Contents lists available at ScienceDirect

Journal of Mathematical Psychology

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

Prior sensitivity in theory testing: An apologia for the Bayes factor

Wolf Vanpaemel

Department of Psychology, University of Leuven, Tiensestraat 102, B-3000 Leuven, Belgium

ARTICLE INFO

Article history:
Received 11 February 2010
Received in revised form
8 July 2010
Available online xxxx

Keywords:
Greek letter syndrome
Model evaluation
Model building
Prior
Posterior
Informative prior
Sensitivity analysis
Bayesian methods
Forgetting function

ABSTRACT

A commonly voiced concern with the Bayes factor is that, unlike many other Bayesian and non-Bayesian quantitative measures of model evaluation, it is highly sensitive to the parameter prior. This paper argues that, when dealing with psychological models that are quantitatively instantiated theories, being sensitive to the prior is an attractive feature of a model evaluation measure. This assertion follows from the observation that in psychological models parameters are not completely unknown, but correspond to psychological variables about which theory often exists. This theory can be formally captured in the prior range and prior distribution of the parameters, indicating which parameter values are allowed, likely, unlikely and forbidden. Because the prior is a vehicle for expressing psychological theory, it should, like the model equation, be considered as an integral part of the model. It is argued that the combined practice of building models using informative priors, and evaluating models using prior sensitive measures advances knowledge.

1. Introduction

The marginal likelihood is often celebrated for providing an automatically built-in Ockham's razor, balancing goodness-of-fit and complexity when evaluating models. However, even advocates of Bayesian methods often bemoan there is a price to be paid for fully taking advantage of the benefits of the marginal likelihood: it is very sensitive to the prior over the model's parameters. For example, in their landmark paper about the Bayes factor, which is the ratio of two marginal likelihoods, Kass and Raftery (1995, p. 792) note that one of the "chief limitations of Bayes factors are their sensitivity to …the choice of priors". The prior sensitivity of the marginal likelihood has not only been lamented in statistics (e.g., Aitkin, 1991; Bayarri & Berger, 2000; Efron, 1986; Gelman, 2008; Kass, 1993; O'Hagan, 1995; Wasserman, 1996), but also in more applied fields like sociology (e.g., Xie, 1999), marketing (e.g., Rossi & Allenby, 2003), economics (e.g., Koop & Potter, 1999), biology (e.g., Anderson, Link, Johnson, & Burnham, 2001), and psychology (e.g., Grünwald, 2000; Liu & Aitkin, 2008; Myung & Pitt, 1997).

This paper attempts to redeem the marginal likelihood and the Bayes factor by highlighting an alternative perspective on the prior and on prior sensitivity in model evaluation. The key argument of the paper is that, if models are quantitatively instantiated theories, the prior can be used to capture theory and should therefore be considered as an integral part of the model. The viewpoint that the prior can be used as a vehicle for expressing theory implies that a model evaluation measure should be sensitive to the prior. Only a model evaluation measure that is sensitive to the prior is informed by all aspects of the model that capture theory and provides a complete test of the theory embodied in the model. Thus, the marginal likelihood is an appropriate measure for evaluating psychological models precisely *because* of its sensitivity to the prior.

The paper starts with a classification of commonly used measures for model evaluation, with the distinction of interest being the one between prior measures (which are always sensitive to the prior) and posterior measures (which can be insensitive to the prior). Next, it is illustrated how the prior can be used to capture psychological theory. Giving the prior the careful attention it deserves when building a model often results in an informative prior, indicating which parameter values are allowed and which are not, and which are likely and which are not. The following section argues that prior measures are preferable over posterior measures for evaluating psychological models. Finally, it is highlighted that it is critical to check the robustness of conclusions against arbitrary and ad hoc assumptions, but that the prior does not necessarily reflect an arbitrary assumption and therefore should not always be subjected to a sensitivity analysis.

2. Prior and posterior measures of model evaluation

At face value, many quantitative measures for model evaluation seem to differ widely (see, e.g., Pitt & Myung, 2002; Shiffrin,

E-mail address: wolf.vanpaemel@psy.kuleuven.be.
URL: http://www.ppw.kuleuven.be/concat.

0022-2496/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmp.2010.07.003

Lee, Kim, & Wagenmakers, 2008, for overviews). However, in spirit, most of them are largely similar. In one way or the other, most measures evaluate a model by assessing the *fit* or *match* between observed, human data and generated, model-based data (i.e., the predictions made by the model). This model evaluation strategy can be broken down in two distinct steps: making the predictions (i.e., generating the model-based data) and comparing the predictions to the observed data (i.e., assessing the fit between the model-based data and the human data). Model evaluation measures can differ in each of these steps. Different possibilities concerning the second step include the squared error or the deviance. This paper focuses on different approaches concerning the first step: making the predictions.

2.1. Making predictions

For a model that is a point hypothesis, containing no free parameters, making a prediction is straightforward, and quantitative model evaluation is fairly easy. Just measuring the discrepancy between the unambiguous model predictions and the empirical data will do the job. Unfortunately, models in this sense are very rare. Most often, psychologists deal with models containing one or more free parameters. Rather than a single prediction, a parameterized model makes several predictions. Intuitively, one can most usefully imagine that at each different parameter value the model makes a different prediction.

The abundance of predictions of a parameterized model makes evaluating a model a non-trivial task. Model evaluation involves deciding exactly which predictions should be compared to the empirical data or, equivalently, deciding which parameter values should be used for generating the model-based data. This decision involves at least two separate issues.

