ARTICLE IN PRESS

Journal of Mathematical Psychology ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

Using Bayes factors to test the predictions of models: A case study in visual working memory

Arthur Kary, Robert Taylor, Chris Donkin *

School of Psychology, University of New South Wales, Kensington NSW 2052, Australia

HIGHLIGHTS

- Whether slots or resource models best account for the capacity of visual working memory is debated.
- High-threshold (slots) and signal detection (resource) models contrasted using ROC curves.
- Informative priors are used, so that the slots and resource models make sensible predictions.
- Bayes factors used to quantify the match between the models' predictions and the observed data.

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Visual working memory
Bayes factors
Signal detection
Short-term memory

ABSTRACT

A critical property of Bayesian model selection, via Bayes factors, is that they test the predictions made by models. Such predictions are a joint function of the likelihood of the model, and the prior distributions placed on the parameters of the model. Prior distributions that are informed by previous data lead to more constrained predictions, and result in Bayes factors that test more specific versions of the models under question. We present a case study applying two models of visual working memory to a series of experiments. We outline a process by which the posterior distributions from previous experiments are used to define and update prior distributions for each subsequent experiment. For each experiment we obtain Bayes factors that test the predictions that these models make, and update our beliefs about the relative likelihood of each model.

© 2015 Elsevier Inc. All rights reserved.

For any two models under consideration, the Bayes factor is the ratio of the marginal likelihoods for each model (Jeffreys, 1961). In short, it tells us which model is more likely to have generated a given set of data, and by how much. The marginal likelihood of observed data, D, under a given model, M, that has parameters, θ, is given by

$$p(D|M) = \int p(D|\theta, M)p(\theta|M)d\theta$$

where $p(D|\theta, M)$ is the probability of the observed data set given a particular set of parameters in a model (usually known as the likelihood function), and $p(\theta|M)$ is the prior probability of that set of parameters. The marginal likelihood, therefore, tells us how likely a model is to have generated a given set of data, based on the average of all possible sets of parameters, weighted by the prior probability of each set of parameters. Stated another way, it tells us how likely it is that a model, specified prior to data collection, would have generated the observed data.

Bayes factors have a number of advantages over traditional methods of model selection. For one, Bayes factors take into account the functional form complexity of models. More complex models make wider ranges of predictions. While being able to predict a large range of data may appear useful, the result is that any given observed data set is not likely *a priori*. Therefore, we prefer models that are restricted in the range of predictions that they make, provided that the observed data are consistent with those predictions. Since the Bayes factor is based on the marginal likelihood of each model, it naturally incorporates this notion of parsimony by rewarding models that have compressed prediction spaces and are also consistent with observed data.

Crucially, the Bayes factor is also governed by the prior distributions placed on parameters. In the calculation of the marginal likelihood, the likelihood of observed parameters are weighted by their prior probability. Since these prior probabilities are defined before data is observed, the marginal likelihood, and hence the Bayes factor, gives a measure of the models ability to *predict* observed data. The incorporation of priors into the Bayes factor also means that

* Corresponding author.
E-mail address: christopher.donkin@gmail.com (C. Donkin).

http://dx.doi.org/10.1016/j.jmp.2015.07.002
0022-2496/© 2015 Elsevier Inc. All rights reserved.

Please cite this article in press as: Kary, A., et al., Using Bayes factors to test the predictions of models: A case study in visual working memory. Journal of Mathematical Psychology (2015), http://dx.doi.org/10.1016/j.jmp.2015.07.002

the prior distributions are a critical part of the definition of any model (Vanpaemel, 2011; Vanpaemel & Lee, 2012).

In what follows, we use Bayes factors to evaluate the relative likelihood of two models of visual working memory. We will test standard versions of slots and resource models. The data we use are from three change-detection experiments reported in Donkin, Tran, and Nosofsky (2015). We will calculate Bayes factors for each experiment. The Bayes factors from each experiment will allow us to update our beliefs about the relative plausibility of the slots and resource models.

Our aim is to provide a case study in using informed prior distributions to test the predictions of models. As such, we begin by sacrificing some data to construct informed priors. We will then show that the model with informed priors make more constrained, and arguably more sensible predictions than when we use vague priors. The posterior distributions for each experiment will then be used as prior distributions for subsequent experiments, as per the expression "yesterday's posterior is today's prior for tomorrow's data". We begin with a brief overview of visual working memory.

1. Visual working memory

There is an ongoing controversy in the study of visual working memory (VWM) regarding its capacity and how items are stored in memory (Luck & Vogel, 2013; Ma, Husain, & Bays, 2014). The discrete slots view argues that a limited number of items are stored with high fidelity (Luck & Vogel, 1997), while continuous resource theorists argue that memory can be distributed flexibly across items, with no set limit on the number of items that can be held in memory (Wilken & Ma, 2004).

The change detection task is a classic experimental design still used to investigate the capacity of VWM (Cowan, 2001; Pashler, 1988; Phillips, 1974). In these tasks, participants observe an array of items which they have been instructed to remember, and are then presented with a test array shortly after. The test array is either the same as the study array, or has changed. The change detection task was used by Luck and Vogel (1997) to advocate for a slots-based theory of VWM after demonstrating that performance decreased once more than 3 to 4 items were present in the study array, regardless of item complexity (Awh, Barton, & Vogel, 2007, Barton, Ester, & Awh, 2009, Vogel, Woodman, & Luck, 2001, but see Oberauer & Eichenberger, 2013 and Wilken & Ma, 2004).

In a recent development, Rouder et al. (2008) demonstrated that a quantitative implementation of the slots theory provided a good account of choice probability data in change detection tasks. Following Wilken and Ma (2004), Rouder and colleagues plotted the proportion of correct *change* responses (hits) as a function of the incorrect *change* responses (false alarms)—the receiver operating characteristic (ROC) curve. They then showed that models based on slots theory, being high-threshold models, make precise predictions about the shape of ROC curves (Green & Swets, 1966). Specifically, unlike the curvilinear ROC curves expected under SDT, a slots model predicts linear ROC curves. Rouder et al. found that the empirical data were more consistent with the linear ROC prediction and outperformed the signal-detection resource model (but see Wilken & Ma, 2004).

Donkin et al. (2015) replicated and expanded on Rouder et al.'s (2008) results in a series of four experiments. They manipulated two independent variables across their experiments—set size and change proportion. The set size manipulation involved changing the number of items in the study array across trials, whilst the change proportion manipulation adjusted the number of trials in a block on which items changed between study and test. Set size manipulations affect the overall difficulty of the task, where increases in the number of items lead to a decrease in the hit rate and an increase in the false alarm rate. Increasing the proportion of change trials, on the other hand, increases both the participants' hit rate and false alarm rate. The concurrent manipulation of these two variables result in ROC curves that allow for contrast of the predictions made by the slots and resource models.

