14 Model Comparison and the Principle of Parsimony

Joachim Vandekerckhove — University of California, Irvine
Dora Matzke — University of Amsterdam
Eric-Jan Wagenmakers — University of Amsterdam

14.1 Introduction

At its core, the study of psychology is concerned with the discovery of plausible explanations for human behavior. For instance, one may observe that "practice makes perfect": as people become more familiar with a task, they tend to execute it more quickly and with fewer errors. More interesting is the observation that practice tends to improve performance such that most of the benefit is accrued early on, a pattern of diminishing returns that is well described by a power law (Logan, 1988; but see Heathcote, Brown, & Mewhort, 2000). This pattern occurs across so many different tasks (e.g., cigar rolling, maze solving, fact retrieval, and a variety of standard psychological tasks) that it is known as the "power law of practice". Consider, for instance, the lexical decision task, a task in which participants have to decide quickly whether a letter string is an existing word (e.g., *sunscreen*) or not (e.g., *tolphin*). When repeatedly presented with the same stimuli, participants show a power law decrease in their mean response latencies; in fact, they show a power law decrease in the entire response time distribution, that is, both the fast responses and the slow responses speed up with practice according to a power law (Logan, 1992).

The observation that practice makes perfect is trivial, but the finding that practice-induced improvement follows a general law is not. Nevertheless, the power law of practice only provides a descriptive summary of the data and does not explain the reasons why practice should result in a power law improvement in performance. In order to go beyond direct observation and statistical summary, it is necessary to bridge the divide between observed performance on the one hand and the pertinent psychological processes on the other. Such bridges are built from a coherent set of assumptions about the underlying cognitive processes—a theory. Ideally, substantive psychological theories are formalized as quantitative models (Busemeyer & Diederich, 2010; Lewandowsky & Farrell, 2010). For example, the power law of practice has been explained by instance theory (Logan, 1992, 2002). Instance theory stipulates that earlier experiences are stored in memory as individual traces or instances; upon presentation of a stimulus, these instances race to be retrieved, and the winner of the race initiates a response. Mathematical analysis shows that, as instances are added to memory, the finishing time of the winning instance decreases as a power function. Hence, instance theory provides a simple and general explanation of the power law of practice.

For all its elegance and generality, instance theory has not been the last word on the power law of practice. The main reason is that single phenomena often afford different competing explanations. For example, the effects of practice can also be accounted for by Rickard's component power laws model (Rickard, 1997), Anderson's ACT-R model (Anderson et al., 2004), Cohen et al.'s PDP model (Cohen, Dunbar, & McClelland, 1990), Ratcliff's diffusion model (Dutilh, Vandekerckhove, Tuerlinckx, & Wagenmakers, 2009; Ratcliff, 1978), or Brown and Heathcote's linear ballistic accumulator model (Brown & Heathcote, 2005, 2008; Heathcote & Hayes, 2012). When various models provide competing accounts of the same data set, it can be difficult to choose between them. The process of choosing between models is called model comparison, model selection, or hypothesis testing, and it is the focus of this chapter.

A careful model comparison procedure includes both qualitative and quantitative elements. Important qualitative elements include the plausibility, parsimony, and coherence of the underlying assumptions, the consistency with known behavioral phenomena, the ability to explain rather than describe data, and the extent to which model predictions can be falsified through experiments. Here we ignore these important aspects and focus solely on the quantitative elements. The single most important quantitative element of

model comparison relates to the ubiquitous tradeoff between parsimony and goodness-of-fit (Pitt & Myung, 2002). The motivating insight is that the appeal of an excellent fit to the data (i.e., high descriptive adequacy) needs to be tempered to the extent that the fit was achieved with a highly complex and powerful model (i.e., low parsimony).

The topic of quantitative model comparison is as important as it is challenging; fortunately, the topic has received—and continues to receive—considerable attention in the field of statistics, and the results of those efforts have been made accessible to psychologists through a series of recent special issues, books, and articles (e.g., Grünwald, 2007; Myung, Forster, & Browne, 2000; Pitt & Myung, 2002; Wagenmakers & Waldorp, 2006). Here we discuss several procedures for model comparison, with an emphasis on minimum description length and the Bayes factor. Both procedures entail principled and general solutions to the tradeoff between parsimony and goodness-of-fit.

The outline of this chapter is as follows. The first section describes the principle of parsimony and the unavoidable tradeoff with goodness-of-fit. The second section summarizes the research of Wagenaar and Boer (1987) who carried out an experiment to compare three competing multinomial processing tree models (MPTs; Batchelder & Riefer, 1980); this model comparison exercise is used as a running example throughout the chapter. The third section outlines different methods for model comparison and applies them to Wagenaar and Boer's MPT models. We focus on two popular information criteria, the AIC and the BIC, on the Fisher information approximation of the minimum description length principle, and on Bayes factors as obtained from importance sampling. The fourth section contains conclusions and take-home messages.

14.2   The Principle of Parsimony

Throughout history, prominent philosophers and scientists have stressed the importance of parsimony. For instance, in the Almagest—a famous 2nd-century book on astronomy—Ptolemy writes: "We consider it a good principle to explain the phenomena by the simplest hypotheses that can be established, provided this does not contradict the data in an important way." Ptolemy's principle of parsimony is widely known as Occam's razor (see Box 14.1); the principle is intuitive as it puts a premium on elegance. In addition, most people feel naturally attracted to models and explanations that are easy to understand and communicate. Moreover, the principle also gives ground to reject propositions that are without empirical support, including extrasensory perception, alien abductions, or mysticism. In an apocryphal interaction, Napoleon Bonaparte asked Pierre-Simon Laplace why the latter's book on the universe did not mention its creator, only to receive the curt reply "I had no need of that hypothesis".

However, the principle of parsimony finds its main motivation in the benefits that it bestows those who use models for prediction. To see this, note that empirical data are assumed to be composed of a structural, replicable part and an idiosyncratic, non-replicable part. The former is known as the signal, and the latter is known as the noise (Silver, 2012). Models that capture all of the signal and none of the noise provide the best possible predictions to unseen data from the same source. Overly simplistic models, however, fail to capture part of the signal; these models underfit the data and provide poor predictions. Overly complex models, on the other hand, mistake some of the noise for actual signal; these models overfit the data and again provide poor predictions. Thus, parsimony is essential because it helps discriminate the signal from the noise, allowing better prediction and generalization to new data.

