Why higher working memory capacity may help you learn: Sampling, search, and degrees of approximation

Kevin Lloyd
Max Planck Institute for Biological Cybernetics
Max-Planck-Ring 8, 72076 Tübingen, Germany

Adam Sanborn
Department of Psychology, University of Warwick
Gibbet Hill Road, Coventry, CV4 7AL, UK

David Leslie
Department of Mathematics and Statistics, Lancaster University
Lancaster, LA1 4YF, UK

Stephan Lewandowsky
School of Psychological Science, University of Bristol
Clifton, BS8 1TU, UK

Abstract

Algorithms for approximate Bayesian inference, such as those based on sampling (i.e., Monte Carlo methods), provide a natural source of models of how people may deal with uncertainty with limited cognitive resources. Here, we consider the idea that individual differences in working memory capacity (WMC) may be usefully modeled in terms of the number of samples, or "particles", available to perform inference. To test this idea, we focus on two recent experiments that report positive associations between WMC and two distinct aspects of categorization performance: the ability to learn novel categories, and the ability to switch between different categorization strategies ("knowledge restructuring"). In favor of the idea of modeling

WMC as a number of particles, we show that a single model can reproduce both experimental results by varying the number of particles — increasing the number of particles leads to both faster category learning and improved strategy-switching. Furthermore, when we fit the model to individual participants, we found a positive association between WMC and best-fit number of particles for strategy switching. However, no association between WMC and best-fit number of particles was found for category learning. These results are discussed in the context of the general challenge of disentangling the contributions of different potential sources of behavioral variability.

1 Introduction

How to deal with uncertainty arising from noisy and incomplete information is a ubiquitous challenge for natural and artificial agents alike. Bayesian statistics provides a rigorous system for representing and reasoning about such uncertainty, yielding a principled method for updating beliefs in the light of new evidence (Bernardo & Smith, 1994). Human behavior is often well described in terms of Bayesian inference, from "low level" sensorimotor (Körding & Wolpert, 2004) and perceptual (Yuille & Kersten, 2006) phenomena, to "high level" competencies, such as causal reasoning (Griffiths & Tenenbaum, 2005), category learning (Sanborn, Navarro, & Griffiths, 2010), and predictions about future everyday events (Griffiths & Tenenbaum, 2006; reviews include Chater & Oaksford, 2008; Sanborn & Chater, 2016; Tenenbaum, Kemp, Griffiths, & Goodman, 2011).

How humans frequently — though by no means always (e.g., Tversky & Kahneman, 1974) — achieve this consistency with Bayesian principles is less clear. 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 (Gelfand & Smith, 1990; Robert & Casella, 2004), which allow for the practical application of Bayesian methods in complex domains.

The practical success of these techniques has naturally led to an interest in whether they also tell us something about how people reason under uncertainty. That is, they provide one source of hypotheses about the nature of the psychological and neural mechanisms that underlie how people process probabilistic information (Chater & Oaksford, 2008; Doya, Ishii, Pouget, & Rao, 2007). Since the aim of these algorithms is to approximate the normative solution to a computational problem — i.e., to approximate Bayesian inference — they have been called *rational process models* when considered as candidate psychological mechanisms (Griffiths, Vul, & Sanborn, 2012; Sanborn et al., 2010). This distinguishes them from traditional process models in cognitive psychology, which are typically rich in postulated psychological mechanisms but often poor in terms of normative foundations (cf. Anderson, 1990).

Importantly, Monte Carlo methods can in principle approximate probabilistic inference arbitrarily well when sufficient time and memory is available, thereby providing a benchmark for ideal performance. At the same time, these methods display systematic deviations from the normative solution when resources are limited. Such "qualitative fingerprints" associated with different species of approximation may then be particularly illuminating when considering human cognition, where it is generally assumed that information processing capacity is limited (Daw, Courville, & Dayan, 2008; Gigerenzer & Goldstein, 1996; Kahneman, 2003; Simon, 1982).

One such limitation has long been associated with working memory (Cowan, 2001; Miller, 1956), defined in cognitive psychology as the memory system responsible for temporary storage and manipulation of task-relevant information (Baddeley, 1992; Baddeley & Hitch, 1974). Individual differences in working memory capacity (WMC), such as measured in the complex span paradigm (Daneman & Carpenter, 1980), have been found to predict performance on a variety of cognitive tasks, including conventional intelligence tests (Conway, Jarrold, Kane, Miyake, & Towse, 2007). Indeed, WMC may account for up to one half of the variance in general intelligence (Conway, Kane, & Engle, 2003).

However, the exact nature of the WMC limitation that underpins such individual differences remains the subject of debate, with proposals variously emphasizing decay of representations (e.g., Baddeley, Thompson, & Buchanan, 1975), resource constraints (e.g., Just & Carpenter, 1992), or interference (e.g., Oberauer & Kliegl, 2006; see Oberauer, Farrell, Jarrold, & Lewandowsky, 2016 for a recent discussion). Indeed, opinions continue to differ as to whether working memory is best conceptualized as discrete, e.g., comprising a limited number of "slots", or as a more continuous "resource" that can be flexibly distributed across representations in memory (Ma, Husain, & Bays, 2014; Suchow, Fougnie, Brady, & Alvarez, 2014).

Our approach in the current work is to consider WMC limitations within the broader context of probabilistic inference, asking whether WMC may be usefully modeled as a constraint on the amount of *inferential* resources available. The implication is that at least in tasks involving uncertainty, enhanced performance in individuals with higher WMC may be attributable to an ability to better approximate "ideal" Bayesian solutions.

To begin to explore this idea, we focus on recent experiments showing positive associations between WMC and performance on category learning tasks (Lewandowsky, 2011; Lewandowsky, Yang, Newell, & Kalish, 2012; Sewell & Lewandowsky, 2011, 2012). This focus is motivated by two considerations. Firstly, category learning tasks are well characterized as probabilistic inference problems, requiring participants to reason about possible underlying category structures. Even when the mapping between stimuli and category labels is deterministic, participants face epistemic uncertainty regarding the nature of this mapping. Normative solutions to such problems, as well as how these solutions may be practically approximated — notably via Monte Carlo methods — have received substantial attention (Anderson, 1990; Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Sanborn et al., 2010). We build on this previous work here. Secondly, WMC appears to be positively associated with two distinct aspects of categorization: the ability to acquire novel categories (i.e., category learning; Lewandowsky, 2011), and the ability to flexibly switch between different categorization strategies (sometimes referred to as "knowledge restructuring"; Sewell & Lewandowsky, 2012). Previous work has explored how such positive associations may arise in formal category learning models (Lewandowsky, 2011; Sewell & Lewandowsky, 2011, 2012) but has treated these aspects of categorization separately, and via different models and mechanisms; the possibility that WMC may influence both category learning and knowledge restructuring via a single mechanism has not been explored, and we seek such a common mechanism in the present article.

The key assumptions of the current work are that individuals approximate Bayesian solutions to category learning problems by sampling from probability distributions (i.e., via Monte Carlo inference) and, more importantly, that an individual's WMC directly translates into how many samples, or hypotheses, they are able to represent at one time. We show that this simple equating of WMC with the number of active hypotheses allows us to reproduce the positive associations between WMC and both aspects of categorization performance — category learning and knowledge restructuring — with a single mechanism. Before describing the modeling approach and results in detail, we briefly summarize the basic ideas behind Monte Carlo methods and the target experimental results.

1.1 Monte Carlo as a psychological mechanism

In the Bayesian paradigm, background knowledge gives rise to a constrained set of candidate hypotheses $\mathcal{H}$ for the true state of nature, and to associated degrees of belief $P(h)$ in each candidate in the set $h \in \mathcal{H}$. The sum of all beliefs about the true state of nature is fixed to 1. Such "prior" beliefs are updated in the light of observed data d to yield "posterior" beliefs $P(h|d)$ via Bayes' theorem,

$$P(h|d) = \frac{P(d|h)P(h)}{\sum_{h' \in \mathcal{H}} P(d|h')P(h')},$$

where the likelihood $P(d|h)$ quantifies how expected the data are under each candidate hypothesis.

As we will describe in detail below, for our purposes the state of nature is the true category structure that participants are required to learn; the set of candidate hypotheses is the space of all possible category structures that a participant is assumed to be able to generate; and the observed data are the particular category instances presented to participants that they must categorize and for which they subsequently receive feedback about the correct category label.

While Bayes' theorem is simple to write down, it leads to complex practical issues such as the source of the prior distribution, the choice of likelihood function, and how to compute and summarize the posterior distribution if the hypothesis space $\mathcal{H}$ is very large — such as when $\mathcal{H}$ is the space of all possible categories.

In Monte Carlo methods, the basic idea is to approximate the target distribution $P(h|d)$ by drawing samples from it. In other words, one represents $P(h|d)$ with a set of samples $\{h^{(i)}\} \sim P(h|d)$ from that distribution, each randomly selected with a frequency proportional to its probability in the full distribution.

In the case where beliefs are updated sequentially as new information arrives — as in the experiments we consider below, where participants receive feedback trial by trial — one attempts to approximate a *sequence* of target distributions, and so we are more specifically interested in the idea of *sequential Monte Carlo*, or "particle filtering" (Doucet, de Freitas, & Gordon, 2001). As we will describe in more detail, one way of promoting a good approximation to posterior distributions in this instance is to propose local changes to a current hypothesis h, and to accept or reject the proposed variant h' as a function of its posterior probability. This latter process can be thought of in terms of continuous exploration, or *search*, of the hypothesis space for regions of high probability.

These two characteristics of Monte Carlo inference — representation by a limited number of hypotheses, and inference as involving an active process of exploration, or search, of the posterior — draw parallels with working memory, which is typically characterized not only as limited in capacity but also as *active* memory (Baddeley, 1992). In other words, if WMC is the number of hypotheses that one can actively maintain and manipulate at a given time, and if these latter processes can be cast in terms of probabilistic inference, then a possible analogy between working memory processes and Monte Carlo inference presents itself.

Of course, the idea that *sampling* plays a role in psychological mechanisms has a long tradition in psychology (Busemeyer, 1985; Estes, 1950; Restle, 1962; Stewart, Chater, & Brown, 2006), though not typically in the context of approximating Bayesian inference. More recent work has explicitly considered sample-based inference as a possible psychological mechanism (recent reviews include Griffiths et al., 2012; Suchow, Bourgin, & Griffiths, 2017). For example, Vul and Pashler (2008) argued that the "wisdom of crowds" effect, where the error of a judgment averaged over individuals is substantially smaller than the average error of individual judgments, is consistent with individuals using only a limited number of samples to form estimates (cf. Lewandowsky, Griffiths, & Kalish, 2009). Other work has focused on apparent suboptimalities displayed in people's sensitivity to the ordering of information when they must update their beliefs over time. Such order effects have been successfully captured by models employing sequential inference with limited samples in a variety of domains, including change detection (Brown & Steyvers, 2009), garden path effects in sentence processing (Levy, Reali, & Griffiths, 2008), and category learning (Sanborn et al., 2010).

1.2 Working memory capacity and category learning

Despite the central importance of both working memory and categorization in cognition, until recently the relationship between these abilities received scant attention.

The nature of this relationship is of interest not only to provide further constraints on adequate theories of these faculties, but also in light of recent arguments for the existence of multiple categorization systems that rely to differing degrees on distinct memory systems. One salient hypothesis is that category learning tasks that can be solved with relatively simple, verbalizable rules ("rule-based" tasks) rely especially on working memory, while tasks with solutions that generally defy description in terms of simple rules ("information-integration" tasks) do not (Ashby & Maddox, 2005, 2011; Ashby & O'Brien, 2005).

In contrast to this proposal, recent studies have found a positive association between WMC and category learning performance, regardless of whether the categorization task is rule-based (Lewandowsky, 2011) or based on information-integration (Lewandowsky et al., 2012). Interestingly, WMC has also been found to be positively associated with a somewhat distinct aspect of categorization, namely the ability to flexibly switch between different categorization strategies (Sewell & Lewandowsky, 2012) — a capacity that the authors refer to as "knowledge restructuring". These apparently disparate findings, which we describe next, form the target of the current work.

1.2.1 A positive association between WMC and category learning

Lewandowsky (2011) used a battery of four working memory tasks (memory updating, operation span, sentence span, and spatial short-term memory tasks — refer to the original paper for further detail and references) to measure the WMC of participants before testing their category learning performance on the six classical problem types of Shepard, Hovland, and Jenkins (1961) (henceforth "SHJ"). Each problem type involves learning to assign each of a set of 8 stimuli to category A or B based on their values on 3 binary dimensions (Fig. 1A); half of the stimuli are assigned to category A, and the other half to category B. There are 72 possible assignments that satisfy these conditions, but these reduce to 6 "types" assuming interchangeability of dimensions and labels (Fig. 1B). The problem types vary with respect to the number of stimulus dimensions that are relevant for classification. For example, in a Type I problem, only a single dimension is relevant; in a Type VI problem, by contrast, all 3 dimensions are relevant.

Consistent with the classical results, Lewandowsky found that the average trend of participants was to learn a Type I problem fastest, a Type VI problem the slowest, with Types II–V clustered in between (Fig. 1C). Crucially, structural equation modeling of WMC and category learning measures also revealed that WMC was positively related to category learning performance in each problem type (see Lewandowsky, 2011 for details). In Figure 1D, we replot the data to show the overall proportion of errors for each problem type given the median split of participants into high- and low-WMC groups based on their WMC scores. There is a clear trend for high-WMC participants to make fewer errors on each type of problem. Entering errors into a 2 (WMC: low, high) × 6 (Problem: I, II, III, IV, V, VI) × 12 (Block: 1–12) repeated measures ANOVA confirmed that high-WMC participants were more accurate than low-WMC participants ($F(1, 111) = 13.63, p < .01$), with no significant interactions between WMC and the other factors. Low-WMC participants made significantly more errors on each problem type, with the exception of Type IV.

1.2.2 A positive association between WMC and knowledge restructuring

Sewell and Lewandowsky (2012) found that higher WMC (where WMC was assessed using the same battery of measures as in Lewandowsky, 2011) was associated not only with better category learning performance, consistent with the findings of Lewandowsky (2011), but also with an improved ability to switch between categorization strategies when instructed to do so — an ability assumed to reflect knowledge restructuring (Sewell & Lewandowsky, 2011).

Like the SHJ problems, the basic task in the studies by Sewell and Lewandowsky (2012) was to learn to assign stimuli to category A or B. Here, stimuli were rectangles that varied with respect to 3 features (height, the position a vertical bar located along their base, and color). Stimuli were assigned to category A or B depending on their position in stimulus space (Fig. 2A). Height and bar offset were continuous dimensions, whereas color could take only one of 2 values (e.g., blue or red). Training stimuli (filled circles, Fig. 2A) were clustered into two separate regions of category space, with categories arranged so that partial category boundaries (solid lines, Fig. 2A) could not be integrated in a coherent manner — i.e, neither partial boundary could be extended in a way that allowed accurate classification of training stimuli in the other cluster, thereby encouraging co-ordination of multiple partial rules (for fuller discussion, see Sewell & Lewandowsky, 2012).