All of the experiments used a standard change detection task, in which a study array of N color squares are presented, removed for a short period, and a single test color presented in one of the study locations. The test item was either the same as the item previously presented in that location (a *same* trial), or was an item not previously presented in the study array (a *change* trial). The participant indicates whether the test item was the same or had changed, and receives feedback on their performance. The experiments differ in the manipulation of two factors: the number of items in a study array, N, and the proportion of trials in a block on which an item changed from study to test. The number of study items in an array was randomized across trials, while change proportion was, by definition, blocked. Table 1 contains the details of the design for each experiment.

For each set size condition, i, and change proportion condition, j, we observe H_{ij} change responses from $n_{ij}^{(c)}$ change trials, and F_{ij} change responses from $n_{ij}^{(s)}$ same trials. We assume that these hit and false alarm trials are distributed according to a binomial distribution

$$h_{ij} \sim Binomial(h_{ij}, n_{ij}^{(c)})$$
$$f_{ij} \sim Binomial(f_{ij}, n_{ij}^{(s)}). \quad (1)$$

We now outline the slots and resource model predictions for the hit and false alarm rates h_{ij} and f_{ij}.

1.1. Slots

According to the slots model, there are two types of responses in a change-detection task: if an item is in memory, the response is based on the contents of memory; if an item has failed to make it into memory, then a guess is made. The probability that any given item, from a set of N items, makes it into memory depends on the number of items that can be stored in memory, or the capacity k. The probability an item is in memory, d, is given by $d = \min(1, \frac{k}{N})$, where the min function ensures that the probability an item is in memory does not exceed 1. For simplicity, it is assumed that when items are encoded into memory, responses are made without error. If an item is not in memory, then the participant must guess whether the item has changed or remained the same. The probability that the observer will guess *change* is denoted as g. Together, the probability of a hit response is given by $d + (1 - d)g$ and the probability of a false alarm is $(1 - d)g$.

The slots model, as defined thus far, predicts perfect performance for set sizes below capacity. Since distraction, or incorrect button presses, cause participants to make errors even when the task is trivially easy, Rouder et al. (2008) explicitly modeled errors due to inattention. They assumed that the observer pays attention with probability a, and thus proceeds as previously defined. However, with probability $1 - a$ the observer is inattentive, and therefore guesses *change* with probability g.

Turning now to the two manipulations in the experiments, we see that the slots model provides a natural account of changing set sizes, since larger N yields a smaller probability that the test item is in memory. The change proportion manipulation is assumed to influence the way in which an observer will guess. As the proportion of change trials increases, the probability that a participant will guess change should also increase. As such, we estimate g separately for each change proportion condition. As applied to the current experiments, the predicted hit and false

Please cite this article in press as: Kary, A., et al., Using Bayes factors to test the predictions of models: A case study in visual working memory. Journal of Mathematical Psychology (2015), http://dx.doi.org/10.1016/j.jmp.2015.07.002

Table 1
Designs for Experiment 1, Experiment 2, and Experiment 4 from Donkin et al. (2015).

Experiment	Set sizes	Change probabilities	Participants	Trials
1	3, 5, 8	0.3, 0.5, 0.7	96	18–42
2	2, 5, 8	0.15, 0.3, 0.5, 0.7, 0.85	20	24–136
4	1, 2, 3, 4, 6, 8	0.5	30	84

alarm rates in the slots model for the ith set size and jth change proportion condition are given by

$$h_{ij} = a(d_i + (1 - d_i)g_j) + (1 - a)g_j$$
$$f_{ij} = a(1 - d_i)g_j + (1 - a)g_j. \quad (2)$$

1.2. Resource

Signal detection theory is used to implement the resource model of visual working memory capacity. When a location is probed with a test item, the item will evoke a particular level of distinctiveness, x.[1] The level of distinctiveness is then compared to a criterion β, which the participant uses to decide whether to respond that the test item is the same or has changed from study. The distinctiveness evoked by a test item that is the same as the study item is assumed to follow a standard Normal distribution (that distinctiveness varies from trial-to-trial encapsulates the idea of a continuous resource). Test items that have changed from study will evoke distinctiveness according to a Normal distribution with mean d' and variance 1. As such, when a test item evokes an amount of distinctiveness the participant compares the likelihood of that distinctiveness under those two distributions:

$$LR(x) = \frac{\phi(x - d')}{\phi(x)}$$

where ϕ is the probability density function of the standard Normal distribution. If the likelihood ratio is greater than criterion β, then the observer responds *change* and otherwise responds *same*.

We assume that as the number of items to remember increases, any one item is given less memory, and as such is less distinctive from same items. Accordingly, d' is expected to decrease as set size increases in the experiments. Additionally, when the number of trials on which a test item changes from study to test is varied, participants are assumed to require more or less distinctiveness from memory in order to respond *change*. The change proportion manipulation is therefore expected to influence the criterion parameter of the signal detection model, β. Combining these assumptions, the signal detection model is defined by the following two equations

$$h_{ij} = \Phi\left(\frac{d'_i}{2} - \frac{\log \beta_j}{d'_i}\right)$$
$$f_{ij} = \Phi\left(\frac{-d'_i}{2} - \frac{\log \beta_j}{d'_i}\right) \quad (3)$$

where Φ is the cumulative distribution function of the standard Normal distribution.

1.3. A cautionary note

For this case study, we use models and experiments that do not necessarily represent the state-of-the-art in the area. Our choice of models was based on those used in Donkin et al. (2015) and Rouder et al. (2008). However, the models we fit here may be too simple to account for visual working memory under all possible conditions. For example, the data from continuous recall tasks have required more complex alternative versions of slots and resource models (van den Berg, Awh, & Ma, 2014; van den Berg, Shin, Chou, George, & Ma, 2012). There also exist simpler versions of the resource model, such as the sample-size model (Sewell, Lilburn, & Smith, 2014). In addition, change detection experiments that vary the size of the change between study and test (Keshvari, van den Berg, & Ma, 2013), and use response time in addition to choice proportion (Donkin, Nosofsky, Gold, & Shiffrin, 2013), have also proven useful in distinguishing between models.

1.4. Prior distributions

The choice of prior distributions on model parameters is critical to model selection via Bayes factors. As such, we endeavor to specify reasonable prior distributions. Before we look at the data from our experiments, we have little information about what values the parameters of the slots and resource models should take. We can place only very vague prior distributions on the parameters of both models. For example, going into Experiment 1, we have little idea what value d' could take for any given set size, or what criterion β that participants will use for a given change proportion manipulation. We could make some relatively sensible guesses. For example, we could assume that d' will decrease with set size, since we expect that performance will worsen as set size, or memory load, increases. We could also set our prior distributions such that people do not require more evidence for a change response when they are completing blocks of trials with fewer change trials.

Rather than calculating Bayes factors with vaguely informative prior distributions, we instead sacrifice half of the participants in our first experiment in order to build up informative prior distributions for the parameters of both models. We use the first 48 participants from Experiment 1 of Donkin et al. (2015) to obtain posterior distributions for the parameters of the slots and resource models. These posterior distributions are then used to specify informative prior distributions with which we will calculate Bayes factors for the remaining 48 participants.