The first issue concerns *how many* different parameter values should be used to make the predictions. Some measures, such as the Percentage of Variance Accounted For (PVAF), the Maximum Likelihood Criterion (MLC: e.g., Myung, 2003) and its nephews the Akaike Information Criterion (AIC: Akaike, 1973; Wagenmakers & Farrell, 2004) and the Bayesian Information Criterion (BIC: Schwarz, 1978; Wagenmakers, 2007), consider the model's prediction at a single parameter value only. Other measures instead rely on a broad range of predictions across different parameter values. Examples of such measures include the marginal likelihood (Jeffreys, 1935; Kass & Raftery, 1995; Lee & Wagenmakers, 2005; Myung & Pitt, 1997), the Deviance Information Criterion (DIC: Myung, Karabatsos, & Iverson, 2005; Spiegelhalter, Best, Carlin, & van der Linde, 2002), the Posterior Likelihood Ratio (PLR: Aitkin, Boys, & Chadwick, 2005) and the Posterior Predictive Loss Criterion (PPLC: Gelfand & Ghosh, 1998). Typically, the overall fit of the model is taken to be the average fit across all of the parameter values considered.

The second issue concerns *which* (set of) parameter value(s) should be used to generate the model-based data. One approach is to consider all parameter values that are allowed by the model. A prominent example of this approach is the marginal likelihood. An alternative approach is to only consider those parameter value(s) that could have generated the empirically observed data. Measures adopting this approach include PVAF, MLC, AIC, BIC, DIC, PLR, and PPLC. Note that in the first approach the predictions are made without having a look at the empirical data, whereas the second approach is data-informed in the sense that the predictions are based on the observed data.

The first distinction, of single versus multiple predictions, has often been characterized as one between *local* versus *global* (e.g., Navarro, Pitt, & Myung, 2004; Pitt, Kim, Navarro, & Myung, 2006; Pitt, Myung, Montenegro, & Pooley, 2008). This paper is concerned with the second distinction, of data-uninformed versus data-informed predictions. In line with the terminology used in the statistical literature, a method relying on data-uninformed predictions will be referred to as *prior*, and a method making use of data-informed predictions will be referred to as *posterior*.[1] Prior refers to the fact that the predictions are made before having seen the data, while posterior indicates that the predictions are made after having seen the data.[2]

2.2. Semi-posterior measures

There is a sense in which a posterior measure of model evaluation uses the same data twice, in both steps of the model evaluation. In the first step, the observed data are used to select the parameter values that generate the model's predictions. In the second step, the predictions are compared to the very same data used to make the predictions. This is not necessarily a bad practice, but it highlights the possibility of yet a third approach.

In this approach, which can be called *semi-posterior*, the empirical data are split up in two parts, and each part is used in one of the two model evaluation steps only. The first part of the data is used to find the parameter values that could have generated the (first part of the) data. The predictions are made based on these data-informed parameter values. Rather than comparing these predictions to the first part of the data again, they are compared to the second part of the data. Thus, making the predictions relies on the *calibration data*, evaluating the predictions relies on the *validation data*.[3]

The best known example of the semi-posterior approach is cross-validation (CV: Stone, 1974). The generalization criterion (GC: Busemeyer & Wang, 2000; Mosier, 1951) is a variation to CV in the sense that, unlike CV, it requires the calibration and validation data to be collected using different experimental designs. In principle, both CV and GC can be applied in both a local and a global fashion, but most applications seem to rely on a single prediction.

2.3. Blurring the boundaries

Model evaluation measures differ in (at least) two dimensions: (1) do they measure the (mis)match between the validation data and a single model prediction (*locally*, based on a single parameter value) or between the validation data and several model predictions (*globally*, based on multiple parameter values); and (2) are the predictions made without calibration data (*prior*) or with reliance on calibration data (*posterior* if the calibration data are the same as the validation data or *semi-posterior* if they are different)? Table 1 shows a schematic classification of different commonly used measures for model evaluation according to these distinctions.

The classification presented in Table 1 is somewhat of a simplification, in the sense that it ignores the fact that some of the measures are sometimes used as a mixture of a prior and a posterior measure. For example, some applications of the MLC rely on a prediction at a parameter value that is selected based on observed data, but from within a certain a priori defined range (e.g., Myung, 2003). Used in this fashion, MLC is not purely

[1] This terminology has not always been used consistently. For example, Liu and Aitkin (2008) somewhat confusingly use "global" and "local" to refer to "prior" and "posterior" methods.

[2] A prediction made after having seen the data does not map well onto our common sense, everyday notion of what a prediction is. In everyday language, prediction tends to refer to, in this terminology, a prior prediction. A posterior prediction, made after having seen the data, can be called a postdiction.

[3] From this perspective, a posterior measure corresponds to the extreme case where the calibration data and the validation data are the same.

Please cite this article in press as: Vanpaemel, W. Prior sensitivity in theory testing: An apologia for the Bayes factor. Journal of Mathematical Psychology (2010), doi:10.1016/j.jmp.2010.07.003

Table 1
Classification of common quantitative measures for model evaluation.

	Local	Global
Prior	–	Marginal likelihood
Posterior	PVAF, MLC, AIC, BIC	DIC, PLR, PPLC
Semi-posterior	CV, GC	–

Note: PVAF is the Percentage of Variance Accounted For; MLC is the Maximum Likelihood Criterion; AIC is the Akaike Information Criterion; BIC is the Bayesian Information Criterion; DIC is the Deviance Information Criterion; PLR is the Posterior Likelihood Ratio; PPLC is the Posterior Predictive Loss Criterion; CV is the Cross-Validation; GC is the Generalization Criterion.

posterior, since the prediction is not exclusively determined by the observed data, but it is also not purely prior, since the prediction is made after having seen the data. Similarly, some applications of the marginal likelihood assume a uniform prior over a range that is chosen to include most of the mass of the likelihood function (e.g., Lee, 2004). Used in this fashion, the marginal likelihood is not purely prior, since the "prior" is chosen by looking at the observed data, but it is also not purely posterior, since it considers parameter values that did not generate the data.

