\Delta CS = 0$). Upper inset: with more particles ($L = 100$), such failures are unlikely (3000 simulation runs; $b = 0.9, a_0 = b_0 = 1, \alpha = 0.95, \beta = 1$).

larger coefficient indicates a steeper learning curve). Figure 3C shows that the learning rate does increase with more particles, though the effect is small beyond ≈ 20 particles.

Note that even without fitting the model parameters, the basic SHJ pattern of results — Type I easiest, Type VI hardest, and Types II-V clustered in between — is reproduced. Briefly, this results from the preference for simpler, or more parsimonious, hypotheses that arises naturally within the Bayesian framework. An advantage for the Type II problem relative to types III-V is not produced by the model here, but we note that any such advantage was extremely marginal in Lewandowsky (2011), and that the effect may arise only under specific conditions (cf. Kurtz, Levering, Stanton, Romero, & Morris, 2013).

Knowledge Restructuring

Sewell and Lewandowsky (2012) found a positive association between WMC and knowledge restructuring. In the model, increasing the number of particles also has a beneficial effect on the average degree of knowledge restructuring (Fig 4A), with an increased probability of being able to successfully switch strategy (Fig 4B).

This result arises from an enhanced ability to accurately represent the posterior distribution with a greater number of particles. Recall that strategy-switching was modelled by a change in posterior distribution, driven by the different priors underlying the distinct strategies; a simple way to track this change was by reweighting particles according to the new distribution (Eq. (9)). However, the success of this will depend on how well the particle set covers the support of the updated distribution. With a sufficiently large number of particles, at least some should be allocated to (previously) lower probability regions; if the new strategy corresponds to such a region, then appropriate reweighting can be applied. However, with a decreasing number of particles, representation of the posterior distribution may be so impoverished that such regions of low probability may not contain any particles at all, and so switching is not immediately possible.

Discussion

Experiments suggest that higher WMC benefits learning of novel categories (Lewandowsky, 2011) and the ability to coordinate different category representations or response strategies (Sewell & Lewandowsky, 2012). We framed such tasks in terms of inference, where individuals seek to infer the most probable category structure(s) given their prior assumptions and experimental observations/instructions. Further, we assumed that individuals approximate inference by representing and manipulating in working memory a relatively small number of hypotheses — samples, or 'particles' — about possible category structures. Our principal hypothesis was that by linking WMC with the number of such particles, we would observe similarly positive effects of higher WMC on performance. Simulation results were consistent with this hy-

pothesis: more particles in the model enhanced both category learning performance and the ability to switch between different categorization strategies.

These effects respectively arise due to increased search efficiency and what we might call 'representational adequacy'. Conceptualized in terms of search for more probable categories, the more resources (i.e., particles) available to search this space — i.e., the greater the number of hypotheses that one can entertain and manipulate within working memory — then the more likely it is that one will quickly discover good solutions, a process which draws natural parallels with the broader topic of problem-solving (Hambrick & Engle, 2003; Newell & Simon, 1972). Furthermore, a greater number of particles generally means that the posterior distribution over categories is more accurately represented — including those assigned lower probability — and this pluralism means that the model can more easily express alternative hypotheses when instructed to switch strategy, as operationalized by a reweighting of particles. This source of flexibility may also be relevant to so-called 'insight' problem-solving (Murray & Byrne, 2005; Ohlsson, 1992).

The current work is preceded by a number of related lines of research. The HyGene model (Dougherty, Thomas, & Lange, 2010; Thomas, Dougherty, Sprenger, & Harbison, 2008), which emphasizes the importance of hypothesis generation and testing, includes the assumption that the number of hypotheses that can be entertained at a given time is limited by working memory constraints. Similarly, in their study of 'garden path' effects in sentence processing, Levy, Reali, and Griffiths (2008) suggested that difficulties in parsing such sentences correctly may be explained by constraints on the resources (i.e., number of particles) available for incremental parsing; their demonstration that a decreasing number of particles increases the probability of parse failure is exactly analogous to the mechanism suggested here in relation strategy-switching.