2. Building informative priors

The design of Experiment 1 from Donkin et al. (2015) can be found in Table 1. For reference, the top panel of Fig. 1 plots the hit and false alarm rates aggregated across all participants from their respective experiments. As expected, we observe that as set size increases, hit rates decrease and false alarm rates increase, and as change proportion increases, so too do both hit and false alarm rates.

2.1. Hierarchical models

Since a number of participants completed each of our experiments, we must decide how to model individual differences. The standard approach is to either estimate an independent set of parameters for each individual, or assume that all individuals share a single set of parameters. The approach we take here is hierarchical

[1] The signal in signal detection theory is often referred to as 'familiarity'. Here, we refer to the signal as the distinctiveness between the test item and the contents of memory. This label is arbitrary, but reflects the choice to use correct 'change' responses as 'hits'.

Please cite this article in press as: Kary, A., et al., Using Bayes factors to test the predictions of models: A case study in visual working memory. Journal of Mathematical Psychology (2015), http://dx.doi.org/10.1016/j.jmp.2015.07.002

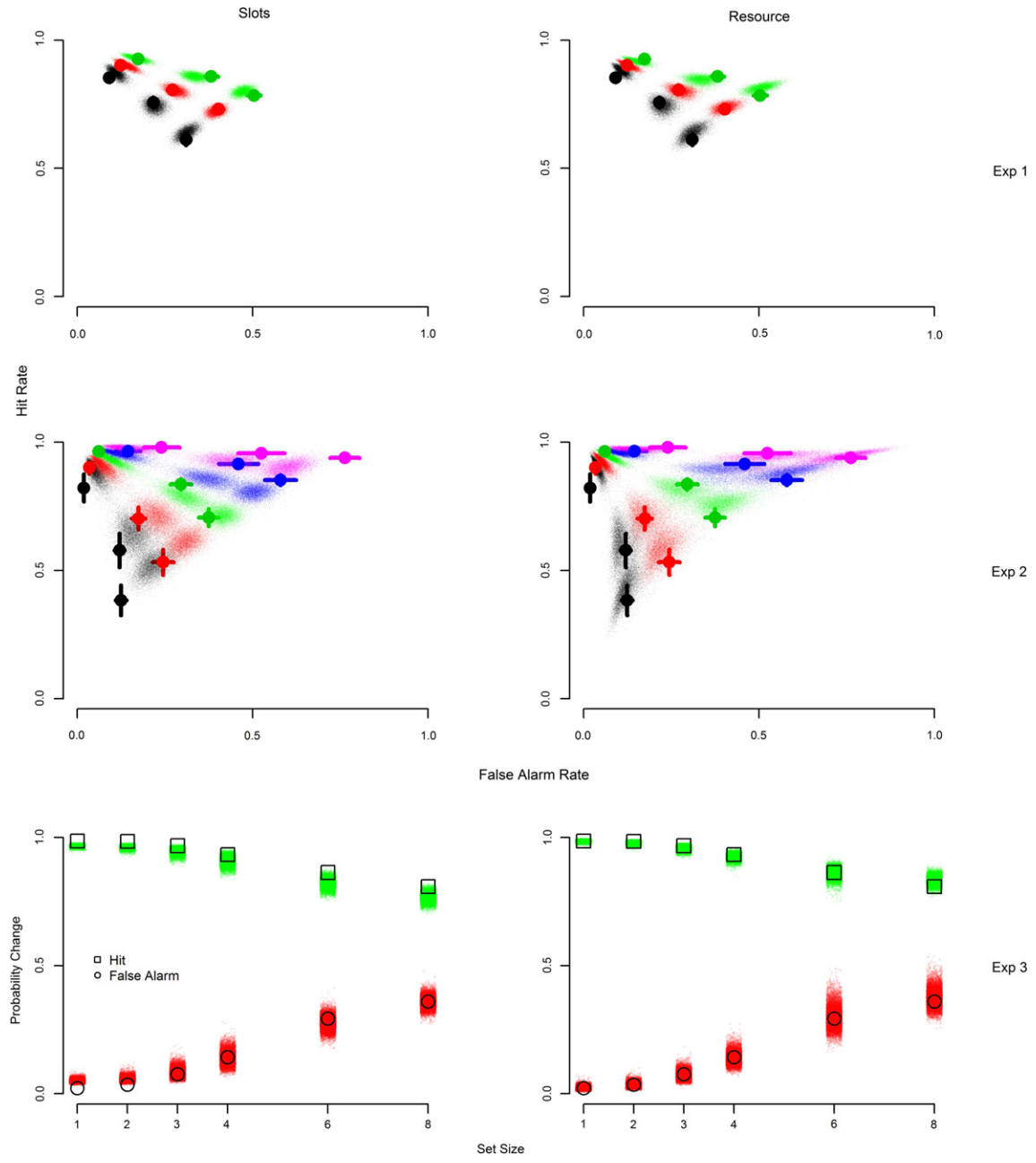

Fig. 1. Observed average hit and false alarm rates for each set size and change proportion condition in Experiments 1, 2, and 3. For Experiments 1 and 2, the observed data are plotted as colored circles, with error bars representing standard errors of the mean. Posterior predictives for the average hit and false alarm rates were also generated from the slots and resource models, and are plotted as semi-transparent dots. The colors of the points indicate the change proportion condition. For Experiment 3, data are plotted as open characters, with posterior predictives overlaid in color. Note: Exp = Experiment. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

modeling, which assumes that individual participants are drawn from some population-level distribution (Lee, 2011). Morey (2011) outlined hierarchical versions of the slots and resource models (see also Morey & Morey, 2011). Here, we assume that all population-level parameters are distributed normally, $N(\mu, \sigma)$. For example, in the slots model, we assume that a capacity parameter for individual m is distributed[2]

$$k_m \sim Normal(\mu^k, \sigma^k).$$

Note that when we draw participant-level parameters from the population-level Normal distributions, we truncate the distributions such that $a_m, g_{j,m} \in [0, 1]$, $k_m \in [0, 8]$, and $\beta_m \in (0, \infty)$.

Our hierarchical slots and resource models are defined by a series of μ and σ parameters, each of which have prior distributions that must be specified. We define vague prior distributions for μ parameters for the first half of the participants in Experiment 1, who are being sacrificed to build informative priors. We set prior distributions for μ^a, μ^g, and μ^k as uniform over their viable ranges: $\mu^a \sim U(0, 1)$, $\mu^g \sim U(0, 1)$, and $\mu^k \sim U(0, 8)$. We set the $\mu^{d'}$ and μ^β parameters from the resource model as $U(0, 5)$. The prior distributions for all σ parameters were set as $U(0.01, 10)$.

Posterior distributions for all parameters were obtained by taking 27,000 samples using 6 chains of 4500 samples, after 500 sam-

[2] Note that capacity is a discrete quantity. However, k is usually estimated as a continuous quantity, approximating variability in capacity across trials. Though it is less than ideal, we follow this standard practice.