The fact that the marginal likelihood can be used as a mixture of a prior and a posterior measure underscores that the classification in Table 1 of the marginal likelihood as a prior measure indicates its common use, rather than its nature. In fact, one could not only consider a prior marginal likelihood (i.e., without calibration data) but also a posterior marginal likelihood (i.e., where the calibration data and the validation data are the same) and different sorts of semi-posterior marginal likelihoods (i.e., where the calibration and validation data differ).[4] A unifying perspective on these and other variations to the marginal likelihood, as well as investigations of the asymptotic behavior of these measures, is provided by Gelfand and Dey (1994). Similarly, measures that are typically used in a posterior fashion, such as DIC, could as well be used in a prior or semi-posterior fashion. Since yet little is known about the performance of such variations, their good performance should not be taken for granted and deserves careful study.

3. Formalizing psychological theory

3.1. Priors express theory

One risk of relying on quantitative measures as the ones discussed in the previous section is that it is easy to lose sight of the purposes for which models are evaluated. It is, however, critical to recognize that model evaluation serves different goals, depending on the type of model that is evaluated. It is useful to distinguish between two types of formal models that are currently in use in psychology. The distinction of interest concerns whether or not the model implements a psychological theory (see, e.g., Kruschke, 2010; Taagepera, 2007).

Descriptive models are generally devoid of any psychological theory. Most off-the-shelf statistical models, such as regression models and generalized linear models are of this type. The goal of evaluating such a theory-free, generic model is to extract information available in the *data*. For example, researchers have evaluated formal statistical models to investigate whether typically developing children outperform children with Attention Deficit Hyperactivity Disorder on the Wisconsin Card Sorting Test (Geurts, Verté, Oosterlaan, Roeyers, & Sergeant, 2004; Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010). In this approach, the model is used for *data analysis* (e.g., Gelman, Carlin, Stern, & Rubin, 2004; Rouder, Speckman, Sun, Morey, & Iverson, 2009; Wetzels, Raaijmakers, Jakab, & Wagenmakers, 2009).

Explanatory models, in contrast, quantitatively instantiate theories. Most psychological process models are explanatory, detailing, for example, the processes and mechanisms that the human brain uses to learn, perceive, make decisions, and so on. The goal of evaluating such a theory-laden model is to provide information about the *model* and the theory it represents. Model and data are brought in contact with each other with the purpose of theory testing—assessing to what extent the psychological theory formalized in the model can be supported.[5] For example, researchers have evaluated formal psychological models to investigate whether people learn a category by abstracting information from the encountered category examples (see Nosofsky, 1992; Vanpaemel & Storms, 2010, for overviews). In this approach, the data are used for *model analysis* (e.g., Pitt et al., 2006).

When building a formal psychological model (for example, by formalizing a verbal theory), typically most effort is devoted to come up with the model equation—a function that describes how psychological variables give rise to behavior on a task. Generally, the behavior the model attempts to represent is not fundamentally understood, and the model builder is forced to represent the psychological variables in the model equation by free parameters. Although free parameters are included in the model because the model builder is faced with unknown variables, it is a mistake to assume that free parameters are *completely* unknown. Psychological theorizing might not be advanced enough to fix each free parameter to a single value, but often at least some knowledge, theory, assumptions, or intuitions about the variables represented by the parameters exists.

Crucially, existing theory about parameters is not easily expressed in the model equation. Often, this sort of knowledge should be captured in the *range* of the parameters, indicating which parameter values are allowed and which are forbidden, and in the *distribution* over the parameters, indicating which parameter values of the allowable parameter variation are likely and which are unlikely. Since, like the model equation, the parameter range and distribution are specified before the data are observed, they are often referred to as the prior range and the prior distribution, or jointly, the prior.

In sum, theorists interested in formally instantiating a theory do not only have the model equation at their disposition, but can also use the range and the distribution over the parameters to formalize theoretical knowledge, intuitions or assumptions that are otherwise difficult to incorporate in a model. To support the claim that the model equation is not the only vehicle for formalizing psychological theory but that also the prior can, I focus on two simple examples considered by Liu and Aitkin (2008) in a recent criticism on the Bayes factor.

3.2. Decision maker

Liu and Aitkin (2008) discuss two hypotheses about a decision maker. The first hypothesis states that the decision maker is indifferent between two alternatives, which is formally expressed

[4] Within this semi-posterior approach, one can further distinguish between calibration and validation data coming from the same experimental design or from different experimental designs. In statistics, the same-design semi-posterior counterpart of the prior Bayes factor is known as the partial Bayes factor (O'Hagan, 1995), and its posterior counterpart is known as the posterior Bayes factor (Aitkin, 1991).

[5] It is meaningless to ask if a psychological model is true or not. A model that instantiates a verbal theory is, at best, a relatively rough approximation to an infinitely complex reality. Busemeyer and Diederich (2009, p. 6) note that all psychological models are "deliberately constructed to be simple representations that only capture the essentials of the cognitive systems. Thus, we know, a priori, that all models are wrong in some details".

as the rate underlying the decision being equal to $\frac{1}{2}$. The second hypothesis assumes that the decision maker can be biased, which is formally expressed by the rate being anywhere between 0 and 1. Consider now the common situation where one of the two alternatives is correct and the other is incorrect (imagine, for instance, a recall memory test). A third hypothesis that in this context might deserve interest is that the decision maker performs better than chance. This hypothesis is formally expressed by the rate being anywhere between $\frac{1}{2}$ and 1.

Crucially, the model equations of these three models are identical. The different theoretical positions between an indifferent, a biased, and an above-chance decision maker are expressed by means of the parameter range. The biased hypothesis imposes the common sense boundaries of zero and one. The above-chance hypothesis cuts this range in half. And the indifference hypothesis restricts the range to a single value. This is a very simple example of how theory can be expressed outside of the model equation.