Goodness-of-fit

"From the earliest days of statistics, statisticians have begun their analysis by proposing a distribution for their observations and then, perhaps with somewhat less enthusiasm, have checked on whether this distribution is true. Thus over the years a vast number of test procedures have appeared, and the study of these procedures has come to be known as goodness-of-fit" (D'Agostino & Stephens, 1986, p. v).

Frame 14.1: Occam's razor.

Occam's razor (sometimes *Ockham's*) is named after the English philosopher and Franciscan friar Father William of Occam (c.1288-c.1348), who wrote "Numquam ponenda est pluralitas sine necessitate" (plurality must never be posited without necessity), and "Frustra fit per plura quod potest fieri per pauciora" (it is futile to do with more what can be done with less). Occam's metaphorical razor symbolizes the principle of parsimony: by cutting away needless complexity, the razor leaves only theories, models, and hypotheses that are as simple as possible without being false. Throughout the centuries, many other scholars have espoused the principle of parsimony; the list predating Occam includes Aristotle, Ptolemy, and Thomas Aquinas ("it is superfluous to suppose that what can be accounted for by a few principles has been produced by many"), and the list following Occam includes Isaac Newton ("We are to admit no more causes of natural things than such as are both true and sufficient to explain their appearances. Therefore, to the same natural effects we must, so far as possible, assign the same causes."), Bertrand Russell, Albert Einstein ("Everything should be made as simple as possible, but no simpler"), and many others.

In the field of statistical reasoning and inference, Occam's razor forms the foundation for the principle of minimum description length (Grünwald, 2000, 2007). In addition, Occam's razor is automatically accommodated through Bayes factor model comparisons (e.g., Jeffreys, 1961; Jefferys & Berger, 1992; MacKay, 2003). Both minimum description length and Bayes factors feature prominently in this chapter as principled methods to quantify the tradeoff between parsimony and goodness-of-fit. Note that parsimony plays a role even in classical null-hypothesis significance testing, where the simpler null hypothesis is retained unless the data provide sufficient grounds for its rejection.

The *goodness-of-fit* of a model is a quantity that expresses how well the model is able to account for a given set of observations. It addresses the following question: Under the assumption that a certain model is a true characterization of the population from which we have obtained a sample, and given the best fitting parameter estimates for that model, how well does our sample of data agree with that model?

Various ways of quantifying goodness-of-fit exist. One common expression involves a Euclidean distance metric between the data and the model's best prediction (the least squared error or LSE metric is the most well-known of these). Another measure involves the likelihood function, which expresses the likelihood of observing the data under the model, and is maximized by the best fitting parameter estimates (Myung, 2000).

Parsimony

Goodness-of-fit must be balanced against model complexity in order to avoid overfitting—that is, to avoid building models that well explain the data at hand, but fail in out-of-sample predictions. The principle of parsimony forces researchers to abandon complex models that are tweaked to the observed data in favor of simpler models that can generalize to new data sets.

A common example is that of polynomial regression. Figure 14.1 gives a typical example. The observed data are the circles in both the left and right panels. Crosses indicate unobserved, out-of-sample data points to which the model should generalize. In the left panel, a quadratic function is fit to the 8 observed data points, whereas the right panel shows a 7$^{\text{th}}$ order polynomial function fitted to the same data. Since a polynomial of degree 7 can be made to contain any 8 points in the plane, the observed data are perfectly captured by the best fitting polynomial. However, it is clear that this function generalizes poorly to the unobserved samples, and it shows undesirable behavior for larger values of x.

In sum, an adequate model comparison method needs to discount goodness-of-fit with model complexity.
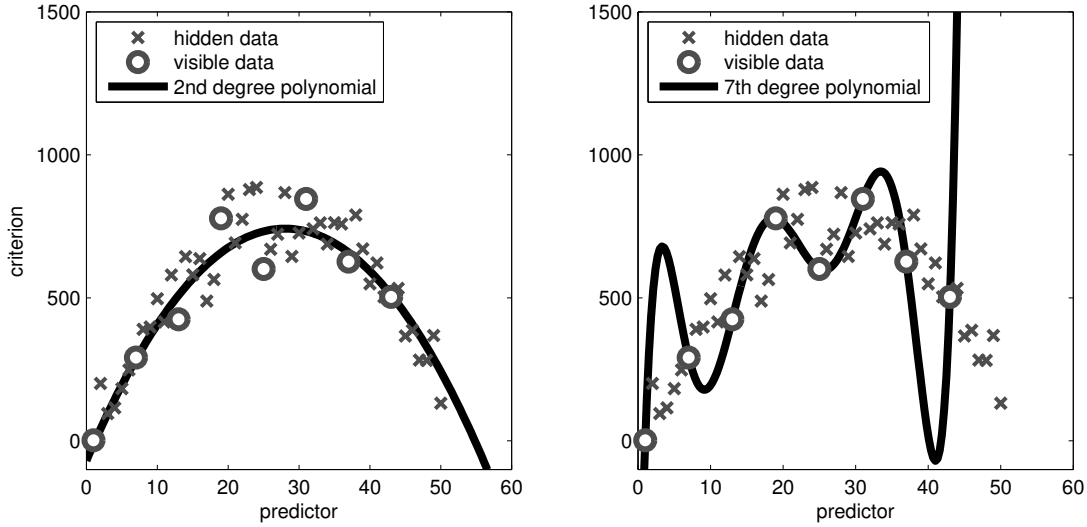
Figure 14.1: A polynomial regression of degree d is characterized by $\hat{y} = \sum_{i=0}^{d} a_i x^i$. This model has $d + 1$ free parameters a_i; hence, in the right panel, a polynomial of degree 7 perfectly accounts for the 8 visible data points. This 7th order polynomial, however, accounts poorly for the out-of-sample data points.

But how exactly can this be accomplished? As we will describe shortly, several model comparison methods are currently in vogue; all resulting from principled ideas on how to obtain *measures of generalizability*[1], meaning that these methods attempt to quantify the extent to which a model predicts unseen data from the same source (cf. Figure 14.1). Before outlining the details of various model comparison methods we now introduce a data set that serves as a working example throughout the remainder of the chapter.