Please cite this article in press as: Kary, A., et al., Using Bayes factors to test the predictions of models: A case study in visual working memory. Journal of Mathematical Psychology (2015), http://dx.doi.org/10.1016/j.jmp.2015.07.002

ples of burn-in. We applied the standard checks for convergence of chains, and autocorrelation within chains, and found no problems for any of the model parameters. We used JAGS to fit the models to our data (Plummer, 2003).[3]

2.2. Creating individual-level priors from population-level posteriors

We now have posterior distributions for each of the model parameters. However, these posterior distributions are at the population level, while we wish to obtain individual-level Bayes factors. As such, we need to convert the information contained in the population-level posteriors to represent what we expect for the parameter values for individual participants.

Our population-level parameters are the means, μ, and standard deviations, σ, of Normal distributions. Individual-level parameters are governed by those Normal distributions; For example, $k_m \sim N(\mu^k, \sigma^k)$. One could simply take the most likely combination of population-level mean and standard deviations, and use the resultant Normal distributions as the prior distributions for individuals. However, such an approach would ignore our uncertainty in the population-level parameters.

Our approach is to use the posterior distribution of μ and σ parameters to generate individual-level priors. Our fitting process yields 27,000 posterior samples for the μ and σ parameter for each of the parameters of the slots and resource models. Now, for each model parameter (e.g., k), we take each posterior sample for μ and σ and generate 1000 samples from the resultant Normal distribution. These 1000 values represent the range of individual-level parameter values we would expect given that particular posterior sample. The result of this process is that for each model parameter, we have 27,000,000 possible individual-level parameter values, as weighted by their posterior probability. We truncate these individual-level parameter values to be constrained within their appropriate ranges (e.g., $a \in [0, 1]$).

Finally, we must decide on a distribution to characterize the shapes of the individual-level parameters. We used Beta distributions for a and g, which are constrained to be between 0 and 1, and Normal distributions for k, d', and β. We obtained maximum-likelihood estimates of the parameters of these Beta and Normal distributions, and use those estimates to define our prior distributions. The exact specifications of these prior distributions are given in Table 2.

The priors for each parameter in each model are independent of one another. As such, any correlations between the parameters of the slots and resource models are discarded. Our assumption of independence is largely out of computational convenience, and it is worth noting that our approximation limits our current efforts to discriminate between the two models.

2.3. Prior predictives

We now have informative prior distributions for our slots and resource models, which will make more constrained predictions for the remaining participants in Experiment 1. To see this, consider the left and right panels of Fig. 2, which plot prior predictives for vaguely specified and informative prior distributions, respectively. We construct these plots of prior predictives by drawing 1 million sets of parameters from the prior distributions outlined in Table 2. We use these randomly-sampled parameters to generate a set of predicted hit and false alarm rates for all set size and change proportion conditions in the experiment (i.e., h_{ij} and f_{ij}, respectively). The predicted hit and false alarm rates from each model are then plotted in Fig. 2.

Table 2
Prior distributions for the slots and resource models used for calculating Bayes factors.

Expt.	Slots		Resource	
	Parameter		Parameter	
1	k	N(3.09,0.91)	d'_3	N(2.73,0.75)
	a	B(16.08,2.04)	d'_5	N(1.57,0.62)
	$g_{0.3}$	B(3.46,4.00)	d'_8	N(0.96,0.33)
	$g_{0.5}$	B(11.26,7.20)	$\beta_{0.3}$	N(1.11,0.37)
	$g_{0.7}$	B(15.40,5.64)	$\beta_{0.5}$	N(0.83,0.18)
			$\beta_{0.7}$	N(0.65,0.15)
2	k	N(2.98,1.01)	d'_2	N(4.03,0.80)[a]
	a	B(37.5,5.17)	d'_5	N(1.52,0.64)
	$g_{0.15}$	B(3.18,5.39)[a]	d'_8	N(0.91,0.33)
	$g_{0.3}$	B(4.41,5.13)	$\beta_{0.15}$	N(1.27,0.31)[a]
	$g_{0.5}$	B(13.46,9.43)	$\beta_{0.3}$	N(1.10,0.30)
	$g_{0.7}$	B(12.69,5.36)	$\beta_{0.5}$	N(0.87,0.16)
	$g_{0.85}$	B(5.04,4.09)[a]	$\beta_{0.7}$	N(0.70,0.16)
			$\beta_{0.85}$	N(0.57,0.31)[a]
4	k	N(2.68,1.05)	d'_1	N(5.89,0.88)[a]
	a	B(39.08,4.44)	d'_2	N(3.24,0.81)
	$g_{0.5}$	B(12.99,9.00)	d'_3	N(2.61,0.79)
			d'_4	N(1.92,0.872)[a]
			d'_6	N(1.38,0.872)[a]
			d'_8	N(1.00,0.35)
			$\beta_{0.5}$	N(0.84,0.17)

Note: Expt. = Experiment. N and B correspond to Normal and Beta distributions, respectively.
[a] Indicates extrapolated prior distributions.

It is clear from the left panel of Fig. 2 that neither the slots nor the resources model make particularly sensible predictions about hit and false alarm rates when the prior distributions are vaguely specified (i.e., using the prior distributions used for the first 48 participants in Experiment 1). For example, as we move across the three columns in the left panel of Fig. 2, we see that the predictions for hit and false alarms remain identical, even though the proportion of change trials is varying. That is, the model predicts that participants will be invariant to this manipulation of bias. Similarly, in the left panel, we see that the resource model with vague priors predicts no influence of set size on hit and false alarm rates; The model predicts that increasing the load on memory will have no effect on performance. Since Bayes factors assess the ability of models to predict the observed data, the models with vague priors represent uninteresting versions of slots and resource models.

The right panel of Fig. 2 shows that the models do make more specific predictions when we use informative prior distributions. For example, in the right panel of Fig. 2 we see that the models now predict that both hit and false alarm rates will increase with the proportion of change trials in the experiment. In addition, the resource model with informed priors makes the sensible prediction that performance will worsen as set size increases. We can now use Bayes factors to compare the slots and resource models on their ability to predict the data from the remaining 48 participants from Donkin et al.'s (2015) Experiment 1.[4]

3. Bayes factors for Experiment 1

To calculate Bayes factors, we need the marginal likelihood of each model for each individual. The simplest way to estimate the marginal likelihood of a model is to sample parameters from prior distributions, evaluate the likelihood for each set of

[3] The code and data used in all analyses in this paper are available on the corresponding author's website, as should be standard.

[4] It is worth noting that prior distributions that place a simple order constraint on certain parameters could have also yielded relatively sensible predictions. Often times, this is an appropriate alternative to data-informed priors.