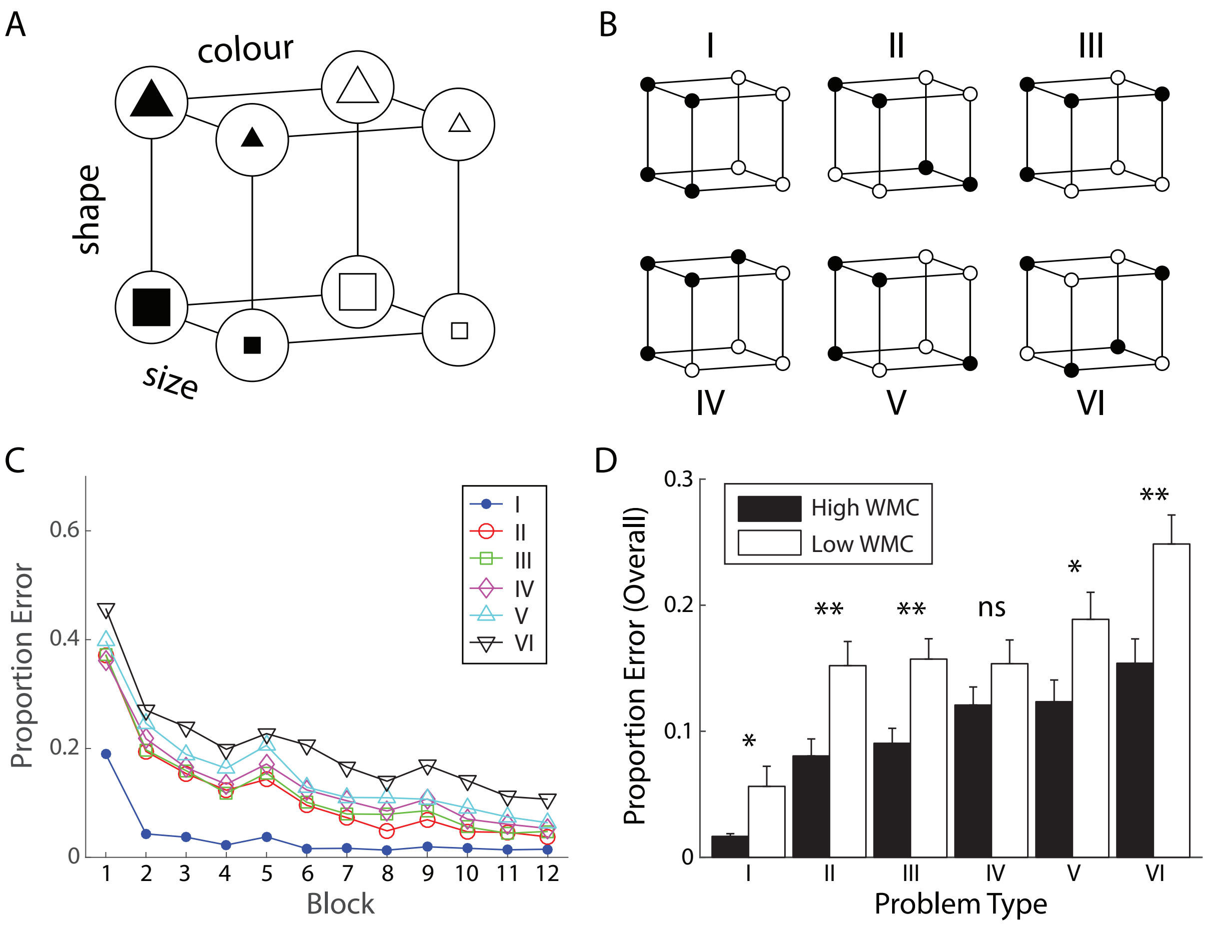

Figure 1: **The 6 category learning problem types of Shepard et al. (1961).** (A) Each one of 8 stimuli is defined by its unique combination of values on three dimensions (e.g., color, size, and shape) that correspond to the edges of the cube. (B) In each problem type, 4 stimuli are assigned to category A (filled circles), and the remaining 4 stimuli are assigned to category B (open circles). (C) Learning curves for each problem type, averaged over all participants, measured by Lewandowsky (data replotted from Lewandowsky, 2011). (D) Overall proportion of errors for high- and low-WMC participants (median split by WMC score) for each problem type. Error bars represent $+1SE$.

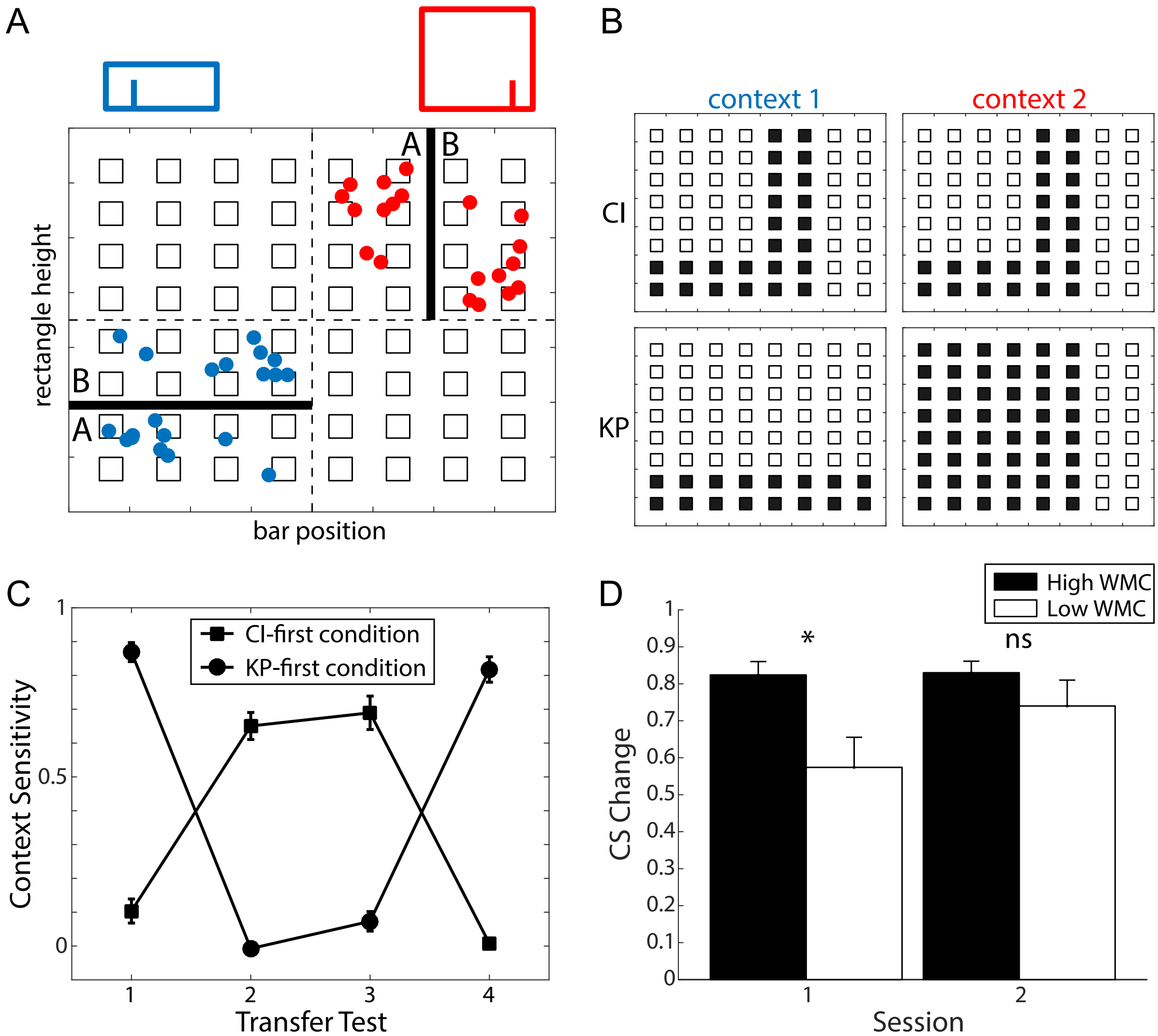

Figure 2: **Knowledge restructuring task of Sewell and Lewandowsky (2012).** (A) Experimental stimuli. These were rectangles (two examples shown at top) that varied with respect to their height, position of a vertically-oriented bar along their base, and color (e.g., blue or red). Stimuli were assigned to category A or B depending on their position in stimulus space. Filled circles denote training stimuli, open squares denote test stimuli, and solid lines indicate the partial rule boundaries. (B) Ideal response profiles associated with the context-insensitive (CI; top row) and knowledge-partitioning (KP; bottom row) categorization strategies. Shading indicates the probability with which a test stimulus should be classified as belonging to category A (darker color indicates a higher probability). Ideal performance in the different contexts (i.e., test stimulus presented in blue or red) is shown in the left and right columns of panels, respectively. (C) Context sensitivity across all transfer tests for knowledge-partitioning (KP)-first and context-insensitive (CI)-first conditions. Error bars indicate ± 1 SEM. (D) Mean absolute change in context sensitivity (CS) for participants with WMC scores in the top and bottom quartiles ("High" and "Low" WMC, respectively) for Session 1 (i.e., between transfer tests 1 and 2) and Session 2 (i.e., between transfer tests 3 and 4). Error bars indicate $+1SE$. Figures A–C after Sewell and Lewandowsky (2012).

Importantly, equally good categorization performance in this task could be obtained by learning any one of a number of different strategies. For example, a participant could use the color of the rectangle to decide whether height (for blue rectangles) or bar position (for red rectangles) predicted category A or B — this was named a *knowledge-partitioning* (KP) strategy. Alternatively, a participant could attend to whether bar position was to the left or right of center in order to then diagnose category membership based on either height or, again, bar position — thereby ignoring the color dimension entirely. This latter was named a *context-insensitive* (CI) strategy.

The crucial experimental manipulation was to encourage a participant, using verbal instruction, to first learn one of these 2 strategies — by hinting that the problem could be solved using bar position (for a participant assigned to the "CI-first" experimental group) or color (for a participant assigned to the "KP-first" experimental group) — before giving the participant an unexpected instruction to switch to using the alternative strategy. The degree to which participants' predictions conformed to a CI or KP strategy could be assessed via their generalization performance on a set of test stimuli (open squares, Fig. 2A), since generalization performance should be either insensitive (CI strategy) or sensitive (KP strategy) to the color of the presented stimuli (Fig. 2B). On the basis of their generalization pattern, participants were assigned a "context sensitivity" score, summarizing the degree to which their performance best conformed to a CI (context sensitivity close to 0) or KP (context sensitivity close to 1) strategy.

Regardless of whether participants were encouraged to use a CI or KP strategy in the first instance, they were able to shift between strategies without any training on the novel strategy (Fig. 2C), an ability assumed to reflect knowledge restructuring (Sewell & Lewandowsky, 2011). More importantly for our purposes, however, was the finding of a significant positive correlation between WMC and the extent of knowledge restructuring, the latter being measured in terms of the absolute change in context sensitivity in each test session (see Sewell & Lewandowsky, 2012, for full details of the structural equation modeling approach and results). Figure 2D shows the average change in context sensitivity for participants with WMC scores in the top and bottom quartiles, for Session 1 (i.e., changes between transfer tests 1 and 2) and Session 2 (i.e., changes between transfer tests 3 and 4). Entering these change scores into a 2 (WMC: low, high) $\times$ 2 (Condition: CI-first, KP-first) $\times$ 2 (Session: 1,

2) repeated measures ANOVA confirmed a main effect of WMC on change in context sensitivity ($F(1, 47) = 4.42, p < .05$). High-WMC participants had significantly higher changes in context sensitivity in Session 1 ($t(48) = 2.81, p < .01$), though not in Session 2 ($t(48) = 1.17, \mathrm{ns}$); we defer discussion of this, and further subtleties of the experimental results, until later (see Discussion).

The results of Sewell and Lewandowsky (2012) thus suggest that WMC supports not just standard category learning but also the flexible application of different categorization strategies.

2 Modeling approach

The hypothesis of the current study was that by equating working memory capacity (WMC) with the number of samples available for inference in a Bayesian category learning model, positive associations between WMC, category learning, and knowledge restructuring would naturally arise, consistent with the experimental findings.

Our model can be described as comprising three parts: 1) a model of how participants are assumed to *represent* categories, specified in terms of an explicit process whereby categories can be constructed (i.e., a "generative model"); 2) a procedure by which participants are assumed to *infer* categories in light of their prior assumptions and the experimental stimuli; and 3) a means for translating participants' beliefs about categories into *choice*, i.e., a prediction of the category label associated with a stimulus before receiving feedback about the true label.

2.1 Category representation

Many representational formats for categories have been discussed in the literature, including rules (Bruner, Goodnow, & Austin, 1956; Goodman et al., 2008; Nosofsky, Palmeri, & McKinley, 1994), prototypes (Posner & Keele, 1968; Rosch, 1973), exemplars (Kruschke, 1992; Medin & Schaffer, 1978; Nosofsky, 1986), or some mixture of these (Anderson, 1991; Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Love, Medin, & Gureckis, 2004). In the current work, we chose to work within the framework of *classification and regression tree* (CART) models (Breiman, Friedman, Olshen, & Stone, 1984), which can be considered a type of rule-based representation. This choice was largely pragmatic. Firstly, CART models offer an intuitive format for the categories used in the experimental tasks of interest, which are readily described in terms of simple, verbalizable rules (i.e., "rule-based", in the terms of Ashby & Maddox, 2005) and that also suggest an ordering on rules (particularly the task of Sewell & Lewandowsky, 2012; see below). Secondly, as we will describe, these models are amenable to a Bayesian formulation (Chipman, George, & McCulloch, 1998), which is obviously crucial for our purposes.

Most broadly, CART models (Breiman et al., 1984) provide a flexible method for specifying the conditional distribution of a response variable (e.g., a category label) given a collection of input predictors (e.g., stimulus features). In the experiments we consider, category labels are always binary, $y \in \{A, B\}$, and each stimulus to be categorized is represented by a p-dimensional feature vector $\mathbf{x} = (x_1, x_2, \ldots, x_p)$.[1] The models work by recursively partitioning the input space into axis-aligned cuboids — imagine making a series of axis-aligned "slices" through the input space — and applying a simple conditional model to each region; the sequence of partitions on the input space can be represented as a binary tree (Fig. 3A).

Formally, a binary tree structure T consists of a hierarchy of nodes $\eta \in \mathsf{T}$. Nodes with children, or leaves, are referred to as *internal* nodes, while nodes without children are referred to as *leaf* nodes (Fig. 3A, right). The set of internal nodes for T is denoted I_{T}, and the set of leaves is denoted L_{T}. Each internal node $\eta \in I_{\mathsf{T}}$ has exactly two children, called the left child η_L and right child η_R. Each node is associated with a block $B(\eta) \subseteq \mathbb{R}^p$ of the input space as follows (cf. Fig. 3A, left): the root node is associated with the entire input space, while each further internal node splits its block into two parts by selecting a single dimension $\kappa(\eta) = \{1, \ldots, p\}$ and location $\tau(\eta)$ so that

$$
\begin{aligned}
B(\eta_L) &= B(\eta) \cap \{\mathbf{x} : x_{\kappa(\eta)} \leq \tau(\eta)\} \quad \text{and} \\
B(\eta_R) &= B(\eta) \cap \{\mathbf{x} : x_{\kappa(\eta)} > \tau(\eta)\}.
\end{aligned}
$$

The block of input space associated with a node η is determined by the ranges on each dimension j that 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*.

In addition to a decision tree $\mathcal{T}$ with K leaf nodes, a CART model has a parameter $\Theta = (\theta_1, \theta_2, \ldots, \theta_K)$, which 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

[1] In both experiments, $p = 3$, but we use the more general notation for presentation purposes.