Example: Competing Models of Interference in Memory

For an example model comparison scenario, we revisit a study by Wagenaar and Boer (1987) on the effect of misleading information on the recollection of an earlier event. The effect of misleading postevent information was first studied systematically by Loftus, Miller, and Burns (1978); for a review of relevant literature see Wagenaar and Boer (1987) and references therein.

   Wagenaar and Boer (1987) proposed three competing theoretical accounts of the effect of misleading postevent information. To evaluate the three accounts, Wagenaar and Boer set up an experiment and introduced three quantitative models that translate each of the theoretical accounts into a set of parametric assumptions that together give rise to a probability density over the data, given the parameters.

---

[1]This terminology is due to Pitt and Myung (2002), who point out that measures often referred to as "model fit indices" are in fact more than mere measures of fit to the data—they combine fit to the data with parsimony and hence measure generalizability. We adopt their more accurate terminology here.

Abstract Accounts

Wagenaar and Boer (1987) outlined three competing theoretical accounts of the effect of misleading postevent information on memory. Loftus' *destructive updating* model (DUM) posits that the conflicting information replaces and destroys the original memory. A *coexistence* model (CXM) asserts that an inhibition mechanism suppresses the original memory, which nonetheless remains viable though temporarily inaccessible. Finally, a *no-conflict* model (NCM) simply states that misleading postevent information is ignored, except when the original information was not encoded or already forgotten.

Experimental Design

The experiment by Wagenaar and Boer (1987) proceeded as follows. In Phase I, a total of 562 participants were shown a sequence of events in the form of a pictorial story involving a pedestrian-car collision. One picture in the story would show a car at an intersection, and a traffic light that was either red, yellow, or green. In Phase II, participants were asked a set of test questions with (potentially) conflicting information: Participants might be asked whether they remembered a pedestrian crossing the road when the car approached the "traffic light" (in the consistent group), the "stop sign" (in the inconsistent group) or the "intersection" (the neutral group). Then, in Phase III, participants were given a recognition test about elements of the story using picture pairs. Each pair would contain one picture from Phase I and one slightly altered version of the original picture. Participants were then asked to identify which of the pair had featured in the original story. A picture pair is shown in Figure 14.2, where the intersection is depicted with either a traffic light or a stop sign. Finally, in Phase IV, participants were informed that the correct choice in Phase III was the picture with the traffic light, and were then asked to recall the color of the traffic light. By design, this experiment should yield different response patterns depending on whether the conflicting postevent information destroys the original information (destructive updating model), only suppresses it temporarily (coexistence model), or does not affect the original information unless it is unavailable (no-conflict model).
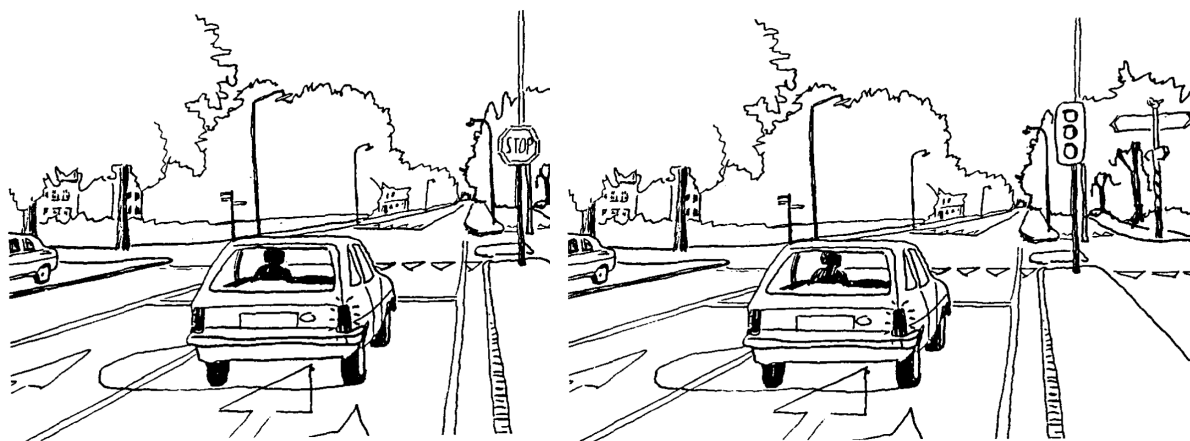

Figure 14.2: A pair of pictures from the third phase (i.e., the recognition test) of (Wagenaar & Boer, 1987, reprinted with permission), containing the critical episode at the intersection.

Table 14.1: Parameter point estimates from Wagenaar and Boer (1987).

	p	c	q	d	s
No-conflict model (NCM)	0.50	0.57	0.50	n/a	n/a
Destructive updating model (DUM)	0.50	0.57	0.50	0.00	n/a
Coexistence model (CXM)	0.55	0.55	0.43	n/a	0.20

Concrete Models

Wagenaar and Boer (1987) developed a series of MPT models (see Box 14.2) to quantify the predictions of the three competing theoretical accounts. Figure 14.3 depicts the no-conflict MPT model in the inconsistent condition. The figure is essentially a decision tree that is navigated from left to right. In Phase I of the collision narrative, the traffic light is encoded with probability p, and if so, the color is encoded with probability c. In Phase II, the stop sign is encoded with probability q. In Phase III, the answer may be known, or may be guessed correctly with probability $1/2$, and in Phase IV the answer may be known or may be guessed correctly with probability $1/3$. The probability of each path is given by the product of all the encountered probabilities, and the total probability of a response pattern is the summed probability of all branches that lead to it. For example, the total probability of getting both questions wrong is $(1 - p) \times q \times 2/3 + (1 - p) \times (1 - q) \times 1/2 \times 2/3$. We would then, under the no-conflict model, expect that proportion of participants to fall in the response pattern with two errors.

The destructive updating model (Figure 2 in Wagenaar & Boer, 1987) extends the three-parameter no-conflict model by adding a fourth parameter d: the probability of destroying the traffic light information, which may occur whenever the stop sign was encoded. The coexistence model (Figure 3 in Wagenaar & Boer, 1987), on the other hand, posits an extra probability s that the traffic light is suppressed (but not destroyed) when the stop sign is encoded. A critical difference between the latter two is that a destruction step will lead to chance accuracy in Phase IV if every piece of information was encoded, whereas a suppression step will not affect the underlying memory and lead to accurate responding. Note here that if $s = 0$, the coexistence model reduces to the no-conflict model, as does the destructive updating model with $d = 0$. The models only make different predictions in the inconsistent condition, so that for the consistent and neutral conditions the trees are identical.

