Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Mathematical Psychology 48 (2004) 28–50
Journal of Mathematical Psychology
http://www.elsevier.com/locate/jmp

Assessing model mimicry using the parametric bootstrap

Eric-Jan Wagenmakers,[a,*] Roger Ratcliff,[a] Pablo Gomez,[b] and Geoffrey J. Iverson[c]

[a] *Northwestern University, USA*
[b] *De Paul University, USA*
[c] *University of California at Irvine, USA*

Received 25 March 2003; revised 7 November 2003

Abstract

We present a general sampling procedure to quantify *model mimicry*, defined as the ability of a model to account for data generated by a competing model. This sampling procedure, called the parametric bootstrap cross-fitting method (PBCM; cf. Williams (J. R. Statist. Soc. B 32 (1970) 350; Biometrics 26 (1970) 23)), generates distributions of differences in goodness-of-fit expected under each of the competing models. In the data informed version of the PBCM, the generating models have specific parameter values obtained by fitting the experimental data under consideration. The data informed difference distributions can be compared to the observed difference in goodness-of-fit to allow a quantification of model adequacy. In the data uninformed version of the PBCM, the generating models have a relatively broad range of parameter values based on prior knowledge. Application of both the data informed and the data uninformed PBCM is illustrated with several examples.

1. Introduction

For many psychological phenomena under scientific study, there exist several mutually exclusive explanations. In some cases these competing explanations are formalized as quantitative models, and this allows a comparison between the competitor explanations based on their descriptive accuracy or goodness-of-fit (GOF) for the observed data. However, GOF cannot be the only criterion for model selection, as the principle of parsimony dictates that a relatively simple model is to be preferred over a complex model when the latter yields only a marginal gain in GOF. Therefore, a solution to the problem of model selection requires a quantification of the tradeoff between GOF and parsimony.

Among the traditional methods for model selection are the likelihood ratio test (LRT; Wilks, 1938), Akaike's information criterion (AIC; e.g., Akaike, 1973; Burnham & Anderson, 2002; Parzen, Tanabe, & Kitagawa, 1998; Wagenmakers & Farrell, in press), and Schwarz' Bayesian information criterion (BIC; e.g., Raftery, 1995; Schwarz, 1978). The LRT is based on a null-hypothesis framework and is almost exclusively used for nested models (but see Vuong, 1989; Golden, 1995, 2003, for a generalization to nonnested and possibly misspecified models). AIC and BIC can be applied to nested as well as nonnested models, but both these methods do not take the functional form of the model parameters into account because they define complexity solely as a function of the *number* of free parameters (e.g., Djurić, 1998; Myung & Pitt, 1997).

In this article, we argue that a choice between models should also be based on a measure that indicates the relative flexibility of the models. Specifically, we advocate examination of the extent to which the candidate models can mimic each other. Throughout this article, we will discuss the case of two nonnested models, A and B. Now suppose we generate replicate/simulated data sets from the models and find that model B is better able to account for data patterns that are in fact generated by model A than vice versa. Under certain specific conditions, the observed difference in GOF (ΔGOF) may then be in need of reinterpretation or correction as a result of bias due to mimicry. Such a bias correction reduces the tendency to prefer the 'chameleon model' B, which is able to give a relatively good account for data it did not in fact generate, in favor of the more idiosyncratic model A, which is relatively poor at accounting for data it did not

*Corresponding author. Department of Developmental Psychology, University of Amsterdam, Roetersstraat 15, Room 1001, NL-1018 WB Amsterdam, The Netherlands. Fax: +31-20-639-02-79.
E-mail address: ej@northwestern.edu (E.-J. Wagenmakers).

0022-2496/$ - see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmp.2003.11.004

generate. Another way of saying this is that we want to know how much evidence for or against model B is provided by an observed difference in GOF between model A and B. Assessment of the diagnosticity of a given difference in GOF is the focus of this article.

The issue of model mimicry has received a lot of attention recently (e.g., Collyer, 1985; Massaro, 1988, 1998; Navarro, Pitt, & Myung, 2003; Navarro, Myung, Pitt, & Kim, 2003; Ratcliff, 1988a; Ratcliff & Smith, in press; Van Zandt & Ratcliff, 1995; see also Townsend, 1972), as has the issue of model selection (e.g., Myung, Forster, & Browne, 2000; Pitt, Myung, & Zhang, 2002). Model mimicry is traditionally studied by means of confusion matrices (i.e., a confusion matrix shows the percentage of correctly and incorrectly recovered models for sets of simulated data). The present work uses the parametric bootstrap method to construct difference distributions of GOF under the hypothesis that model A is correct and under the hypothesis that model B is correct. These difference distributions allow a more informative quantification of model mimicry than is usually provided by confusion matrices.

We would like to stress that the proposed method to quantify model mimicry is very general and easy to implement. All that is needed is a program for estimating parameters from data, and a program for generating simulated data from the estimated models. This means that a quantification of model mimicry is possible even for the most complex nonlinear and nonnested psychological models such as production rule systems (e.g., Anderson & Lebiere, 1998) or connectionist models.

The outline of the paper is as follows. First, we will describe the proposed method for assessing model mimicry in some detail. This first version of this method depends on the observed data, and is labeled 'data informed' or 'local'. We illustrate the use of this method by an application to two models for information integration. Next, we discuss assumptions and extensions of the method. In particular, we will compare the proposed method both to Bayesian methods for assessing model adequacy and to Bayesian methods of model selection. This comparison leads to the formulation of a second version of the method. This second version does *not* depend on the observed data and is labeled 'data uninformed' or 'global'. Finally, we will illustrate how the proposed method can be used to aid experimental design by estimating the number of observations needed to discriminate between two competing models.

2. The data informed parametric bootstrap cross-fitting method

The bootstrap resampling method was introduced by Efron (1979). Since then, the bootstrap has been subjected to detailed study (e.g., Davidson & Hinkley, 1997; Diciccio & Romano, 1988; Efron & Tibshirani, 1986, 1993; Hall, 1988, 1992; Hinkley, 1988; Horowitz, 2001). As summarized by Hall and Wilson (1991), "The nonparametric bootstrap is a particularly versatile tool for data analysis. Its good performance in many important statistical problems has been established by theoretical analysis, by simulation study, and by application to real data." (p. 757). The bootstrap is regularly used to estimate standard errors when analytic approximations are not available or are unreliable (Golden, 1995). The broad scope of the bootstrap and its straightforward implementation make it very attractive for use in many psychological applications. So far, however, use of the bootstrap in psychology has been somewhat limited (but see Wichmann & Hill, 2001) and this is perhaps due to the fact that the method is not taught in standard statistics courses. Before outlining the data informed PBCM for the assessment of model mimicry, we will first briefly describe the bootstrap technique (for an excellent introduction to the bootstrap see Efron & Tibshirani, 1993).

Let $\mathbf{x}$ denote an i.i.d. sequence[1] of n observations, $\mathbf{x} = (x_1, x_2, ..., x_n)$. These observations originate from an unobserved probability distribution F, $\mathbf{x} \sim F$, which is estimated by the observed (empirical) distribution $\hat{F}$, $\hat{F}$ assigning probability $1/n$ to every observed value (i.e., each observation is assigned equal probability). Now suppose we are interested in a parameter θ (e.g., the mean or the correlation coefficient) which is a function of F. Generally, θ based on F is approximated by $\hat{\theta}$ calculated from $\hat{F}$ (i.e., the plug-in principle, Efron & Tibshirani, 1993, Chap. 4). The accuracy of this approximation is given by the standard error (or by confidence intervals). For many statistics of interest standard errors cannot be calculated using a mathematical formula. It is obviously of great importance to obtain a quantitative indication of statistical accuracy, and this is where the bootstrap comes in.

The basic idea of the bootstrap is that $\hat{F}$, the empirical distribution, is the best (i.e., maximum likelihood) estimator of the true distribution F. An estimation of the sampling variability for $\hat{\theta}$ can then be obtained by repeatedly sampling from $\hat{F}$. Hence, the bootstrap estimate of standard error for a parameter $\hat{\theta}$ from $\hat{F}$ is calculated as follows. Sample with replacement from the original data set $\hat{F} \rightarrow \mathbf{x} = (x_1, x_2, ..., x_n)$ to obtain a bootstrap sample $\mathbf{x}^*$ of size n (a star indicates 'bootstrap'). For instance, if $\mathbf{x} = (x_1, x_2, x_3)$, then $\mathbf{x}^*$ could be (x_1, x_1, x_3). The bootstrap sample $\mathbf{x}^*$ is analyzed as if it

[1] Throughout this article we assume all observations to be independent and identically distributed (i.e., i.i.d.). When successive observations are not independent, one option is to filter out the dependencies (e.g., by fitting a time series model) and bootstrap the whitened residuals (e.g., Efron & Tibshirani, 1986).

were the observed data $\mathbf{x}$, that is, the statistic of interest, $\hat{\theta}^*$, is calculated using $\mathbf{x}^*$. A new bootstrap sample is then selected from $\mathbf{x}$, and a new value of $\hat{\theta}^*$ is obtained. This procedure is repeated M times, resulting in M values of $\hat{\theta}^*$. The bootstrap estimate of standard error, se_{boot}, for a parameter $\hat{\theta}$ from $\hat{F}$ is then given by the standard deviation of the M values of $\hat{\theta}^*$. For an accurate assessment of standard error, M usually varies from 25 up to 200 (for details see Efron & Tibshirani, 1993, pp. 50–53). The computation of confidence intervals with good coverage is more complicated, and also requires larger M (i.e., roughly in the order of $M = 1000$; see Andrews & Buchinsky, 2000, and Davidson & MacKinnon, 2000, for a discussion on how to choose the number of bootstrap samples).

The foregoing has briefly described the *nonparametric* bootstrap, in which resampling is done on the observed data. An alternative procedure is to generate replicate or simulated data sets from a parametric model that has first been fitted to the observed data. These artificial data sets are termed parametric bootstrap samples, since they are obtained not by sampling from the observed data, but instead by 'sampling' from a parametric model. This procedure is known as the *parametric* bootstrap (or earlier as a member of the class of Monte Carlo methods, cf. Atkinson, Bower, & Crothers, 1965; Bush & Mosteller, 1955; Metropolis & Ulam, 1949; Press, Flannery, Teukolsky, & Vetterling, 1986, pp. 529–538; von Neumann, 1951; see Ratcliff (1979, 1993) and Ratcliff & Tuerlinckx (2002) for applications to psychological phenomena).

After the above preliminaries, we are now ready to outline the data informed PBCM. Suppose two models, A and B, have been estimated from a data set $\mathbf{x} = (x_1, x_2, ..., x_n)$ by optimizing the values for their parameters. Model fitting might be accomplished by maximum likelihood estimation (MLE; e.g., Stuart & Ord, 1991), or a method such as least-squares minimization. The fitting procedure will result in a set of best fitting or most likely parameters $\hat{\theta}_A$ for model A, and $\hat{\theta}_B$ for model B. Each model also has a GOF value, and this yields a difference in GOF between models A and B for the observed data: $\Delta GOF_{AB} = GOF_A - GOF_B = \delta_{AB}$. As mentioned earlier, the diagnosticity of the evidence that δ_{AB} yields for model A over model B is affected by the extent to which models A and B can mimic each other. The data informed PBCM is a method to quantify such model mimicry and consist of the following stages:

1. Sample with replacement from the observed data $\mathbf{x}$ to obtain a nonparametric bootstrap sample $\mathbf{x}^*$.
2. Fit both models A and B to the nonparametric sample $\mathbf{x}^*$, resulting in MLE parameter vectors $\hat{\theta}^*_A$ and $\hat{\theta}^*_B$.
3. Apply the parametric bootstrap to both models. That is, generate a simulated data series D from model A and model B using $\hat{\theta}^*_A$ and $\hat{\theta}^*_B$, respectively.
4. Treat the simulated data series under model A, i.e., $D(\hat{\theta}^*_A)$, and the simulated data series generated under model B, i.e., $D(\hat{\theta}^*_B)$, exactly as if they had been the observed data. That is, fit model A and model B to $D(\hat{\theta}^*_A)$, thereby obtaining parameter estimates and, more importantly, a difference in GOF given that the data were generated by model A: $\Delta GOF^*_{AB}|A = GOF^*_A - GOF^*_B$. The data simulated under model B, i.e., $D(\hat{\theta}^*_B)$ are also fitted by model A and B, and this also yields a difference in GOF: $\Delta GOF^*_{AB}|B = GOF^*_A - GOF^*_B$.
5. Repeat steps 1–4 M times. This yields a distribution of differences in GOF under model A, and a distribution of differences in GOF under model B. In the work presented here, $M = 1000$.

Thus, the data informed PBCM combines the nonparametric bootstrap (step 1) with the parametric bootstrap (step 3). Fig. 1 illustrates the procedure.

Step 1, the nonparametric bootstrap, is included in the procedure to account for uncertainty in parameter estimation. By definition, the likelihood of observing the data is highest when $\theta = \hat{\theta}$. However, other values of $\hat{\theta}$ can also be considered, although their likelihood will be smaller than the one for the point estimate. When only $\hat{\theta}$ is used to generate data for the data informed PBCM there is the risk of drawing conclusions from the results which do not hold for other plausible values of $\hat{\theta}$. More specifically, ignoring the role of sampling variability in parameter estimation can lead to overly optimistic predictions and assessments, especially when the distribution of parameter values is not highly peaked around its maximum value—a situation that is especially likely when only few data are available and the models under consideration are nonlinear. Aitchison & Dunsmore (1975, pp. 227–234) discuss and illustrate this

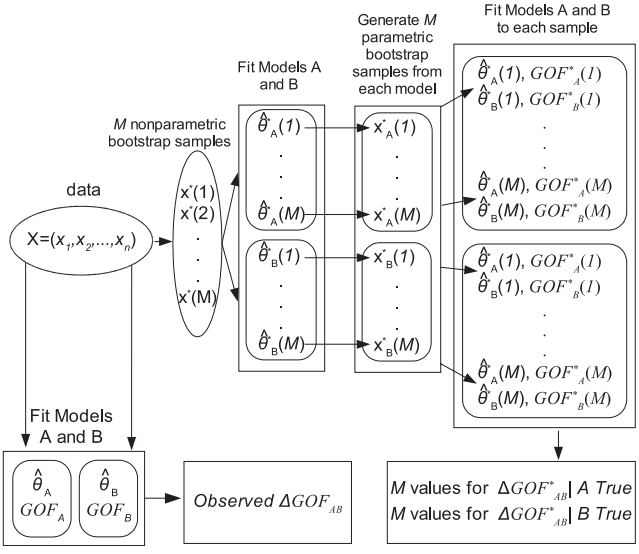


Fig. 1. The data informed parametric bootstrap cross-fitting method (PBCM) for model mimicry (see text for details).

point in more detail (see also Bollback, 2002, p. 1179). Thus, sampling variability in $\hat{\theta}$ needs to be taken into account by considering a set of plausible values around the point estimate $\hat{\theta}$.[2]

The method used here to obtain the sampling distribution of $\hat{\theta}$ is the nonparametric bootstrap (i.e., step 1 above). This is not only consistent with the bootstrap approach advocated in this paper, but the nonparametric bootstrap is also attractive because it works even when the distribution of $\hat{\theta}$ is not normal—maximum likelihood theory only guarantees *asymptotic* normality. A second method to obtain the sampling distribution of $\hat{\theta}$ is by means of the parametric approach. The parametric method can perhaps be more readily applied to data-sparse situations than can the nonparametric method. Also, the parametric method might be less sensitive to measurement noise than the nonparametric method. This increase in robustness comes at the cost of assuming that the fitted model is approximately correct. Here we will use the nonparametric approach, but it should be noted that both methods, parametric and nonparametric, have their merits.