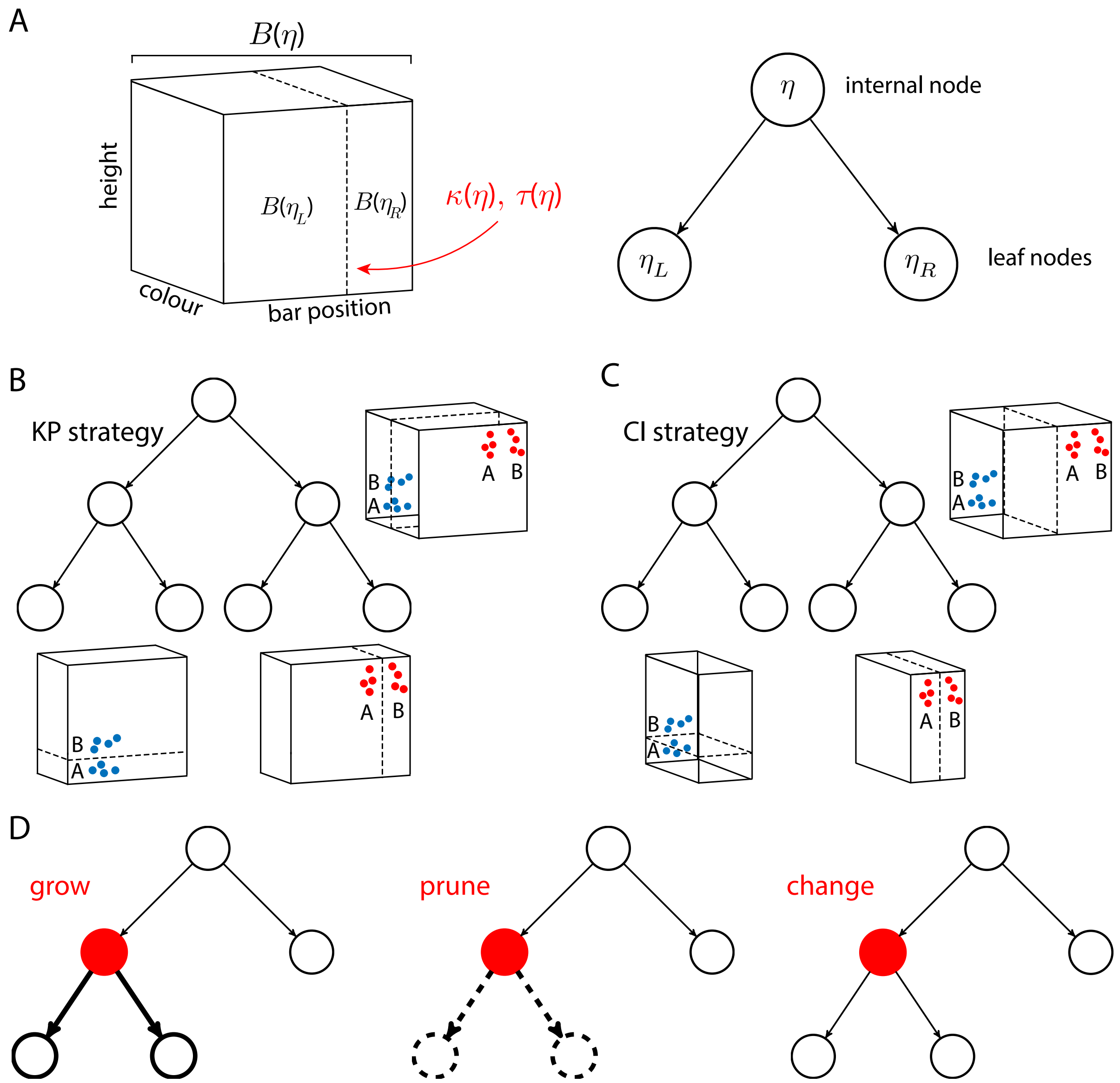

Figure 3: **Representing categories with a classification tree.**

(A) Consider the stimulus space of Sewell and Lewandowsky (2012), which comprises 3 stimulus dimensions (color, height, and bar position) and can be represented as a cube (left). A single partition of this space into 2 subspaces can be achieved by selecting one of the stimulus dimensions (here, bar position) and splitting the space on that dimension at a particular location. This partitioning can be represented by a simple binary tree (right). The root node η (which is also an "internal" node) is associated with the full stimulus space $B(\eta)$. In this example, node η is split on the dimension corresponding to bar position ($\kappa(\eta)$ = bar position) at a location $\tau(\eta)$. This partitions the input space into two blocks, $B(\eta_L)$ and $B(\eta_R)$, associated with the "leaf" nodes η_L and η_R. (B) Tree corresponding to a knowledge-partitioning (KP) strategy; the initial split is on the color dimension. (D) Tree corresponding to a context-insensitive (CI) strategy; the initial split is on the bar position dimension. (E) In the model, proposed modifications to trees may be of 3 types, each involving the initial random selection of a node (shaded red): *grow* selects a leaf node for expansion (i.e., splitting); *prune* selects an internal node and renders it a leaf node by deleting all nodes below it; and *change* selects an internal node and assigns it a new rule (i.e., a splitting dimension and location).

$f(y|\theta_k)$ 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 furthermore, 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. As we will make more precise below, for us, the parameter θ_k is the probability that a stimulus within the kth leaf node has category label A.

This provides a general framework for representing categories, but we require a more detailed specification for the experiments of interest. We now do this for the categorization task used by Sewell and Lewandowsky (2012), described above. The SHJ tasks employed in Lewandowsky (2011) are simpler and are straightforwardly modeled with only minor modifications.

In the Sewell–Lewandowsky task, the stimulus on each trial t comprised a 3-dimensional input $\mathbf{x}_t = (x_{t,1} = \text{bar position}_t \in \mathbb{R},\ x_{t,2} = \text{height}_t \in \mathbb{R}^+,\ \ x_{t,3} = \text{color}_t \in \{\text{blue} = 0, \text{red} = 1\})$.[2] On training trials, participants made a category prediction before observing the binary category label $y_t \in \{A, B\}$. The "ideal" knowledge-partitioning (KP) and context-insensitive (CI) strategies which participants were encouraged to learn and deploy can be naturally represented in tree form (Figs 3B,C).

In the Bayesian framework, we need to specify some prior beliefs about the state of nature. In the current case, the relevant prior beliefs concern category structure which, by modeling assumption, can be formalized as a prior distribution on decision trees. Such a prior can be imposed implicitly by specifying a stochastic process for generating such trees. Following Chipman et al. (1998), we set the prior probability of a node η in tree structure T being split into children nodes to be

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

where d_η denotes the depth of the node (the depth of the root node is zero), and $\alpha < 1$ and $\beta \geq 0$ are parameters controlling expected tree size. Under this specification,

[2] Of course, in reality, bar position and height were much more restricted than indicated — we mean only to emphasize by the use of $\mathbb{R}$ that these are continuous variables.

the probability p_{SPLIT} is a decreasing function of node depth, and decreases more steeply for large β (cf. Figure 3 of Chipman et al., 1998). In all simulations, we fix $\alpha = 0.95$ and $\beta = 1$, which gives a prior mean on the number of terminal nodes ≈ 3.7 (Chipman et al., 1998), but results are essentially identical for other reasonable parameterizations.

In addition to a prior on tree structure T achieved through a prior on a node's probability of splitting, we need to specify the prior probability of a node η splitting on each stimulus dimension $\kappa(\eta) = \{1, \ldots, p\}$ and location $\tau(\eta)$. We generally assume that the probability of splitting on each dimension is equal, i.e.,

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

Conditional on the choice of dimension, a split location is assumed to be drawn uniformly from the node's range on the relevant dimension:

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

However, consideration of the information given to participants at the outset of Sewell and Lewandowsky's experiment leads us to a slightly different prior for the root node η_0. In particular, in the experiment, participants were initially told that stimulus color (KP-first condition) or bar position (CI-first condition) reliably indicated whether height or bar position was diagnostic of stimulus category. We assume that this information is reflected in the prior probability of splitting the root node η_0 on a particular dimension. Thus, we introduce a "bias" parameter b to indicate that splits of the root node η_0 on one dimension should be regarded as much more likely than on the others. Letting j^* indicate the dimension highlighted by instruction, we can write this prior probability as

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

Setting $b < 1$, which would give nonzero probability to alternative splits at the root, might reflect incomplete confidence in the experimenter's instructions, for example.

In addition, participants were not only guided to a particular initial dimension — bar position or color — but effectively also to an initial split location. Thus, in the KP-first condition, attention was drawn to the color of the stimulus, while in the CI-first condition, participants were explicitly told that the relevant feature was whether the bar was to the left or right of centre. We therefore assume that split locations for the highlighted dimension at the root node are known. Note that the question of split location is actually irrelevant in the case of the (binary) color dimension since all split locations on $(0, 1)$ are equivalent in terms of the resulting partition. However, this dimension can be treated as continuous for ease of presentation and without consequence for modeling outcomes.

The preceding specifies a simple prior distribution on decision trees $p(\mathcal{T})$ that can be summarized as a process of deciding whether to split each node and, if so, selecting a splitting dimension and location. To complete the model specification, we also require a likelihood model $p(y_{1:t}|\mathbf{x}_{1:t}, \mathcal{T})$ that gives the conditional probabilities of stimulus labels given the tree structure. In this case, we simply 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,

$$\theta_k \overset{iid}{\sim} \text{Beta}(a_0, b_0). \tag{7}$$

Standard analytical simplification for this beta-binomial model yields the marginal likelihood

$$p(y_{1:t}|\mathcal{T}, \mathbf{x}_{1:t}) = \left(\frac{\Gamma(a_0 + b_0)}{\Gamma(a_0)\Gamma(b_0)}\right)^K \prod_{k=1}^{K} \frac{\Gamma(n_{kA}^t + a_0)\Gamma(n_{k\cdot}^t - n_{kA}^t + b_0)}{\Gamma(n_{k\cdot}^t + a_0 + b_0)}, \tag{8}$$

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 homogeneous labels (i.e., with labels that are either mostly A or mostly B). These are exactly the partitions that constitute "good" solutions to the categorization problem.

2.2 Inference

Given the model specified above, we assume that participants seek to represent the sequence of posterior distributions over possible trees $\{p(\mathcal{T}|\mathbf{x}_{1:t}, y_{1:t})\}_{t=1}^T$ as they successively predict and receive information about stimulus labels over trials.

Generally, a brute force procedure of enumerating all possible trees, a space which dramatically increases in size with t, is not a plausible model of how participants perform inference. Instead, we assume that people's beliefs are represented by a relatively small number of samples from these posterior distributions which can be updated over time. In other words, we model participants as performing *particle filtering* (Daw & Courville, 2008; Doucet et al., 2001; Sanborn, Griffiths, & Navarro, 2006).

As mentioned above, two aspects of the inference process which we now describe draw parallels with working memory. Firstly, similar to the idea that there is 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 samples/particles which correspond to particular tree structures — that can be entertained at a given time. Secondly, similar to the notion that working memory is *active* (Baddeley, 1992), involving the manipulation rather than merely passive storage of items, we assume that inference involves a continuing 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 a given trial t, a participant's beliefs are represented by a small set of L possible trees $\{\mathcal{T}^{(l)}\}_{l=1}^{L}$ with associated weights $\{w_t^{(l)}\}_{l=1}^{L}$ proportional to their posterior probability. 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):

$$\begin{aligned} w_{t+1}^{(l)} &\propto w_t^{(l)} \frac{p(\mathcal{T}^{(l)}|\mathbf{x}_{1:t+1}, y_{1:t+1})}{p(\mathcal{T}^{(l)}|\mathbf{x}_{1:t}, y_{1:t})} \\ &\propto w_t^{(l)} \frac{p(y_{1:t+1}|\mathcal{T}^{(l)}, \mathbf{x}_{1:t+1})}{p(y_{1:t}|\mathcal{T}^{(l)}, \mathbf{x}_{1:t})} \\ &= w_t^{(l)} p(y_{t+1}|\mathcal{T}^{(l)}, \mathbf{x}_{t+1}, y_{1:t}). \end{aligned} \tag{9}$$

As standard within particle filtering methods (Doucet et al., 2001), this reweighting process can be alternated with a *resampling* stage in which very unlikely trees, i.e., those with very low weights, are discarded to be 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).

Additionally, this resampled particle set can then be "rejuvenated" (Chopin, 2002; Gilks & Berzuini, 2001), 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 of interest, propose transformations of trees from a Markov chain transition kernel $q_{t+1}(\cdot|\mathcal{T}^{(l)})$ and accept or reject these proposals such that we retain the 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 (Fig. 3D):

1. GROW: Randomly select a leaf node, then draw a splitting dimension and location from the prior (Equations (3) and (4)). Not permitted if the split leads to an empty node (i.e., a partition with no assigned data points).
2. PRUNE: Randomly select an internal node, then turn it into a leaf node by deleting all nodes below it. Not permitted if the tree comprises only the root node.
3. CHANGE: Randomly select an internal node, then randomly reassign it a splitting dimension and location by a draw from the prior. Not permitted if the reassigned split is inconsistent with splits of nodes below the selected node.

This proposed tree $\mathcal{T}^{(l)*}$ is then accepted with probability

$$\alpha(\mathcal{T}^{(l)}, \mathcal{T}^{(l)*}) = \min\left\{1, \frac{p(\mathcal{T}^{(l)*}|\mathbf{x}_{1:t+1}, y_{1:t+1})/q_{t+1}(\mathcal{T}^{(l)*}|\mathcal{T}^{(l)})}{p(\mathcal{T}^{(l)}|\mathbf{x}_{1:t+1}, y_{1:t+1})/q_{t+1}(\mathcal{T}^{(l)}|\mathcal{T}^{(l)*})}\right\}, \tag{10}$$

as per the standard Metropolis-Hastings algorithm (Gelman, Carlin, Stern, & Rubin, 2004). This simple "resample-move" algorithm (Chopin, 2002; Gilks & Berzuini, 2001) is summarized in Algorithm 1.

Why might the number of samples/particles be expected to influence category learning? The basic intuition comes from viewing the category learning process as one of *search* (Fig. 4). In particular, "good" category structures are those that partition stimuli into regions with homogeneous labels (A or B), and these are the category structures that have high posterior probability. In the sample-based

Algorithm 1 Resample-Move.

Draw L sample trees from the prior $p(\mathcal{T})$ and initialize all weights to $w_0^{(l)} = 1/L$.
for each trial $t = 1, 2, \ldots$ **do**
  Update each particle's weight $w_t^{(l)} \propto w_{t-1}^{(l)} \times p(y_t|\mathcal{T}^{(l)}, \mathbf{x}_t, y_{1:t-1})$.
  Resample particles proportional to their updated weights $\{w_t^{(l)}\}_{l=1}^{L}$.
  Reset each of the (resampled) particle's weights to $w_t^{(l)} = 1/L$.
  for each particle $l = 1, 2, \ldots, L$ **do**
    Propose a new tree $\mathcal{T}^{(l)*} \sim q_t(\cdot|\mathcal{T}^{(l)})$.
    Accept the proposal with probability $\alpha(\mathcal{T}^{(l)}, \mathcal{T}^{(l)*})$ (as in Eq.(10)).
  end for
end for

inference procedure we consider, the population of particles will seek out regions of high posterior probability, and the rate at which these regions are found may plausibly depend on the number of particles.

So far, we have suggested a particle filtering scheme for representing a sequence of posterior distributions over category structures, where that structure is assumed to be specified by a classification tree. However, we have not yet addressed the issue of *strategy switching*. Thus, in the Sewell–Lewandowsky experiment, participants were able to immediately switch between different categorization strategies when instructed to do so, and in the absence of further training.