Previous Conclusions

After fitting the three competing MPT models, Wagenaar and Boer (1987) obtained the parameter point estimates in Table 14.1. Using a χ^2 model fit index, they concluded that "a distinction among the three model families appeared to be impossible in actual practice" (p. 304), after noting that the no-conflict model provides "an almost perfect fit" to the data. They propose, then, "to accept the most parsimonious model, which is the no-conflict model." In the remainder of this chapter, we re-examine this conclusion using various model comparison methods.

Three Methods for Model Comparison

Many model comparison methods have been developed, all of them attempts to address the ubiquitous tradeoff between parsimony and goodness-of-fit. Here we focus on three main classes of interrelated methods: (1) AIC and BIC, the most popular information criteria; (2) minimum description length; (3) Bayes factors. Below we provide a brief description of each method and then apply it to the model comparison problem that confronted Wagenaar and Boer (1987).

Figure 14.3: Multinomial processing tree representation of the inconsistent condition according to the no-conflict model (adapted from Wagenaar & Boer, 1987).

Information Criteria

Information criteria are among the most popular methods for model comparison. Their popularity is explained by the simple and transparent manner in which they quantify the tradeoff between parsimony and goodness-of-fit. Consider for instance the oldest information criterion, AIC ("an information criterion"), proposed by Akaike (1973, 1974a):

$$\text{AIC} = -2 \ln \boldsymbol{p}\left(y \mid \hat{\theta}\right) + 2k. \tag{14.1}$$

The first term $\ln \boldsymbol{p}\left(y \mid \hat{\theta}\right)$ is the log maximum likelihood that quantifies goodness-of-fit, where y is the data set and $\hat{\theta}$ the maximum-likelihood parameter estimate; the second term $2k$ is a penalty for model complexity,

<div style="border:1px solid green;">

Frame 14.2: Popularity of multinomial processing tree models.

Multinomial processing tree models (Batchelder & Riefer, 1980; Chechile, 1973; Chechile & Meyer, 1976; Riefer & Batchelder, 1988) are psychological process models for categorical data. MPT models are used in two ways: as a psychometric tool to measure unobserved cognitive processes, and as a convenient formalization of competing psychological theories. Over time, MPTs have been applied to a wide range of psychological tasks and processes. For instance, MPT models are available for recognition, recall, source monitoring, perception, priming, reasoning, consensus analysis, the process dissociation procedure, implicit attitude measurement, and many other phenomena. For more information about MPTs, we recommend the review articles by Batchelder and Riefer (1999), Batchelder and Riefer (2007, pp. 24–32), and Erdfelder et al. (2009). The latter review article also discusses different software packages that can be used to fit MPT models. Necessarily missing from that list is the recently developed R package MPTinR (Singmann & Kellen, in press) with which we have good experiences. As will become apparent throughout this chapter, however, our preferred method for fitting MPT models is Bayesian (Chechile & Meyer, 1976; Klauer, 2010; Lee & Wagenmakers, in press; Matzke, Dolan, Batchelder, & Wagenmakers, in press; Rouder, Lu, Morey, Sun, & Speckman, 2008; Smith & Batchelder, 2010).

</div>

measured by the number of adjustable model parameters k. The AIC estimates the expected information loss incurred when a probability distribution f (associated with the true data-generating process) is approximated by a probability distribution g (associated with the model under evaluation). Hence, the model with the lowest AIC is the model with the smallest expected information loss between reality f and model g, where the discrepancy is quantified by the Kullback-Leibler divergence $I(f, g)$, a distance metric between two probability distributions (for full details, see Burnham & Anderson, 2002). The AIC is unfortunately not *consistent*: as the number of observations grows infinitely large, AIC is not guaranteed to choose the true data generating model. In fact, there is cause to believe that the AIC tends to select complex models that overfit the data (O'Hagan & Forster, 2004; for a discussion see Vrieze, 2012).

Another information criterion, the BIC ("Bayesian information criterion") was proposed by Schwarz (1978):

$$\text{BIC} = -2 \ln \boldsymbol{p}\left(y \mid \hat{\theta}\right) + k \ln n. \tag{14.2}$$

Here, the penalty term is $k \ln n$, where n is the number of observations.[2] Hence, the BIC penalty for complexity increases with sample size, outweighing that of AIC as soon as $n \geq 8$. The BIC was derived as an approximation of a Bayesian hypothesis test using default parameter priors (the "unit information prior"; see below for more information on Bayesian hypothesis testing, and see Raftery, 1995, for more information on the BIC). The BIC is consistent: as the number of observations grows infinitely large, BIC is guaranteed to choose the true data generating model. Nevertheless, there is evidence that in practical applications the BIC tends to select simple models that underfit the data (Burnham & Anderson, 2002).

Now consider a set of candidate models, $\mathcal{M}_i, i = 1, ..., m$, each with a specific IC (AIC or BIC) value. The model with the smallest IC value should be preferred, but the extent of this preference is not immediately apparent. For better interpretation we can calculate IC model weights (Akaike, 1974b; Burnham & Anderson, 2002; Wagenmakers & Farrell, 2004); First, we compute, for each model i, the difference in IC with respect to the IC of the best candidate model:

$$\Delta_i = \text{IC}_i - \min \text{IC}. \tag{14.3}$$

---

[2]Note that for hierarchical models, the definition of sample size n is more complicated (Pauler, 1998; Raftery, 1995).
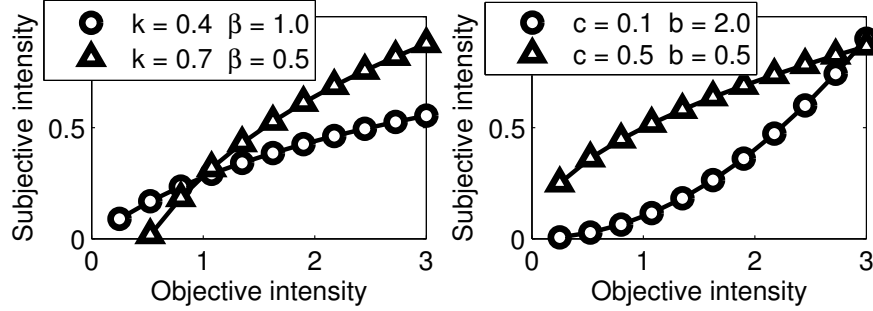
Figure 14.4: Two representative instances of Fechner's law (left) and Steven's law (right). While Fechner's law is restricted to nonlinear functions that level off as stimulus intensity increases, Steven's law can additionally take shapes with accelerating slopes.