Thus, the data informed PBCM yields two distributions of ΔGOF^*_{AB}, one derived under the assumption that model A is true, and one derived under the assumption that model B is true. The diagnosticity of δ_{AB} (i.e., the observed difference in GOF) can then be quantitatively assessed with reference to these two distributions. When δ_{AB} has a higher probability under the distribution of ΔGOF^*_{AB} (model A is true) than under the distribution of ΔGOF^*_{AB} (model B is true) this suggests that model A is more adequate than model B. A literature search revealed that this procedure (without taking parameter uncertainty into account) was first suggested and applied as a model selection tool for nonnested models by Williams (1970a, b). For nested mixture models, the parametric bootstrap has been used to construct the difference distribution under the null-hypothesis (e.g., Feng & McCulloch, 1996; McLachlan, 1987).[3]

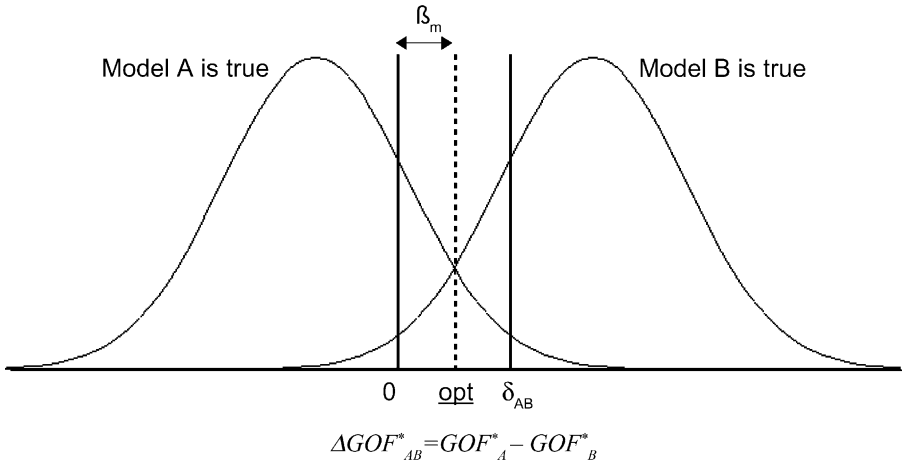
At this point it is useful to stress that the conclusions from this method are conditional on the specific data of interest. That is, it is important to realize that models A and B are not generic models, but rather specific instantiations that use parameter values that provide the best fit of the model to the data at hand. For a different set of data, the best fitting models A and B are likely to have different parameter values, and this may in turn lead to different difference distributions. The procedure discussed here focuses on specific instantiations of models (''model tokens'') instead of generic instantiations of models (''model types''). We will discuss this issue in more detail later, but until that point the reader should be aware that the method discussed so far is only informative with respect to the specific data set and the specific models that give the best fit to those data. Thus, only under the provision that the observed data are the only plausible data is it allowed to interpret δ_{AB} as ''evidence''. Henceforth, specific instantiations of models (i.e., models whose parameter values are obtained by fitting the data of interest) will be denoted model tokens, whereas the generic instantiations will be denoted model types. Also, whenever the need for the distinction arises, the PBCM applied to model tokens and model types will be termed 'data informed PBCM' and 'data uninformed PBCM', respectively. Note that generic instantiations are not *necessarily* uninformative with respect to parameter values—if prior research has pointed to specific regions of the parameter space that are relevant, these selective areas could be used in a 'data uninformed PBCM' analysis. The important point, elaborated on later, is that assessment of model mimicry based on model tokens incorporates information from the experiment itself, whereas an assessment of model mimicry based on model types is based on general prior information that is not directly dependent on the specific experimental data that are the focus of the modeling enterprise. For now, the focus is on the PBCM for assessment of mimicry between model tokens (i.e., 'data informed PBCM').

We will first discuss what kind of information can be extracted from the two difference distributions and the observed difference δ_{AB}. Next, we will show an example in which the data informed PBCM is applied to real data. Fig. 2, top panel, shows two fictitious difference distributions. Conclusions from these distributions with respect to δ_{AB} follow from signal detection theory (e.g., Green & Swets, 1966). Let $P(x = \Delta GOF^*_{AB}|A\ true)$ be denoted $P_A(x)$, the probability of observing a difference in GOF of x when A is the true data-generating model token. Similarly, we denote $P(x = \Delta GOF^*_{AB}|B\ true)$ by $P_B(x)$.

The relative likelihood that the data originate from model token A rather than from model token B, given an observed difference in GOF, can be quantified by the division of the estimated heights of the difference distributions at the observed value δ_{AB}: $P(\delta_{AB}|A\ true)/P(\delta_{AB}|B\ true) = P_A(x)/P_B(x)$. We define the optimal decision criterion (*opt*) to be the criterion that maximizes the probability of a correct binary classification (i.e., either 'token A is true' or 'token B is true') and this criterion is given by $P_A(opt)/P_B(opt) = 1$. In other

[2] Parameter variability in psychological models can also be explicitly accounted for (cf. Ratcliff & Rouder, 1998; Van Zandt & Ratcliff, 1995).

[3] It should be pointed out that nonparametric bootstrap methods (i.e., resampling from the data instead of from the fitted model) cannot be applied to goodness-of-fit measures in a straightforward fashion (Bollen & Stine, 1992; Wagenmakers, Farrell, & Ratcliff, in press). The main obstacle is that $\hat{F}$ almost certainly violates the null-hypothesis, even if $\hat{F}$ was in fact generated under the null-hypothesis.



Nominal Criterion

		Recovered Model A	Recovered Model B
Generating Model	A	.70	.30
	B	.05	.95

Optimal Criterion

		Recovered Model A	Recovered Model B
Generating Model	A	.90	.10
	B	.10	.90

Fig. 2. Top panel: two hypothetical difference distributions obtained by the data informed PBCM. Bottom panel: confusion matrices constructed by using the nominal criterion (left) and by using the optimal criterion (right), with reference to the difference distributions in the top panel.

words, at the optimal decision criterion the "evidence" favors both model tokens to the same extent—this is the point on the x-axis where the distributions cross.[4]

As can be seen from Fig. 2, top panel, the nominal criterion of $\Delta GOF_{AB} = 0$ is not optimal. Because model token B is able to provide a better account for data generated by model token A than vice versa, use of the nominal criterion would lead to a bias in favor of model token B. Assume that a binary decision is required. According to the nominal criterion, the decision 'token A is true' will be made when $\delta_{AB} = GOF_A - GOF_B < 0$ (Assuming that low values of GOF are to be preferred, as is the case for negative log likelihood or residual squared error). That is, observed values to the left of the criterion lead to 'A' classifications, and observed values to the right of the criterion lead to 'B' classifications. Fig. 2, bottom left panel, shows the resulting probability of correct classification when the nominal criterion is used. The fact that the nominal criterion is not optimal results in an asymmetric confusion matrix: data from model token B are less likely to be erroneously classified as originating from model token A ($p = 0.05$) than vice versa ($p = 0.30$). The overall probability of correct classification is $(0.70 + 0.95)/2 = 0.825$. Fig. 2, bottom right panel, shows the confusion matrix when the optimal criterion *opt* is used instead of the nominal criterion. Using the optimal criterion, the confusion matrix is now symmetrical (i.e., both models are equally confusable), and the overall probability of correct classification is $(0.90 + 0.90)/2 = 0.90$ which is considerably higher than 0.825. We will term the difference between the nominal criterion and the optimal criterion the mimicry bias β_m.

We would like to stress here that a plot of the difference distributions provides more information than the confusion matrix. For instance, the hypothetical example above illustrated that a biased criterion leads to an asymmetric confusion matrix. The reverse, however, is not necessarily true. Fig. 3, top panel, shows two difference distributions that yield a symmetric confusion matrix. That is, the optimal criterion $opt \rightarrow P_A(opt)/P_B(opt) = 1$ is used (in this case, the optimal criterion equals the nominal criterion $\Delta GOF_{AB} = 0$). Fig. 3, bottom panel, shows the same difference distributions, the only change being that the variance of $P_B(x)$ is now twice that of $P_A(x)$. In this case the optimal criterion no longer equals the nominal criterion, as is apparent from the figure. More importantly, the optimal criterion maximizes the overall probability of correct classification but will nonetheless lead to an asymmetric confusion matrix. Thus, asymmetric confusion matrices can originate either from a suboptimal (i.e., biased) decision criterion or from difference distributions that

[4] A different criterion would try to optimize classification performance under the constraint that the probability of an error is the same for both models. Visual inspection of the difference distributions will be informative as to whether this criterion is more useful than the more traditional criterion that optimizes overall classification performance. Also, note that an asymmetrical payoff structure can change the setting of the 'optimal' criterion.

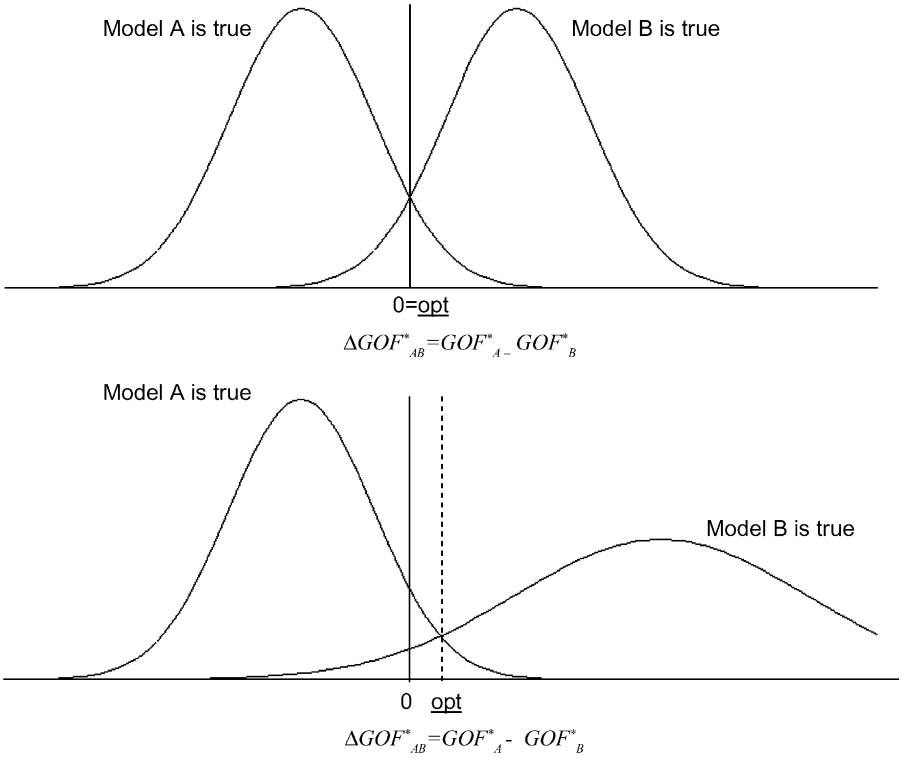


Fig. 3. Hypothetical difference distributions obtained by the data informed PBCM. Top panel: the two distributions have equal variance. Bottom panel: the two distributions have unequal variance, causing the confusion matrix to be asymmetrical despite the fact that the decision criterion is set optimally.

have unequal variances. Only in the former case can classification performance be improved upon.

We now illustrate the use of the data informed PBCM with a popular example, and then turn to a discussion of the underlying assumptions.

2.1. Example 1. Two models of information integration: FLMP vs. LIM

In everyday life, most humans obtain perceptual information through five sensory modalities. This raises the question in what way the possibly conflicting information from the various sources is combined to result in subjective experience. The two models discussed in this section provide different answers as to how information from distinct modalities is blended or integrated. For details on the many theoretical and empirical issues involved we refer the reader to the recent monograph by Massaro (1998).

In the experimental paradigm that is most often used to study information integration, participants are simultaneously exposed to auditory information (i.e., via a headset) and visual information (i.e., via a computer monitor). The visual input is in the form of a computer-generated face that is designed to mimic facial expressions used in speech production. For example, the movement of the computer-generated face can be consistent with the production of the syllable /ba/, /da/, or various levels between these two extremes. The auditory input also provides information that can also be consistent with the syllable /ba/, /da/, or various levels between. The participant has to decide whether his combined audiovisual experience for the produced syllable was consistent with either /ba/ or /da/. Obviously, this task is easiest when the information from the visual and auditory modalities is consistent (i.e., visual and auditory information are not in conflict) and unambiguous (i.e., clearly a /ba/ or clearly a /da/). By factorially combining the auditory and visual evidence for /ba/ vs. /da/, it is possible to study how information from these two modalities merges to create subjective experience.