Please cite this article in press as: Kary, A., et al., Using Bayes factors to test the predictions of models: A case study in visual working memory. Journal of Mathematical Psychology (2015), http://dx.doi.org/10.1016/j.jmp.2015.07.002

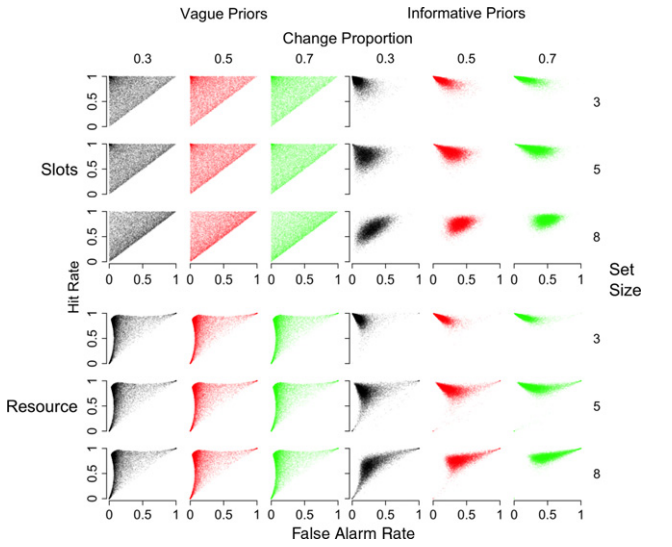

Fig. 2. Prior predictives for Experiment 1 with vague priors (left panel) and informative priors (right panel). Note that the models with informed priors make more constrained predictions than the models with vague priors. The colors of the points reflect the change proportion condition. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

parameters, and average the resultant likelihoods. However, this method is inefficient, and so we instead used importance sampling (see Vandekerckhove, Matzke, & Wagenmakers, 2015 for a clear explanation of importance sampling).

The basic idea of importance sampling is to use posterior distributions to increase the efficiency of sampling parameters that are used to calculate a marginal likelihood. We drew $N = 10,000$ samples from an importance distribution, $g(\theta)$. The importance distribution for each parameter was a mixture, 80% of which was the posterior distribution for that parameter, and 20% was a uniform distribution that spanned the range that the parameter can take.

The likelihood for each parameter set, $p(D|\theta_i, M)$ is evaluated using Eqs. (1) and (2) for the slots model, and Eqs. (1) and (3) for the resource model. The likelihood of each parameter set is evaluated under the prior distribution $p(\theta_i|M)$ and the importance distribution $g(\theta_i)$, and the marginal likelihood is given by

$$p(M|D) = \frac{1}{N}\sum_{i=1}^{N}\frac{p(D|\theta_i, M)p(\theta_i|M)}{g(\theta_i)}.$$

The ratio of the marginal likelihoods for the slots and resource models gives us a Bayes factor, BF.

The top plot in Fig. 3 shows the log of the Bayes factors for the second 48 participants in Experiment 1. The data from the majority of participants are more likely under the slots model than the resource model. Interestingly, many of the participants do not provide a large deal of support for either model. Only the data from 10 of the 48 participants is more than 10 times more likely under the slots model than the resource model, and only one participant's data is more than 10 times more likely under the resource model.

We chose to sacrifice half of the data from our first experiment in order to derive a more informative Bayes factor. Therefore, we have calculated a *partial* Bayes factor, which only gives the relative likelihood that the two models generated the latter half of the data from Donkin et al.'s (2015) Experiment 1, and is conditional on this particular division of the data. Methods do exist for attempting to remove the contribution of the division and generalize to the entire data set, such as the Intrinsic Bayes factor and the Fractional Bayes Factor (Berger & Pericchi, 1996, O'Hagan, 1995; and see Mulder, 2014a,b for alternative approaches to using data to construct priors). We chose not to implement these methods due to their computational cost.

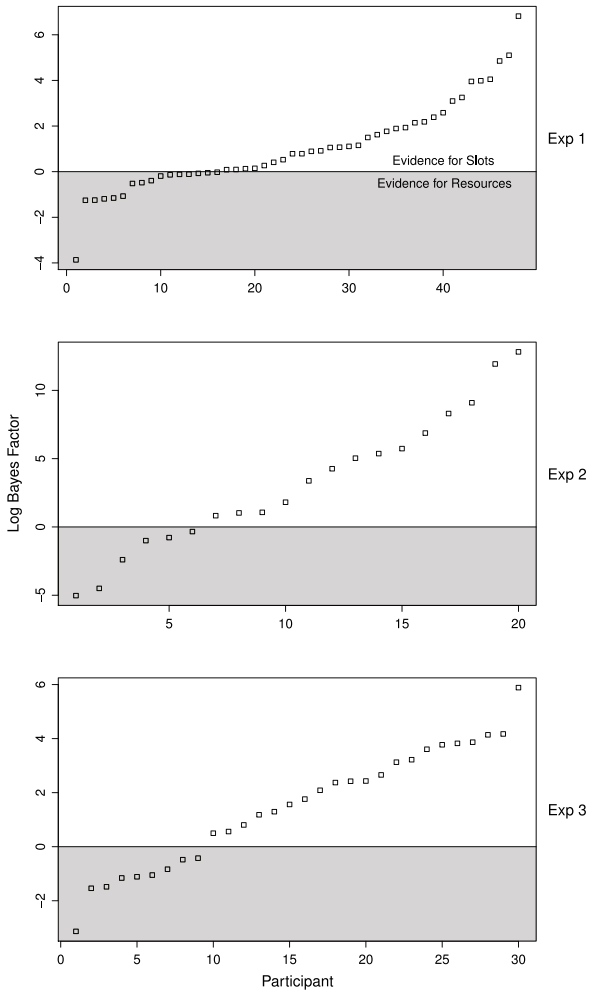

Fig. 3. The logarithm of each individual's Bayes factor are plotted. Positive log Bayes factors indicate support for the slots model. Note: Exp = Experiment.

4. Updating our priors

We now want to update our prior distributions so that we can calculate Bayes factors for a second experiment. We use the remaining 48 participants to update our beliefs regarding the likely parameter values for each model. To do this, we obtain posterior distributions for the hierarchical versions of the slots and resource models for all 96 participants in Donkin et al.'s (2015) Experiment 1. Note that this is equivalent to estimating the posterior distributions for the remaining 48 participants, while using the (joint) posterior distributions from the first 48 participants as prior distributions.

We use the same procedure as for the first 48 participants to estimate posterior distributions for all 96 participants, using JAGS with the same number of chains, samples, burn-in, and thinning as outlined earlier. Both the slots and resource models provide a good account of the observed data, as shown by the posterior predictives plotted in the top panel of Fig. 1. The posterior predictives were generated by drawing 9,000 values from the posterior distribution of each parameter, and calculating the predicted hit and false alarm rates for each set of parameters. It is reassuring to know that the models are able to fit the data well, since the Bayes factor only provides a relative measure of model performance.