This step is taken to increase numerical stability, but it also serves to emphasize the point that only differences in IC values are relevant. Next we obtain the model weights by transforming back to the likelihood scale and normalizing:

$$w_i = \frac{\exp\left(-\Delta_i/2\right)}{\sum_{m=1}^{M} \exp\left(-\Delta_m/2\right)}. \tag{14.4}$$

The resulting AIC and BIC weights are called Akaike weights and Schwarz weights, respectively. These weights not only convey the relative preference among a set of candidate models (i.e., they express a degree to which we should prefer one model from the set as superior), but also provide a method to combine predictions across multiple models using model averaging (Hoeting, Madigan, Raftery, & Volinsky, 1999).

Both AIC and BIC rely on an assessment of model complexity that is relatively crude, as it is determined entirely by the number of free parameters but not by the shape of the function through which they make contact with the data. To illustrate the importance of the functional form in which the parameters participate, consider the case of Fechner's law and Steven's law of psychophysics. Both of these laws transform objective stimulus intensity to subjective experience through a two-parameter nonlinear function.[3] According to Fechner's law, perceived intensity $\Psi_W(I)$ of stimulus I is the result of the logarithmic function $k\ln(I + \beta)$. Steven's law describes perceived intensity as an exponential function: $\Psi_S(I) = cI^b$. While both laws have the same number of parameters, Steven's is more complex as it can cover a larger number of data patterns (see Fig. 14.4).

Application to Multinomial Processing Tree Models

In order to apply AIC and BIC to the three competing MPTs proposed by Wagenaar and Boer (1987), we first need to compute the maximum log likelihood. Note that the MPT model parameters determine the predicted probabilities for the different response outcome categories (cf. Figure 14.3 and Box 14.2); these predicted probabilities are deterministic parameters from a multinomial probability density function. Hence, the maximum log likelihood parameter estimates for an MPT model produce multinomial parameters that maximize the probability of the observed data (i.e., the occurrence of the various outcome categories).

Several software packages exist that can help find the maximum log likelihood parameter estimates for MPTs (e.g. Singmann & Kellen, in press). With these estimates in hand, we can compute the information

---

[3]For a more in-depth treatment, see Townsend (1975).

Table 14.2: AIC and BIC for the Wagenaar & Boer MPT models.

	log likelihood	k	AIC	wAIC	BIC	wBIC
No-conflict model (NCM)	-24.41	3	54.82	0.41	67.82	0.86
Destructive updating model (DUM)	-24.41	4	56.82	0.15	74.15	0.04
Coexistence model (CXM)	-23.35	4	54.70	0.44	72.03	0.10

Note: k is the number of free parameters.

criteria described in the previous section. Table 14.2 shows the maximum log likelihood as well as AIC, BIC, and their associated weights (wAIC and wBIC; from Equation 14.4).

Interpreting wAIC and wBIC as measures of relative preference, we see that the results in Table 14.2 are mostly inconclusive. According to wAIC, the no-conflict model and coexistence model are virtually indistinguishable, though both are preferable to the destructive updating model. According to wBIC, however, the no-conflict model should be preferred over both the destructive updating model and the coexistence model. The extent of this preference is noticeable but not decisive.

Minimum Description Length

The minimum description length principle is based on the idea that statistical inference centers around capturing regularity in data; regularity, in turn, can be exploited to compress the data. Hence, the goal is to find the model that compresses the data the most (Grünwald, 2007). This is related to the concept of Kolmogorov complexity—for a sequence of numbers, Kolmogorov complexity is the length of the shortest program that prints that sequence and then halts (Grünwald, 2007). Although Kolmogorov complexity cannot be calculated, a suite of concrete methods are available based on the idea of model selection through data compression. These methods, most of them developed by Jorma Rissanen, fall under the general heading of minimum description length (MDL; Rissanen, 1978, 1987, 1996, 2001). In psychology, the MDL principle has been applied and promoted primarily by Grünwald (2000), Grünwald, Myung, and Pitt (2005), and Grünwald (2007), as well as Myung, Navarro, and Pitt (2006), Pitt and Myung (2002), and Pitt, Myung, and Zhang (2002).

Here we mention three versions of the MDL principle. First, there is the so-called *crude two-part code* (Grünwald, 2007); here, one sums the description of the model (in bits) and the description of the data encoded with the help of that model (in bits). The penalty for complex models is that they take many bits to describe, increasing the summed code length. Unfortunately, it can be difficult to define the number of bits required to describe a model.

Second, there is the Fisher information approximation (FIA; Pitt et al., 2002; Rissanen, 1996):

$$\text{FIA} = -\ln \boldsymbol{p}\left(y \mid \hat{\theta}\right) + \frac{k}{2}\ln\left(\frac{n}{2\pi}\right) + \ln \int_{\Theta} \sqrt{\det\left[I(\theta)\right]}\,d\theta, \tag{14.5}$$

where $I(\theta)$ denotes the Fisher information matrix of sample size 1 (Ly, Verhagen, & Wagenmakers, in preparation). $I(\theta)$ is a $k \times k$ matrix whose $(i,j)^{th}$ element is

$$I_{i,j}(\theta) = E\left(\frac{\partial \ln \boldsymbol{p}\left(y \mid \theta\right)}{\partial \theta_i} \frac{\partial \ln \boldsymbol{p}\left(y \mid \theta\right)}{\partial \theta_j}\right),$$

where $E()$ is the expectation operator.

Note that FIA is similar to AIC and BIC in that it includes a first term that represents goodness-of-fit, and additional terms that represent a penalty for complexity. The second term resembles that of BIC, and the third term reflects a more sophisticated penalty that represents the number of distinguishable probability

Table 14.3: Minimum description length values for the Wagenaar & Boer MPT models.