The first model that provides a quantitative account of information-integration is the Fuzzy Logical Model of Perception (FLMP; e.g., Massaro, 1998; Massaro, Cohen, Campbell, & Rodriguez, 2001; Massaro & Friedman, 1990; Oden & Massaro, 1978; see also Movellan & McClelland, 2001). Let α_i denote the level of auditory information consistent with /da/, and let β_j denote the level of visual information consistent with /da/, $0 \leqslant \{\alpha_i, \beta_j\} \leqslant 1$. The FLMP probability of responding /da/ given these two sources of information, that is,

$P_{\mathrm{FLMP}}(/\mathrm{da}/|\alpha_i,\beta_j)$, is

$$P_{\mathrm{FLMP}}(/da/|\alpha_i,\beta_j)=\frac{\alpha_i\beta_j}{\alpha_i\beta_j+(1-\alpha_i)(1-\beta_j)}. \quad (1)$$

The degree of auditory and visual support for /ba/ is given by $(1-\alpha_i)$ and $(1-\beta_j)$, respectively. Eq. (1) follows from basic probability theory and combines the available information in a manner that optimizes the probability of correct classification. Crowther, Batchelder, and Hu (1995) have shown that the FLMP equation can be rewritten as an item-response Rasch model equation. As we will see later, the FLMP predicts that the evidence from one modality will have more effect to the extent that the evidence from the other modality is ambiguous.

The FLMP has been shown to be consistent with many empirical data sets, and this is sometimes taken to suggest a universal principle of information integration (Massaro, 1998; see also Movellan & McClelland, 2001). In addition, the fits of the FLMP to empirical data are almost always superior to that of competitor models. Here we will focus on one particularly simple competitor to the FLMP. This simple model is the Linear Integration Model (LIM; Anderson, 1981), in which

$$P_{\mathrm{LIM}}(/da/|\alpha_i,\beta_j)=\frac{\alpha_i+\beta_j}{2}. \quad (2)$$

Thus, in the LIM model the decision is based on the arithmetic average of the support provided by the auditory and visual information. In addition, there exist several other competitor models to FLMP (Massaro, 1998), including the TRACE model (e.g., McClelland & Elman, 1986) and the weighted average model (cf. Pitt, Kim, & Myung, 2003), in which $P(/da/|\alpha_i,\beta_j)=w\alpha_i+(1-w)\beta_j$. The latter model reduces to LIM when $w=\frac{1}{2}$. We chose to illustrate the PBCM using FLMP and LIM for two reasons. First, several recently published articles on model complexity have also used the FLMP vs. LIM example (e.g., Cutting, 2000; Myung & Pitt, 1997; Pitt et al., 2002) and this allows us to compare the PBCM to other methods such as minimum description length (e.g., Pitt et al., 2002). It is important to note here that the FLMP vs. LIM debate has centered around the issue of whether the superior fit of FLMP is due to excess flexibility rather than inherent correctness (e.g., Cutting, Bruno, Brady, & Moore, 1992; Dunn, 2000; Massaro, 1998; Massaro et al., 2001). In other words, it could be argued that the FLMP will provide a relatively good fit to the data, even when these data were in fact generated by LIM.

Second, FLMP and LIM have an equal number of free parameters: in both models, each level of auditory and visual support for /da/ (i.e., α_i and β_j, respectively) is associated with an estimated parameter. Because the penalty term of AIC and BIC depends critically on the number of free parameters, a difference in AIC or BIC between FLMP and LIM is completely determined by the difference in descriptive accuracy. Hence, AIC, BIC, and unpenalized log likelihood are in complete agreement with respect to the amount of statistical evidence for FLMP vs. LIM. The difference distributions obtained from the PBCM will thus be based on differences in maximum log likelihood (i.e., log likelihood ratios). Note that the presence of an additive penalty term would only shift the difference distributions along the x-axis, preserving both their shape and the distance between them.

2.2. Experimental design and model fitting

As an example of how the data informed PBCM can be applied to real data, we analyzed a database of experimental results, kindly made available by Massaro through the World Wide Web (i.e., http://mambo.ucsc.edu/psl/8236/). The experiment required participants to distinguish between the syllables /da/ and /ba/, given auditory support α_i and visual support β_j for /da/ (auditory and visual support for /ba/ is $1-\alpha_i$ and $1-\beta_j$, respectively). Both auditory and visual support was varied on five levels. Hence, both FLMP and LIM estimate 10 parameters: $\{\alpha_1,\alpha_2,\alpha_3,\alpha_4,\alpha_5\}$ for auditory support and $\{\beta_1,\beta_2,\beta_3,\beta_4,\beta_5\}$ for visual support for /da/.

In the experimental design, the five levels of auditory support for /da/ and the five levels of visual support for /da/ were factorially combined to yield $5\times5=25$ conditions. In addition to these 25 'mixed' conditions, 10 'pure' conditions were present; in the pure conditions only auditory support or only visual support was presented, but not both. Thus, a total of 35 conditions was obtained. Note that if the experiment only included the 25 'mixed' conditions, the FLMP would be identified with 9 parameters instead of 10 (Crowther et al., 1995; for a note on the identifiability problem for LIM see Navarro et al., 2003, footnote 1). For each condition, the datum of interest is the probability of responding /da/, based on 24 binary judgments (for further details see Massaro, 1998). To avoid averaging artifacts (cf. Estes, 2002; Massaro, 1998, pp. 132–135), we will examine two typical sample participants, #24 and, later, #15.

The data from Massaro's (1998) participant #24 were fit both by FLMP and by LIM using maximum likelihood estimation (MLE). More specifically, both FLMP and LIM predict the proportion of binary (/da/ or /ba/) classifications for every experimental condition. To assess goodness-of-fit, the probability density can be computed from the predicted proportion using the binomial probability density function $pdf(\text{binomial})=\hat{p}_i^{k_i}(1-\hat{p}_i)^{n-k_i}\binom{n}{k_j}$, where k_i is the number of /da/ responses out of n observations, and $\hat{p}_i$ is the predicted probability for condition i. The overall log likelihood

for FLMP over all $i = 1, ..., 35$ conditions is given by

$$\ell_{\mathrm{FLMP}} \propto \sum_{i=1}^{35} \log\left[\hat{p}_{i,\mathrm{FLMP}}^{k_i}(1-\hat{p}_{i,\mathrm{FLMP}})^{n-k_i}\binom{n}{k_i}\right], \qquad (3)$$

where $\hat{p}_{i,\mathrm{FLMP}}$ is given by Eq. (1). ℓ_{FLMP} can be obtained by replacing $\hat{p}_{i,\mathrm{FLMP}}$ by $\hat{p}_{i,\mathrm{LIM}}$ as given by Eq. (2). We also fit FLMP and LIM using the root mean squared deviation (RMSD), which is defined as $RMSD = \sqrt{\sum_{i=1}^{N}(\hat{p}_i - p_i)^2/N}$, where $N = 35$ is the number of predicted data points and p_i is the observed probability for condition i (e.g., Massaro, 1998). The results based on RMSD are very similar to the results from the more principled MLE method reported here.

For both FLMP and LIM, the MLE parameter values and their standard errors are presented in Table 1. The standard errors were estimated by taking 1000 nonparametric bootstrap samples from the data. Specifically, when the observed probability of responding /da/ in condition x, p_x, equaled 0.8, resampling was done by drawing an integer number between 0 and 24 from a binomial distribution, $p_x^* \sim Bin(p_x = 0.8, n = 24)$. Table 1 shows that the parameter estimates are more reliable for extreme support values than for intermediate support values, a regularity that is also present for binomial models (i.e., a binomial model with parameter p has $se(\hat{p}) = \sqrt{p(1-p)/n}$, which is maximal when $p = 1 - p = \frac{1}{2}$). Also, the average parameter estimates for the nonparametric bootstrap samples equal the MLE point estimates almost exactly. This indicates that the nonparametric bootstrap procedure does not lead to a systematic overestimation or underestimation (i.e., the procedure is unbiased, Efron & Tibshirani, 1993, Chap. 10).

The fit of FLMP (i.e., $\ell_{\mathrm{FLMP}} = -38.0$) was much better than the fit of LIM (i.e., $\ell_{\mathrm{LIM}} = -141.6$). Fig. 4 shows the data, and the predicted values for FLMP (top panel) and LIM (bottom panel). The data clearly show that the effect of visual support is largest when the auditory support is ambiguous. In other words, the data from the mixed condition are shaped somewhat like an American football, and this pattern is captured by FLMP. In contrast, LIM predicts that the effect of a difference in visual support is constant across the levels of auditory support, a prediction that the data do not support.

2.3. Assessment of model flexibility: preliminary comments

In the following, we study to what extent the better fit of FLMP is due to excess flexibility. As mentioned at the

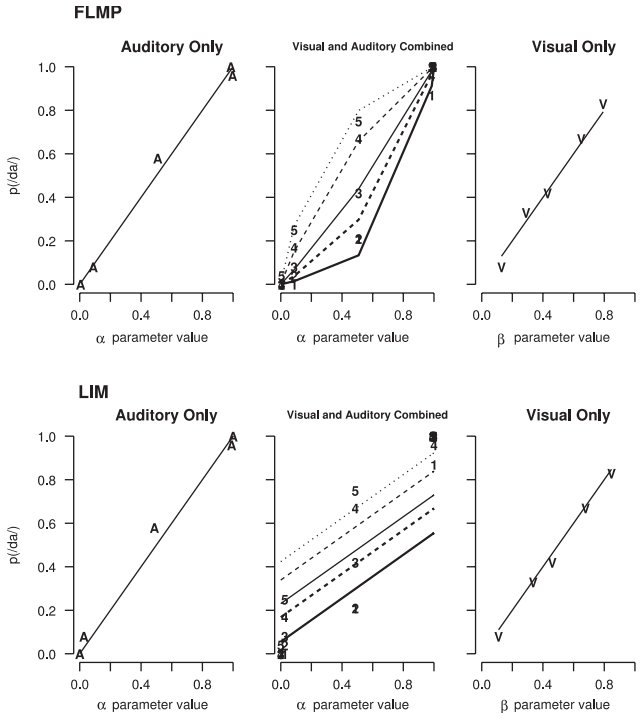


Fig. 4. Data from Massaro's (1998) participant #24. Top panel: FLMP fit. Bottom panel: LIM fit. The numbers in the middle panels indicate the level of visual support for the syllable /da/, ranging from 1 (low support) to 5 (high support). The lines give the model prediction. See text for details.

Table 1
Parameter point estimates, nonparametric bootstrap mean parameter estimates and nonparametric standard errors for FLMP and LIM, fitted to Massaro's (1998) Participant #24

	α_1	α_2	α_3	α_4	α_5	β_1	β_2	β_3	β_4	β_5
FLMP										
Point	0.01	0.09	0.51	0.99	1.00	0.13	0.29	0.43	0.65	0.79
Mean	0.00	0.09	0.51	0.99	1.00	0.13	0.29	0.43	0.65	0.79
SE	0.01	0.03	0.06	0.01	0.00	0.04	0.06	0.07	0.07	0.05
LIM										
Point	0.00	0.03	0.49	1.00	0.99	0.11	0.34	0.46	0.68	0.85
Mean	0.00	0.03	0.49	1.00	0.99	0.10	0.33	0.46	0.68	0.85
SE	0.00	0.02	0.07	0.00	0.01	0.06	0.07	0.07	0.07	0.06

Note: Point = parameter point estimate that maximizes log likelihood for the observed data. Mean = average parameter value for 1000 nonparametric bootstrap samples. SE = Standard error for each parameter based on 1000 nonparametric bootstrap samples (see text for details).

beginning, various model selection methods aim to discount goodness-of-fit of flexible models by penalizing lack of parsimony. Since FLMP and LIM each have 10 free parameters, differences in both AIC and BIC will be determined by differences in GOF and hence favor FLMP over LIM for this particular data set.

In contrast to AIC and BIC, the method of minimum description length (MDL; e.g., Grünwald, 2000; Pitt et al., 2002; Rissanen, 1996, 2001) and Bayesian model selection (BMS; e.g., Kass & Raftery, 1995; Myung & Pitt, 1997; Wasserman, 2000) are sensitive to the functional form of the model parameters. Both BMS and MDL are among the most promising and modern model selection methods (see also Hastie, Tibshirani, & Friedman, 2001, Chaps. 7 and 8). From an MDL analysis of FLMP and LIM, Pitt et al. (2003) concluded that FLMP is more complex than LIM, although the authors nevertheless preferred FLMP.

Before turning to the results obtained by applying the data informed PBCM, we first briefly discuss earlier methods that are similar to the data informed PBCM. Suppose two nonnested models, A and B, need to be distinguished using a null-hypothesis testing framework. That is, let H_0 be the hypothesis that the data $\mathbf{x}$ are distributed according to model A. A significance test could then be based on a comparison of the observed difference in GOF (e.g., log likelihood ℓ) to the expected value of the GOF difference given that H_0 holds (Cox, 1962). The test statistic T_A is then defined as

$$T_A = \{\ell_A(\hat{\theta}_A) - \ell_B(\hat{\theta}_B)\} - E_A\{\ell_A(\hat{\theta}_A) - \ell_B(\hat{\theta}_B)\}, \quad (4)$$

where E_A denotes the expectation given that model A is the data-generating model. When this null-hypothesis is true, T_A is normally distributed with mean zero. When the null-hypothesis is false, T_A should be negative, because the expectation of a better fit for model A than for model B, given that model A is the data-generating model, is an overestimate when model A did not, in fact, generate the data. This procedure is very similar to the data informed PBCM. Note that Cox's method is phrased in a null-hypothesis testing framework. One model is assigned the role of null-hypothesis, and it is either rejected or not rejected. In the case of two models, separate analyses could be done, one based on model A taking on the role of null-hypothesis, and one based on model B taking on this role.