3.3. Retention

Liu and Aitkin (2008) further discuss a study by Lee (2004), who focuses on five retention functions, each relying on two parameters, b and m, to specify the proportion of correct recall at time t, $p_c(t, b, m)$. The model equations are given by

$$p_c(t, b, m) = b - mt, \tag{1}$$

$$p_c(t, b, m) = \frac{1}{mt + b}, \tag{2}$$

$$p_c(t, b, m) = b \exp(-mt), \tag{3}$$

$$p_c(t, b, m) = b - m \ln t, \tag{4}$$

and

$$p_c(t, b, m) = bt^{-m}, \tag{5}$$

for the linear, hyperbolic, exponential, logarithmic, and power model, respectively. Further, Lee (2004) assumes a uniform distribution over the following intervals:

$$m \in [0, 2] \tag{6}$$

and

$$b \in [0, 2]. \tag{7}$$

The intuitions captured in these models are at odds with basic common sense. In particular, the models can predict proportions of correct recall that exceed one, which does not seem to make a lot of sense. For example, when $t = 0$, the power and logarithmic models predict, irrespective of the exact value of b and m, the proportion of correct recall to be infinite. Further, when $t = 0$, the exponential model makes the impossible prediction that, for any value of b above one, the proportion of correct recall is greater than one.[6]

Consider now the following redefinitions of the five models:

$$p_c(t, b, m) = \max(0, b - mt), \tag{8}$$

$$p_c(t, b, m) = \frac{b}{mt + 1}, \tag{9}$$

$$p_c(t, b, m) = b \exp(-mt), \tag{10}$$

$$p_c(t, b, m) = \max(0, b - m \ln(t + 1)), \tag{11}$$

$$p_c(t, b, m) = b(t + 1)^{-m}, \tag{12}$$

with

$$m > 0, \tag{13}$$

and

$$b \in [0, 1]. \tag{14}$$

The predictions of these models accord with common sense, as the proportion of correct recall p_c cannot exceed one at any time, nor can it drop below zero, irrespective of whether the retention was actually measured at that time.

Importantly, the models in Eq. (8) to (14) do more than simply avoiding impossible predictions. In particular, the range of m expresses a basic psychological intuition: Restricting m to positive values assures that $\frac{\partial p_c(t,b,m)}{\partial t} < 0$, for all values of b and m. The first derivative being negative implies that the models assume decreasing retention curves. Thus, the range of m, in Eq. (13), formally captures the basic psychological assumption that people forget.

Also the range of b can be used to express a psychologically meaningful position. To see this, it is useful to consider the model equations at the boundary condition of $t = 0$ (i.e., immediately after the study phase). For all models, we find $p_c(0, b, m) = b$, indicating that b can be interpreted as the initial proportion of correct recall. In this light, changing Eq. (14) to $b \in [\frac{1}{2}, 1]$ expresses the psychological intuition that people have actually learned something during the study phase and will perform above chance.

3.4. Informative priors as an antidote to the Greek letter syndrome

Model builders often suffer from what can be called the *Greek letter syndrome* (Lindley, 1999). Sufferers do not consider the meaning of a parameter and thus ignore the existing theory about the variable the parameter represents. For them, the parameter is just a (often Greek, in the retention example Roman) letter. It is, however, crucial to appreciate that the free parameters in explanatory models correspond to variables with a theoretical interpretation. Because parameters are theoretically meaningful, there are often intuitions about which parameters values are allowed, likely, unlikely, or forbidden before data have been observed. This kind of information can be expressed in the parameter prior, in the form of a range and a distribution.[7] Thus, rather than regarded as a nuisance necessary to get the Bayesian modeling machinery going, the prior should be embraced by model builders as an additional opportunity to express theory.

The realization that the prior expresses psychological theory brings about an increased responsibility for the model builder. When building a model, all aspects of the model, not just the model equation, should be given careful thought. Rather than being lazy and imprecise about the parameter range and distribution, model builders cannot shirk the responsibility of thoughtfully specifying the prior, articulating their ideas and intuitions about the psychological variables in the model. At the very least, the prior should be such that it gives zero prior weight to impossible predictions, such as proportions smaller than zero or larger than one. At best, the prior should be chosen to reflect psychological theory and intuitions, such as the assumption that people forget or that people perform better than chance after a study phase. Not having any psychological theory about the variables represented by the parameters, or not caring to translate it results in a prior with a broad range and a diffuse distribution. In contrast, using the range and distribution to express theory gives rise to a prior with a narrow range and a sharply peaked distribution. Such a prior is known as an informative prior. Thus, the status of the prior as a bearer of theory calls for informative priors, whenever possible.

[6] See also Myung (2003, Equation 13) for a similar model definition allowing impossible predictions.

[7] The simple examples given above illustrate how theory can be captured in the range. In a similar spirit, the distribution of parameters can also translate theoretical assumptions. For example, Vanpaemel & Lee (submitted for publication) demonstrate how the idea of optimal attention allocation, which is one of the core psychological assumptions of the Generalized Context Model of category learning (Nosofsky, 1986), can be captured using the prior distribution over the model's attention parameters.

In conclusion, I have argued that, when dealing with content-rich models, the prior captures theory. This implies that specifying an informative prior is a crucial part of the model building process, as crucial as constructing the model equation. One might object that model builders do not know whether the theory expressed in their prior is correct. This is definitely the case, but exactly the same observation holds for the model equation. The goal of building a formal model is to make it possible to quantitatively evaluate whether intuitions and theories are viable. For theorists who know that their ideas are correct, there is no point in engaging in theory testing. Another reasonable concern with the position that priors are vehicles for expressing theory is that the specification of the prior is a delicate task. But again, so is the specification of the model equation. No formal guidelines about how to capture theory into a prior exist, just like there are no formal guidelines about how to capture theory into a model equation. Model building crucially depends on the skill and the creativity of the modeler and cannot be automated.

4. Prior sensitivity in theory testing

Prior and posterior measures differ in their sensitivity to the prior. In a prior measure, the parameter values that generate the model's predictions are drawn from the prior distribution over the parameters. Consequently, the predictions, and hence the model evaluation relying on these predictions, are, by design, sensitive to the prior. In a posterior measure, in contrast, the prediction-generating parameter values are drawn from the posterior distribution, which is the data-updated version of the prior distribution. In principle, the posterior distribution, and hence the posterior predictions and the model evaluation relying on these predictions, is affected by the exact choice of the prior. However, as data provide sufficient information, they overwhelm the prior, and the posterior is hardly influenced by the prior. Thus unlike prior measures, posterior measures are not always sensitive to the prior.