We model such switches as a simple *reweighting* operation on the set of trees. Take the specific example where a participant has initially been encouraged 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

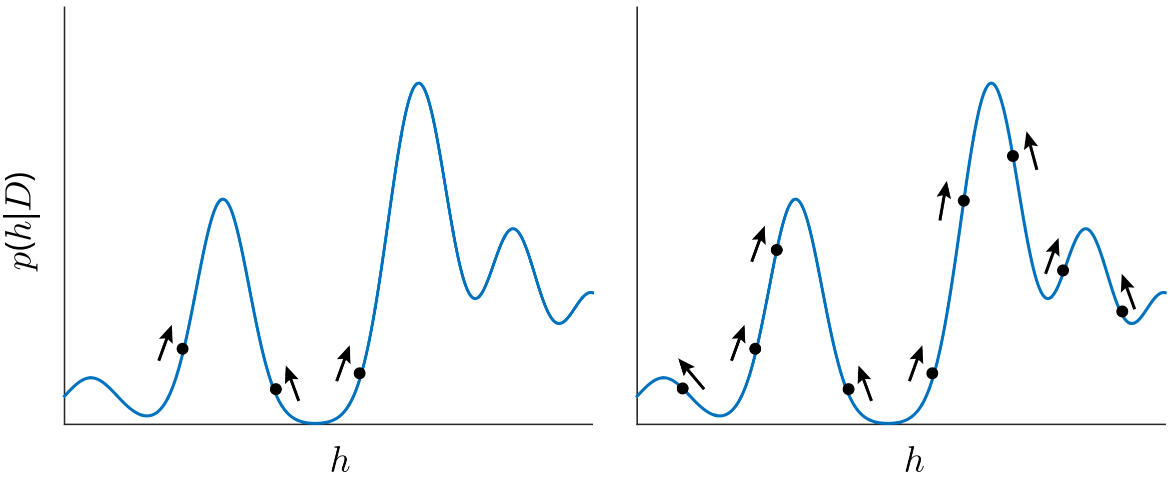

Figure 4: **Category learning as search.**

In the formulation here, category learning is conceptualized as a process of search for category structures $h \in \mathcal{H}$ that have a high posterior probability, $p(h|D)$, given both the prior distribution on category structures and the observed data, D. In the sample-based inference procedure considered, this search is enacted by a particle set (black circles) whose positions may be changed through the acceptance of proposed local changes to the corresponding category structure. Proposals that result in a category structure with higher posterior probability (arrows) will be accepted more often. With a larger number of particles (right), this search may be more efficient, in that high probability structures will be discovered more quickly.

$$w_t^{(l)+} \propto w_t^{(l)-} \frac{p_{KP}(\mathcal{T}^{(l)}|\mathbf{x}_{1:t}, y_{1:t})}{p_{CI}(\mathcal{T}^{(l)}|\mathbf{x}_{1:t}, y_{1:t})}, \tag{11}$$

which, under the specified model, becomes particularly simple:

$$w_t^{(l)+} \propto \begin{cases} w_t^{(l)-} \times (\frac{1-b}{2})/b & \text{if } \kappa(\eta_0) = \text{bar position,} \\ w_t^{(l)-} & \text{if } \kappa(\eta_0) = \text{height,} \\ w_t^{(l)-} \times b/(\frac{1-b}{2}) & \text{if } \kappa(\eta_0) = \text{color.} \end{cases} \tag{12}$$

To switch in the reverse direction — from the KP to CI strategy — the appropriate reweighting involves the ratio $p_{CI}(\mathcal{T}^{(l)}|\mathbf{x}_{1:t}, y_{1:t})/p_{KP}(\mathcal{T}^{(l)}|\mathbf{x}_{1:t}, y_{1:t})$, with the appropriate alterations made to Equation 12.

Again, why might a greater number of particles improve ability to switch between strategies? Consider the cartoon example in Figure 5A, depicting the posterior probability $P(h|D)$ of different possible category structures $h \in \mathcal{H}$ given a stimulus set D. In this example, two particular category structures, h_1 and h_2, are most probable, and equally so, and we can think of these as being two equally valid categorization strategies, as in the Sewell–Lewandowsky task. Again, this probability distribution will be represented by a set of particles with locations (i.e., particular category structures) drawn from this distribution, along with corresponding weights that are proportional to the posterior probabilities of those locations.

Now assume that the effect of an instruction to use a particular strategy is to increase the posterior probability of category structures that accord with that strategy, in this case those in the region of h_1 (Fig. 5B). Such a change in posterior distribution, driven by the different priors underlying the distinct strategies, is exactly what we assumed when suggesting that strategy-switching is mediated by a reweighting of particles (see above). Depending on the number of particles available, how well this collection of particles represents the true posterior distribution — especially in regions of lower probability — may differ. With a sufficiently large number of particles, at least some particles should be allocated to regions of lower probability, such as around h_2 (Fig. 5B, upper). However, with a decreasing number of particles, representation of the posterior distribution may become impoverished to the extent that such regions of low probability may not contain any particles at all (Fig. 5B, lower). In other words, the shift in "mental set" associated with a switch in categorization strategy is here implemented by a change in posterior distribution; the participant's immediate ability to represent this change is assumed to depend in some sense on the diversity of the current hypothesis set.

The possible relevance to knowledge restructuring is what these different degrees of approximation to the true posterior may entail when instructed to switch categorization strategy. Intuitively, if fewer resources have been devoted to representing alternative strategies in the first place, however unlikely, then it may be more difficult to entertain these alternatives when instructed to do so. In our particular formulation of the switching process, we considered a simple formulation in which the immediate effect of an instruction to switch strategy is that the locations of the particles remain the same, but the relative weightings of particles are updated according to the new posterior distribution (Fig. 5C). In particular, if there are particles located in the region of h_2, these will immediately be updated (Fig. 5C, upper), and the new categorization strategy can be immediately deployed. By contrast, if there are no particles located in the region of h_2, no up-weighting can occur and the alternative strategy is initially unavailable (Fig. 5C, lower).

2.3 Choice

We have so far described a process for performing inference (i.e., particle filtering) under an assumed generative model for the structure of categories (i.e., CART). What is still missing is a model of how participants finally generate a guess about a stimulus' category label before they receive feedback in the form of the true label. We consider two possible choice rules: one in which a participant chooses the category label with the highest probability ("maximum-probability rule"), and another in which a participant chooses a category label stochastically in accord with their probabilities ("probability-matching rule"). Since there is no explore-exploit dilemma in the categorization tasks we consider — full information about the correct label is always received, regardless of choice — participants should always select the label the think is most likely (i.e., maximum-probability rule). On the other hand, given that probability-matching behavior has sometimes been observed in this domain (e.g., Estes, Campbell, Hatsopoulos, & Hurwitz, 1989; Gluck & Bower, 1988), we considered it possible that participants also used this strategy, despite it being suboptimal in the tasks considered.

From the above, a sample-based approximation to the predictive probability that a stimulus $\mathbf{x}_{t+1}$ has label $y_{t+1} = A$ is given by

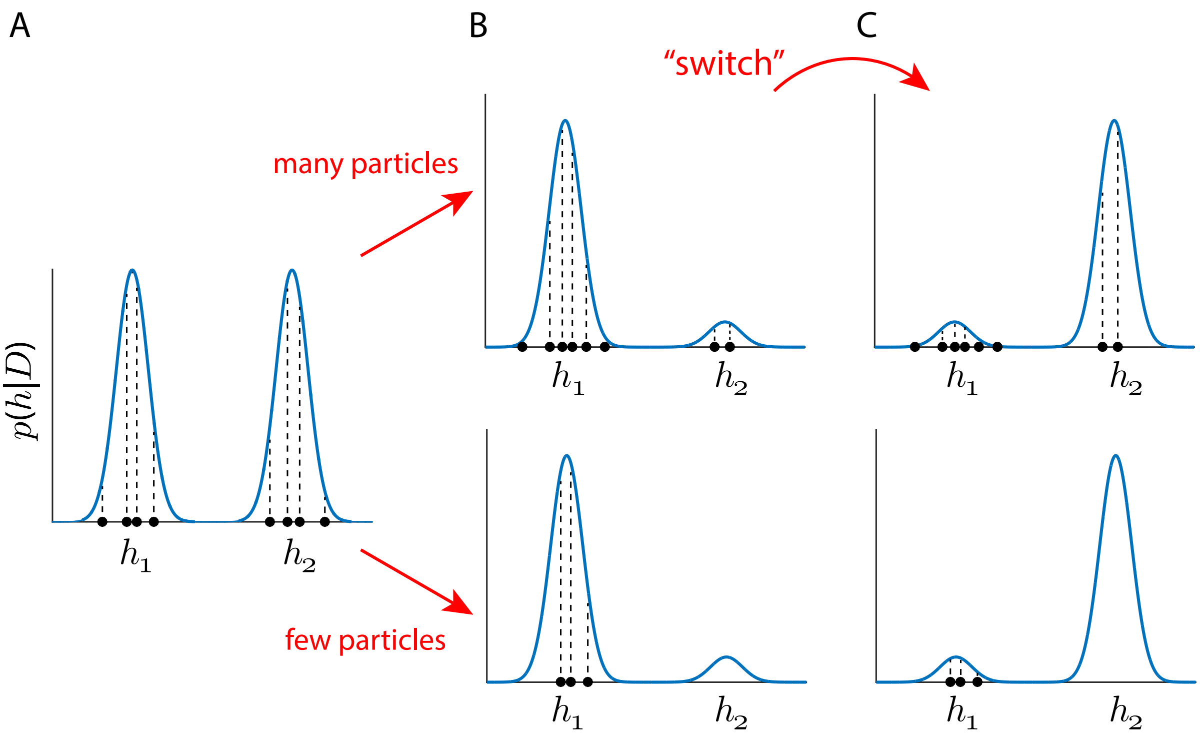

Figure 5: **Particle diversity and flexibility of behavior.**

Cartoon of how different numbers of particles affect the model's ability to switch between different categorization strategies. (A) Given the observed data D, comprising a set of stimuli and their category labels, there is a posterior distribution $P(h|D)$ over the set of possible category structures $h \in \mathcal{H}$. Here, two particular category structures h_1 and h_2 are equally probable, and can be considered as two equally valid categorization strategies. The distribution can be approximated by a set of particles, where each particle has a particular location (circles), corresponding to a category structure h, and a weight, which is proportional to the posterior probability (vertical, dashed lines). (B) The instruction to use a particular strategy is conceptualized as biasing the posterior distribution so that particular category structures are more probable, in this case category structures in the region of h_1. Whether regions of lower probability are represented in the approximation depends on the number of particles: if there are many particles, some are likely to be located in regions of lower probability, such as around h_2 (upper); if there are fewer particles, there may be no particles in this region (lower). (C) The instruction to switch strategy is conceptualized as leading to a change in the posterior distribution, and a corresponding change in the particle weights (upper); however, in the case of fewer particles, there may be no particles immediately available to represent the change in distribution (lower).

$$
\begin{aligned}
p(y_{t+1} = A|\mathbf{x}_{1:t+1}, y_{1:t}) &= \sum_{\mathcal{T}} p(y_{t+1} = A|\mathbf{x}_{1:t+1}, y_{1:t}, \mathcal{T}) p(\mathcal{T}|\mathbf{x}_{1:t}, y_{1:t}) \\
&\approx \frac{1}{L} \sum_{l=1}^{L} p(y_{t+1} = A|\mathbf{x}_{1:t+1}, y_{1:t}, \mathcal{T}^{(l)}) \\
&= \frac{1}{L} \sum_{l=1}^{L} \mathbb{E}_{\theta_k|\mathbf{x}_{1:t+1}, y_{1:t}, \mathcal{T}^{(l)}}[\theta_k], \qquad (13)
\end{aligned}
$$

noting that

$$
\begin{aligned}
p(y_{t+1} = A|\mathbf{x}_{1:t+1}, y_{1:t}, \mathcal{T}^{(l)}) &= \int p(y_{t+1} = A|\mathbf{x}_{1:t+1}, y_{1:t}, \theta_k, \mathcal{T}^{(l)}) p(\theta_k|\mathbf{x}_{1:t+1}, y_{1:t}, \mathcal{T}^{(l)}) d\theta_k \\
&= \int \theta_k \; p(\theta_k|\mathbf{x}_{1:t+1}, y_{1:t}, \mathcal{T}^{(l)}) d\theta_k \\
&= \mathbb{E}_{\theta_k|\mathbf{x}_{1:t+1}, y_{1:t}, \mathcal{T}^{(l)}}[\theta_k].
\end{aligned}
$$

Equation (13) simply says that an approximation to the predictive probability in this case 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 $\mathcal{T}^{(l)}$. For the leaf model used in the current case, the posterior mean is given by

$$
\mathbb{E}_{\theta_k|\mathbf{x}_{1:t+1}, y_{1:t}, \mathcal{T}^{(l)}}[\theta_k] = \frac{n^t_{kA} + a_0}{n^t_{k\cdot} + a_0 + b_0}, \qquad (14)
$$

where, again, n^t_{kA} and $n^t_{k\cdot}$ 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.

The deterministic maximum-probability rule would choose the category label with the highest predictive probability, but more generally we consider the ϵ-greedy form

$$
P_{t+1}(A) = (1 - \epsilon) \mathbb{1}_{\widetilde{p}(y_{t+1}=A) > \widetilde{p}(y_{t+1}=B)} + 0.5\epsilon, \qquad (15)
$$

where $P_{t+1}(A)$ is the probability of guessing category A on trial $(t+1)$, $\widetilde{p}(y_{t+1})$ is shorthand for the sample-based approximation given in Eq. (13), ϵ is the probability of guessing a category label according to the flip of a fair coin, and $\mathbb{1}_{\cdot}$ is the indicator function. In other words: choose the most probable label with probability $(1 - \epsilon)$, or with probability ϵ simply flip a coin. When $\epsilon = 0$, we recover the deterministic case.

The probability-matching rule takes the slightly different form

$$P_{t+1}(A) = (1-\epsilon)\widetilde{p}(y_{t+1} = A) + 0.5\epsilon, \tag{16}$$

so that the probability of guessing a category label is a linear combination of its predictive probability $\widetilde{p}(y_{t+1})$ (again, using shorthand for the probability given in Eq. (13)) and the guessing rate ϵ; strict probability-matching is obtained when $\epsilon = 0$.

Given that sample-based inference will itself tend to introduce stochasticity, we should comment on the addition of a guessing rate ϵ, which, for $\epsilon > 0$, will provide an additional source of variability. Briefly, our motivation was simply the (common) observation that model fit was improved by including this parameter; the behavior of participants tended to exhibit levels of variability beyond what our model would generate with $\epsilon = 0$, even with a single particle. As such, ϵ captures our ignorance about such variability, which may arise from sources distinct from sample-based inference (e.g., lapses in attention, lack of motivation, etc.). Of course, the price to be paid for this improvement in fit, as we will see below, is that apportioning responsibility for behavioral variability to different components of the model — inference versus choice — becomes all the more difficult.

2.4 Model-fitting and analysis

Models of varying degrees of complexity were fit to the data by finding the combination of the parameters of our category-learning model (described above) that maximized the likelihood of the observed sequence of category predictions. Models varied in the number of parameters to be fit, lying on a spectrum from the simplest case, which required that all participants be fit by a single set of parameters, to the most complex case, in which each participant was fit with a separate set of parameters. Formally, denoting an observed sequence of predictions over T trials by $c_{1:T}$ and the full set of parameters by $\Phi = \{L, b, \alpha, \beta, a_0, b_0, \epsilon\}$ (see Table 1), the general aim was to find the (free) parameters Φ that maximized the probability

$$p(c_{1:T}|\Phi, \mathbf{x}_{1:T}, y_{1:T-1}) = \prod_{t=1}^{T} p(c_t|\Phi, \mathbf{x}_{1:t}, y_{1:t-1}),$$

with the trial-by-trial probabilities extracted from Eq. (15) or Eq. (16), as appropriate.

Best-fit parameters for a given model were defined as those maximizing the average likelihood in a grid search. The grid was defined as follows: number of particles L logarithmically spaced on the interval $[1, 100]$, yielding thirty-four values; guessing rate uniformly-spaced $\epsilon \in (0, 0.02, 0.04, \ldots, 0.2)$; and shape $a_0 \in (0.01, 0.1, 0.5, 1)$. In the knowledge-restructuring case, we also included three possible values of bias, $b \in \{0.5, 0.75, 0.9\}$. The grid values were chosen to reflect our a priori assumptions about plausible parameter values. That is, we expected participants to be more plausibly modeled as instantiating relatively few particles (hence the logarithmic scale), and as expressing noise levels in the lower range (hence the upper limit of 0.2 on the guessing rate ϵ). The choice of comparatively finely-spaced ϵ values was motivated by the expectation that L and ϵ would at least partly trade off with each other, so effort was made to make the resolution of these parameters comparable in order to minimize the possibility of bias. In addition, we included the case where the number of particles was set to a much larger number ($L = 10,000$); this was to provide a comparison model that approximated the full posterior distribution much more closely than when the number of particles was more restricted.