	Complexity	FIA	wFIA
No-conflict model (NCM)	6.44	30.86	0.44
Destructive updating model (DUM)	7.39	31.80	0.17
Coexistence model (CXM)	7.61	30.96	0.39

distributions that a model can generate (Pitt et al., 2002). Hence, FIA differs from AIC and BIC in that it also accounts for functional form complexity, not just complexity due to the number of free parameters. Note that FIA weights (or Rissanen weights) can be obtained by multiplying FIA by 2 and then applying Equations 14.3 and 14.4.

The third version of the MDL principle discussed here is normalized maximum likelihood (NML; Myung et al., 2006; Rissanen, 2001):

$$\text{NML} = \frac{\boldsymbol{p}\left(y \mid \hat{\theta}(y)\right)}{\int_X \boldsymbol{p}\left(x \mid \hat{\theta}(x)\right) dx}. \tag{14.6}$$

This equation shows that NML tempers the enthusiasm about a good fit to the observed data y (i.e., the numerator) to the extent that the model could also have provided a good fit to random data x (i.e., the denominator). NML is simple to state but can be difficult to compute; for instance, the denominator may be infinite and this requires further measures to be taken (for details see Grünwald, 2007). Additionally, NML requires an integration over the entire set of possible data sets, which may be difficult to define as it depends on unknown decision processes in the researchers (Berger & Berry, 1988). Note that, since the computation of NML depends on the likelihood of data that *might have occurred but did not*, the procedure violates the *likelihood principle*, which states that all information about a parameter θ obtainable from an experiment is contained in the likelihood function for θ for the given y (Berger & Wolpert, 1988).

Application to Multinomial Processing Tree Models

Using the parameter estimates from Table 14.1 and the code provided by Wu, Myung, and Batchelder (2010), we can compute the FIA for the three competing MPT models considered by Wagenaar and Boer (1987).[4] Table 14.3 displays, for each model, the FIA along with its associated complexity measure (the other one of its two constituent components, the maximum log likelihood, can be found in Table 14.2). The conclusions from the MDL analysis mirror those from the AIC measure, expressing a slight disfavor for the destructive updating model, and approximately equal preference for the no-conflict model versus the coexistence model.

Bayes Factors

In Bayesian model comparison, the posterior odds for models $\mathcal{M}_1$ and $\mathcal{M}_2$ are obtained by updating the prior odds with the diagnostic information from the data:

$$\frac{\boldsymbol{p}(\mathcal{M}_1 \mid y)}{\boldsymbol{p}(\mathcal{M}_2 \mid y)} = \frac{\boldsymbol{p}(\mathcal{M}_1)}{\boldsymbol{p}(\mathcal{M}_2)} \times \frac{\boldsymbol{m}(y \mid \mathcal{M}_1)}{\boldsymbol{m}(y \mid \mathcal{M}_2)}. \tag{14.7}$$

Equation 14.7 shows that the change from prior odds $\boldsymbol{p}(\mathcal{M}_1)/\boldsymbol{p}(\mathcal{M}_2)$ to posterior odds $\boldsymbol{p}(\mathcal{M}_1 \mid y)/\boldsymbol{p}(\mathcal{M}_2 \mid y)$ is given by the ratio of marginal likelihoods $\boldsymbol{m}(y \mid \mathcal{M}_1)/\boldsymbol{m}(y \mid \mathcal{M}_2)$ (see below for the definition of the marginal likelihood). This ratio is known as the *Bayes factor* (Jeffreys, 1961; Kass & Raftery, 1995). The

---

[4]Analysis using the `MPTinR` package by Singmann and Kellen (in press) gave virtually identical results. Technical details for the computation of the NML for MPTs are provided in Appendix B of Klauer and Kellen (2011).

Table 14.4: Evidence categories for the Bayes factor BF_{12} (based on Jeffreys, 1961).

Bayes factor BF_{12}			Interpretation
	$>$	100	Extreme evidence for $\mathcal{M}_1$
30	—	100	Very Strong evidence for $\mathcal{M}_1$
10	—	30	Strong evidence for $\mathcal{M}_1$
3	—	10	Moderate evidence for $\mathcal{M}_1$
1	—	3	Anecdotal evidence for $\mathcal{M}_1$
	1		No evidence
1/3	—	1	Anecdotal evidence for $\mathcal{M}_2$
1/10	—	1/3	Moderate evidence for $\mathcal{M}_2$
1/30	—	1/10	Strong evidence for $\mathcal{M}_2$
1/100	—	1/30	Very Strong evidence for $\mathcal{M}_2$
	$<$	1/100	Extreme evidence for $\mathcal{M}_2$

log of the Bayes factor is often interpreted as the weight of evidence provided by the data (Good, 1985; for details see Berger & Pericchi, 1996; Bernardo & Smith, 1994; Gill, 2002; O'Hagan, 1995).

Thus, when the Bayes factor $BF_{12} = \boldsymbol{m}(y \mid \mathcal{M}_1)/\boldsymbol{m}(y \mid \mathcal{M}_2)$ equals 5, the observed data y are 5 times more likely to occur under $\mathcal{M}_1$ than under $\mathcal{M}_2$; when BF_{12} equals 0.1, the observed data are 10 times more likely under $\mathcal{M}_2$ than under $\mathcal{M}_1$. Even though the Bayes factor has an unambiguous and continuous scale, it is sometimes useful to summarize the Bayes factor in terms of discrete categories of evidential strength. Jeffreys (1961, Appendix B) proposed the classification scheme shown in Table 14.4. We replaced the labels "not worth more than a bare mention" with "anecdotal", "decisive" with "extreme", and "substantial" with "moderate". These labels facilitate scientific communication but should be considered only as an approximate descriptive articulation of different standards of evidence.