Even among adherents of Bayesian methods, the sensitivity of prior measures to the choice of the prior has been a serious source of concern (e.g., Kass & Raftery, 1995; Myung & Pitt, 1997). However, in the light of the argument of the previous section that in theory-laden models the prior can capture psychological theory, the prior sensitivity of prior measures is unproblematic and even desirable. In particular, in this section I argue that when priors are judiciously constructed to capture psychological theory, being sensitive to the prior is an asset rather than a nuisance. To provide a complete test of the theory formalized in a psychological model, a model evaluation measure should be sensitive to all the different aspects of the model that express psychological theory: the model equation, the parameter range, and the parameter distribution.

4.1. *Prior measures are unproblematic*

To illustrate the issue of prior sensitivity in model evaluation, consider the exponential retention model assuming $[0, 2]$ as the range for b. When a prior measure is used to evaluate this model, it will be evaluated on the impossible prior predictions it generates, such as a proportion of correct recall larger than one. Since the observed proportion will, of course, be smaller than one, the prior prediction will not be confirmed and the model will be penalized. Thus when the assumption of exponential retention is tested by means of a formal model that assumes $[0, 2]$ as the range for b, the prior measure's sensitivity to the prior leads to an unfair test of this assumption.

When the same model is evaluated using a posterior measure and data are sufficiently informative, the only predictions that are used in the evaluation are those that can generate the observed data. So despite the fact that the model can make impossible predictions, the impossible predictions are most likely not considered when the model is evaluated, and the model is not penalized for its impossible, and thus wrong, predictions. Thus when the assumption of exponential retention is tested by means of a formal model that assumes $[0, 2]$ as the range for b, the posterior measure's insensitivity to the prior leads to a fair test of this assumption.

In this light, it might seem that prior measures are problematic and should be abandoned in favor of posterior measures (e.g., Liu & Aitkin, 2008). There is, however, a second remedy to make sure that a model is not unfairly penalized for making impossible predictions. Quite simply, it involves making sure the model does not make impossible predictions. As mentioned earlier, if the range of b is restricted to $[0, 1]$ rather than to $[0, 2]$, the exponential retention model stops making impossible predictions. When this model is used to represent the assumption that people forget exponentially, thehe prior measure no longer provides an unfair test of this assumption. Thus, when a prior is cautiously picked to translate plausible theory, a prior measure of model evaluation is not problematic at all.

4.2. *Prior measures are desirable*

As a second example, consider two different exponential retention models, one that assumes that people forget and one that is agnostic about whether people actually forget. The different theoretical positions taken by these models are expressed by their priors. The first model assumes an informative prior by restricting m to positive values only. The second model, in contrast, allows m to be any real number.

Suppose that sufficiently informative data are observed, such that the prior has been overwhelmed by the data. Further, suppose that the observed proportions of correct recall decrease over time, indicating that people forget. In this situation, prior and posterior model evaluation measures provide different conclusions. Using a posterior measure, both models perform similarly to each other. There is no preference for the informative model, which assumes that people forget, or for the other model, which is mute about forgetting. This state of affairs contrasts markedly with prior measures. Using a prior measure, the model assuming the observed behavior (i.e., people forget) is rewarded for its correct intuition, and is preferred over the model that did not take a stand.

In sum, the posterior measure is insensitive to one of the core assumptions of the retention models: whether or not people forget. Although both models instantiate a different theory, they perform alike under a posterior measure. The posterior measure provides an incomplete test of the theory embodied in the models. The prior measure, in contrast, provides a complete test, as it is highly sensitive to the theoretical position the retention models adopt about retention.

4.3. *Prior measures encourages precise models*

The previous example brings to light that a theorist using a posterior measure to evaluate a model can take the prior range of any parameter, whatever its psychological meaning, to equal the full real line and the prior distribution to be any vague distribution, such as the uniform distribution. By observing data and updating this prior, the correct range and distribution of the parameters is found, and these are used for assessing the fit between model and data. No pressure is laid on the theorist to come up with a theoretically motivated range and distribution; all the hard work is done by the data. In contrast, when a model is evaluated using a prior measure, the theorist needs to be much more judicious about the priors. Assigning a non-zero or high prior weight to parameter values that lead to bad fits will damage the model's

Please cite this article in press as: Vanpaemel, W. Prior sensitivity in theory testing: An apologia for the Bayes factor. Journal of Mathematical Psychology (2010), doi:10.1016/j.jmp.2010.07.003

average performance. The upside is that if these high-weighted parameter values generate predictions that are confirmed by the data, the model is rewarded for the risk it took.

Posterior measures provide theorists with no incentive to think carefully about their parameters and formalize precise intuitions about them. With sufficiently informative data, it is possible to get away with imprecise intuitions about the parameters. The retention theorist can gain nothing by positing that retention increases or decreases. Worse yet, the theorist can only lose by doing so: The more vague a model is about its parameter values, the more likely it is to perform well on a posterior measure. Thus, posterior measures encourage vague and empty models. Prior measures, in contrast, encourage theorists to think carefully and to take a stand. The retention theorist is forced to think about whether retention actually increases, decreases, or can do both. Since prior measures encourage the practice of building models with specific, strong assumptions, prior measures increase knowledge and advance science.

4.4. Conclusion

Both examples illustrated that posterior measures can be insensitive to the prior. Prior measures, in contrast, are, by design, always sensitive to the prior, regardless of the number of data. In the first example, the prior insensitivity of the posterior measure was helpful, since the theory captured in the prior was absurd—it allowed proportions to be smaller than zero and larger than one. However, the observation that the retention models formalized in Eq. (1) to (7) make prior predictions that are clearly false should not be taken as an argument to abandon prior measures in favor of posterior measures. Instead, it should be taken as an encouragement to be more thoughtful when formalizing a prior.