Since the estimate of the likelihood was generally less reliable with fewer particles (due to greater variability in the algorithm's behavior), the number of simulation runs was chosen so that an "effective" number of particles would be constant, thereby facilitating a fair comparison between the fits of different numbers of particles. We set the effective number of particles to 1000, so that the number of simulation runs was determined by rounding to the nearest integer the result of $1000/L$ (i.e., the 1-particle case was run 1000 times, the 100-particle case was run 10 times, etc.).

As mentioned in Section 2.3, we additionally compared two different choice models. Modulo the effect of the guessing rate ϵ, either a stimulus was deterministically assigned to the most likely category (maximum-probability choice rule), or it was probabilistically assigned to a category in proportion to that category's predictive probability (probability-matching choice rule).

In evaluating the fit of different models, we used the Bayesian information criterion (BIC) to select the best-fitting model (Schwarz, 1978); that is, we chose the model M for which the quantity $\text{BIC} \equiv -\log(P(D|M, \hat{\Phi}_M)) + \frac{1}{2}k\log(n)$ was minimized, where $P(D|M, \hat{\Phi}_M)$ is the value of the likelihood function (see above) given the maximum likelihood estimate $\hat{\Phi}_M$ of the model parameters, k is the number of estimated parameters in the model, and n is the number of data points (i.e., the number of trials).

To assess relationships between best-fitting model parameters and participants' WMC scores, we used two methods. The simplest was simply to measure the correlation between these and determine whether the correlation was significantly different from zero. While this method is straightforward, the strength of the correlation can be reduced by both imprecision in estimating the best-fitting model parameters, as well as tradeoffs between parameters in fitting the data. While these issues cannot be entirely avoided, we developed a second measure to mitigate them which involved estimating a function that mapped WMC scores to a particular parameter of interest as part of the fitting procedure. To do so, we used BIC scores to compare slope-intercept models (in which the parameter of interest was a linear function of the individual WMC scores) against intercept-only models (in which the parameter was fixed across participants and thus did not depend on WMC scores). In cases in which there is a relationship between a parameter and WMC scores the slope-intercept model should perform better as the slope helps to capture that relationship. Our second measure helps address imprecision in estimating parameters because the parameters fit in the slope-intercept model are the best-fitting values that are consistent with a relationship with WMC, so if the individual parameters are somewhat imprecise but still consistent with a relationship to WMC, then the slope-intercept model would still perform best. Additionally, because of the concern of parameter tradeoffs in fitting the data, we allowed the other parameters in both the slope-intercept and intercept-only models to freely vary, so that these other parameters could trade off against the linear relationship between the parameter of interest and WMC in the way that allowed the best fit to the data. (ADAM HERE?) When comparing details of model fit with a participant's WMC score, we always used for the latter the average of that participant's scores over the battery of working memory tasks used in Lewandowsky (2011) and Sewell and Lewandowsky (2012).

Table 1: Model parameters. See text for details.

Parameters	
Fixed	Free
$\alpha = 0.95$	L: number of particles
$\beta = 1$	b: bias
$b_0 = a_0$	a_0: Beta shape parameter
	ϵ: random guessing rate

3 Results

3.1 Category learning

3.1.1 Simulations

Both Lewandowsky (2011) and Sewell and Lewandowsky (2012) found that working memory capacity (WMC) was positively correlated with category learning performance, such that participants with higher WMC tended to make fewer categorization errors. We hypothesized that a greater number of particles would have a similar effect because, 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 at local maxima, with a higher number of particles (Fig. 4). Here, we focus on simulating the classical SHJ tasks used by Lewandowsky (2011). Since we always found that the probability-matching choice rule yielded better fits to the data than the maximum-probability rule (see Table 2 below), the simulation results always reflect use of the former.

Figure 6A shows the overall average error rate for simulations as the number of particles is increased from 1 to 20 while keeping other parameter values fixed ($a_0 = 1, \epsilon = 0$); each data point represents 113 simulation runs, where each simulation run uses a stimulus sequence of 192 trials observed by one of the 113 participants in Lewandowsky (2011). For each problem type, increasing the number of particles does indeed lead to a decrease in the average proportion of errors, though the size of this effect is rather modest and quickly asymptotes (Note that the x-axis here indicates the number of particles — not block number, as in Fig. 1C).

Note that even without attempting to fit the parameters of the model, the ordering of error rates produced by the model for the different problem types conforms to the basic SHJ pattern of results — Type I easiest and Type VI hardest, with Types II–V clustered in between. Briefly, this is because of the so-called "automatic Occam's razor", which refers to a preference for simpler, or more parsimonious, hypotheses, and which arises naturally within the Bayesian framework (Goodman et al., 2008; MacKay, 2003).

It is also interesting to note that the difference in the simulated error rates between the Type II problem and, for example, Type IV increases — up to a point — as the number of particles grows. An advantage in learning Type II relative to Type IV problems has been reported in the experimental literature (e.g., Nosofsky, Gluck, Palmeri, McKinley, & Glauthier, 1994; Shepard et al., 1961), though this has not always been found, as in Lewandowsky (2011) (cf. Kurtz, Levering, Stanton, Romero, & Morris, 2013). Given our basic hypothesis that WMC reflects number of particles, this simulation result prompts the question of whether Type II advantage depends on WMC.

To investigate this further, we revisited the data of Lewandowsky (2011), splitting participants into low- and high-WMC groups according to a median split of WMC scores and entering blockwise error rates into a 2 (WMC: low, high) $\times$ 2 (Problem: II, IV) $\times$ 12 (Block: 1–12) repeated measures ANOVA. In addition to a main effect of WMC ($F(1, 111) = 7.65, p < .01$), we found a significant 3-way interaction between WMC*Problem*Block ($F(11, 1221) = 2.19, p = .01$). High-WMC participants performed significantly better in terms of proportion correct on Type II ($M = 0.92, SD = 0.10$) than on Type IV ($M = 0.88, SD = 0.11$; $t(56) = 2.19, p = .02$), while low-WMC participants did not perform significantly differently on these two problem types (Type II: $M = 0.85, SD = 0.14$; Type IV: $M = 0.85, SD = 0.14$, t(55)=0.06, n.s.). Learning curves are shown in the Appendix (Fig. S1). This result is consistent with our basic hypothesis, as we expect a Type II advantage to appear, or become stronger, with more particles (i.e., higher WMC).

The effect of a larger number of particles across problem types is further illustrated in Figure 6B, where we compare the overall error proportions for the extreme case of 1 particle vs. 100 particles. A larger number of particles reduces the error rate for each problem type, and in a manner that qualitatively resembles that observed in the experimental data when participants are grouped according to WMC

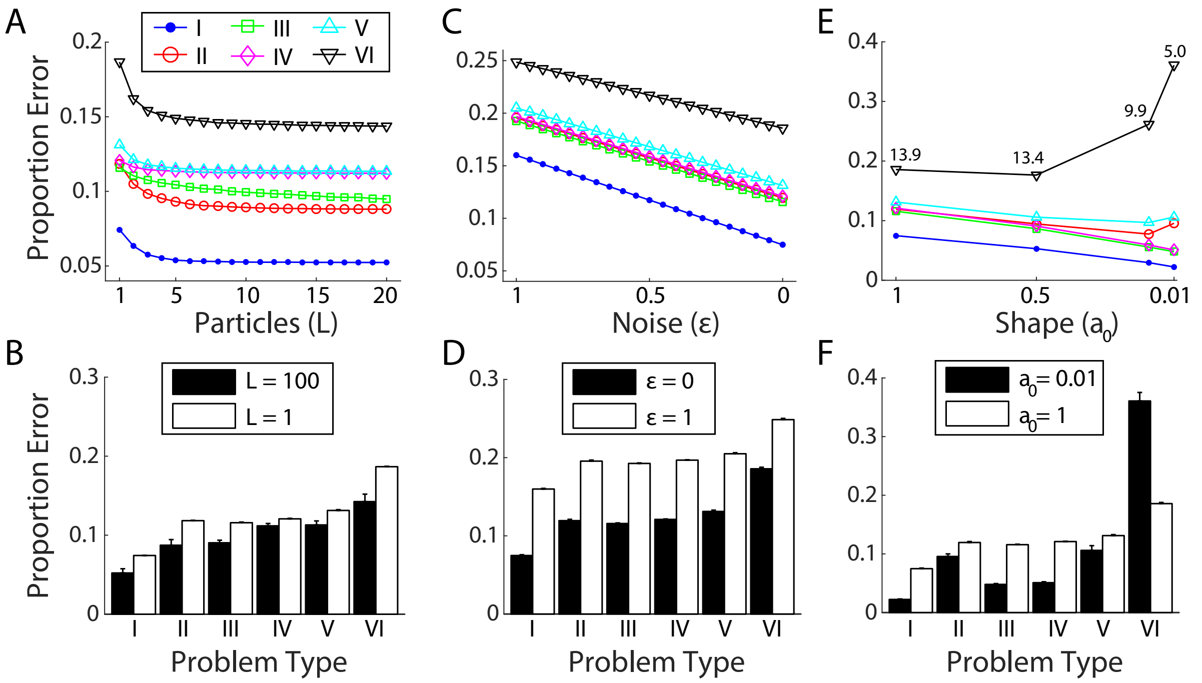

Figure 6: **Effect of model parameters on category learning in the SHJ problems.** Overall proportion of errors (i.e., averaged across blocks) for each problem type as each parameter is varied. (A;B) Number of particles L; other parameters fixed $a_0 = 1, \epsilon = 0$. (C;D) Noise ϵ; other parameters fixed $a_0 = 1, L = 1$. (E;F) Shape a_0; other parameters fixed $L = 1, \epsilon = 0$. Numbers in (E) for Problem VI indicate the average number of nodes in the final classification tree. Each data point represents an average of 113 simulation runs; error bars in lower panels indicate $+1SD$. Note the reversed x-axes for Figures C and E.

score (cf. Fig. 1D). Indeed, a rank-ordering of problem types by the extent to which performance is better for higher WMC/particles revealed a significant positive correlation (Spearman's rank-order correlation $r_s(4) = .94, p < .05$). In other words, the problem types that show greatest difference between high- and low-WMC participants tend also to be those where an increased number of particles also makes the most difference (from greatest to smallest advantage, the experimental pattern follows the order VI,II,III,I,V,IV; our simulations follow the order VI,II,III,V,I,IV).

We also examined the effect on performance of varying the other free parameters (i.e., guessing rate ϵ and shape a_0). Figure 6C shows, unsurprisingly, that the proportion of errors decreases linearly as ϵ decreases. Since this rate of decrease is essentially uniform across problem types, the amount of improvement in each problem type is roughly the same (Fig. 6D). A simple inverse association between WMC and guessing rate therefore fails to capture differential effects of WMC on performance of the problem types (Spearman's rank-order correlation $r_s(4) = -.37, p = .50$, n.s.).

Decreasing a_0 generally leads to a lower error rate — recall that a higher a_0 entails a higher tolerance for "mixed" categories (i.e., instances of both A and B; cf. Section 2.1) — with Problem Type VI proving a notable exception (Figs 6E,F). Briefly, what happens in the latter case is that the model becomes increasingly intolerant of the intermediate tree manipulations necessary to reach a more satisfactory solution; this can be observed, for example, in the decreasing average number of nodes in the final tree as a_0 is decreased (Fig. 6E). A simple inverse association between WMC and shape therefore does a worse job compared to particles at capturing differential effects of WMC on performance of the different problem types (Spearman's rank-order correlation, $r_s(4) = -.83, p = .06$, n.s.).

3.1.2 Model-fitting

In fitting model parameters, we compared a number of possibilities ranging from the case where all participants were constrained to share a single set of parameters (Model 1; least flexible) to the case where each participant was free to have a different set of parameters for each problem type (Model 7; most flexible). Models of intermediate complexity included the cases where two of the three free parameters $\{L, a_0, \epsilon\}$ were fixed across subjects, while the other free parameter was allowed to vary between subjects (Model 3: vary L; Model 4: vary ϵ; Model 5:

vary a_0). Since the probability-matching choice rule always fit better than the maximum-probability choice rule (compare numbers without and with parentheses, respectively, in Table 2), we restrict our attention to the results of the former case.

In terms of Bayesian information criterion (BIC), the model in which the number of particles L and shape a_0 were fixed across subjects, while guessing rate ϵ was allowed to vary between participants (i.e., Model 4), was found to fit best (Table 2). By contrast, fit for the model in which the number of particles L was allowed to vary between participants, with a_0 and ϵ fixed (Model 3), was comparatively poor. The comparison model — with a large number (10,000) of particles — resulted in poorer fit both when we allowed shape a_0 and noise ϵ to vary between subjects (NLL= 41624, BIC= 42955), and when only noise was allowed to vary between subjects (NLL= 42469, BIC= 43141). The probability-matching choice rule yielded a better fit than the maximum-probability choice rule in all models.

Table 2: Model comparison, SHJ tasks. We compared model fit under different constraints of the number of parameters. Model 1: single set of parameters $\{L, a_0, \epsilon\}$ fixed across all participants and problem types. Model 2: single set of parameters per problem type, fixed across participants. Model 3: different number of particles L per participant, fixed across problems, with $\{a_0, \epsilon\}$ fixed across participants. Model 4: different guessing rate ϵ per participant, fixed across problems, with $\{L, a_0\}$ fixed across participants. Model 5: different shape a_0 per participant, fixed across problems, with $\{L, \epsilon\}$ fixed across participants. Model 6: single set of parameters per participant, fixed across problem types. Model 7: single set of parameters per participant-problem type. Values for the maximum-probability choice rule are shown in parentheses. NLL = negative log likelihood; BIC = Bayesian information criterion.

Model	# free parameters	NLL	BIC
1	3	40801 (45411)	40819 (45429)
2	18	39669 (44131)	39775 (44237)
3	115	40664 (45251)	41342 (45928)
4	115	38569 (41119)	**39246** (41796)
5	115	39336 (45171)	40013 (45848)
6	339	37649 (40827)	39645 (42824)
7	2034	34284 (34915)	46261 (46892)

The upper panels of Figure 7 display the blockwise average learning curves resulting from respectively simulating from Model 3 (vary particles), Model 4 (vary noise), and Model 5 (vary shape) using the best-fit parameters for each. All models produce similar behavior on average, recapitulating the ordering of problem types in the experimental data and the qualitative character of the learning curves (cf. Fig. 1C).

The lower panels of Figure 7 plot each participant's average WMC against their best-fit parameters for each model. When only the number of particles L was allowed to vary between participants (Model 3), WMC and L were positively correlated ($r = .30, p < .01$), which was consistent with our initial hypothesis. Best-fit values of the other parameters (fixed across subjects) were $a_0 = 0.5$ and $\epsilon = 0.04$. However, this model was not found to fit the data best. Furthermore, assuming that a participant's best-fit number of particles is a (linear) function of WMC, we found that an intercept-only model (NLL= 37823, BIC= 39160), with best-fit intercept set to $L = 1$, fit these data better than a slope-intercept model relating these variables (NLL= 37823, BIC= 39166), allowing the other parameters (a_0 and ϵ) to vary freely in both cases.

The best-fitting model allowed the guessing rate ϵ to vary between participants, while fixing the remaining parameters across participants (Model 4). In this case, ϵ was found to be negatively correlated with our aggregate WMC measure ($r = -.30, p < .01$), suggesting that high-WMC participants tended to be less "noisy" in their choices. Best-fit values of the remaining parameters, fixed across subjects, were $a_0 = 0.5$ and $L = 1$. Furthermore, we found that a slope-intercept model (NLL= 38740, BIC= 40083) fit these data better than an intercept-only model (NLL= 39188, BIC= 40525). The best-fit slope was $\beta_1 = -0.6$, supporting an inverse relationship between WMC and variability in behavior.