Bayes factors negotiate the tradeoff between parsimony and goodness-of-fit and implement an automatic Occam's razor (Jefferys & Berger, 1992; MacKay, 2003; Myung & Pitt, 1997). To see this, consider that the marginal likelihood $\boldsymbol{m}(y \mid \mathcal{M}_{(\cdot)})$ can be expressed as $\int_{\Theta} \boldsymbol{p}(y \mid \theta, \mathcal{M}_{(\cdot)})\boldsymbol{p}(\theta \mid \mathcal{M}_{(\cdot)})\, \mathrm{d}\theta$: an average across the entire parameter space, with the prior providing the averaging weights. It follows that complex models with high-dimensional parameter spaces are not necessarily desirable—large regions of the parameter space may yield a very poor fit to the data, dragging down the average. The marginal likelihood will be highest for parsimonious models that use only those parts of the parameter space that are required to provide an adequate account of the data (Lee & Wagenmakers, in press). By using marginal likelihood the Bayes factor punishes models that hedge their bets and make vague predictions. Models can hedge their bets in different ways: by including extra parameters, by assigning very wide prior distributions to the model parameters, or by using parameters that participate in the likelihood through a complicated functional form. By computing a weighted average likelihood across the entire parameter space, the marginal likelihood (and, consequently, the Bayes factor) automatically takes all these aspects into account.

Bayes factors represent "the standard Bayesian solution to the hypothesis testing and model selection problems" (Lewis & Raftery, 1997, p. 648) and "the primary tool used in Bayesian inference for hypothesis testing and model selection" (Berger, 2006, p. 378), but their application is not without challenges (Box 14.3). Below we show how these challenges can be overcome for the general class of MPT models. Next we compare the results of our Bayes factor analysis with those of the other model comparison methods using Jeffreys weights (i.e., normalized marginal likelihoods).

> **Frame 14.3:** Two challenges for Bayes factors.
>
> Bayes factors (Jeffreys, 1961; Kass & Raftery, 1995) come with two main challenges, one practical and one conceptual. The practical challenge arises because Bayes factors are defined as the ratio of two marginal likelihoods, each of which requires integration across the entire parameter space. This integration process can be cumbersome and hence the Bayes factor can be difficult to obtain. Fortunately, there are many approximate and exact methods to facilitate the computation of the Bayes factor (e.g., Ardia, Baştürk, Hoogerheide, & van Dijk, 2012; Chen, Shao, & Ibrahim, 2002; Gamerman & Lopes, 2006); in this chapter we focus on BIC (a crude approximation), the Savage-Dickey density ratio (applies only to nested models) and importance sampling. The conceptual challenge that Bayes factors bring is that the prior on the model parameters has a pronounced and lasting influence on the result. This should not come as a surprise: the Bayes factor punishes models for needless complexity, and the complexity of a model is determined in part by the prior distributions that are assigned to the parameters. The difficulty arises because researchers are often not very confident about the prior distributions that they specify. To overcome this challenge one can either spend more time and effort on the specification of realistic priors, or else one can choose default priors that fulfill general desiderata (e.g., Jeffreys, 1961; Liang, Paulo, Molina, Clyde, & Berger, 2008). Finally, the robustness of the conclusions can be verified by conducting a sensitivity analysis in which one examines the effect of changing the prior specification (e.g., Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011).

Application to Multinomial Processing Tree Models

In order to compute the Bayes factor we seek to determine each model's marginal likelihood $\boldsymbol{m}(y \mid \mathcal{M}_{(\cdot)})$. As indicated above, the marginal likelihood $\boldsymbol{m}(y \mid \mathcal{M}_{(\cdot)})$ is given by integrating the likelihood over the prior:

$$\boldsymbol{m}(y \mid \mathcal{M}_{(\cdot)}) = \int \boldsymbol{p}\left(y \mid \theta, \mathcal{M}_{(\cdot)}\right) \boldsymbol{p}\left(\theta \mid \mathcal{M}_{(\cdot)}\right) \mathrm{d}\theta. \tag{14.8}$$

The most straightforward manner to obtain $\boldsymbol{m}(y \mid \mathcal{M}_{(\cdot)})$ is to draw samples from the prior $\boldsymbol{p}(\theta \mid \mathcal{M}_{(\cdot)})$ and average the corresponding values for $\boldsymbol{p}(y \mid \theta, \mathcal{M}_{(\cdot)})$:

$$\boldsymbol{m}(y \mid \mathcal{M}_{(\cdot)}) \approx \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{p}\left(y \mid \theta_i, \mathcal{M}_{(\cdot)}\right), \qquad \theta_i \sim \boldsymbol{p}(\theta). \tag{14.9}$$

For MPT models, this brute force integration approach may often be adequate. An MPT model usually has few parameters, and each is conveniently bounded from 0 to 1. However, brute force integration is inefficient, particularly when the posterior is highly peaked relative to the prior: in this case, draws from $\boldsymbol{p}(\theta \mid \mathcal{M}_{(\cdot)})$ tend to result in low likelihoods and only few chance draws may have high likelihood. This problem can be overcome by a numerical technique known as *importance sampling* (Hammersley & Handscomb, 1964).

In importance sampling, efficiency is increased by drawing samples from an importance density $\boldsymbol{g}(\theta)$ instead of from the prior $\boldsymbol{p}(\theta \mid \mathcal{M}_{(\cdot)})$. Consider an importance density $\boldsymbol{g}(\theta)$. Then,

$$\begin{aligned}
\boldsymbol{m}(y \mid \mathcal{M}_{(\cdot)}) &= \int \boldsymbol{p}\left(y \mid \theta, \mathcal{M}_{(\cdot)}\right) \boldsymbol{p}\left(\theta \mid \mathcal{M}_{(\cdot)}\right) \frac{\boldsymbol{g}(\theta)}{\boldsymbol{g}(\theta)} \mathrm{d}\theta \\
&= \int \frac{\boldsymbol{p}\left(y \mid \theta, \mathcal{M}_{(\cdot)}\right) \boldsymbol{p}\left(\theta \mid \mathcal{M}_{(\cdot)}\right)}{\boldsymbol{g}(\theta)} \boldsymbol{g}(\theta) \mathrm{d}\theta \\
&\approx \frac{1}{N} \sum_{i=1}^{N} \frac{\boldsymbol{p}\left(y \mid \theta_i, \mathcal{M}_{(\cdot)}\right) \boldsymbol{p}\left(\theta_i \mid \mathcal{M}_{(\cdot)}\right)}{\boldsymbol{g}(\theta_i)}, \qquad \theta_i \sim \boldsymbol{g}(\theta).
\end{aligned} \tag{14.10}$$

Frame 14.4: Importance sampling for MPT models using the Beta mixture method.