We use the posterior distributions from Experiment 1 as prior distributions for Experiment 2 from Donkin et al. (2015). Note that the design for Experiment 2 was not identical to that of Experiment 1—the smallest set size in Experiment 2 was two, rather than three,

Please cite this article in press as: Kary, A., et al., Using Bayes factors to test the predictions of models: A case study in visual working memory. Journal of Mathematical Psychology (2015), http://dx.doi.org/10.1016/j.jmp.2015.07.002

and there were 2 additional proportion change conditions, 0.15 and 0.85 (see Table 1). As such, Experiment 1 does not directly yield informative priors for these new conditions.

4.1. Creating prior distributions for new conditions

We extrapolated prior distributions for parameters associated with the new conditions in Experiment 2 using the posterior distributions we observed in Experiment 1. We first extrapolated priors for the population-level μ and σ parameters, and then use them to generate priors for the individual-level parameters. Our approach was to extrapolate sensible mean values for the prior distributions for μ and σ parameters, and to set the standard deviation of the μ and σ parameters to be double that of the observed posterior distributions from Experiment 1.

We start with the prior for the $\mu^{d'_2}$ parameter in the set size 2 condition. We chose to use a normal distribution as a prior for the $\mu^{d'_2}$ parameter—$\mu^{d'_2} \sim Normal(A, B)$. To set A, we take the mean of the posterior distributions for the $\mu^{d'}$ parameters for the set size 3, 5, and 8 conditions. We then extrapolate A from the fit of a power-law function through these mean values of $\mu^{d'_3}$, $\mu^{d'_5}$ and $\mu^{d'_8}$. We set B to be twice the value of the largest standard deviation of the posterior distributions for $\mu^{d'_3}$, $\mu^{d'_5}$, and $\mu^{d'_8}$.

We use a Gamma distribution as a prior for the $\sigma^{d'_2}$ parameter—$\sigma^{d'_2} \sim Gamma(A, B)$. We set A to be the largest of the A parameters of the posteriors of $\sigma^{d'_3}$, $\sigma^{d'_5}$ and $\sigma^{d'_8}$. We set B as twice the maximum of the B parameters of the posterior distributions of $\sigma^{d'_3}$, $\sigma^{d'_5}$ and $\sigma^{d'_8}$ parameters. We used the same process to extrapolate the $\sigma^{g_{0.15}}$, $\sigma^{g_{0.85}}$, $\sigma^{\beta_{0.15}}$, and $\sigma^{\beta_{0.85}}$.

We outline how we extrapolated the $\mu^{g_{0.15}}$ parameter, and note that the remaining parameters were extrapolated in the same manner. We chose to use a beta distribution to characterize the priors for the $\mu^{g_{0.15}}$ and $\mu^{g_{0.85}}$ parameters. We used a normal distribution for the $\mu^{\beta_{0.15}}$ and $\mu^{\beta_{0.85}}$ parameters.

For the $\mu^{g_{0.15}}$ parameter, we first took the mean of the posterior distribution of all of the μ^{g} parameters. The difference between the means of the μ^{g} parameters was approximately 0.12. We set the mean of the $\mu^{g_{0.15}}$ parameter to be 0.09 ($\frac{3}{4}$ of 0.12) less than the mean of the $\mu^{g_{0.3}}$ change proportion condition. We use $\frac{3}{4}$ because the difference between the new change proportion condition (0.15) and the next largest change proportion condition (0.30) is 75% of the difference between the old change proportion conditions (0.3, 0.5, and 0.7). The standard deviation of the $\mu^{g_{0.15}}$ parameter was set at twice the average of the standard deviation of the posterior distributions of the μ^{g} parameters.

We now have prior distributions for the population-level parameters, and use these to generate individual-level parameters. For example, to generate the prior distributions for the individual-level d'_2 parameter, we sample 100,000 values from the prior distributions we had generated for $\mu^{d'_2}$ and $\sigma^{d'_2}$. We then used a beta distribution to characterize the individual-level parameters for the g parameters, and normal distributions for the β and d' parameters. The result of this process is shown in Table 2.

5. Bayes factors for Experiment 2

To obtain the marginal likelihood of the slots and resource models for each participant in Experiment 2, we again use importance sampling to estimate marginal likelihoods. The center panel of Fig. 3 plots the log of the Bayes factors for each individual in Experiment 2. Again, we see that the majority of individual's data are more likely under the slots model than the resource model. This time, there is more certainty in the conclusions we draw from this data. In particular, the data from 10 out of the 20 participants are more than 10 times likely under the slots model than the resource model, and only 3 participants are more than 10 times likely under the resource model.

6. Another update to our priors

We again obtained posterior distributions for the slots and resource model parameters using hierarchical versions of each model. We can use as the posterior distributions for μ and σ parameters from Experiment 1 as prior distributions for Experiment 2. For reference, we again plot the average hit and false alarm rates from Donkin et al.'s (2015) Experiment 2 in the center row of Fig. 1. We also plot posterior predictives for the slots and resource models. Note that the resource model fits all of the data well, while the slots model seems to struggle to account for differences across the change proportion conditions (e.g., the black, red, and blue data points lie away from posterior predictives).

One may wonder why the Bayes factors favor the slots model over the resource model, while Fig. 1 seems to indicate the resource model is more appropriate. The simplest explanation is that aggregate data do not necessarily represent the behavior of individuals. In addition, the resource model is more flexible than the slots model, and so should be able to better fit empirical data than the slots model. The Bayes factor, however, does not evaluate how well the model can account for the data once it is observed, but represents the ability of the models to predict the observed data. Fig. 1 suggests that the resource model is better able to *postdict* the data, while the Bayes factors tell us that the slots model is better able to predict our data.

We use the posterior distributions estimated for the second experiment to update our priors going into the third and final experiment. We now calculate Bayes factors for Experiment 4 in Donkin et al. (2015). In this experiment, only set size was manipulated. The exact design is outlined in Table 1. In order to set prior distributions for all conditions in Donkin et al.'s Experiment 4, we again had to extrapolate the prior distributions for a number of parameters in the resource model—d'_1, d'_4, and d'_6. We used the same process to extrapolate, and this time interpolate, priors for these parameters as we did for d'_2. Note that we used a power-law function fitted to the means of the previously observed posterior distributions for d'_2, d'_3, d'_5, and d'_8 parameters. The resultant prior distributions are shown in Table 2.

7. Bayes factors for Experiment 3

Bayes factors were calculated in the same way as for Experiments 1 and 2. The bottom plot in Fig. 3 plots the log of the Bayes factors for each individual in this third experiment. Again, most participants are more likely under the slots model than the resource model. Again, we see some certainty in our conclusions, where 13 out of 30 participants were more than 10 times more likely under the slots model, while only 1 participant is more than 10 times more likely under the resource model.

8. Updating our priors for future experiments

Though we do not fit any more data sets in this manuscript, we can still update our prior distributions in light of the data observed in the third experiment we analyzed. We must first estimate posterior distributions for the model parameters for this third data set. To do so, we set the prior distributions of the hierarchical slots and resource models based on the posterior distributions we obtained from the first and second data sets.