There are a number of avenues for future investigation. We have focused on qualitative effects here, but fitting the model to individual participants will be necessary for a more quantitative assessment; the obvious prediction is that high-WMC individuals should tend to be fit best by a larger number of particles. Decomposing the relative contributions of particular features of the model, such as resampling, should also be explored, and quality of fit directly compared with 'single-particle' approaches (e.g., Bramley, Dayan, Griffiths, & Lagnado, 2017). How the approach fares in domains beyond category learning is also of clear interest. More generally, Monte Carlo methods provide a rich source of ideas for psychological models — exploring how such methods may succeed or fail to illuminate aspects of human cognition is a substantial task for future research.

Acknowledgments

This work was supported by the Gatsby Charitable Foundation (KL), and by EPSRC grant EP/I032622/1 (DL).

References

Baddeley, A. (1992). Working memory. *Science*, *255*, 556-559.

Bramley, N., Dayan, P., Griffiths, T., & Lagnado, D. (2017). Formalizing Neurath's ship: Approximate algorithms for online causal learning. *Psychological Review*, *124*(3), 301–338.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth.

Chater, N., & Oaksford, M. (Eds.). (2008). *The probabilistic mind: Prospects for Bayesian cognitive science*. Oxford University Press.

Chipman, H., George, E., & McCulloch, R. (1998). Bayesian CART model search. *Journal of the American Statistical Association*, *93*(443), 935-948.

Chopin, N. (2002). A sequential particle filter method for static models. *Biometrika*, *89*(3), 539–552.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–114.

Dougherty, M., Thomas, R., & Lange, N. (2010). Toward an integrative theory of hypothesis generation, probability judgment, and hypothesis testing. In B. Ross (Ed.), *The psychology of learning and motivation* (Vol. 52, pp. 299–342). Burlington: Academic Press.

Gordon, N., Salmond, D., & Smith, A. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *Radar and Signal Processing, IEE Proceedings F*, *140*(2), 107–113.

Griffiths, T., Vul, E., & Sanborn, A. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science*, *21*(4), 263-268.

Hambrick, D., & Engle, R. (2003). The role of working memory in problem solving. In J. Davidson & R. Sternberg (Eds.), *The psychology of problem solving* (pp. 176–206). Cambridge University Press.

Just, M., & Carpenter, P. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, *99*, 122–149.

Kurtz, K., Levering, K., Stanton, R., Romero, J., & Morris, S. (2013). Human learning of elemental category structures: revising the classic result of Shepard, Hovland, and Jenkins (1961). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(2), 552–572.

Levy, R., Reali, F., & Griffiths, T. (2008). Modeling the effects of memory on human online sentence processing with particle filters. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in Neural Information Processing Systems 21*.

Lewandowsky, S. (2011). Working memory capacity and categorization: Individual differences and modeling. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(3), 720-738.

Murray, M. A., & Byrne, R. M. (2005). Attention and working memory in insight problem solving. In B. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the xxvii annual conference of the cognitive science society* (pp. 1571–1575). Lawrence Erlbaum Associates.

Newell, A., & Simon, H. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Ohlsson, S. (1992). Information processing explanations of insight and related phenomena. In M. Keane & K. Gilhooly (Eds.), *Advances in the psychology of thinking*. London: Harvester-Wheatsheaf.

Sewell, D., & Lewandowsky, S. (2012). Attention and Working Memory Capacity: Insights From Blocking, Highlighting, and Knowledge Restructuring. *Journal of Experimental Psychology: General*, *141*(3), 444-469.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied*, *75*(13), 1–42.

Thomas, R., Dougherty, M., Sprenger, A., & Harbison, J. (2008). Diagnostic hypothesis generation and human judgment. *Psychological Review*, *115*(1), 155–185.