Williams (1970a) first suggested the use of the parametric bootstrap to obtain two difference distributions, one under the hypothesis that model token A is true and the other under the hypothesis that model token B is true. The objective is then to associate the observed difference in GOF with one of these two distributions. Williams (1970b) applied this procedure to choose between two regression token models for enzyme synthesis. It is probably due to lack of computational resources that Williams did not attempt to construct the two difference distributions by generating a large number of data series from each token model. Instead, only $M = 10$ parametric bootstrap samples were generated from each token model. Both models were then fit to each parametric bootstrap sample, and the difference between the maximized likelihoods was computed. Next, the mean and standard deviation for each of the two difference distributions were obtained as approximations of the 'true' difference distributions. As a result, each model had associated with it an estimated difference distribution that provides information about the expected difference in GOF given that the model has generated the data. In order to associate the observed difference in likelihood with one of the two distributions, two boundaries were determined, b_A associated with the difference distribution f_A (i.e., consisting of data generated by model A) and b_B associated with the difference distribution f_B. Specifically, $b_A = \max\{\hat{\mu}_A + 2\hat{\sigma}_A, \max(\mathbf{x}^*)\}$, and $b_B = \min\{\hat{\mu}_B - 2\hat{\sigma}_B, \min(\mathbf{x}^*)\}$, where $\hat{\mu}_A$, $\hat{\mu}_B$, and $\hat{\sigma}_A$, $\hat{\sigma}_B$ are the means and standard deviations of the difference distributions f_A and f_B, respectively (due to the order of subtraction, $\hat{\mu}_A < \hat{\mu}_B$), and $\mathbf{x}^*$ is one of the $M = 10$ bootstrap samples. The observed ΔGOF_{AB} was then compared to the boundaries b_A and b_B. Williams (1970b, p. 28) used the following guidelines:

If	$\Delta GOF_{AB} < b_A$,	$\Delta GOF_{AB} < b_B$	the true model is A
	$\Delta GOF_{AB} > b_A$	$\Delta GOF_{AB} > b_B$	the true model is B
	$\Delta GOF_{AB} > b_A$	$\Delta GOF_{AB} < b_B$	neither model is true
	$\Delta GOF_{AB} < b_A$	$\Delta GOF_{AB} > b_B$	model A and B cannot be discriminated.

Again, it is crucial to realize that these conclusions are only valid when one restricts attention to the observed data and hence to the specific instantiations of the models.

2.4. Application of the data informed PBCM to FLMP and LIM

We now apply what might be considered an extended version of this idea to the data from Fig. 4. It should be noted that Massaro et al. (2001, pp. 7–8) used a very similar procedure, but they did not discuss their results in any detail. Recall that in order to calculate standard errors for the FLMP and LIM parameters, $M = 1000$

nonparametric bootstrap samples were used. Each of these nonparametric samples yielded estimates $\hat{\theta}^*_{\text{FLMP:NP}}$ and $\hat{\theta}^*_{\text{LIM:NP}}$ (NP stands for nonparametric). To incorporate parameter uncertainty in the construction of the difference distributions, $M = 1000$ parametric bootstrap samples were generated from FLMP and LIM by using the $M = 1000$ nonparametric parameter vectors $D_{\text{FLMP}} = \{\hat{\theta}^*_{\text{FLMP:}NP}(1), \hat{\theta}^*_{\text{FLMP:}NP}(2), ..., \hat{\theta}^*_{\text{FLMP:}NP}(M)\}$ and $D_{\text{LIM}} = \{\hat{\theta}^*_{\text{LIM:}NP}(1), \theta^*_{\text{LIM:}NP}(2), ..., \theta^*_{\text{LIM:}NP}(M)\}$. Each nonparametric parameter vector (i.e., each element of D_{FLMP} and D_{LIM}) was used once to generate data by first computing, from the best fitting parameters (i.e., $\hat{\theta}^*_{\text{FLMP:}NP}(i)$ and $\hat{\theta}^*_{\text{LIM:}NP}(i)$), the probability of responding /da/ in each condition (according to the model). Next, sampling was done from a binomial distribution, $Bin(p_i, n = 24)$, where p_i is the probability of responding /da/ for condition i. Thus, M data sets were generated from FLMP, and M data sets were generated from LIM, taking parameter uncertainty into account.[5]

Each generated data set was then fit both by FLMP and by LIM, and the difference in log likelihood calculated. In fitting FLMP and LIM back to the generated data, the starting values for the $\{\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5\}$ parameter vector were $\{0.01, 0.25, 0.50, 0.75, 0.99, 0.01, 0.25, 0.50, 0.75, 0.99\}$, and each parameter was constrained to take on values between 0 and 1 only. The difference distributions, each based on $M = 1000$ differences in log likelihood—one given token FLMP was the generating model, one given token LIM was the generating model—are shown in Fig. 5. The figure shows the bar graph and a Gaussian kernel smoothing estimate (e.g., Silverman, 1986; Van Zandt, 2002).

For this particular data set, Fig. 5 allows several conclusions:

1. Based on the nominal criterion of no difference in log likelihood, all simulated data sets are correctly classified. In other words, the entire distribution f_{LIM} lies to the left of the nominal criterion $\ell_{\text{FLMP}} - \ell_{\text{LIM}} = 0$, and the entire distribution f_{FLMP} lies to the right of this criterion. This observation is in agreement with that of Massaro et al., 2001.
2. The observed difference in log likelihood is located near the middle of the f_{FLMP} distribution. This fact, in combination with the previous observation of perfect discrimination, indicate that the data are much more likely under the token version of FLMP than under the token version of LIM (cf. Cox, 1962; Williams, 1970a, b).

[5] Incorporating parameter uncertainty did not lead to qualitative changes in the difference distributions. This indicates that the pattern of results holds over a range of slightly different data sets and parameter estimates, at least for the models and data sets examined here.

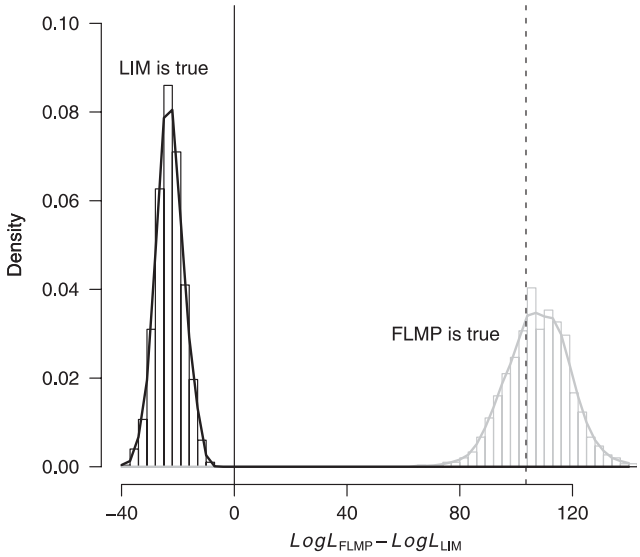


Fig. 5. Difference distributions obtained by the data informed PBCM applied to the experimental data from Fig. 4 (participant #24). The dashed vertical line indicates the difference in log likelihood that was observed for the participant. The bin-width for the bar graph and the kernel estimate was 3.

3. The difference distributions are not symmetrical around the nominal criterion of $\Delta\ell = 0$. This indicates that LIM is relatively poor at accounting for FLMP-patterns, whereas FLMP is better (but still poor) at accounting for LIM-patterns. This analysis shows that FLMP is more flexible than LIM in terms of mimicry (cf. Pitt et al., 2002, 2003).

In sum, results from the data informed PBCM demonstrate that, for this particular data set, FLMP is much to be preferred over LIM, despite the fact that FLMP is better able to mimic LIM than vice versa. Note that the criterion that optimizes overall classification performance ranges anywhere between a difference in log likelihood of about 0–70—when the difference distributions do not overlap, as is the case here, precise location of an optimal criterion is difficult. The aim of this example was not to argue for or against FLMP (or LIM). Rather, the data informed PBCM illustrates how consideration of the difference distributions can increase understanding of the mimicry problem.

The data set analyzed here was fairly typical. Another pattern that often appeared in the database of 82 participants (from Massaro, 1998) was characterized by similarly located difference distributions as depicted in Fig. 5, but with the observed difference in log likelihood located just outside the f_{FLMP} distribution, toward f_{LIM} but clearly above the nominal criterion of $\Delta\ell = 0$. This pattern of results is illustrated by applying the same method of analysis to Massaro's participant #15. Fig. 6 shows the maximum likelihood fit of FLMP (top panel) and LIM (bottom panel). Again, the fit of FLMP

($\ell_{\mathrm{FLMP}} = -56.7$) is better than that of LIM ($\ell_{\mathrm{LIM}} = -59.5$), although the difference in GOF is much less pronounced than it was for participant #24. The data give the impression of being noisy and idiosyncratic. Table 2 shows the MLE parameter estimates and their bootstrap standard errors. Fig. 7 shows the results of applying the data informed PBCM. It is evident that the observed likelihood ratio falls precisely in between the two distributions of likelihood ratios.

The pattern of results shown in Fig. 7 could be judged to be more consistent with token FLMP than with token LIM, as would be the case when only the nominal criterion were to be used. However, we believe the more appropriate labeling of such a pattern would be 'ambiguous'. The fact that the observed $\Delta\ell$ falls outside the FLMP generated difference distribution strongly suggests that something is amiss either in the data or in the model. This interpretation is consistent with that of Cox (1962) and Williams (1970b), who argued that such a pattern of results casts doubt on both models. For an in-depth analysis of the FLMP vs. LIM debate we refer the interested reader to Massaro (1998). Thus, the data informed PBCM can be used to assess the relative adequacy of the models under consideration. In particular, Fig. 7 shows that the PBCM is able to classify data as inconclusive or ambiguous, despite the fact that the nominal criterion yields perfect model recovery.

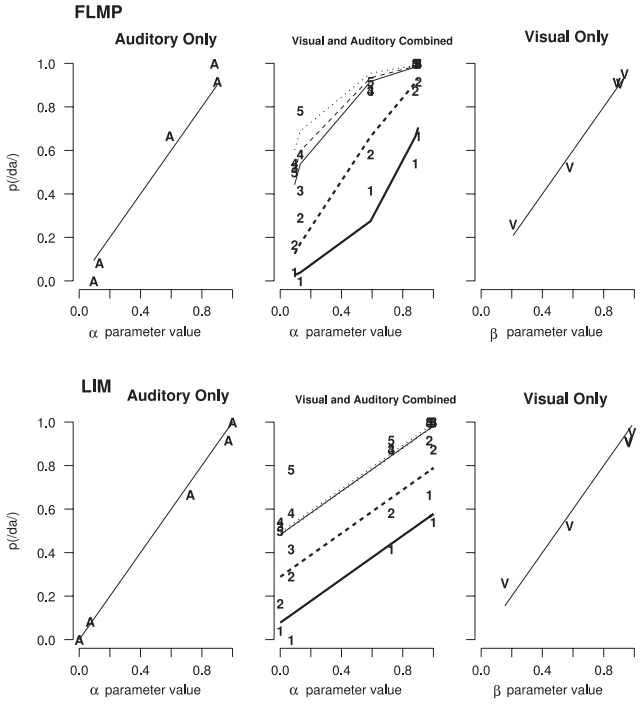


Fig. 6. Data from Massaro's (1998) participant #15. Top panel: FLMP fit. Bottom panel: LIM fit. The numbers in the middle panels indicate the level of visual support for the syllable /da/, ranging from 1 (low support) to 5 (high support). The lines give the model prediction (see text for details).

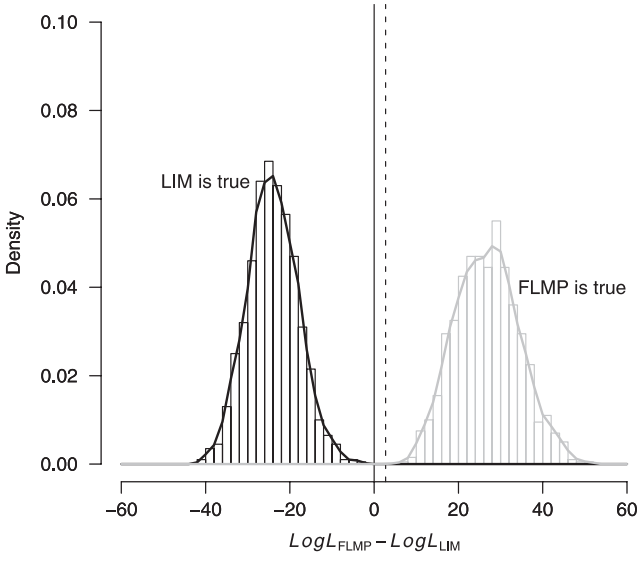


Fig. 7. Difference distributions obtained by the data informed PBCM applied to the experimental data from Fig. 6 (participant #15). The dashed vertical line indicates the difference in log likelihood that was observed for the participant. The bin-width for the bar graph and the kernel estimate was 2.

Table 2
Maximum likelihood parameter point estimates, nonparametric bootstrap mean parameter estimates and nonparametric standard errors for FLMP and LIM, fitted to Massaro's (1998) Participant #15

	α_1	α_2	α_3	α_4	α_5	β_1	β_2	β_3	β_4	β_5
FLMP										
Point	0.09	0.13	0.59	0.88	0.90	0.21	0.58	0.88	0.91	0.94
Mean	0.09	0.13	0.59	0.88	0.90	0.21	0.57	0.88	0.90	0.93
SE	0.02	0.03	0.06	0.03	0.03	0.05	0.06	0.03	0.03	0.02
LIM										
Point	0.00	0.07	0.72	1.00	0.97	0.15	0.58	0.96	0.96	0.99
Mean	0.00	0.07	0.72	1.00	0.97	0.16	0.57	0.96	0.96	0.98
SE	0.00	0.05	0.06	0.00	0.02	0.05	0.07	0.03	0.03	0.02

Note: Point = parameter point estimate that maximizes log likelihood for the observed data. Mean = average parameter value for 1000 nonparametric bootstrap samples. SE = Standard error for each parameter based on 1000 nonparametric bootstrap samples. See text for details.

3. Assumptions and extensions

Up to this point we have outlined the data informed PBCM, and illustrated its use by an application to two models of information integration. We now turn to a more detailed discussion of the underlying assumptions, advantages, and limitations of the data informed PBCM. This discussion will ultimately lead to a similar but different version of the PBCM, a version that does not depend on the observed data.

3.1. Data informed PBCM as a frequentist implementation of the Bayesian posterior predictive distributions

To recapitulate, the PBCM for model tokens generates parametric bootstrap samples from a token model with MLE parameter vector $\hat{\theta}$. Uncertainty in parameter estimation can be taken into account using the nonparametric bootstrap for data $\mathbf{x}$, yielding slightly different parameters $\hat{\theta}^*$ for every $\mathbf{x}^*$. The bootstrap samples generated from the model can be considered future or replicate data, given that the model under consideration is true.