We obtained posterior distributions for the data from Experiment 4 of Donkin et al. (2015). The bottom panel of Fig. 1 plots the average data from this experiment, and the posterior predictives for the slots and resource models confirm that both models again provided a good fit.

Over the course of these three experiments, we have acquired considerable information about the parameters of the slots and

Please cite this article in press as: Kary, A., et al., Using Bayes factors to test the predictions of models: A case study in visual working memory. Journal of Mathematical Psychology (2015), http://dx.doi.org/10.1016/j.jmp.2015.07.002

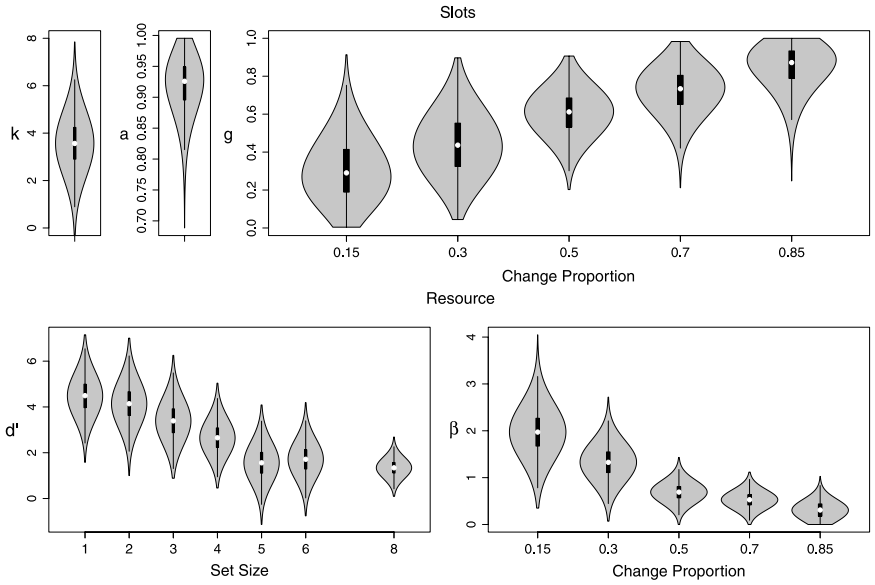

Fig. 4. Prior distributions for all of the model parameters based on the three experiments we analyzed.

resource models of the change detection task. Fig. 4 contains a plot of what we would use prior distributions for the individual-level parameters of both the slots and resource models. The specification of these distributions, and the population-level μ and σ parameters can be found in code available on the corresponding authors' website. These prior distributions contribute to the current definition of these slots and resource models for the change detection task.

9. General discussion

9.1. Informative prior distributions

One of the most interesting results we observed was that the final experiment provided more support for the resource model than did the original study. The original interpretation, based on a landscaping analysis, was that all participants were better fit by a slots model. This difference likely arose because our analysis used informed prior distributions, which leads us to an interesting aspect of the Bayes factor method. As we observe more data, our prior distributions grow more informed. As a result, models that initially make vague predictions, become increasingly constrained in their predictions. In other words, models become simpler as prior distributions become more informed.

The constraint on predictions offered by informed priors will benefit complex models more than simple models. We see that quite clearly in our final experiment. The resource model estimates a separate d' parameter for each of the 6 set sizes in the experiment, where the slots model requires only 2 parameters. Regardless, for some individuals, the resource model is preferred by the Bayes factor. The informed prior distributions for the d' parameters yielded predictions that were more consistent with the observed data than the strict predictions of the slots model.

9.2. A note on extrapolating our knowledge

In the current analysis, we faced situations where we had to extrapolate or interpolate our knowledge about parameters that we had some information about to parameters that we had no information about when moving between experiments. We relied on reasonable assumptions for our extrapolation. For example, we assumed that participants would respond *change* more often in a change proportion condition of 0.85 than a change proportion condition of 0.7. In addition, we increased the variability in the extrapolated prior distributions, reflecting the additional uncertainty we have about the parameter. More variable prior distributions make less specific predictions for those conditions, and so maintain a degree of flexibility.

It is likely that in many cases the definition of prior distributions will rely on such extrapolation from previous data. For example, in experiments that involve a manipulation that has not yet been applied to that particular domain. However, it is rare for a manipulation to be truly novel. Further, it is rare that the experimenter does not have at least an ordinal prediction for the impact a manipulation will have on model parameters. Such intuitions can be implemented into the models via prior distributions, and can often leverage information from previous data.

9.3. Experimental design

We should clarify that we are not advocating that experiments are designed to be as similar to previous experiments as possible. Of course, progress will come through experiments designed to discriminate between models. It is worth noting that the use of prior predictives can aid in the development of such constraints, as they help define and then contrast the predictions made by the models. Also, though we aim for qualitative contrasts between models, it is exceedingly rare for any debate over experimental data to not come down to model selection. Rather, it is typical that despite best-laid plans, we end up contrasting models on their ability to account for data from multiple, and similar, experiments. The method we outline is ideal for such situations.

9.4. An iterative Bayesian treatment of model complexity

We can unpack the advantages of an iterative Bayesian analysis in model selection across designs a little more if we consider what an experimental design actually is. One working definition would be that an experimental design in psychology is an attempt to estimate the nature of a cognitive process under controlled conditions. To achieve a level of control, experimental designs have to be limited. In a typical experiment, we are measuring a cognitive process under certain conditions defined by the manipulations of the experiment. Since any experiment, by design, is only going

Please cite this article in press as: Kary, A., et al., Using Bayes factors to test the predictions of models: A case study in visual working memory. Journal of Mathematical Psychology (2015), http://dx.doi.org/10.1016/j.jmp.2015.07.002

to be testing the process under restricted conditions, we are only going to be able to sample from a certain part of our process models' prediction spaces using any one experimental design. This can be why different experiments purporting to estimate the same process can yield strikingly different results. It is not necessarily the case that one experiment gives a more true characterization of a process, it can also be that the same process operating under different conditions can yield different estimates of the parameters modeling that process. We argue that an iterative Bayesian approach is uniquely equipped to deal with this possibility.

At this point, we again return to simple vs. complex models. A complex model is the sort of model that we need if the same process can behave differently depending on the conditions of estimation. A simple model might be built on the basis of one class of experimental designs, and can explain the data of that class very well. But when we change the design, a simple model may no longer make good predictions. A complex model that explained the data well in the first class of experiments, and can also explain the data well for new designs is a useful model to have, and one that would ultimately be favored by an iterative Bayesian approach. If the data generating process is itself complex, and over time we are able to estimate the range of that complexity through multiple experimental designs, then iterative Bayes factors will ultimately lead us to believe in a more complex data-generating model, provided that it is a good model and that we are actually measuring the same data generating process across designs. If the model is poor, or our experiments are estimating different processes, then the model will not make good predictions across the designs and the Bayes factors will advise us not to believe it. The general message here is that it might not be wise to reject a complex model if we are only relying on similar experimental designs, where we never get to test if that complexity is useful or explanatory. Bayes factors with informed prior predictives give us the tools to test whether a model's complexity is a good thing in the long run.