In the second example, the prior insensitivity of the posterior measure was unwanted, since it implied that the posterior measure was insensitive to one of the core psychological assumptions of the models under consideration. The posterior measure provided an incomplete test, and encouraged vague and weak models. Thanks to its sensitivity to the prior, the prior measure provided a complete test and encouraged precise, strong models.

A model evaluation measure that is always sensitive to all theoretical assumptions embodied in the model seems to contrast favorably with one that is not guaranteed to be sensitive to all of the theory. A model evaluation measure that encourages theorists to be precise and to articulate their stance seems to be more attractive than one that encourages theorists to be vague and theoretically empty. For these reasons, prior measures are to be preferred over posterior measures when evaluating psychological models.

5. Sensitivity analyses in theory testing

Because the marginal likelihood is known to be highly sensitive to the prior, it is often stressed that it is important to check that conclusions based on the marginal likelihood are not too sensitive to the choice of the prior, in the form of a sensitivity analysis (Kass & Raftery, 1995; Myung & Pitt, 1997; Sinharay & Stern, 2002). I agree that sensitivity analyses to arbitrary assumptions are critical when engaging in model evaluation. However, the recommendation of performing sensitivity analyses to the prior when using the marginal likelihood seems to rest on three assumptions that need correction: the prior is always an unavoidable source of arbitrariness; only users of prior measures should perform a sensitivity analysis to the prior; and the prior is the major (or even only) source of arbitrariness. In this section, I clarify that as the prior, just like the model equation, expresses psychological theory, it is not necessarily arbitrary and thus not necessarily requires a sensitivity analysis. Further, I argue that much more than users of prior measures, users of posterior measures should perform sensitivity analyses to the prior. Finally, I indicate that many other aspects of theory testing involve arbitrary assumptions and that, much more than the prior, these aspects warrant a sensitivity analysis.

5.1. Sensitivity to arbitrary assumptions in model building

Model building – formalizing theory and intuitions into a formal model – is a non-trivial undertaking and often requires making ad hoc or arbitrary assumptions. The reason is that the theory the formal model is designed to represent is not always precise enough to give rise to a single and unique formal model. Arbitrary assumptions can arise in any place where theory is expressed: the range, the distribution, and the model equation. For example, different theorists have all attempted to evaluate "the" hyperbolic model of retention, but a quick glance at the literature shows that at least three different model equations have been used to formalize the assumption that retention occurs hyperbolically (Cavagnaro, Pitt, & Myung, 2009; Lee, 2004; Navarro et al., 2004). Just like the model equation, the prior can also reflect ad hoc or arbitrary assumptions. For example, as noted earlier, when evaluating the retention models, Lee (2004) picked two as the upper bound of the m parameter, without any theoretical, psychologically motivated justification for this choice.[8] As there was no principled reason the upper bound of m was not, say, three, the choice of two is clearly an arbitrary one.

Arbitrariness in model building should inspire caution. Ideally, conclusions should be independent of any arbitrariness involved in the models, and it is the responsibility of the researcher to check the robustness of the results against arbitrary decisions. Consequently, researchers interested in evaluating the assumption of hyperbolic retention should consider different model equations translating this assumption, unless they have non-arbitrary, well-motivated reasons to prefer one model equation over the others. By the same token, if the prior involves an arbitrary decision, a sensitivity analysis should be performed. For example, the upper bound of m is an arbitrary choice and the sensitivity of the conclusion to this choice should be assessed.[9]

However, it is a mistake to assume that a prior is inevitably ad hoc. For example, choosing zero as the lower bound of m is not a theoretically-empty, arbitrary whim. Rather, it represents a clear theoretical position: It translates the core assumption that people forget, as negative values of m would allow increasing retention curves. Likewise, picking $[\frac{1}{2}, 1]$ as the range for b is not an arbitrary decision. It is a theoretically-motivated choice, because it translates the assumption that after a study phase people will perform above chance (by the lower bound), and will not correctly classify more items than they are tested on (by the upper bound). If the prior involves a theoretically motivated decision, a sensitivity analysis is not required.[10]

Further, relying on a posterior model evaluation measure does not exempt the researcher from performing a sensitivity analysis to the prior. Posterior measures may be less influenced by the

[8] Lee (2004) motivated this choice by sneak-peeking at the data rather than by theoretical considerations.

[9] Fig. 1 in Lee (2004) provides such an analysis, as it indicates that extending the range of m beyond two does not result in a significant increase of mass of any of the models.

[10] Obviously, the fact that the range of a parameter is non-arbitrary should not be taken to mean that this particular range is correct, or beyond scrutiny. Indeed, the underlying theory it expresses might be wrong. Non-arbitrary just means that it expresses a (possibly wrong) theory.

prior than prior measures, but they are not totally insensitive to the prior, especially when data sets are small. Moreover, I suspect that most users of posterior measures do not consider the prior as a vehicle for expressing theory—if they did they would rely on prior measures. Without a commitment to the prior as a bearer of theory, the prior necessarily corresponds to an atheoretical and hence arbitrary decision and a sensitivity analysis to the prior is necessary. Thus it seems that especially the users of posterior measures, rather than the users of prior measures, should make sure that their conclusions do not depend too heavily on the choice of the prior.

5.2. Sensitivity to arbitrary assumptions outside model building

The above discussion focused on arbitrariness in model building. Of course, theory testing involves more that only this single step, and all the additional steps are also prone to ad hoc assumptions. The conclusions based on both prior and posterior measures can be sensitive to these arbitrary assumptions, and ideally, their robustness should be checked.

First, theory testing involves a host of technical decisions, which are at best governed by rules of thumb. Examples include the choice of the optimization or sampling algorithm, the starting points of the algorithm, the number of MCMC chains, and the number of (burned-in and recorded) samples in a chain. To the extent that these decisions are arbitrary, they should be subject to a sensitivity analysis.