In the model in which only shape a_0 was allowed to vary between participants (Model 5), WMC and a_0 were also significantly negatively correlated ($r = -.35, p < .01$). Best-fit values of the other parameters (fixed across subjects) were $L = 1$ and $\epsilon = 0.03$. Here, a slope-intercept model (NLL= 38329, BIC= 39671) fit better than an intercept-only model (NLL= 39147, BIC= 40484), with best-fit slope $\beta_1 = -0.7$.

While model comparison did not support a model allowing a unique set of parameters (L, a_0, ϵ) for each participant (Model 6), this was the second-best fitting model and it was of interest to examine how the free parameters might trade off against each other. The only significant correlations found were a negative correlation between best-fit shape and number of particles ($r = -.28, p < .01$), and a positive correlation between shape and guess rate ($r = .19, p < .05$).

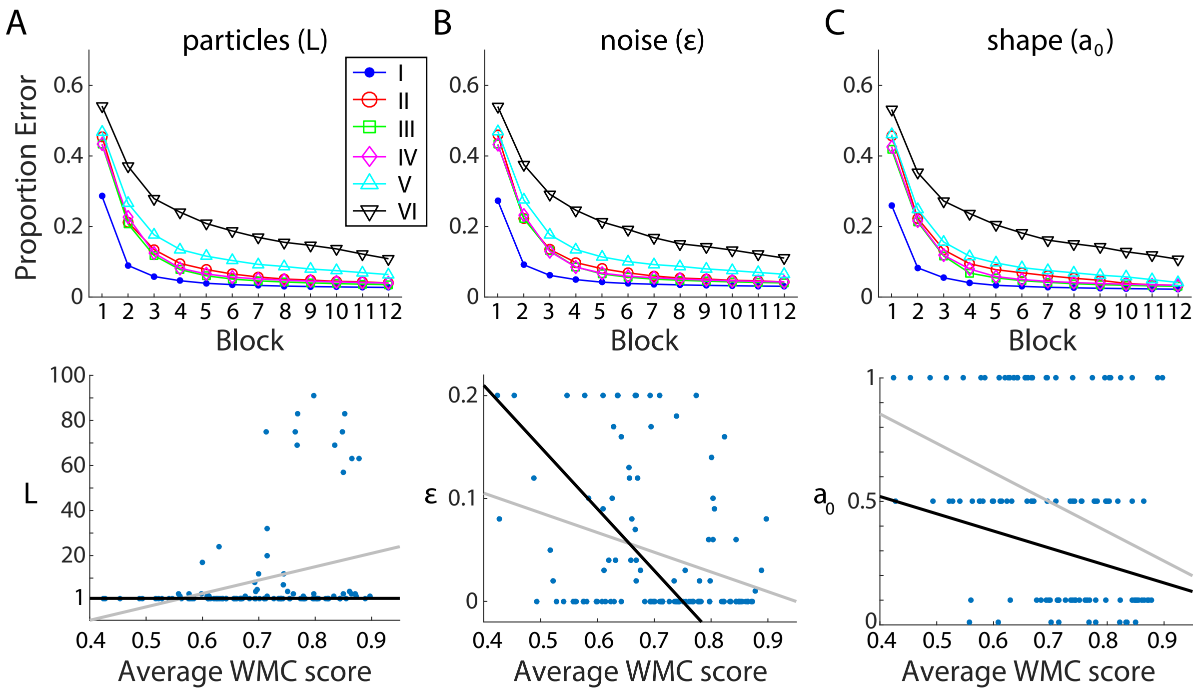

Figure 7: **SHJ model-fitting results.**

Average behavior of best-fit parameters (upper) and scatterplot of average working memory capacity (WMC) against best-fit parameters (lower) for (A) Model 3 (vary particles L, fix a_0, ϵ); (B) Model 4 (vary noise ϵ, fix a_0, L); (C) Model 5 (vary shape a_0, fix ϵ, L). Lower panels: line of least squares (grey); regression line for best-fit intercept-slope/intercept-only model (black).

3.2 Knowledge restructuring

3.2.1 Simulations

Sewell and Lewandowsky (2012) found a positive association between WMC and knowledge restructuring, as measured by an individual's ability to switch between different categorization strategies. We hypothesized that a greater number of particles would also give rise to this effect since a greater diversity of hypotheses could be represented, leading to an enhanced ability to flexibly shift between representations with changes in task demands (Fig. 5). As for our simulations of model performance in the SHJ task, we report results in which the probability-matching choice rule is used, since it always yielded better fits to the data (see Table 3 below).

Figure 8 shows the effect of varying the number of particles L on the degree of context sensitivity (CS) change between test sessions (all other model parameters were kept fixed: $b = 0.9, a_0 = 1, \epsilon = 0$). Averaging over simulation runs, we observe that the extent of CS-change increases gradually with the number of particles, regardless of whether the model initially learns a context-insensitive (CI; Fig. 8A, left) or knowledge-partitioning (KP; Fig. 8A, right) strategy. This graded effect predominantly reflects the effect of averaging over CS changes which are of "all or none" character — switch or no switch — where the probability of switching increases with the number of particles (Fig. 8B). Interestingly, the empirical CS-change scores also display some degree of bimodality, though this is not to the same extent, nor does the degree of bimodality notably differ between high- and low-WMC participants (see Appendix, Fig. S2A). Analogous to the increase in successful switching that we observe in simulations, it is also the case that participants' probability of making a successful switch (defined as for simulations, i.e., a change in CS between test sessions that crosses 0.5) increases on average with higher WMC (see Appendix, Fig. S2B).

3.2.2 Model-fitting

As for the SHJ tasks, we fit models of different complexity to the data. In the knowledge restructuring task, we found that allowing each participant to have their own set of parameters fit the data better in terms of BIC than simpler, less flexible models (Table 3). As in the SHJ case, the comparison model, with $L = 10,000$ particles, always resulted in a poorer fit, and the probability-matching choice rule yielded a better fit than the maximum-probability choice rule in all models (Table 3).

Figures 9A–C show aspects of behavior of the best model using the best-fitting parameters for each participant. Figure 9A shows that the average changes in context sensitivity between transfer tests of the model qualitatively resemble the empirical data (cf. Fig 1C). Similarly, Figure 9B confirms that the model generalizes its categorization behavior to test stimuli in a strategy-dependent manner that closely resembles the "ideal" response profiles (cf. Fig 1B), and the generalization patterns of participants (see Figure 7 in Sewell & Lewandowsky, 2012). A median

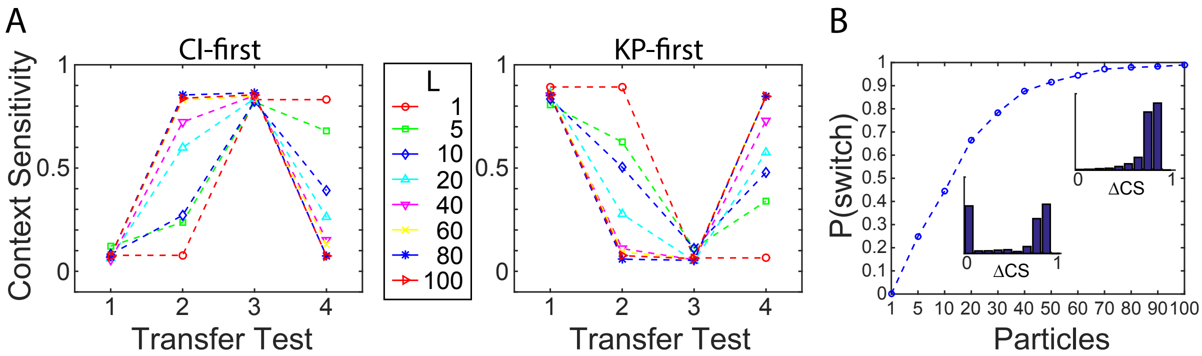

Figure 8: **A greater number of particles leads to improved strategy switching.** (A) In both the context-sensitive (CI)-first (left) and knowledge-partitioning (KP)-first (right) condition, increasing the number of particles L leads to a greater change in context sensitivity (CS) score on average when prompted to change strategy. Average CS scores from 1500 simulation runs per condition. (B) The effect arises because the probability of successfully switching between strategies, $P(\text{switch})$, increases with more particles. A successful switch is here defined as a change in context sensitivity between test sessions, ΔCS, which "crosses" a score of 0.5. Lower inset: with fewer particles ($L = 20$), it will frequently occur that the model completely fails to switch (i.e., $\Delta CS = 0$), as visible from the distribution over change values ΔCS. Upper inset: with more particles ($L = 100$), such failures are very unlikely. Switch probabilities and distributions are from 3000 simulation runs. All other parameter values were fixed: $b = 0.9, a_0 = 1, \epsilon = 0$.

Table 3: Model comparison, knowledge restructuring task. We compared model fit under different constraints of the number of parameters. Model 1: single set of parameters $\{L, b, a_0, \epsilon\}$ fixed across all participants. Model 2: different number of particles L per participant, with $\{b, a_0, \epsilon\}$ fixed across participants. Model 3: different bias b per participant, with $\{L, a_0, \epsilon\}$ fixed across participants. Model 4: different shape a_0 per participant, with $\{L, b, \epsilon\}$ fixed across participants. Model 5: different noise ϵ per participant, with $\{L, b, a_0\}$ fixed across participants. Model 6: single set of parameters per participant. Values for the maximum-probability choice rule are shown in parentheses. NLL = negative log likelihood; BIC = Bayesian information criterion.

Model	**# free parameters**	**NLL**	**BIC**
1	4	24535 (25051)	24557 (25074)
2	103	23366 (23833)	23950 (24416)
3	103	23837 (24641)	24421 (25224)
4	103	23826 (24421)	24409 (25005)
5	103	22915 (23924)	23499 (24507)
6	400	21165 (21709)	**23431** (23975)

split of best-fit parameters according to number of particles also leads to a pattern of changes in context sensitivity that resembles that of participants when grouped by WMC scores: simulations using greater numbers of particles show larger CS changes (compare Figs 9C and 1D).

When we examined the relationships between individuals' average WMC scores and best-fitting parameters (Fig. 9D), we found that there was no significant correlation between WMC and best-fit number of particles ($r(98) = .09, p = .38$, n.s.). However, this correlation analysis is affected by tradeoffs between parameters, which would likely act to reduce the correlation coefficient. A more robust analysis comes from comparing a slope-intercept model, in which WMC is assumed to be linearly related to the number of particles, to an intercept-only model, where the number of particles is assumed to be fixed and independent of WMC; the other parameters are free to vary, as this analysis is less affected by parameter tradeoffs (the same analysis was applied to the SHJ results, above). A slope-intercept model (NLL=22045, BIC=23756) was found to fit this relationship better than an intercept-only model

(NLL=22078, BIC=23786), but the best-fitting slope was small, suggesting a rather weak effect ($\beta_1 = 9$; black line in Fig. 9D).

As in the SHJ tasks, we found a significant negative correlation between WMC and guessing rate ϵ ($r(98) = -.26, p < .01$; Fig. 9E). A slope-intercept model with slope $\beta_1 = -0.3$ (NLL=22153, BIC=23863) fit better than an intercept-only model (NLL=22326, BIC=24032).

We found no significant correlation between WMC and shape a_0 ($r(98) = -.03, p = .75$, n.s.; Fig. 9F). A slope-intercept model (NLL=21473, BIC=23184), with slope $\beta_1 = -0.4$, was found to fit this relationship better than an intercept-only model (NLL=21496, BIC=23201).

Finally, there was a significant correlation between WMC and bias b ($r(98) = .27, p < .01$; Fig. 9G). A slope-intercept model (NLL=21459, BIC=23170), with slope $\beta_1 = 0.9$, was found to fit this relationship better than an intercept-only model (NLL=21489, BIC=23194).

As in the SHJ case, it was of interest to examine how these best-fitting parameters potentially traded off against each other. We found a negative correlation between the number of particles L and the guessing rate ϵ ($r(98) = -.40, p < .01$). We also found that bias b was positively correlated with number of particles L ($r(98) = .44, p < .01$), and negatively correlated with the guessing rate ($r(98) = -.59, p < .01$). Other correlations were not significant.

4 Discussion

Dealing with the world's many uncertainties in a consistent and principled manner presents a formidable computational challenge. That humans routinely do so despite necessarily finite cognitive resources is an impressive feat. Algorithms for approximate Bayesian inference provide one natural source of ideas for how this may be achieved. Thus, one suggestion has been that people may approximate Bayesian computations by representing and manipulating a set of samples drawn according to the relevant probability distributions (Sanborn & Chater, 2016), i.e., by implementing Monte Carlo inference (Gelfand & Smith, 1990; Gordon et al., 1993). Such methods admit a spectrum of degrees of approximation, from essentially ideal performance given plentiful computational resources (e.g., a large number of samples), to much coarser approximations when such resources are scarce (e.g., few sam-

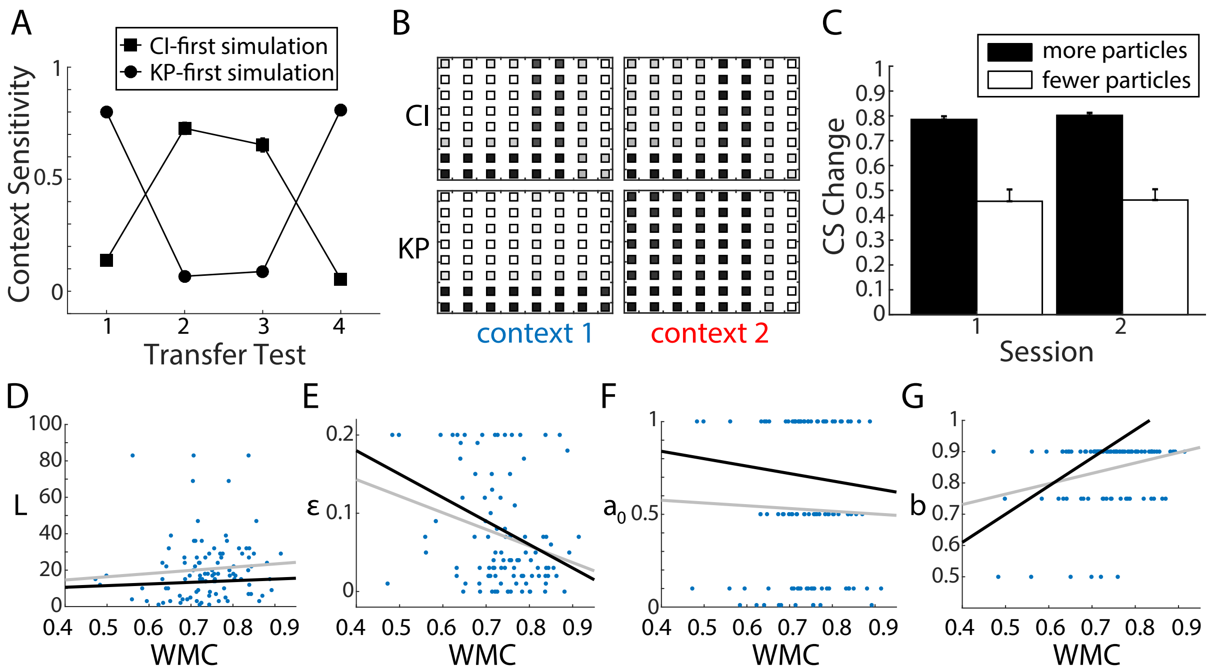

Figure 9: **Knowledge-restructuring model-fitting results.**

(A) Simulated average changes in context sensitivity ($\pm 1SE$; obscured by markers) for CI-first (squares) and KP-first (circles) conditions. (B) Simulated average probabilities of categorizing a test stimulus as an instance of category A in the CI-first (upper) and KP-first (lower) conditions in the first transfer test. Darker shading indicates a higher probability. (C) Simulated average change ($+1SE$) in context sensitivity (CS) given a median split of the best-fit parameters for all participants ranked in terms of numbers of particles. (D–G) Scatter plots of average WMC scores vs. best-fitting parameters, with lines of least squares (grey) and regression lines for best-fit intercept-slope models (black): (D) number of particles L; (E) guessing rate ϵ; (F) shape a_0; (G) bias b.

ples). In the current work, we considered constraints on working memory capacity (WMC) in the context of probabilistic inference, asking whether parallels may be drawn between WMC limitations and resource-constrained, approximate Bayesian inference. In particular, we hypothesized that variations in task performance that correlate with WMC would be captured by assuming that WMC directly reflects the number of samples, or "particles", available to perform inference.