In this section we will briefly point out a striking similarity between the data informed PBCM and what is known as Bayesian posterior predictive p-values (BPP; e.g., Berkhof, van Mechelen, & Gelman, in press; Bollback, 2002; Gelman, Goegebeur, Tuerlinckx, & van Mechelen, 2000; Meng, 1994; Rubin, 1984, Section 5). We are grateful to In Jae Myung (personal communication) for pointing this out to us. To anticipate, the data informed PBCM can be interpreted as a frequentist implementation of BPP. The use of the Bayesian posterior predictive p-values is usually advocated as a method to assess model adequacy, although similar Bayesian ideas have been proposed as methods for model selection (e.g., Aitkin, 1991; Laud & Ibrahim, 1995).

Before proceeding it is useful to introduce certain Bayesian ideas and establish some notation. One of the characteristic features of a Bayesian analysis is that it requires the specification of plausible parameter values prior to the observation of the data. The prior density for θ is updated or altered by the impact of observed data $\mathbf{x}$ and the result is a posterior distribution of θ:

$$p(\theta|\mathbf{x}) = \frac{p(\theta)p(\mathbf{x}|\theta)}{p(\mathbf{x})}, \tag{5}$$

where $p(\mathbf{x}) = \int_\Theta p(\theta)p(\mathbf{x}|\theta)\,d\theta$ is the normalizing constant (for an early application to psychology see Edwards, Lindman, & Savage, 1963). The predictive density for unseen data $\mathbf{x}^{\text{new}}$, after observing data $\mathbf{x}$, is then given by (e.g., Aitchison & Dunsmore, 1975)

$$p(\mathbf{x}^{\text{new}}|\mathbf{x}) = \int_\Theta p(\mathbf{x}^{\text{new}}|\theta)p(\theta|\mathbf{x})\,d\theta. \tag{6}$$

This predictive density can be used as a diagnostic tool to assess whether the proposed model performs adequately or not. It is relatively straightforward to evaluate $p(\mathbf{x}^{\text{new}}|\mathbf{x})$ using Monte Carlo techniques. The first step in a Monte Carlo approximation of $p(\mathbf{x}^{\text{new}}|\mathbf{x})$ is to sample a set of i parameter vectors, θ_i, $i = 1, ..., I$, from $p(\theta|\mathbf{x})$, the posterior distribution of θ. In the second step, each parameter vector θ_i is then used to generate a replicated data set (or future data set) $\mathbf{x}_i^{\text{new}}$. The entire collection of replicate data sets $\{\mathbf{x}_1^{\text{new}}, \mathbf{x}_2^{\text{new}}, ..., \mathbf{x}_I^{\text{new}}\}$ is then representative of the posterior predictive distribution $p(\mathbf{x}^{\text{new}}|\mathbf{x})$. Next, a goodness-of-fit measure is computed both for the observed data $\mathbf{x}$ and for the replicate data sets $\{\mathbf{x}_1^{\text{new}}, \mathbf{x}_2^{\text{new}}, ..., \mathbf{x}_I^{\text{new}}\}$. The GOF measure may be absolute or relative—an example of the latter category is the likelihood ratio test statistic (Rubin & Stern, 1994). A comparison of the distribution of GOF values for the replicate data sets versus the GOF value for the observed data can then be used to assess model adequacy. This assessment can be quantified by the tail area probability of the replicate distribution associated with the observed GOF value (i.e., a small tail area probability indicates a deviation between the model and the observed data).

Thus, both the data informed PBCM and BPP generate distributions of expected GOF values from the models under consideration. These distributions can be used to assess model mimicry and model adequacy. The difference between the data informed PBCM and BPP lies in the fact that the generating model for the BPP is explicitly Bayesian. The data informed PBCM generates simulated data from a model whose parameterization is determined by sampling from a distribution of parameter values that is obtained using the nonparametric bootstrap. The BPP, in contrast, generates simulated data by sampling from the posterior distribution for the parameter values. Therefore, the data informed PBCM and the BPP will yield exactly the same result if the nonparametric bootstrap distribution of model parameters is identical to the Bayesian posterior distribution of model parameters. This identity condition is closely approximated for multinomial models (e.g., Hastie et al., 2001; Rubin, 1981). In such circumstances the use of the nonparametric distribution has practical advantages over the use of the Bayesian posterior distribution:

> (…) the bootstrap distribution represents an (approximate) nonparametric, noninformative posterior distribution for our parameter. But this bootstrap distribution is obtained painlessly—without having to formally specify a prior and without having to sample from the posterior distribution. Hence we might think of the bootstrap distribution as a "poor man's" Bayes posterior. By perturbing the data, the bootstrap approximates the Bayesian effect of

perturbing the parameters, and is typically much simpler to carry out. (Hastie et al., 2001, p. 236)

3.2. Model mimicry versus model selection

The PBCM provides information about model mimicry. For overlapping difference distributions, calculation of β_m gives the criterion that optimizes overall classification performance when the choice is between two models, given that one of the models is the true data-generating model. The criterion β_m is determined by the two difference distributions. Suppose that for each parametric sample a model selection criterion based on penalized log likelihood would be calculated instead of log likelihood alone (e.g., AIC or BIC, both having an additive penalty term for the number of free parameters). The penalty term would have the same influence for every generated sample, regardless of which model generated the sample. Hence, an additive penalty term shifts the two difference distributions along the x-axis, preserving shape and relative distance. Note that such a shift changes β_m by a constant, but does not influence the discrete selection probabilities when the optimal criterion is used (i.e., the confusion matrix is left intact). The robustness of PBCM against additive penalty terms highlights an important difference between it and model selection methods, and this is illustrated by the next example.

3.3. Example 2. Nested models

The data informed PBCM allows the calculations of an optimal criterion β_m for choosing between two specific models, given one of these models is true and both models are a priori equally likely. To get a feeling for what this means, consider the case of two nested models. Model A is $N(\mu, 1)$ and model B is $N(\mu, \sigma)$. The data are standard noise, $N(0, 1)$, consistent with the simple model A. We generated $n = 1000$ observations from $N(0, 1)$, and applied the PBCM (cf. Fig. 1). The MLE estimate for μ from model A equals $\hat{\mu}_A \approx -0.066$, and for model B $\hat{\mu}_B \approx -0.066$ and $\hat{\sigma}_B \approx 0.950$. The fit of model B ($\ell_B = -136.73$) was slightly better than that of model A ($\ell_A = -136.99$). The value of $\hat{\sigma}_B$ is close to the true value of 1, as expected. Surely, the minimal improvement in log likelihood does not warrant selection of the more complex model B—this is confirmed by AIC, BIC, and the likelihood ratio test. For these model selection methods, the decision is an easy one.

In contrast, the problem is extraordinarily difficult for the data informed PBCM. The data informed PBCM tries to answer the question "are the observed data generated by token model A (i.e., parameter $\hat{\mu}_A \approx -0.066$), or by token model B (i.e., parameters $\hat{\mu}_B \approx -0.066$ and $\hat{\sigma}_B \approx 0.950$)?" Because token models A and B generate very similar data patterns, the choice of discriminating between the token models is hard. Fig. 8 shows the two difference distributions, each based on $M = 1000$ parametric bootstrap samples generated from the MLE point estimates for the parameters. The difference distributions overlap considerably, and $\beta_m = 0.50$.

This example clearly shows that the data informed PBCM is biased in favor of complex models. The root of this problem is that the observed data have been used twice: once to determine the best fitting parameters, and again to assess the reasonableness of the model (the same problem is well known in the case of the BPP analysis, e.g., the discussion following Aitkin, 1991, and Thurman, 2001). By focusing on the parameter estimates from the data, the data informed PBCM effectively ignores the additional complexity associated with regions of parameter space that are extremely unlikely given the observed data.

Thus, the data informed PBCM ignores a priori considerations about model complexity, *and therefore, unlike AIC and BIC, does not assess model generalizability*. When the difference distributions overlap completely, this indicates perfect mimicry and zero discriminability, regardless of how complex the models under consideration are. In contrast, general model selection methods such as AIC and BIC focus on discounted goodness-of-fit, which can usually be

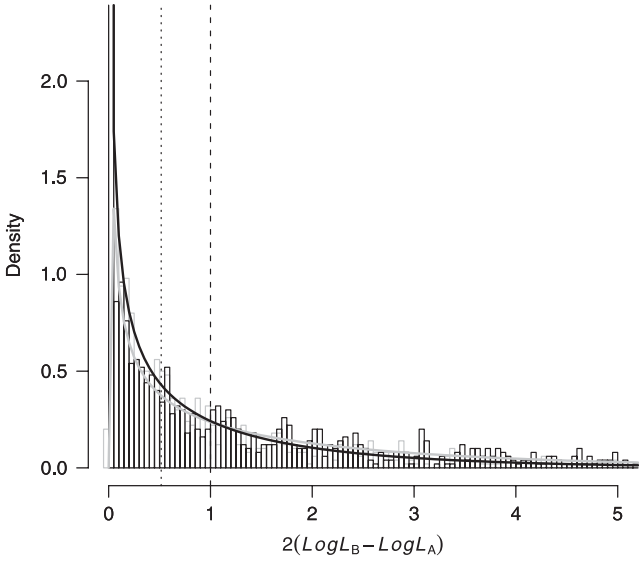


Fig. 8. Difference distributions obtained by the data informed PBCM applied to $N(0, 1)$ data fitted by two nested models: model A (i.e., $N(\mu, 1)$) and model B (i.e., $N(\mu, \sigma)$). The theoretical distributions of two times the log likelihood ratio for data generated from model A and from model B are the $\chi^2_{df=1}$ and the noncentral $\chi^2_{df=1}$ distribution, respectively. The noncentrality parameter was estimated to be 0.55. The observed difference in two times the likelihood ratio is given by the dotted vertical line. The optimal criterion β_m, given by the dashed vertical line, was calculated from the empirical distribution functions, not from the $\chi^2_{df=1}$ distributions.

expressed as an absolute number. However, these methods do not assess model adequacy or model mimicry with respect to a specific data set. In general, complex models will be able to mimic simpler models. It can therefore be argued that to some extent model selection methods implicitly take mimicry into account. However, increasing complexity does not necessarily increase the ability to mimic (Myung, 2002).

Because the data informed PBCM and model selection methods address related but different questions, we believe model evaluation should be based on applying both techniques. This is especially true for models of cognition, where the issue of mimicry is particularly pronounced (e.g., Van Zandt & Ratcliff, 1995). Future work could investigate the possibility of combining model mimicry and model selection methods. For instance, the MDL method of model selection (e.g., Rissanen, 1996, 2001) has an interpretation in differential geometry (e.g., Pitt et al., 2002)—an MDL measure of model mimicry could be developed based on the overlap of models in the space of predicted probability distributions (e.g., Myung, 2002).

Another option is to extend the data informed PBCM to incorporate an a priori preference for the simpler model. One way in which to accomplish preference in the data informed PBCM is to base a binary decision not on the criterion β_m (i.e., the criterion that optimizes overall classification performance), but on an adjusted criterion β_a, such that $P_A(\beta_a)/P_B(\beta_a) = \kappa$. When $\kappa = 1$ we have the optimal criterion. When $\kappa > 1$ or $\kappa < 1$, β_a is non-optimal and favors token model B or token model A, respectively. This adjustment is similar to what happens according to signal detection theory when the payoff matrix is asymmetrical (i.e., one response is more desirable or more likely than the other). The criterion bias $\beta_m - \beta_a$ can be considered a prior that influences the evidence ratio away from the more complex model. It is, however, not immediately clear how to determine this prior in a principled manner.

We believe the most principled method to punish the more complex model is to consider not just the parameter values that were determined by fitting the model to a specific set of data, but to consider the entire (plausible) range of parameter values. This data uninformed version of PBCM (i.e., 'global' PBCM) is explored next.

3.4. Types versus tokens in model selection

As mentioned previously, the data informed PBCM can only discriminate between models to the extent that they are functionally different (i.e., generate different sets of data). If a model is structurally more complex, but functionally identical to a simpler model, the data informed PBCM selection will be potentially misleading. Such a situation can arise for instance in nested models when the additional parameters of the complex model have relatively little impact. Intuitively, we would like to punish the more complex model because of its additional free parameters—only if these free parameters are important should they be included in the model. The following example shows how this can be accomplished in a Bayesian framework. We are indebted to Rich Shiffrin (personal communication) for suggesting this example and its interpretation.

3.5. Example 3. Coin tossing

A coin, randomly drawn from an urn that contains an equal number of fair and unfair coins, needs to be classified as fair or unfair. Let the extent to which the unfair coin is biased be uniformly distributed, ranging from $\theta = 0$ (i.e., the coin always lands tails) to $\theta = 1$ (i.e., the coin always lands heads). The coins is tossed $n = 100$ times. In other words, the data vector of length n is $\mathbf{x} = (x_1, x_2, ..., x_n)$, and each i.i.d. element is either 1 (for heads) or 0 (for tails), $x_i \in \{0, 1\}$. How do we make an optimal choice between the fair coin model F and the unfair coin model U? Model F says that $\mathbf{x}$ is generated from a binomial distribution with parameter $\theta = \frac{1}{2}$, whereas model U states that $\theta \in [0, 1]$, $\theta \neq \frac{1}{2}$. In this example, the prior probability of the coin being fair before observing a single datum is $\frac{1}{2}$, which means that the posterior odds and the Bayes factor (Kass & Raftery, 1995) coincide:

$$\Phi = \frac{p(\mathbf{x}|F)}{p(\mathbf{x}|U)} = \frac{L\left(\frac{1}{2}\right)}{\int_0^1 L(\theta)p(\theta)\,\mathrm{d}\theta}, \tag{7}$$

where L is the likelihood function and $p(\theta)$ is a uniform prior (cf. Wasserman, 2000).

An optimal choice compares the likelihood of the data under the F model to the *average* likelihood under the U model. If $\Phi > 1$, the F model is more plausible; if $\Phi < 1$, the U model is more plausible. The probability of observing $\mathbf{x}$ under model F is $\frac{1}{2}^n$ (i.e., each sequence of heads and tails is equally likely). When we take the prior to be the flat $Beta(1, 1)$ distribution, the average likelihood under the U model after observing h heads and t tails out of n tosses becomes $Beta(h + 1, t + 1) = \int_0^1 \theta^h (1 - \theta)^t \,\mathrm{d}\theta = \dfrac{h!t!}{(n + 1)!}$. The odds that the coin is fair is then given by $\Phi = \dfrac{2^{-n}(n + 1)!}{h!t!}$. For $h = 59$, $h = 60$, and $h = 61$ the odds of the coin being fair, based on $n = 100$ tosses, are about 1.60, 1.10, and 0.72, respectively.