Secondly, theory testing often involves the choice of a likelihood function. For example, for the retention models, both a Gaussian and a Binomial likelihood seem justifiable (see Lee, 2004; Myung, 2003, for both uses), and the use of either of these can involve an arbitrary decision. Even settling for a Gaussian likelihood leaves one with a choice of the variance. As this choice most likely involves some degree of arbitrariness, a check of the robustness of the conclusion against different choices for the variance is necessary.[11]

Finally, theory testing also involves experimental design, data collection, truncation, and transformation, which all involve various ad hoc decisions. For example, with the retention models, there seems to be little theoretical ground to decide whether or not the retention interval should be scaled to lie between zero and one, or should instead reflect the exact measurements. Since scaling or not influences the conclusion, this decision either calls for a theoretical argument of why one choice is preferable over the other or for a sensitivity analysis. Further, the range, number and spacing of the time points, the difficulty of the task, and the exact experimental methodology (e.g., free recall vs cued recall) all involve more or less ad hoc decisions. Consequently, these choices should be motivated or their effects on the conclusion should be checked. A motivation for choices of experimental design can derive from design optimization (Myung & Pitt, 2009). A sensitivity analysis can consist of performing a meta-analysis using a range of data sets spanning a wide variety of conditions (see, for examples, Lee, 2004; Navarro et al., 2004; Rubin & Wenzel, 1996).

6. Letting the model speak for itself

Progress in psychology is becoming increasingly reliant on formal models. The increasing popularity of formal models makes it all the more important to have a clear understanding of the roles and purposes of models. An important distinction between formal models is their theoretical content. Psychology, like most empirical sciences, relies on generic, theory-free models, as well as on hand-crafted theory-rich models, designed to represent a psychological theory. The basic claim of this paper is that in these latter psychological models, parameters are not just random, unknown numbers. Rather, they represent psychological variables about which knowledge, expectations, assumptions, theory, or intuitions exist before data are observed. This kind of information can be expressed by carefully specifying the prior range and prior distribution over the parameters, indicating the values parameters are likely to take, and the relationship between them. In its capacity of representing theory, the prior is not an arbitrary whim. The prior is an integral part of the model that serves the exact same function as the model equation: Formally translating the theorist's assumptions to make them amenable to quantitative test.

This paper explored the implications of the viewpoint that the prior can be used as a vehicle for psychological theory, both in model building and in model evaluation. In particular, I have stressed that, when building a model, the prior needs to be given much more careful thought than is currently practiced and that, when evaluating a model, sensitivity to the prior is necessary.

As a bearer of psychological theory, the prior should be carefully chosen and, ideally, be informative. Model builders restricting their attention only to the model equation do only half of the work. They should also devote their full attention to the range (indicating which parameter values are allowed and which are not) and to the distribution (indicating which parameter values are likely and which are not) of the parameters, rather than being happy with a perfunctory definition of the prior. Thinking of a parameter as a psychologically meaningful variable and translating psychological theory about this variable in an informative parameter prior leads to a strong and precise model.

As a bearer of psychological theory, the prior should influence the model evaluation. Model evaluation measures that are only sensitive to the model equation do only half of the work. The evaluation of a model should be informed by all the assumptions the model is making, not just by the assumptions that happen to be expressed in the model equation. Only a model evaluation measure that is also sensitive to the prior provides a complete test of the theory embodied in the model (see Vanpaemel, 2009, for the related observation that also a measure of model complexity is only complete if it is sensitive to the model equation, the prior range and the prior distribution).

This paper investigated two main approaches to model evaluation. In prior measures, such as the marginal likelihood, the parameter values used to make predictions are drawn from the prior, while in posterior measures, the parameter values are drawn from the posterior. Prior measures are, by design, highly sensitive to the prior, whereas posterior measures are not always guaranteed to be sensitive to the prior. Recognizing that quantitative model evaluation measures should be sensitive to the prior, this difference in prior sensitivity implies that prior measures are more suitable for evaluating psychological models than posterior measures.

In data analysis, priors have been severely criticized, out of a desire to let the data speak for themselves. Using a prior adds "too much model", and biases the (data) analysis. This paper does not deal with data analysis, but rather with model analysis. In terms of how models and data can interact, model analysis is the exact opposite of data analysis. In data analysis, a model is used to learn something about the data. In model analysis, the relationship between models and data is reversed: The data are used to learn something about the model. Perhaps ironically, my criticism on the use of posteriors in model analysis is the mirror image of the criticism on the use of priors in data analysis. Using a posterior adds "too much data", and biases the (model) analysis. In model analysis, we should let the models speak for themselves.

[11] An example of such a sensitivity analysis is provided in Figs. 4 and 5 of Lee (2004), showing how inference changes across a range of choices for the variance in the Gaussian likelihood.

Please cite this article in press as: Vanpaemel, W. Prior sensitivity in theory testing: An apologia for the Bayes factor. Journal of Mathematical Psychology (2010), doi:10.1016/j.jmp.2010.07.003

Acknowledgments

The research reported in this article was funded by Grant FWO G.0513.08. I wish to thank Michael Lee, Francis Tuerlinckx, Liz Bonawitz, Gert Storms, Charles Liu and an anonymous reviewer, whose comments have substantially improved the article. The usual disclaimer applies.

References

Aitkin, M. (1991). Posterior Bayes factors. *Journal of the Royal Statistical Society. Series B*, *53*, 111–142.

Aitkin, M., Boys, R., & Chadwick, T. (2005). Bayesian point null hypothesis testing via the posterior likelihood ratio. *Statistics and Computing*, *15*, 217–230.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. Petrov, & B. Csaki (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest: Academiai Kiado.

Anderson, D., Link, W., Johnson, D., & Burnham, K. (2001). Suggestions for presenting the results of data analyses. *The Journal of Wildlife Management*, *65*, 373–378.

Bayarri, M. J., & Berger, J. O. (2000). *P* values for composite null models. *Journal of the American Statistical Association*, *95*, 1127–1142.

Busemeyer, J. R., & Diederich, A. (2009). *Cognitive modeling*. Sage.