To test this, we focused on experiments that suggest a positive association between WMC and two apparently disparate aspects of categorization: (a) the ease with which novel categories are learned (Lewandowsky, 2011); and (b) the ability to switch between different categorization strategies (Sewell & Lewandowsky, 2012). We saw that such categorization tasks can be considered probabilistic inference problems in which individuals seek to infer the most probable category structure(s) given their prior assumptions and what they subsequently observe. We assumed that individuals approximate inference by representing and manipulating in working memory a relatively small number of hypotheses (samples/particles) about the possible underlying category structures. The number of hypotheses an individual is able to entertain at a given time was assumed to depend on their WMC.

Support for our principal hypothesis was decidedly mixed. On the one hand, we provided a "proof of concept" that increasing the number of particles in our algorithm could both hasten category learning and improve switching performance, at least on average. In simulations of the SHJ problem types, we also found that the degree to which increasing the number of particles differentially improved performance in the problem types was closely matched to the manner in which higher WMC is differentially associated with improved performance in these problem types; this pattern was not matched as well by changes in other parameters. Furthermore, when the model was fit to individuals' behavior in the knowledge-restructuring experiment of Sewell and Lewandowsky (2012), linear regression between WMC and number of particles suggested a positive — albeit rather weak — relationship. On the other hand, when the model was fit to individuals' performance in the SHJ tasks (Lewandowsky, 2011), model comparison did not support a variant in which the number of particles changes as a function of WMC. Rather, the winning model favored setting the number of particles to one, and captured individual variation in performance through the guessing-rate, or "noise", parameter. Possible reasons for this mixed picture are discussed next.

4.1 Limitations

One possible reason for our failure to find a relationship in the SHJ case is the relatively weak effect of varying the number of particles on learning rate. That is, although we demonstrated that increasing the number of particles could hasten category learning in these problems, the effect was subtle — the improvement in learning was relatively small, and generally reached asymptote at a comparatively small number of particles (cf. Fig. 6A).

A second contributory factor to the mixed picture — though we believe our regression analyses mitigate this — is likely the substantial correlations between model parameters. In formulating the category learning model, we included the possibility that various of its parameters — not just number of particles — would show variation when fit to behavior. As our results made clear, the parameters showed substantial correlations, making the job of disentangling their effects more difficult. In the SHJ case, the best-fitting model had a separate noise/guessing-rate ϵ for each participant, with other parameters fixed across participants; both correlation and regression indicated a negative relationship between WMC and ϵ, suggesting that higher WMC participants were less "noisy" in their choices. When we allowed all parameters to vary between individuals (the second best fitting model), we saw that ϵ and shape a_0 were significantly positively correlated, as one might anticipate — recall that a higher a_0 leads to more tolerance of category structures with mixed labels, which would lead to more errors. Furthermore, a_0 was negatively correlated with the number of particles L, which is also expected, since an increasing number of particles tends to reduce the number of errors. However, in this case we found no significant correlation between particles L and ϵ, which we might have expected given their tendencies to decrease and increase errors, respectively. In the knowledge-restructuring experiment, the best model allowed all parameters to vary between participants, and here we did indeed find that L and ϵ were negatively correlated. The fact that the bias parameter b was respectively positively and negatively correlated with L and ϵ also makes sense, since a lower bias would tend to generate more classification errors. Although we haven't demonstrated it here, we expect that b and L would also interact in strategy-switching, in addition to the category learning phase, since a higher bias may require a larger number of particles to ensure that switching occurs reliably.

Clearly, our model has multiple sources of variability, or "noise", that trade off in ways that unfortunately make it difficult to draw strong conclusions from our model-fitting results. Of course, this is not an uncommon scenario, and the challenge of apportioning behavioral variability to different possible sources is a general one. In relation to the latter, it is interesting that we found in all cases that a model with a relatively low number of particles (i.e., in the range of 0–100) fit better than a model with a large number (10,000) of particles. The purpose of the latter was to approximate exact inference more closely, thereby providing a comparison in which noise in the inference process (as opposed to other sources of noise, such as in the choice process) was minimized. The finding therefore lends some support to the idea that inference noise plays a role in accounting for variability in participants' behavior (e.g., Wyart & Koechlin, 2016). However, we would caution against drawing too strong a conclusion here — though we did not see much evidence of floor/ceiling effects in our fitting results, a more decisive comparison would involve an expanded range or parameters (e.g., considering ϵ on the full range $[0, 1]$).

We also found that a probability-matching choice rule always fit the data better than a maximum-probability choice rule. Probability-matching behavior has previously been reported in the categorization literature (Estes et al., 1989; Gluck & Bower, 1988), so this result is perhaps not surprising, even if it is strictly suboptimal in this setting. However, in the context of our model, it is difficult to assign responsibility for probability matching to the inference or choice mechanism, since probability matching could conceivably arise from either separately, or both together. Indeed, since an inference mechanism based on sampling, such as the one we have described, would naturally tend to probability matching under a limited number of samples (cf. Vul, Goodman, Griffiths, & Tenenbaum, 2014), the addition of a probability-matching choice process makes disentangling these separate sources of variability particularly challenging.

Why, then, did we include noise in the choice process at all? Here, the motivation was simply to improve model fit — at least some participants' behavior was more variable than even a severely resource-constrained particle filter (i.e., a single particle). The guessing rate primarily represented our ignorance about variability arising from sources distinct from sample-based inference (e.g., attentional lapses). It is interesting that in both experiments the best-fitting model had guessing rates that were negatively correlated with WMC. This is consistent with observations that an increase in WMC load is accompanied with what look like random responses (e.g., Adam, Vogel, & Awh, 2017; Zhang & Luck, 2008), since we would then expect individuals with lower WMC (as well as higher WMC individuals under increased memory load) to guess more often *because* their capacity is lower. However, given that our starting point was the operationalization of WMC in terms of number of particles L, the fact that we only found a negative correlation between L and ϵ in one of the two experiments is only partially consistent with this.

Another limitation concerns our model's inability to handle particular attentional phenomena. In our presentation of the results of Sewell and Lewandowsky (2012), we briefly highlighted that high-WMC participants displayed significantly greater changes in context sensitivity (CS) in Session 1 but not in Session 2, where low-WMC participants appeared to "catch up"; in our model, by contrast, there is no reason to expect the amount of CS change to vary for different sessions (compare Figs 2D and 9C). At least some of the asymmetry in the human data is likely to arise due to attentional factors that are not included in our model. In particular, Sewell and Lewandowsky (2012) noted that participants in the KP-first condition generally found it easier to switch in Session 1 than participants in the CI-first condition (compare the magnitude of CS change between transfer tests 1 and 2 for the two conditions in Fig. 2C). In their interpretation of this, Sewell and Lewandowsky appealed to dimensional relevance shifts, and specifically to evidence that it is easier to attend to a previously relevant dimension than to a previously ignored dimension (e.g., Kruschke, 1996). Thus, in the CI-first condition, participants initially learn to *ignore* one of the dimensions (color, or "context"), since it is not involved in the CI strategy; this means that it will be harder to switch to the KP strategy, since the latter requires attending to the previously ignored dimension. In the KP-first condition, by contrast, participants initially attend to all stimulus dimensions, so do not have to learn to attend to a previously ignored dimension. A modest augmentation of the current model with a prior that incorporates the assumption that only a subset of stimulus dimensions may be relevant to classification (i.e., a sparsity assumption) would conceivably address the asymmetry between KP-first and CI-first conditions, but presumably not the fact that low-WMC participants appear to catch up with high-WMC participants in Session 2.

4.2 WMC and search efficiency

Category learning in the model proceeded quicker with more samples due to what we might refer to as increased *search efficiency*. Category structures that represent “good” solutions to the category learning problem were those with high posterior probability, and so the inference problem could be thought of in terms of search for such category structures in the hypothesis space (cf. Fig. 4). The more resources available to search this space — the more samples — then the more likely it is that (a) a good solution is discovered at all, and (b) a good solution is discovered quickly. In our simulations, we found that the marginal benefit to learning rate of increasing the number of samples was rapidly diminishing (cf. Fig. 6A), though we expect the point at which this occurs to depend on both the complexity of the problem and the precise details of the inference algorithm.

In more psychological terms, the implication is that the greater the number of hypotheses that one can entertain and manipulate within working memory, the more likely that one will quickly discover good solutions. The idea of exploring a space of solutions is of course well-established in psychology, where problem-solving has long been cast in such terms (Newell & Simon, 1972; Simon, 1983). There, however, the search problem is conventionally defined in terms of finding a path from an initial state to an explicit goal state while minimizing the path cost. This is rather different from search in the present case, which is best described in terms of simple stochastic hill-climbing in the absence of an explicit goal representation or, indeed, a path cost. Nevertheless, the idea that one may have greater or lesser resources with which to search may be fruitful in considering the link between WMC and problem solving more generally (Hambrick & Engle, 2003). Other stochastic sampling algorithms that have been applied to finding action sequences in large search spaces, such as Monte Carlo tree search (Coulom, 2006; Gelly & Silver, 2011), may also be a natural source of inspiration in such settings.

Interestingly, we also found some evidence in the data of Lewandowsky (2011) that WMC may interact with extent of Type II advantage in the SHJ tasks. Additional analysis of the experimental data was prompted by the observation in our model that the degree of Type II advantage appeared to be modulated by the number of particles (cf. Fig. 6A). This is consistent with the recent suggestion, in the context of category learning in older adults, that Type II advantage is modulated by WMC (Rabi & Minda, 2016), though our present model does not speak to the observation that relative performance on Type II and Type IV problems may sometimes reverse (e.g., in older adults — see Badham, Sanborn, & Maylor, 2017; Rabi & Minda, 2016).

4.3 WMC and flexibility

A greater number of samples led not only to faster category learning, but also to an improved ability to switch between categorization strategies. This was due to an increase in what we might call *representational adequacy*. That is, with a greater number of samples, the full posterior distribution over category structures was more accurately represented, encompassing category structures that were assigned lower probability. By representing this greater plurality of category structures, the model could easily express alternative hypotheses when instructed to switch strategy, as operationalized by a reweighting of the current sample/hypothesis set (cf. Fig. 5).

Again, in more psychological terms, the obvious interpretation is that the greater one's ability to entertain a variety of hypotheses, the more flexible one will be. There is evidence that individuals with higher WMC are better at solving so-called "insight" problems, and this may be because such problems are exactly those that require keeping in mind several different possibilities (Gilhooly & Fioratou, 2009; Murray & Byrne, 2005). Indeed, insight problems typically involve inducing task representations in participants which are not conducive to solving the problem, and so require "restructuring" of the initial task representation (Ohlsson, 1992; Weisberg, 1995).

4.4 Related work

The current study is framed by a number of related strands of research. Most pertinently, Lewandowsky and colleagues have themselves previously addressed the experimental results discussed here, though using a rather different modeling approach. Lewandowsky (2011) found that individual differences in category learning performance could be captured by varying only the learning rate of a particular category learning model (ALCOVE; Kruschke, 1992), but did not establish a rationale for why WMC should be related to this parameter. Sewell and Lewandowsky (2011) found that while a "single-module" model such as ALCOVE failed to capture the general ability to fluidly switch between categorization strategies, a "multiple-module" model, such as ATRIUM (Erickson & Kruschke, 1998) — which is able to learn more than one mapping between stimuli and category labels — could do so. However, a mechanism by which such recoordination could take place was not proposed, nor was the issue of why WMC should be related to this ability addressed. In the current work, we provide a model able to capture both experimental results using a single mechanism (i.e., variation in the number of samples), propose a simple mechanism for how recoordination could occur (i.e., importance reweighting), and offer rationales for why WMC may be associated with faster learning (search efficiency) and flexibility (representational adequacy).

Levy et al. (2008) directly anticipate our suggestion that the number of samples used for inference may be equated with WMC in their exploration of "garden path" effects in sentence processing. Briefly, garden path sentences (e.g., "The old man the boat.") are grammatical sentences that people typically fail to parse correctly, at least at first, due to early parts of the sentence tending to promote one (incorrect) interpretation over another. This initial interpretation then leads to subsequent difficulties of comprehension. Levy et al. suggested that difficulties in parsing such sentences correctly — and in particular, the probability of successfully re-parsing the sentence in light of disambiguating information arriving late in the sentence — may be explained by constraints on the resources (i.e., number of samples) available for incremental parsing. They showed that a particle filter model for performing online inference could reproduce these phenomena, with variation of the number of particles altering the strength of the effects. In particular, as the number of particles decreased, the probability that the correct interpretation of the sentence was not represented in the ensemble — leading to parse failure — increased. This is exactly analogous to the mechanism suggested to account for category switching performance in the current work: a lower number of particles makes it less likely that the alternative strategy is represented, meaning that the probability of being able to switch is decreased. However, the current work goes beyond Levy et al. both in expanding the range of phenomena explained (i.e., both switching *and* learning effects) and in actually measuring correlations between best-fit model parameters and WMC scores.

The HyGene model of Dougherty and colleagues (Dougherty, Thomas, & Lange, 2010; Thomas, Dougherty, Sprenger, & Harbison, 2008) is also closely related to the current work. HyGene provides a general framework for diagnostic inference, incorporating processes by which hypotheses may be generated and maintained in working memory. This includes the assumption that working memory processes constrain the number of hypotheses that one can actively maintain, though to the best of our knowledge this framework has not been applied to the domain of category learning that we consider here.

Finally, a number of previous models have considered the category learning problem in Bayesian terms (Anderson, 1991; Goodman et al., 2008; Sanborn et al., 2006, 2010). Notably, both Sanborn et al. (2006, 2010) and Goodman et al. (2008), despite considering rather different category representations, considered sample-based inference to be a particularly good candidate as a psychological mechanism for approximating Bayesian inference. For example, Sanborn et al. found that they were able to replicate a wide range of category learning effects by fitting relatively few samples to experimental data, though individual differences were not explored in that work. Our use of a representation based on classification and regression trees (CART) was primarily driven by pragmatic reasons, in particular what seemed most natural for the tasks concerned, rather than a theoretical commitment to a particular way of representing categories. We expect similar results to be obtained with alternative category representations, such as those used in the Rational Model of Categorization (Anderson, 1990; Sanborn et al., 2010) and Rational Rules (Goodman et al., 2008).

4.5 Future directions

The current work suggests a number of avenues for future investigation. One is to further explore the relative contributions of different components of the inference process. For example, search in the model effectively relies on two processes. The first is resampling, in which particles with lower probability are discarded and particles with higher probability are copied. Intuitively, this should be beneficial for learning since search is then focused on more "promising" (i.e., high probability) regions of hypothesis space. The second process is the proposal and acceptance/rejection of new hypotheses via MCMC moves, leading to local hill-climbing in probability space. A more detailed understanding of how these processes interact, and how they may relate to various psychological phenomena, would be of interest.