This idea is a key reason why we do not wish to generalize far from the conclusions in this current manuscript. We acknowledge that the slots model provides the best account of change detection experiments in which set size is randomized across trials, change proportion is blocked, and the size of the change between study and test items is large. But we believe that this class of experimental designs gives only a limited estimate of the full process underlying visual working memory, and that further experiments may well provide evidence for a resource account. Our prediction is that further experiments with different designs, analyzed with prior predictives derived from the current experiments (or similar experiments), will eventually favor resource models of visual working memory, as we find it implausible that a model as simple as the slots model (at least the version outlined here) can account for the full psychological complexity of VWM. The use of informative prior distributions is important here, as without informed predictions the more complex resource models will always be handicapped compared to a simple model like the slots model when analyzing any one experimental data set alone. This is one advantage of Bayesian analysis—even when we are finding evidence against a model, we can still be accumulating information about it that can prove useful for understanding future experimental output.

9.5. Conclusion: the benefits of prediction

It is a common feature of model selection papers that upon fitting and punishing a series of models, authors tend to describe the best-fitting model as having best *predicted* the data. But, as we have demonstrated here with our comparison of the slots model and signal detection model of visual working memory, prediction is much more than simply fitting models. Using the knowledge that we gain over previous experiments to inform our evaluation of new evidence allows us to avoid giving weight to parts of the parameter space that we would not expect to find data. This gives a more precise estimate of the performance of the models that we are comparing, and can reduce the gap between functionally complex and simple models. In this instance, the slots model outperformed the signal detection model in change detection tasks with set size randomized and change proportion blocked. But this is not the end of the story, and Bayesian inference can help us to tell the rest. An iterative Bayesian approach gives us the means to address more complex problems in science, such as how to discuss and evaluate the results of different experiments as a collective whole. Scientists necessarily make inferences about different theories based on the experiments that they survey. But without Bayesian analysis, they may not be doing so in a principled and consistent way. The difference is simply that Bayesian methods do this kind of evaluation explicitly, systematically, and using the rules of probability.

Acknowledgment

Chris Donkin's contribution to this research was supported by the Australian Research Council (DP130100124; DE130100129).

References

Awh, E., Barton, B., & Vogel, E. K. (2007). Visual working memory represents a fixed number of items regardless of complexity. *Psychological Science*, *18*, 622–628.

Barton, B., Ester, E. F., & Awh, E. (2009). Discrete resource allocation in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *35*, 1359–1367.

Berger, J. O., & Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, *91*, 109–122.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*, 87–185.

Donkin, C., Nosofsky, R. M., Gold, J., & Shiffrin, R. M. (2013). Fixed slots models of visual working memory response times. *Psychological Review*, *120*, 873–902.

Donkin, C., Tran, S., & Nosofsky, R. M. (2015). Landscaping analyses of the ROC predictions of discrete-slots and signal-detection models of visual working memory. *Attention, Perception, & Psychophysics*, *22*, 170–178.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Jeffreys, H. (1961). *Theory of probability*. Oxford, UK: Oxford University Press.

Keshvari, S., van den Berg, R., & Ma, W. J. (2013). No evidence for an item limit in change detection. *PLoS Computational Biology*, *9*.

Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology*, *55*, 1–7.

Luck, S. K., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*, 279–281.

Luck, S. K., & Vogel, E. K. (2013). Visual working memory capacity: from psychophysics and neurobiology to individual differences. *Trends in Cognitive Sciences*, *17*, 391–400.

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

Morey, R. D. (2011). A Bayesian hierarchical model for the measurement of working memory capacity. *Journal of Mathematical Psychology*, *55*, 8–24.

Morey, R. D., & Morey, C. C. (2011). WoMMBAT: A user interface for hierarchical Bayesian estimation of working memory capacity. *Behavior Research Methods*, *43*, 1044–1065.

Mulder, J. (2014a). Bayes factors for testing inequality constrained hypotheses: Issues with prior specification. *British Journal of Mathematical and Statistical Psychology*, *67*, 153–171.

Mulder, J. (2014b). Prior adjusted default Bayes factors for testing (in)equality constrained hypotheses. *Computational Statistics and Data Analysis*, *71*, 418–463.

Oberauer, K., & Eichenberger, S. (2013). Visual working memory declines when more features must be remembered for each object. *Memory & Cognition*, *41*, 1212–1227.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison (with discussion). *Journal of the Royal Statistical Society: Series B*, *57*, 99–138.

Pashler, H. (1988). Familiarity and visual change detection. *Perception & Psychophysics*, *44*, 369–378.

Phillips, W. A. (1974). On the distinction between sensory storage and short-term visual memory. *Perception & Psychophysics*, *16*, 283–290.

Please cite this article in press as: Kary, A., et al., Using Bayes factors to test the predictions of models: A case study in visual working memory. Journal of Mathematical Psychology (2015), http://dx.doi.org/10.1016/j.jmp.2015.07.002

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling.
Rouder, J. N., Morey, R. D., Cowan, N., Zwilling, C. E., Morey, C. C., & Pratte, M. S. (2008). An assessment of fixed-capacity models of visual working memory. *Proceedings of the National Academy of Sciences*, *105*, 5976–5979.
Sewell, D. K., Lilburn, S. D., & Smith, P. L. (2014). An information capacity limitation of visual short-term memory. *Journal of Experimental Psychology: Human Performance and Perception*, *40*, 2214–2242.
Vandekerckhove, J., Matzke, D., & Wagenmakers, E.-J. (2015). Model comparison and the principle of parsimony. In J. R. Busemeyer, J. T. Townsend, Z. J. Wang, & A. Eidels (Eds.), *Oxford handbook of computational and mathematical psychology* (pp. 300–317). Oxford, UK: Oxford University Press.
van den Berg, R., Awh, E., & Ma, W. J. (2014). Factorial comparison of working memory models. *Psychological Review*, *121*, 129–149.
van den Berg, R., Shin, H., Chou, W. C., George, R., & Ma, W. J. (2012). Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences*, *109*, 8780–8785.
Vanpaemel, W. (2011). Constructing informative priors using hierarchical methods. *Journal of Mathematical Psychology*, *55*, 106–117.
Vanpaemel, W., & Lee, M. D. (2012). Using priors to formalize theory: Optimal attention and the generalized context model. *Psychonomic Bulletin & Review*, *19*, 1047–1056.
Vogel, E. K., Woodman, G. F., & Luck, S. J. (2001). Storage of features, conjunctions, and objects in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *27*, 92–114.
Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision*, *4*, 1120–1135.

Please cite this article in press as: Kary, A., et al., Using Bayes factors to test the predictions of models: A case study in visual working memory. Journal of Mathematical Psychology (2015), http://dx.doi.org/10.1016/j.jmp.2015.07.002