Thus, for $n = 100$ and $h = 60$ the posterior odds Φ is fairly close to one, that is, the probability of the coin being fair is about the same as the probability of it being unfair. To study how the data informed PBCM would perform when the evidence is inconclusive, we

constructed the data to be $\mathbf{x} = \{h = 60, t = 40\}$. Next, we applied the nonparametric bootstrap on $\mathbf{x}$ to get a distribution of estimated values for the binomial parameter θ under model U, based on $M = 1000$ bootstrap samples. Note that this nonparametric bootstrap distribution is very similar to the Bayesian posterior distribution for θ (e.g., Rubin, 1981). Next, each of the 1000 θ's is used to generate a new, replicate data set. For model F, a fixed value of $\theta = \frac{1}{2}$ is used to likewise generate $M = 1000$ replicate data sets. Each replicate data set is then fit by the U and F models, and the difference in maximum likelihood between these models is calculated. The U model has one free parameter θ, whereas in the F model θ is fixed at $\frac{1}{2}$. Because θ is free in the U model, the GOF value for U will always be higher than that for F. The optimal criterion β_m calculated using the data informed PBCM was estimated to be somewhere between $h = 55$ and $h = 56$. This result again shows that the data informed PBCM is biased toward selection of the complex model; in the above case of $h = 60$, a data informed PBCM analysis would indicate that the coin is unfair, whereas we know that the evidence is really inconclusive. A similar bias would be evident from a BPP analysis (see previous section).

In addition to the data informed PBCM analysis, we also performed a simulation in which the θ's for generating data from the U model are sampled not from the distribution for θ obtained by the nonparametric bootstrap, but instead from a uniform distribution $\theta \sim Uniform[0, 1]$. From a Bayesian perspective, this means that replicate data sets are generated based on the uniform prior $p(\theta)$ rather than on the posterior distribution for θ. In a simulation, the optimal criterion (i.e., the point where the difference distributions cross) did agree closely with the theoretical value of $h = 60$. This simple example highlights what it is about the data informed PBCM that makes it well-suited for addressing model adequacy, but less suited for general model selection.

In the coin tossing example, the data informed PBCM does not actually test whether the coin is fair or unfair. Instead, the data informed PBCM tests whether the coin is fair or whether it has a specific kind of unfairness—a kind of unfairness that is informed by the data. Thus, after observing $h = 60$, $t = 40$, the data informed PBCM tests model F as $\theta = \frac{1}{2}$ versus a specific instantiation of model U (i.e., a distribution of θ centered around the maximum likelihood value $\theta = 0.6$). Obviously, the fact that model U is inspired by the data will lead to a bias in its favor. If one wants to make general model selection claims that do not just apply to specific instantiations of models, sampling should arguably be done from the uninformed priors (cf. Box, 1980; but see Aitkin, 1991, and Laud & Ibrahim, 1995, for a discussion). One difficulty with sampling from the uninformed priors is how to determine such priors (for reviews see Berger, 1985; Kass & Wasserman, 1996)—indeed, the fact that the uninformed approach places probability mass on parameter values that prove to be extremely implausible is one of the motivations for using the informed approach:

> If the prior is intended to be 'objective', rather than to represent one's subjective belief, why should this objective prior assign weight to values (…) which are untenable given the data, thus reducing the probability of the observed data to zero? If the probability of the observed data goes to zero under the integrated model, this surely means that the *prior assignment* is untenable (Aitkin, 1991, p. 115).

Thus, the data informed version of the PBCM takes the extreme position that the only plausible data are the data that were observed in the experiment. This solves the problem that it is unclear in many models for psychological phenomena what exactly is the range of plausible parameter values. It should be kept in mind though, that the data informed PBCM yields a measure of local mimicry by comparing specific rather than generic instantiations of models.

In sum, applying the PBCM by generating simulated data sets from models that have first been fitted to the data implies testing model tokens, whereas generating simulated data sets from data uninformed models implies testing model types. In model selection, the interest is usually in the latter type of inference. Of course, prior knowledge does not need to be uninformative; if prior experience has taught that θ is, say, $Beta(3, 2)$ distributed, we can certainly use this prior instead of the uniform $Beta(1, 1)$ prior. The crucial point is that in order to select between model types, information from the data under consideration should not be used to constrain or specify the models of interest. The data uninformed PBCM method is sometimes called "landscaping" and is also discussed in detail by Navarro et al. (2003).

To further contrast the data informed versus the data uninformed versions of the PBCM, we now revisit Example 1, models of information integration. Recall that we previously contrasted two models of information integration (i.e., FLMP and LIM), and assessed model mimicry by cross-fitting replicate data sets that were generated by specific, data inspired versions of FLMP and LIM. We now present the results of a simulation in which the parameters of the two models are not based on the data. All other aspects of the simulation remained the same. Since the parameters of FLMP and LIM are probabilities, a first idea might be to let each simulated data set be generated by a new set of randomly (i.e., $\sim Uniform[0, 1]$) determined parameter values. However, the experimental design is such that the amount of support for each level $i + 1$ is always

higher than for level i. Therefore, each of the $M = 1000$ sets of parameter values was constructed by drawing $i = 5$ numbers from a uniform distribution on the interval from 0 to 1, ordering these random values, and then assigning them to the parameters in that order (cf. Myung & Pitt, 1997). This was done separately for FLMP and for LIM, and separately for the auditory parameters and the visual parameters. After the parameter values were assigned in this data uninformed fashion, the PBCM was applied in the usual manner. As can be seen from Fig. 9, the results confirm several conclusions that were based on the data informed PBCM analysis. First, the variability in the difference in maximum log likelihood is much greater for data generated from the FLMP than for data generated from LIM. Second, for the amount of data in the Massaro (1998) experiment, the FLMP and LIM models are not very confusable (i.e., the difference distributions overlap only slightly). Nonetheless, the FLMP is better able to account for LIM-generated data than vice versa. This asymmetry leads to a relatively small adjustment of the optimal criterion. The simulation shows that a difference in log likelihood of zero actually constitutes evidence in favor of LIM, and that a difference in log likelihood of 4 in favor of the FLMP represents the state of inconclusive evidence.

It is interesting to compare the data uninformed PBCM solution to the MDL solution with respect to the FLMP versus LIM problem. Using the MDL method, Navarro et al.(2003) found that the difference in 'geometric complexity' equaled about 1.88. It is important to realize that MDL complexity depends on a number of factors, such as sample size and experimental design (cf. Pitt et al., 2002, pp. 485–486). For comparison purposes, we performed a data uninformed PBCM analysis using the same 2×8 factorial design and the same parameter estimation procedure employed by Navarro et al. (2003). The overall PBCM results were similar to the results shown in Fig. 9, with the exception that LIM model recovery for the 2×8 design was relatively poor when the nominal criterion was used. We performed the data uninformed PBCM analysis for $n = 24$, i.e., the same sample size used by Navarro et al. (2003), and estimated the optimal PBCM criterion to be equal to a difference in log likelihood of about 2.7. Thus, the results from an MDL analysis are relatively close to those from a data uninformed PBCM analysis. Note that the difference in MDL complexity between FLMP and LIM is not affected by sample size, because sample size adds the same amount of complexity for both models (cf. Pitt et al. Eq. (8), noting that both models have the same number of free parameters).

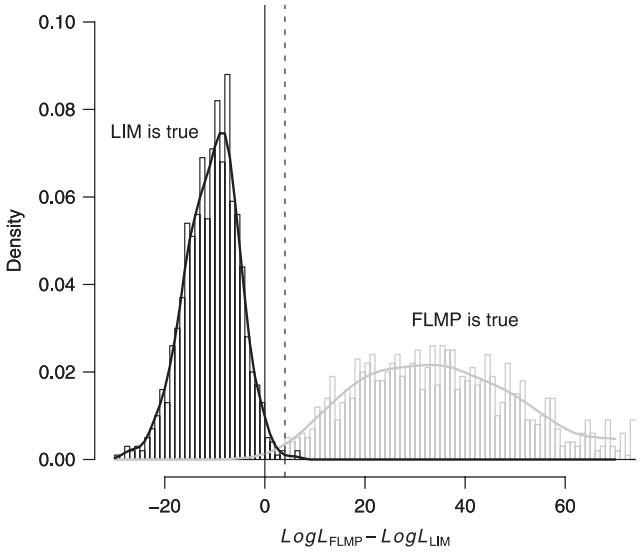


Fig. 9. Difference distributions obtained by the data uninformed PBCM when parameter values are determined semi-randomly instead of by first fitting the models to data. The optimal criterion β_m, indicated by the dashed line, was calculated from the empirical distribution functions, not from the Gaussian kernel estimates. The bin-width for the bar graph and the kernel estimate was 1.

It is interesting that based on the same 2×8 design, Pitt et al. (2002) report a difference in geometric complexity between FLMP and LIM equal to 8.74, obviously much larger than the 1.88 figure reported in Navarro et al. (2003). We suspected that the smaller number from Navarro et al. (2003) might have been due to the fact that Navarro et al. incorporated ordinal constraints on the parameter values, whereas Pitt et al. (2002) might not have done so—this was later confirmed by I. J. Myung (personal communication, November 7, 2003). To further investigate this issue, we performed a data uninformed PBCM analysis for the same design used by Pitt et al. (2002), omitting any ordinal parameter constraints. As expected, the optimal PBCM criterion was now substantially increased, and roughly equaled 11. These simulations show that the incorporation of parameter constraints greatly reduce the a priori difference in model flexibility between FLMP and LIM. In addition, the correspondence between the data uninformed PBCM and MDL in the two cases is remarkable.

Model selection methods such as MDL and BMS aim to minimize prediction error and hence maximize *generalizability*—the better a model approximates the true data-generating process, the higher its predictive value for data that have yet to be observed (cf. Hastie et al., 2001, Chap. 7). The PBCM does not focus directly on model generalizability, but instead addresses the issue of model *mimicry*. Despite the fact that MDL/BMS and the data uninformed PBCM are motivated by different ideas, the fact that in several examples (i.e., coin tossing, FLMP vs. LIM) we have found a relatively close correspondence between these methods warrants further research (e.g., along the lines of the 'partial information Bayes factor', Geweke, 1999a, b Sections 6 and 5; Geweke & McCausland, 2001).

In sum, the data informed version of the PBCM is useful to assess model adequacy and model mimicry for a specific data set. When it is necessary to make more general claims about the plausibility of model types rather than model tokens, the data uninformed version of the PBCM is more appropriate. Since the two methods address different issues and operate at different levels (e.g., types versus tokens, global versus local mimicry) both should ideally be used. We would like to mention that a data uninformed PBCM analysis could very well incorporate prior knowledge about ranges of plausible parameter values and parameter covariance. For instance, much research has been done exploring parameters of the diffusion model by fitting the model to a wide range of data (e.g., Ratcliff, Gomez, & McKoon, in press; Ratcliff & Rouder, 1998, 2000; Ratcliff, Thapar, & McKoon, 2001; Ratcliff & Tuerlinckx, 2002). The data uninformed PBCM could be used by sampling parameter values from the parameter ranges identified by this prior work. For instance, Ratcliff et al. (2001) and Thapar, Ratcliff, and McKoon (2003) present distributions of diffusion model parameters for groups of old and young participants. When a new but related experiment was to be performed with the aim of assessing model mimicry for the diffusion model versus a competitor model (cf. Ratcliff & Smith, in press), these distributions of parameter values could be combined with uniform priors to yield informed prior distributions for the diffusion model. A data uninformed PBCM analysis would then have the diffusion model generate simulated data from these informed priors.

3.6. A priori power analysis using PBCM

Up to this point we have discussed the PBCM as a tool to assess model mimicry. An assessment of model mimicry can have useful practical applications. In particular, the PBCM can provide an estimate of the number of empirical observations that would be required to distinguish two models with a certain probability of success. We now illustrate how such an a priori analysis in terms of statistical power is done with respect to two models that provide different descriptions of how response accuracy increases to asymptote as a function of processing time.

3.7. Example 4. Speed–accuracy curves

A considerable amount of research in psychology has produced data that can be described by curves that initially rise quickly and then slowly approach asymptote, such as learning curves (e.g., Estes, 1956). The functions considered here are speed–accuracy trade-off (SAT) curves and describe the growth of accuracy as a function of time in procedures such as the response signal method (e.g., Ratcliff & Iverson, 1984; Reed, 1976; Wickelgren, 1977). Two specific SAT models are examined: an exponential SAT function and an SAT function derived from the diffusion model (e.g., Ratcliff, 1978 for details) to describe the accumulation of information during retrieval. The two SAT models are of similar shape yet do not mimic each other exactly, and so are identifiably different, given sufficient data.

Dosher (1981) and McElree & Dosher (1989) have fitted both SAT models to data and have found that the exponential SAT curve provides a slightly better fit than the diffusion model SAT curve. Here we will attempt to examine the issue in more detail. Discrimination between these two models is of theoretical interest because the models represent different theoretical interpretations of information processing (see Dosher, 1981; Ratcliff, 1978, 1988b; Usher & McClelland, 2001, pp. 559–568; Wickelgren, 1976).

The equation for the exponential growth to a limit is:

$$d'(t) = A\{1 - \exp[-R^{-1}(t - I)]\}, \quad t > I,$$
$$d'(t) = 0, \quad t \leqslant I, \tag{8}$$

where A is the asymptotic value of d', R is the time constant (for $t = R + I$, the curve is about 2/3 the way to asymptote), and I is the time intercept (i.e., the time at which performance starts to exceed chance). The equation derived from the diffusion model (e.g., Ratcliff, 1978, Eq. (10)) is:

$$d'(t) = \frac{A}{\sqrt{1 + [V/(t - I)]}}, \quad t > I,$$
$$d'(t) = 0, \quad t \leqslant I, \tag{9}$$

where V is a constant rate parameter that determines the rate of approach to asymptote. Now suppose the aim is to perform an experiment to distinguish between the two SAT curves described above. Simulations using the PBCM can provide an indication of how many observations are required to make a choice between the two SAT models with some level of certainty. First, parameter values and values of the independent variables (t) should be chosen that are typical of response signal experiments (e.g., Corbett, 1977; Dosher, 1981; McElree & Carrasco, 1999; McElree & Dosher, 1993; Ratcliff, 1978; Reed, 1976; Wickelgren, Corbett, & Dosher, 1980). In the simulations reported here, the parameters for the exponential SAT function were set to $A = 3$, $I = 320$, and $R = 200$ (cf. Eq. (8)). The parameter values of the diffusion SAT function that provides the best least-squares fit to the exponential SAT function were $A = 3.5$, $I = 346$, and $V = 399$ (cf. Eq. (9)). The information processing times t used in the simulation were $t = (350, 400, 500, 600, 800, 1000, 1500, 2000)$. Fig. 10 shows the noise-free SAT functions for the exponential model and the diffusion model.