Similarly, one could consider alternative conceptualizations of the process by which participants switch between different categorization strategies. We implemented strategy-switching as a simple reweighting operation on particles according to a new target distribution. One consequence of this modeling choice is that it may be impossible — at least for the initial time step — to switch to a new strategy if the corresponding region of hypothesis space is not represented. Though we found some hints of bimodality in the human data, the prospect of such "catastrophic failure" may not seem entirely realistic, so one could imagine exploring modifications such as allowing additional propose-accept/reject steps during this phase.

More generally, it is likely that there is a trade-off between the sophistication of the processes by which individual hypotheses are maintained and manipulated, and the number of such hypotheses that one would need to support. In other words, one could presumably replace a larger number of relatively "dumb" particles/hypotheses with a smaller number of comparatively "smart" particles/hypotheses. Indeed, it has recently been suggested that, at least when considering more global hypotheses about the world where the hypothesis space becomes particularly complex, only one hypothesis would plausibly be represented (Bramley, Dayan, Griffiths, & Lagnado, 2017). How to negotiate this spectrum of possibilities is a pressing challenge.

Clearly, future work should also test whether the current modeling approach can be applied to other category learning tasks and beyond. As mentioned in the Introduction, it has been suggested that category learning tasks which can be solved with relatively simple, verbalizable rules ("rule-based" tasks) are especially reliant on working memory, while tasks with solutions that generally defy description in terms of simple rules ("information-integration" tasks) are not (Ashby & Maddox, 2005, 2011; Ashby & O'Brien, 2005). However, recent results suggest rather that working memory is equally involved in these different types of task (Craig & Lewandowsky, 2012; Lewandowsky et al., 2012). An obvious first step would therefore be to assess whether the current approach can be applied to tasks that are more clearly of the information-integration type.

A broader challenge for rational process models is to find constraints that will help determine more precisely the algorithms that underpin cognition. In the present work, we followed previous suggestions that inference algorithms based on Monte Carlo sampling are promising, but this only weakly constrains the variety of models under consideration. Determining the signatures of particular modeling choices within this larger class, and how these may succeed or fail in matching features of human cognition and behavior, is a substantial task for future research.

Acknowledgments

This work was supported by an EPSRC doctoral training award (KL), ESRC grant ES/K004948/1 (AS), EPSRC grant EP/I032622/1 (DL), and by the Royal Society (SL).

References

Adam, K., Vogel, E., & Awh, E. (2017). Clear evidence for item limits in working memory. *Cognitive Psychology*, *97*, 79–97.

Anderson, J. (1990). *The Adaptive Character of Thought*. Erlbaum.

Anderson, J. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*, 409–429.

Ashby, F., Alfonso-Reese, L., Turken, A., & Waldron, E. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*, 442–481.

Ashby, F., & Maddox, W. (2005). Human Category Learning. *Annual Review of Psychology*, *56*, 149–178.

Ashby, F., & Maddox, W. (2011). Human category learning 2.0. *Annals of the New York Academy of Sciences*, *1224*, 137–161.

Ashby, F., & O'Brien, J. (2005). Category learning and multiple memory systems. *Trends in Cognitive Sciences*, *9*(2), 83–89.

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

Baddeley, A., & Hitch, G. (1974). Working memory. *Psychology of Learning and Motivation*, *8*, 47–89.

Baddeley, A., Thompson, N., & Buchanan, M. (1975). Word length and the structure of short term memory. *Journal of Verbal Learning and Verbal Behavior*, *14*, 575–589.

Badham, S., Sanborn, A., & Maylor, E. (2017). Deficits in category learning in older adults: Rule-based versus clustering accounts. *Psychology and Aging*, *32*(5), 473–488.

Bernardo, J., & Smith, A. (1994). *Bayesian Theory*. Chichester: Wiley.

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.

Brown, S., & Steyvers, M. (2009). Detecting and predicting changes. *Cognitive Psychology*, *58*(58), 49–67.

Bruner, J., Goodnow, J., & Austin, G. (1956). *A Study of Thinking*. Wiley.

Busemeyer, J. R. (1985). Decision making under uncertainty: a comparison of simple scalability, fixed-sample, and sequential-sampling models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(3), 538–564.

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.

Conway, A., Jarrold, C., Kane, M., Miyake, A., & Towse, J. (Eds.). (2007). *Variation in Working Memory*. Oxford University Press.

Conway, A., Kane, M., & Engle, R. (2003). Working memory capacity and its relation to general intelligence. *Trends in Cognitive Sciences*, *7*(12), 547–552.

Coulom, R. (2006). Efficient selectivity and backup operators in Monte-Carlo tree search. In H. van der Herik, P. Ciancarini, & H. Donkers (Eds.), *5th International Conference on Computer and Games* (pp. 72–83). Springer.

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.

Craig, S., & Lewandowsky, S. (2012). Whichever way you choose to categorize, working memory helps you learn. *The Quarterly Journal of Experimental Psychology*, *65*(3), 439–464.

Daneman, M., & Carpenter, P. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*(4), 450–466.

Daw, N., & Courville, A. (2008). The rat as particle filter. In J. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in Neural Infor-*

mation Processing Systems 20 (pp. 369–376). Cambridge, MA: MIT Press.

Daw, N., Courville, A., & Dayan, P. (2008). Semi-rational models of conditioning: The case of trial order. In N. Chater & M. Oaksford (Eds.), *The Probabilistic Mind: Prospects for Bayesian Cognitive Science* (pp. 427–448). New York: OUP.

Doucet, A., de Freitas, N., & Gordon, N. (Eds.). (2001). *Sequential Monte Carlo Methods in Practice*. Springer.

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.

Doya, K., Ishii, S., Pouget, A., & Rao, R. (Eds.). (2007). *Bayesian Brain: Probabilistic Approaches to Neural Coding*. MIT Press.

Erickson, M., & Kruschke, J. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, *127*(2), 107–140.

Estes, W. K. (1950). Toward a statistical theory of learning. *Psychological Review*, *57*(2), 94–107.

Estes, W. K., Campbell, J. A., Hatsopoulos, N., & Hurwitz, J. B. (1989). Base-rate effects in category learning: A comparison of parallel network and memory storage-retrieval models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(4), 556–571.

Gelfand, A., & Smith, A. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*(410), 398–409.

Gelly, S., & Silver, D. (2011). Monte-Carlo tree search and rapid action value estimation in computer Go. *Artificial Intelligence*, *175*, 1856–1875.

Gelman, A., Carlin, J., Stern, H., & Rubin, D. (2004). *Bayesian Data Analysis* (2nd ed.). Chapman & Hall/CRC.

Gigerenzer, G., & Goldstein, D. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*(4), 650–669.

Gilhooly, K., & Fioratou, E. (2009). Executive functions in insight versus noninsight problem solving: An individual differences approach. *Thinking & Reasoning*, *15*(4), 355–376.

Gilks, W., & Berzuini, C. (2001). Following a moving target — Monte Carlo inference for dynamic Bayesian models. *Journal of the Royal Statistical Society Series B*, *63*, 127–146.

Gluck, M., & Bower, G. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General*, *117*(3), 227–247.

Goodman, N., Tenenbaum, J., Feldman, J., & Griffiths, T. (2008). A Rational Analysis of Rule-Based Concept Learning. *Cognitive Science*, *32*(1), 108–154.

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., & Tenenbaum, J. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*, 334–384.

Griffiths, T., & Tenenbaum, J. (2006). Optimal predictions in everyday cognition. *Psychological Science*, *17*, 180–226.

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.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, *58*(9), 697–720.

Körding, K., & Wolpert, D. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*(6971), 244–247.

Kruschke, J. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*, 22–44.

Kruschke, J. (1996). Dimensional Relevance Shifts in Category Learning. *Connection Science*, *8*(2), 225–247.

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.

Lewandowsky, S., Griffiths, T., & Kalish, M. (2009). The wisdom of individuals: Exploring people's knowledge of everyday events using iterated learning. *Cognitive Science*, *33*, 969–998.

Lewandowsky, S., Yang, L.-X., Newell, B., & Kalish, M. (2012). Working memory does not dissociate between different perceptual categorization tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(4), 881–904.

Love, B., Medin, D., & Gureckis, T. (2004). Sustain: a network model of category learning. *Psychological Review*, *111*(2), 309–332.

Ma, W., Husain, M., & Bays, P. (2014). Changing concepts of working memory. *Nature Neuroscience*, *17*(3), 347–356.

MacKay, D. C. (2003). *Information Theory, Inference, and Learning Algorithms.* Cambridge University Press.

Medin, D., & Schaffer, M. (1978). Context Theory of Classification Learning. *Psychological Review*, *85*(3), 207–238.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81–97.

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.

Nosofsky, R. (1986). Attention, Similarity, and the Identification-Categorisation Relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57.

Nosofsky, R., Gluck, M., Palmeri, T., McKinley, S., & Glauthier, P. (1994). Comparing models of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory and Cognition*, *22*(3), 352–369.

Nosofsky, R., Palmeri, T., & McKinley, S. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*(1), 53–79.

Oberauer, K., Farrell, S., Jarrold, C., & Lewandowsky, S. (2016). What limits working memory capacity? *Psychological Bulletin*, *142*(7), 758–799.

Oberauer, K., & Kliegl, R. (2006). A formal model of capacity limits in working memory. *Journal of Memory and Language*, *55*, 601–626.

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.

Posner, M., & Keele, S. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*, 353–363.

Rabi, R., & Minda, J. (2016). Category learning in older adulthood: A study of the Shepard, Hovland, and Jenkins (1961) tasks. *Psychology and Aging*, *31*, 185–197.

Restle, F. (1962). The selection of strategies in cue learning. *Psychological Review*, *69*(4), 329–343.

Robert, C., & Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer.

Rosch, E. (1973). Natural categories. *Cognitive Psychology*, *4*, 328–350.

Sanborn, A., & Chater, N. (2016). Bayesian brains without probabilities.

Trends in Cognitive Sciences, *20*(12), 883–893.

Sanborn, A., Griffiths, T., & Navarro, D. (2006). A more rational model of categorization. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society.*

Sanborn, A., Navarro, D., & Griffiths, T. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, *117*(4), 1144–1167.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Sewell, D., & Lewandowsky, S. (2011). Restructuring partitioned knowledge: The role of recoordination in category learning. *Cognitive Psychology*, *62*, 81–122.

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.

Simon, H. (1982). *Models of Bounded Rationality, Volume 1.* Cambridge, MA: MIT Press.

Simon, H. (1983). Search and reasoning in problem solving. *Artificial Intelligence*, *21*, 7–29.

Stewart, N., Chater, N., & Brown, G. (2006). Decision by sampling. *Cognitive Psychology*, *53*, 1–26.

Suchow, J. W., Bourgin, D. D., & Griffiths, T. L. (2017). Evolution in mind: Evolutionary dynamics, cognitive processes, and bayesian inference. *Trends in Cognitive Sciences*, *21*(7), 522–530.

Suchow, J. W., Fougnie, D., Brady, T. F., & Alvarez, G. A. (2014). Terms of the debate on the format and structure of visual memory. *Attention, Perception, & Psychophysics*, *76*(7), 2071–2079.

Tenenbaum, J., Kemp, C., Griffiths, T., & Goodman, N. (2011). How to Grow a Mind: Statistics, Structure, and Abstraction. *Science*, *331*, 1279–1285.

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

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science*, *185*(4157), 1124–1131.

Vul, E., Goodman, N., Griffiths, T., & Tenenbaum, J. (2014). One and done? optimal decisions from very few samples. *Cognitive Science*, 1–39.

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

Weisberg, R. (1995). Prolegomena to theories of insight in problem solving: A taxonomy of problems. In R. Sternberg & J. Davidson (Eds.), *The nature of insight.* Cambridge, MA: MIT Press.

Wyart, V., & Koechlin, E. (2016). Choice variability and suboptimality in uncertain environments. *Current Opinion in Behavioral Sciences*, *11*, 109–115.

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences*, *10*, 301–308.

Zhang, W., & Luck, S. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*, 233.

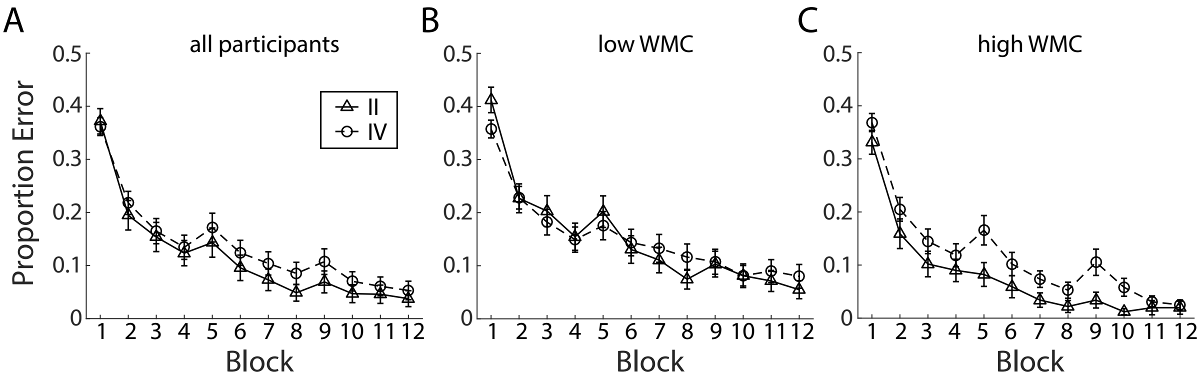

Figure S1: **Interaction of Type II advantage with working memory capacity.** Average learning curves ($\pm 1SE$) for Types II and IV in the experiment of Lewandowsky (2011) for (A) all participants; (B) participants with lower-median WMC scores; and (C) participants with upper-median WMC scores. Only the high WMC participants show a Type II advantage (see main text for statistics).

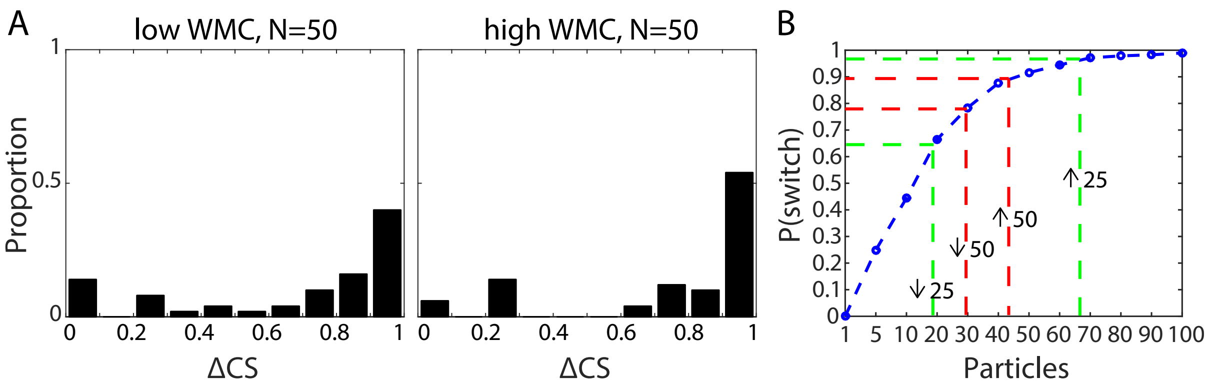

Figure S2: **Participants' context-sensitivity changes and switch probabilities.** (A) Distribution of (absolute) changes in context sensitivity (ΔCS; pooling over both test sessions) for low (left) and high (right) WMC participants. (B) The probability of making a successful switch of categorization strategy goes up with increasing WMC. Mean probabilities of a successful switch were respectively .64, .77, .89, and .96 for participants with WMC scores in the lower quartile ($\downarrow$ 25), lower median ($\downarrow$ 50), upper median ($\uparrow$ 50), and upper quartile ($\uparrow$ 25) of the experimental population. These scores are superimposed, for comparison, on the probability of switching as a function of the number of particles obtained from simulations (cf. Fig. 8B). As in the simulation results, a successful switch is defined as a change in context sensitivity between test sessions, ΔCS, that "crosses" a score of 0.5.