Busemeyer, J. R., & Wang, Y. (2000). Model comparisons and model selections based on generalization criterion methodology. *Journal of Mathematical Psychology*, *44*, 171–189.

Cavagnaro, D. R., Pitt, M. A., & Myung, J. I. (2009). Adaptive design optimization in experiments with people. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems 22* (pp. 234–242).

Efron, B. (1986). Why isn't everyone a Bayesian? *The American Statistician*, *40*, 1–5.

Gelfand, A., & Dey, D. (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society. Series B*, *56*, 501–514.

Gelfand, A., & Ghosh, S. (1998). Model choice: a minimum posterior predictive loss approach. *Biometrika*, *85*, 1–11.

Gelman, A. (2008). Objections to Bayesian statistics. *Bayesian Analysis*, *3*, 445–450.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis*. London: Chapman & Hall.

Geurts, H., Verté, S., Oosterlaan, J., Roeyers, H., & Sergeant, J. (2004). How specific are executive functioning deficits in attention deficit hyperactivity disorder and autism?. *Journal of Child Psychology and Psychiatry*, *45*, 836–854.

Grünwald, P. (2000). Model selection based on minimum description length. *Journal of Mathematical Psychology*, *44*, 133–152.

Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. In *Proceedings of the Cambridge philosophical society* (pp. 203–222).

Kass, R. E. (1993). Bayes factors in practice. *Journal of the Royal Statistical Society. Series D*, *42*, 551–560.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Koop, G., & Potter, S. (1999). Bayes factors and nonlinearity: evidence from economic time series. *Journal of Econometrics*, *88*, 251–281.

Kruschke, J. K. (2010). What to believe: bayesian methods for data analysis. *Trends in Cognitive Sciences*, *14*, 293–300.

Lee, M. D. (2004). A Bayesian analysis of retention functions. *Journal of Mathematical Psychology*, *48*, 310–321.

Lee, M. D., & Wagenmakers, E. J. (2005). Bayesian statistical inference in psychology: comment on Trafimow (2003). *Psychological Review*, *112*, 662–668.

Lindley, D. V. (1999). Comment on "Nested hypothesis testing: the Bayesian reference criterion" by J. M. Bernardo. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics 6* (pp. 122–124). Oxford University Press.

Liu, C. C., & Aitkin, M. (2008). Bayes factors: prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, *53*, 362–375.

Mosier, C. (1951). Problems and designs of cross-validation. *Educational and Psychological Measurement*, *11*, 5–11.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, *47*, 90–100.

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: a Bayesian approach. *Psychonomic Bulletin & Review*, *4*, 79–95.

Myung, J. I., Karabatsos, G., & Iverson, G. J. (2005). A Bayesian approach to testing decision making axioms. *Journal of Mathematical Psychology*, *49*, 205–225.

Myung, J. I., & Pitt, M. A. (2009). Optimal experimental design for model discrimination. *Psychological Review*, *116*, 499–518.

Navarro, D. J., Pitt, M. A., & Myung, I. J. (2004). Assessing the distinguishability of models and the informativeness of data. *Cognitive Psychology*, *49*, 47–84.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57.

Nosofsky, R. M. (1992). Exemplars, prototypes, and similarity rules. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *From learning theory to connectionist theory: essays in honor of William K. Estes, Vol. 1* (pp. 149–167). Hillsdale, NJ: Lawrence Erlbaum.

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

Pitt, M. A., Kim, W., Navarro, D. J., & Myung, J. I. (2006). Global model analysis by parameter space partitioning. *Psychological Review*, *113*, 57–83.

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences*, *6*, 421–425.

Pitt, M. A., Myung, J. I., Montenegro, M., & Pooley, J. (2008). Measuring model flexibility with parameter space partitioning: an introduction and application example.. *Cognitive Science*, *32*, 1285–1303.

Rossi, P. E., & Allenby, G. M. (2003). Bayesian statistics and marketing. *Marketing Science*, *22*, 304–328.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic bulletin & review*, *16*, 225–237.

Rubin, D., & Wenzel, A. (1996). One hundred years of forgetting: a quantitative description of retention. *Psychological Review*, *103*, 734–760.

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

Shiffrin, R. M., Lee, M. D., Kim, W., & Wagenmakers, E. J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, *32*, 1248–1284.

Sinharay, S., & Stern, H. S. (2002). On the sensitivity of Bayes factors to the prior distributions. *The American Statistician*, *56*, 196–201.

Spiegelhalter, D., Best, N., Carlin, B., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society. Series B*, *64*, 583–639.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B*, *36*, 111–147.

Taagepera, R. (2007). Predictive versus postdictive models. *European Political Science*, *6*, 114–123.

Vanpaemel, W. (2009). Measuring model complexity with the prior predictive. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems 22* (pp. 1919–1927).

Vanpaemel, W., & Lee, M.D. Using priors to formalize theory: optimal attention and the generalized context model (submitted for publication).

Vanpaemel, W., & Storms, G. (2010). Abstraction and model evaluation in category learning. *Behavior Research Methods*, *42*, 421–437.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review*, *14*, 779–804.

Wagenmakers, E. J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*, 192–196.

Wagenmakers, E. J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: a tutorial on the Savage-Dickey method. *Cognitive Psychology*, *60*, 158–189.

Wasserman, L. (1996). The conflict between improper priors and robustness. *Journal of Statistical Planning and Inference*, *52*, 1–15.

Wetzels, R., Raaijmakers, J. G. W., Jakab, E., & Wagenmakers, E. J. (2009). How to quantify support for and against the null hypothesis: a flexible WinBUGS implementation of a default Bayesian *t* test. *Psychonomic Bulletin & Review*, *16*, 752–760.

Xie, Y. (1999). The tension between generality and accuracy. *Sociological Methods & Research*, *27*, 428–435.

Please cite this article in press as: Vanpaemel, W. Prior sensitivity in theory testing: An apologia for the Bayes factor. Journal of Mathematical Psychology (2010), doi:10.1016/j.jmp.2010.07.003