In an experimental situation we expect that the SAT functions are not deterministic but rather stochastic.

Assuming additive normal noise with constant variance over time and using the parameter values specified above, 1000 data sets were generated from each model by $d'(t) = f(t) + N(0, \sigma_\varepsilon)$, where $f(t)$ for the exponential SAT curve and the diffusion SAT curve are given by Eqs. (8) and (9), respectively. Thus, each simulated data set consisted of 8 d' values, one for each level of processing time t. Occasional negative d' values can occur, especially when processing time t is very limited. The impact of the noise component was varied on three levels, $\sigma_\varepsilon = 0.25$, $\sigma_\varepsilon = 0.15$, and $\sigma_\varepsilon = 0.10$. When the standard deviation of the noise component is relatively large, we expect the PBCM difference distributions to have large variance and relatively much overlap, making discrimination difficult.

The exponential SAT function and the diffusion SAT function were fit to the 1000 data sets generated by exponential model and to the 1000 data sets generated by the diffusion model using least-squares minimization. Thus, the PBCM was applied, resulting in two difference distributions for each of the three noise levels. Fig. 11 shows these difference distributions in RMSD (cf. footnote 4), from which it is apparent that the overlap of the distributions diminishes as the impact of noise on d' is reduced. Overall, the distributions appear to be roughly symmetrical around the nominal criterion of $\Delta RMSD = 0$. A more precise quantitative analysis is presented in Table 3. The analysis in Table 3 focuses on the probability of correct model recovery. This probability is usually presented in a confusion matrix (cf. Fig. 2, bottom panel) and given by the area of the difference distribution to the left or to the right of the nominal decision criterion. As can be seen from Table 3, for all levels of noise there is an asymmetry in the confusion matrix that indicates the exponential model to be more flexible than the diffusion model when the nominal criterion is used. However, we have already shown that an asymmetric confusion matrix need not imply a biased decision criterion when the variances of the difference distributions are unequal (cf. Fig. 3). Hence, Table 3 also presents the estimated optimal decision criterion β_m. For all three noise levels, the overall probability of correct model recovery increases when the decision criterion is shifted away from the nominal criterion so as to favor the diffusion SAT curve. This means that the nominal criterion $\Delta RMSD = 0$ is biased and that the exponential SAT function is better able to capture SAT data generated by the diffusion model than vice versa.

It should be noted that this bias is not very reliable, as can be seen from the standard errors for both confusion matrix asymmetry and β_m presented in Table 3 (the standard errors were calculated from 500 bootstrap samples of the two difference distributions). Nevertheless, the exponential SAT curve is arguably more flexible than the diffusion SAT curve when all noise levels are simultaneously taken into account. The effect is not very large in an absolute sense, and is unlikely to be of great practical significance.

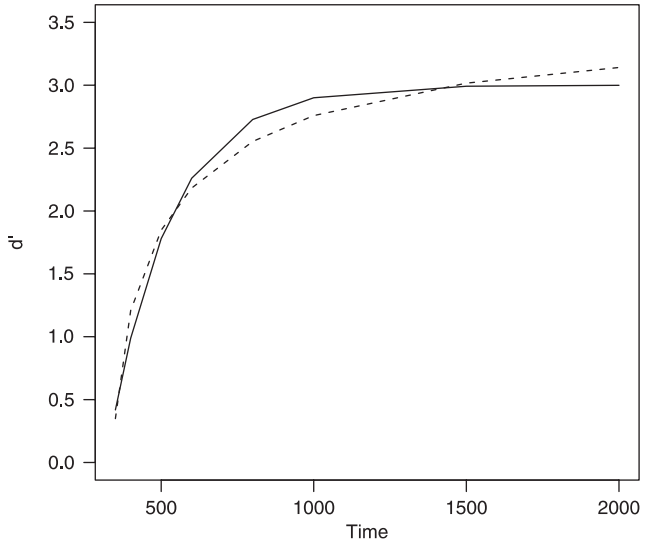


Fig. 10. The noise-free exponential SAT function (e.g., Dosher, 1981; $A = 3$, $I = 320$, $R = 200$ in Eq. (14); the solid line) and the noise-free diffusion SAT function (e.g., Ratcliff, 1978; $A = 3.5$, $I = 346$, $V = 399$ in Eq. (15); the dashed line) for the information processing times used in the simulations presented here, i.e., $t = (350, 400, 500, 600, 800, 1000, 1500, 2000)$. The parameters of the diffusion SAT function were determined by least-squares fitting of the noise-free exponential SAT function.

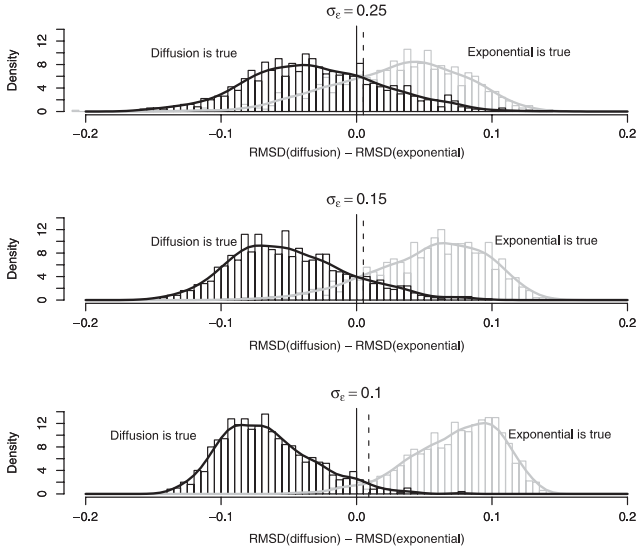


Fig. 11. Difference distributions in RMSD obtained by the PBCM applied to the exponential SAT function and the diffusion SAT function. Top panel: standard deviation of normally distributed noise in d' (i.e., σ_ε) = 0.25. Middle panel: $\sigma_\varepsilon = 0.15$. Bottom panel: $\sigma_\varepsilon = 0.1$. The optimal criteria β_m, indicated by dashed lines, were calculated from the empirical distribution functions, not from the Gaussian kernel estimates. The bin-width for the bar graph and the kernel estimate was 0.005 (see text for details).

Table 3
Probability of correct model recovery for three noise levels, using the nominal criterion or (between brackets) the optimal criterion

σ_ε	Exp.	Diffusion	Asym.	se_{boot} (Asym.)	β_m	se_{boot} (β_m)
0.10	0.969 (0.955)	0.954 (0.975)	0.015	0.008	0.010	0.003
0.15	0.890 (0.873)	0.858 (0.877)	0.032	0.015	0.005	0.006
0.25	0.743 (0.711)	0.728 (0.769)	0.015	0.020	0.005	0.006

Note: σ_ε = standard deviation of the normally distributed noise in d', Exp. = exponential model. The exponential parameters were $A = 3.0$, $R = 200$, $I = 320$. The diffusion model parameters were $A = 3.5$, $V = 399$, $I = 346$. These two sets of parameters were parameters that gave the best fits to the data points generated from the other model. Values of t used were 350, 400, 500, 600, 800, 1000, 1500, and 2000. Asym. = asymmetry in the confusion matrix (i.e., first column minus second column). The bootstrap estimates for standard error, se_{boot}, are based on 500 samples from each difference distribution.

We will now address the question we set out to answer, that is, how many observations are needed to distinguish the exponential SAT curve from the diffusion curve, given certain plausible parameter values and a given noise level? Given that the exponential model is true, the probability of the exponential fitting the data better is 0.969, and given the diffusion model is true the probability of the diffusion model fitting the data better is 0.954 (for $\sigma_\varepsilon = 0.10$). Thus we have a power analysis: if a set of future data is to have the power to discriminate the diffusion and exponential models at a 95% level of confidence, a sample size must be used that is sufficient to produce standard deviations in the data of about 0.1 (for the specific parameter values used here).

A rough calculation can give the sample size: for a d' of 3.1 (A in Eqs. (14) and (15)), hit and false alarm rates are 0.94 and 0.06, respectively. For a standard error of 0.1 in d', the hit and false alarm rate for $d' = 3.2$ would be 0.945 and 0.055. Thus changes in the hit and false alarm rate of 0.005 would be equivalent to a change of 0.1 in d'. Given that $se(\hat{p}) = \sqrt{p(1-p)/n}$ for the binomial, then for a standard error of 0.005 in p with $p = 0.94$, n would have to be about 2250. Thus, 2250 observations per condition would be needed for a single experiment to distinguish the two models with probability 0.95 (or a little better). Similarly, a 0.75 probability of correct model classification is associated with a standard error in d' of about 0.25 (cf. Table 3). Following the above, it can be calculated that about 220 observations per condition would be required to achieve this standard error.

Although this example is only approximate and applies only in the specific parameter domain used above, the example does demonstrate that useful information can be obtained about the ability of experiments to discriminate between models. For example, the diffusion and exponential models have been compared a number of times for goodness-of-fit and the conclusion has been drawn that both models fit the data quite well (e.g., Dosher, 1981). Usually the exponential fits the data a little better, but the above analysis shows that the exponential should fit the data slightly better (e.g., the exponential SAT function fits data generated by the diffusion SAT function slightly better than vice versa). Inspection of the cases in the above simulations in which the exponential fits better than the diffusion model when the diffusion model is true indicates that this occurs when the first point on the function is relatively high by chance. When the first point is relatively low, the diffusion model fits better when the exponential is true. This can be understood by realizing that the rise of the diffusion model is faster than the rise of the exponential: the diffusion model intercepts the abscissa with an infinite slope (i.e., rises perpendicularly at the x-axis). Thus the exponential can deal with a more linear rise while the diffusion model can deal with a more step-like rise. One way to address this problem is to add variability to the starting point of the SAT functions, corresponding to a variable onset of retrieval from memory. Such a variable onset can be achieved by assuming that the non-decision component of response time (i.e., T_{er} in the diffusion model, the time needed for encoding and response processes) varies across trials (e.g., Ratcliff et al., in press; Ratcliff & Smith, in press; Ratcliff & Tuerlinckx, 2002). In this case the diffusion model will no longer have such a steep rise and will probably fit the data just as well as the exponential even when the rise in the data appears more linear. An alternative possibility is that the observed behavior in signal-to-response experiments is a mixture of processes that have terminated (i.e., reached a response boundary in the diffusion model) by the time the signal-to-respond is detected, and guesses that can or cannot be based on partial information (cf. Ratcliff, 1988b).

Regardless of the details for the SAT models under consideration, the SAT simulations presented here demonstrate how the PBCM can be used to estimate the amount of data needed to discriminate two models with a certain probability of success. Such an a priori power analysis will increase the efficiency of experiments designed to provide data that discriminate two models. As noisy data are less informative than 'clean' data, we recommend to use the PBCM for different noise levels, as illustrated in Fig. 11.

4. Discussion

Although the important role of model mimicry in the evaluation of mathematical models of psychological phenomena is widely acknowledged (e.g., Massaro, 1998; Ratcliff, 1988a; Townsend, 1972; Van Zandt & Ratcliff, 1995), no general method has been proposed to quantitatively assess such model mimicry. In this article, we have shown how the parametric bootstrap cross-fitting method (e.g., Williams, 1970a, b) can be used to assess model mimicry, model discriminability, and model adequacy. Uncertainty in parameter estimation can be easily incorporated in the PBCM by taking nonparametric bootstrap samples from the data. Use of the PBCM was illustrated by an application to two models for information integration (i.e., FLMP and LIM). Conceptually, the PBCM is closely related to the diagnostic model check using Bayesian posterior predictive distributions.

The PBCM has a number of distinct advantages. First, the PBCM is a quite general and easy to implement procedure. The importance of this point should not be underestimated. Models in psychology have grown more and more complex in recent years, often making analytic solutions practically impossible. The PBCM is based on simple resampling techniques and does not depend on the availability of analytic solutions, nor is it based on asymptotic approximations that may not hold for limited data. The only requirements for using the PBCM are a program that estimates model parameters and a program that generates simulated data from the model. Thus, the PBCM circumvents many mathematical complications by computer-intensive simulations. The PBCM affords useful insights about the relative flexibility and appropriateness of competing models for a given data set with minimal investments in time and mathematical development. Also, since it is based on a simple resampling scheme, the PBCM can be applied to any kind of GOF measure, and is not constrained to problems that can be solved by maximum likelihood estimation. This opens up the possibility of addressing local and global mimicry properties of models such as connectionist models, random walk or diffusion models for response time and accuracy (e.g., Ratcliff & Smith, in press), and models based on procedural rule systems (e.g., ACT-R, Anderson & Lebiere, 1998).

Second, the PBCM can be used in advance of any data collection. This is of relevance because the PBCM may provide an estimate of how many experimental data are needed to reliably distinguish competing models, as was illustrated by two competing SAT functions. More generally, the PBCM has the potential to help identify the most diagnostic experimental conditions or measures of performance. In a similar fashion, bootstrap methods may be used to assess the relative flexibility of models as a function of their parameter values. That is, certain models may be easily mimicked by other models only for a specific subset of parameter values (cf. Ratcliff & Smith, in press; see Myung, 2002; Navarro et al., 2003; and Pitt et al., 2003, for details on the 'landscaping' technique).

The PBCM as implemented here is only appropriate for pair-wise model comparisons. Therefore, the method is less helpful in problems of variable selection. The main strength of the PBCM lies in the evaluation of a limited set of complex nonnested models. When using the data informed version of the PBCM it is particularly important to realize that it gives a local indication of model mimicry, model discriminability, or model adequacy. A measure of global mimicry can be obtained by sampling either from an uninformative prior distribution for the parameters, or by sampling from a limited range of parameter values that has been determined by an extensive body of prior work. The PBCM should however, at its present state of development, not be used to assess model generalizability. It should also be remembered that the PBCM is based on expected differences in goodness-of-fit provided that one of the competing models is true.

In sum, we believe that both the data informed and the data uninformed versions of the PBCM are useful tools for assessing model adequacy and model mimicry. The generality of the PBCM and the continuous increase in cheap computing power holds considerable promise for future application of the PBCM to complex models of human cognition.

Acknowledgments

Part of this work is closely related to earlier studies presented at the Seventeenth Annual Mathematical Psychology Meeting in Chicago, IL, by Ratcliff and Iverson (1984). We thank In Jae Myung, Mark Steyvers, and Rich Shiffrin for insightful discussions and kind advice. We thank Lourens Waldorp, Richard Golden, and Simon Farrell for extensive comments on an earlier version of this paper, and we are grateful to Trish Van Zandt for making available her Matlab code for the Gaussian kernel estimator. We are grateful to Dan Navarro and an anonymous reviewer for many useful suggestions for improvement.

References

Aitchison, J., & Dunsmore, I. R. (1975). *Statistical prediction analysis*. Cambridge: Cambridge University Press.

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

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov, & F. Caski (Eds.),

Proceeding of the second international symposium on information theory. Budapest: Akademiai Kiado.

Anderson, N. H. (1981). *Foundations of information integration theory*. New York: Academic Press.

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. London: Lawrence-Erlbaum Associates.

Andrews, D. W. K., & Buchinsky, M. (2000). A three-step method for choosing the number of bootstrap repetitions. *Econometrica*, *68*, 23–51.

Atkinson, R. C., Bower, G. H., & Crothers, E. J. (1965). *An introduction to mathematical learning theory*. New York: Wiley.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis*. New York: Springer.

Berkhof, J., van Mechelen, I., & Gelman, A. A Bayesian approach to the selection and testing of latent class models. *Statistica Sinica*, in press.

Bollback, J. P. (2002). Bayesian model adequacy and choice in phylogenetics. *Molecular Biology and Evolution*, *19*, 1171–1180.

Bollen, K. A., & Stine, R. (1992). Bootstrapping goodness of fit measures in structural equation models. *Sociological Methods and Research*, *21*, 205–229.

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society A*, *143*, 383–430.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference*: *A practical information-theoretic approach*. New York: Springer.

Bush, R. R., & Mosteller, F. (1955). *Stochastic models for learning*. New York: Wiley.

Collyer, C. E. (1985). Comparing strong and weak models by fitting them to computer-generated data. *Perception and Psychophysics*, *38*, 476–481.

Corbett, A. T. (1977). Retrieval dynamics for rote and visual image mnemonics. *Journal of Verbal Learning and Verbal Behavior*, *16*, 233–246.

Cox, D. R. (1962). Further results on tests of separate families of hypotheses. *Journal of the Royal Statistical Society B*, *24*, 406–424.

Crowther, C. S., Batchelder, W. H., & Hu, X. (1995). A measurement-theoretic analysis of the fuzzy logical model of perception. *Psychological Review*, *102*, 396–408.

Cutting, J. E. (2000). Accuracy, scope, and flexibility of models. *Journal of Mathematical Psychology*, *44*, 3–19.

Cutting, J. E., Bruno, N., Brady, N. P., & Moore, C. (1992). Selectivity, scope, and simplicity of models: A lesson from fitting judgments of perceived depth. *Journal of Experimental Psychology: General*, *121*, 364–381.

Davidson, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge: Cambridge University Press.

Davidson, R., & MacKinnon, J. G. (2000). Bootstrap tests: How many bootstraps? *Econometric Reviews*, *19*, 55–68.

Diciccio, T. J., & Romano, J. P. (1988). A review of bootstrap confidence intervals. *Journal of the Royal Statistical Society B*, *50*, 338–354.

Djurić, P. M. (1998). Asymptotic MAP criteria for model selection. *IEEE Transactions on Signal Processing*, *46*, 2726–2735.

Dosher, B. A. (1981). The effects of delay and interference: A speed-accuracy study. *Cognitive Psychology*, *13*, 551–582.

Dunn, J. C. (2000). Model complexity: The fit to random data reconsidered. *Psychological Research*, *63*, 174–182.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*, 193–242.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, *7*, 1–26.

Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, *1*, 54–77.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin*, *53*, 134–140.

Estes, W. K. (2002). Traps in the route to models of memory and decision. *Psychonomic Bulletin & Review*, *9*, 3–25.

Feng, Z. D., & McCulloch, C. E. (1996). Using bootstrap likelihood ratios in finite mixture models. *Journal of the Royal Statistical Society B*, *58*, 609–617.

Gelman, A., Goegebeur, Y., Tuerlinckx, F., & van Mechelen, I. (2000). Diagnostic checks for discrete-data regression models using posterior predictive simulations. *Applied Statistics*, *49*, 247–268.

Geweke, J. (1999a). Simulation methods for model criticism and robustness analysis. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics 6* (pp. 275–299). Oxford: Oxford University Press.

Geweke, J. (1999b). Using simulation methods for Bayesian econometric models: Inference, development, and communication. *Econometric Reviews*, *18*, 1–126.

Geweke, J., & McCausland, W. (2001). Bayesian specification analysis in econometrics. *The American Journal of Agricultural Economics*, *83*, 1181–1186.

Golden, R. M. (1995). Making correct statistical inferences using a wrong probability model. *Journal of Mathematical Psychology*, *39*, 3–20.

Golden, R. M. (2003). Discrepancy risk model selection test theory for comparing possibly misspecified or nonnested models. *Psychometrika*, *68*, 229–249.

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

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

Hall, P. (1988). Theoretical comparison of bootstrap confidence intervals. *Annals of Statistics*, *16*, 927–953.

Hall, P. (1992). *The bootstrap and Edgeworth expansion*. New York: Springer.

Hall, P., & Wilson, S. R. (1991). Two guidelines for bootstrap hypothesis testing. *Biometrics*, *47*, 757–762.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning*: *Data mining, inference, and prediction*. New York: Springer.

Hinkley, D. V. (1988). Bootstrap methods. *Journal of the Royal Statistical Society B*, *50*, 321–337.

Horowitz, J. L. (2001). The bootstrap. In J. J. Heckman, & E. E. Leamer (Eds.), *Handbook of econometrics: Vol. 5* (pp. 3159–3228). Amsterdam: Elsevier Science.

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

Kass, R. E., & Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, *91*, 1343–1369.

Laud, P. W., & Ibrahim, J. G. (1995). Predictive model selection. *Journal of the Royal Statistical Society B*, *57*, 247–262.

Massaro, D. W. (1988). Some criticisms of connectionist models of human performance. *Journal of Memory and Language*, *27*, 213–234.

Massaro, D. W. (1998). *Perceiving talking faces*: *From speech perception to a behavioral principle*. Cambridge, MA: MIT Press.

Massaro, D. W., Cohen, M. M., Campbell, C. S., & Rodriguez, T. (2001). Bayes factor of model selection validates FLMP. *Psychonomic Bulletin & Review*, *8*, 1–17.

Massaro, D. W., & Friedman, D. (1990). Models of integration given multiple sources of information. *Psychological Review*, *97*, 225–252.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*, 1–86.

McElree, B., & Carrasco, M. (1999). The temporal dynamics of visual search: Speed-accuracy tradeoff analysis of feature and conjunctive searches. *Journal of Experimental Psychology: Human Perception and Performance*, *25*, 1517–1539.

McElree, B., & Dosher, B. A. (1989). Serial position and set size in short-term memory: The time course of recognition. *Journal of Experimental Psychology: General*, *118*, 346–373.

McElree, B., & Dosher, B. A. (1993). Serial retrieval processes in the recovery of order information. *Journal of Experimental Psychology: General*, *122*, 291–315.

McLachlan, G. J. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Applied Statistics*, *36*, 318–324.

Meng, X.-L. (1994). Posterior predictive *p*-values. *The Annals of Statistics*, *22*, 1142–1160.

Metropolis, N., & Ulam, S. (1949). The Monte Carlo method. *Journal of the American Statistical Association*, *44*, 335–341.

Movellan, J. R., & McClelland, J. L. (2001). The Morton-Massaro law of information integration: Implications for models of perception. *Psychological Review*, *108*, 113–148.

Myung, I. J. (2002). *Comparing models with 'landscaping'*. Presentation for the 1st Annual Summer Interdisciplinary Conference, Squamish (B.C.), Canada, August 2002.

Myung, I. J., Forster, M. R., & Browne, M. W. (Eds.), (2000). Model selection [Special issue]. *Journal of Mathematical Psychology*, *44*(1–2).

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

Navarro, D. J., Myung, I. J., Pitt, M. A., & Kim, W. (2003). Global model analysis by landscaping. *Proceedings of the 25th annual conference of the cognitive science society*.

Navarro, D. J., Pitt, M. A., & Myung, I. J. (2003). Assessing the distinguishability of models and the informativeness of data. Manuscript submitted for publication.

Oden, G. C., & Massaro, D. W. (1978). Integration of featural information in speech perception. *Psychological Review*, *85*, 172–191.

Parzen, E., Tanabe, K., & Kitagawa, G. (Eds.), (1998). *Selected papers of Hirotugu Akaike*. New York: Springer.

Pitt, M. A., Kim, W., & Myung, I. J. (2003). Flexibility versus generalizability in model selection. *Psychonomic Bulletin and Review*, *10*, 29–44.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, *109*, 472–491.

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1986). *Numerical recipes*. Cambridge: Cambridge University Press.

Raftery, A. E. (1995). Bayesian model selection in social research (with discussion). In P. V. Marsden (Ed.), *Sociological methodology* (pp. 111–196). Cambridge: Blackwells.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59–108.

Ratcliff, R. (1979). Group reaction time distributions and an analysis of distribution statistics. *Psychological Bulletin*, *86*, 446–461.

Ratcliff, R. (1988a). A note on mimicking additive reaction time models. *Journal of Mathematical Psychology*, *32*, 192–204.

Ratcliff, R. (1988b). Continuous versus discrete information processing: Modeling accumulation of partial information. *Psychological Review*, *95*, 238–255.

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, *114*, 510–532.

Ratcliff, R., Gomez, P., & McKoon, G. A diffusion model account of the lexical decision task. *Psychological Review*, in press.

Ratcliff, R. & Iverson, G. J. (1984). Methods for investigating parameters spaces of models. Paper presented at the Seventeenth Annual Mathematical Psychology Meeting, Chicago, IL.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, *9*, 347–356.

Ratcliff, R., & Rouder, J. N. (2000). A diffusion model account of masking in two-choice letter identification. *Journal of Experimental Psychology: Human Perception and Performance*, *26*, 127–140.

Ratcliff, R. & Smith, P. L. A comparison of sequential sampling models for two-choice reaction time. *Psychological Review*, in press.

Ratcliff, R., Thapar, A., & McKoon, G. (2001). The effects of aging on reaction time in a signal detection task. *Psychology and Aging*, *16*, 323–341.

Ratcliff, R., Thapar, A., & McKoon, G. (2003). A diffusion model analysis of the effects of aging on brightness discrimination. *Perception and Psychophysics*, *65*, 523–535.

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin and Review*, *9*, 438–481.

Reed, A. V. (1976). List length and the time-course of recognition in immediate memory. *Memory & Cognition*, *4*, 16–30.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, *42*, 40–47.

Rissanen, J. (2001). Strong optimality of the normalized ML models as universal codes and information in data. *IEEE Transactions on Information Theory*, *47*, 1712–1717.

Rubin, D. B. (1981). The Bayesian bootstrap. *Annals of Statistics*, *9*, 130–134.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, *12*, 1151–1172.

Rubin, D. B., & Stern, H. S. (1994). Testing in latent class models using a posterior predictive check distribution. In A. von Eye, & C. Clogg (Eds.), *Latent variable analysis: Applications for Developmental Research*. Thousand Oaks, CA: Sage.

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

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London: Chapman & Hall.

Stuart, A. & Ord, J. K. (1991). *Kendall's advanced theory of statistics*. (Vol. 2). New York: Oxford University Press.

Thapar, A., Ratcliff, R., & McKoon, G. (2003). A diffusion model analysis of the effects of aging on letter discrimination. *Psychology and Aging*, *18*, 415–429.

Thurman, W. N. (2001). Bayesian specification analysis in econometrics: Comment. *The American Journal of Agricultural Economics*, *83*, 1187–1189.

Townsend, J. T. (1972). Some results concerning the identifiability of parallel and serial processes. *British Journal of Mathematical and Statistical Psychology*, *25*, 168–197.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, *108*, 550–592.

Van Zandt, T. (2002). Analysis of response time distributions. In J. T. Wixted (Ed.), *Stevens' handbook of experimental psychology* (pp. 461–516) Vol. 4 (3rd ed.). New York: Wiley Press.

Van Zandt, T., & Ratcliff, R. (1995). Statistical mimicking of reaction time data: Single-process models, parameter variability, and mixtures. *Psychonomic Bulletin and Review*, *2*, 20–54.

von Neumann, J. (1951). Various techniques used in connection with random digits. *National Bureau of Standards, Applied Mathematics Series*, *12*, 36–38.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, *57*, 307–333.

Wagenmakers, E.-J., & Farrell, S. AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, in press.

Wagenmakers, E.-J., Farrell, S., & Ratcliff, R., Naïve nonparametric bootstrap model weights are biased. *Biometrics*, in press.

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, *44*, 92–107.

Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: II. Bootstrap-based confidence intervals and sampling. *Perception and Psychophysics*, *63*, 1314–1329.

Wickelgren, W. A. (1976). Network strength theory of storage and retrieval dynamics. *Psychological Review*, *83*, 466–478.

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica*, *41*, 67–85.

Wickelgren, W. A., Corbett, A. T., & Dosher, B. A. (1980). Priming and retrieval from short-term memory: A speed accuracy trade-off analysis. *Journal of Verbal Learning and Verbal Behavior*, *19*, 387–404.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics*, *9*, 60–62.

Williams, D. A. (1970). In discussion of ''A method for discriminating between models'' by A.C. Atkinson. *Journal of the Royal Statistical Society B*, *32*, 350.

Williams, D. A. (1970a). Discrimination between regression models to determine the pattern of enzyme synthesis in synchronous cell cultures. *Biometrics*, *26*, 23–32.