Importance sampling was invented by Stan Ulam and John von Neumann. Here we use it to estimate the marginal likelihood by repeatedly drawing samples and averaging—the samples are, however, not drawn from the prior (as per Equation 14.9, the brute force method), but instead they are drawn from some convenient density $g(\theta)$ (as per Equation 14.10; Andrieu, De Freitas, Doucet, & Jordan, 2003; Hammersley & Handscomb, 1964). The parameters in MPT models are constrained to the unit interval, and therefore the family of Beta distributions is a natural candidate for $g(\theta)$. The middle panel of Figure 14.5 shows an importance density (dashed line) for MPT parameter c in the no-conflict model for the data from Wagenaar and Boer (1987). This importance density is a Beta distribution that was fit to the posterior distribution for c using the method of moments. The importance density provides a good description of the posterior (the dashed line tracks the posterior almost perfectly) and therefore is more efficient than the brute force method illustrated in the left panel of Figure 14.5, which uses the prior as the importance density. Unfortunately, Beta distributions do not always fit MPT parameters so well; specifically, the Beta importance density may sometimes have tails that are thinner than the posterior, and this increases the variability of the marginal likelihood estimate. To increase robustness and ensure that the importance density has relatively fat tails we can use a Beta mixture, shown in the right panel of Figure 14.5. The Beta mixture consists of a uniform prior component (i.e., the Beta$(1,1)$ prior as in the left panel) and a Beta posterior component (i.e., a Beta distribution fit to the posterior, as in the middle panel). In this example, the mixture weight for the uniform component is $w = 0.2$. Small mixture weights retain the efficiency of the Beta posterior approach but avoid the extra variability due to thin tails. It is possible to increase efficiency further by specifying a multivariate importance density, but the present univariate approach is intuitive, easy to implement, and appears to work well in practice. The accuracy of the estimate can be confirmed by increasing the number of draws from the importance density, and by varying the w parameter.

Note that if $g(\theta) = p(\theta \mid \mathcal{M}_{(.)})$ the importance sampler reduces to the brute force integration shown in Equation 14.9. Also note that if $g(\theta) = p(\theta \mid y, \mathcal{M}_{(.)})$, a single draw suffices to determine $p(y)$ exactly.

In sum, when the importance density equals the prior we have brute force integration, and when it equals the posterior we have a zero-variance estimator. However, in order to compute the posterior, we would have to be able to compute the normalizing constant (i.e., the marginal likelihood), which is exactly the quantity we wish to determine. In practice then, we want to use an importance density that is similar to the posterior, is easy to evaluate, and is easy to draw samples from. In addition, we want to use an importance density with tails that are not thinner than those of the posterior; thin tails cause the estimate to have high variance. These desiderata are met by the *Beta mixture* importance density described in Box 14.4: a mixture between a Beta$(1,1)$ density and a Beta density that provides a close fit to the posterior distribution. Here we use a series of univariate Beta mixtures, one for each separate parameter, but acknowledge that a multivariate importance density is potentially even more efficient as it accommodates correlations between the parameters.

In our application to MPT models, we assume that all model parameters have uniform Beta$(1,1)$ priors. For most MPT models this assumption is fairly uncontroversial. The uniform priors can be thought of as a default choice; in the presence of strong prior knowledge one can substitute more informative priors. The uniform priors yield a default Bayes factor that can be a reference point for an analysis with more informative priors, if such an analysis is desired (i.e., when reliable prior information is available, such as can be elicited from experts or derived from earlier experiments).
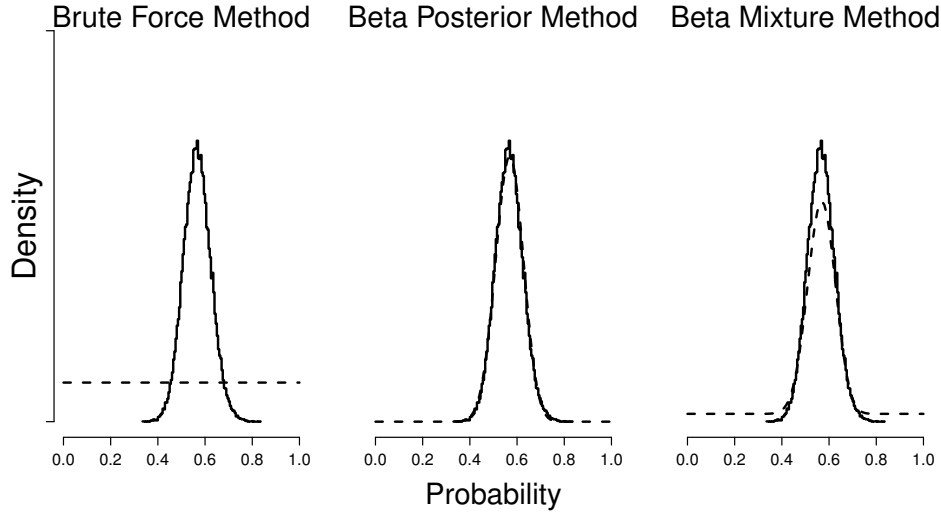
Figure 14.5: Three different importance sampling densities (dashed lines) for the posterior distribution (solid lines) of the c parameter in the no-conflict model as applied to the data from Wagenaar and Boer (1987). Left panel: a uniform Beta importance density (i.e., the brute force method); middle panel: a Beta posterior importance density (i.e., a Beta distribution that provides the best fit to the posterior); right panel: a Beta mixture importance density (i.e., a mixture of the uniform Beta density and the Beta posterior density, with a mixture weight $w = 0.2$ on the uniform component).

Monte Carlo sampling for the posterior distribution   Before turning to the results of the Bayes factor model comparison we first inspect the posterior distributions. The posterior distributions were approximated using Markov chain Monte Carlo sampling implemented in JAGS (Plummer, 2003) and WinBUGS (Lunn, Jackson, Best, Thomas, & Spiegelhalter, 2012).[5] All code is available on the authors' websites. Convergence was confirmed by visual inspection and the $\hat{R}$ statistic (Gelman & Rubin, 1992). The top panel of Figure 14.6 shows the posterior distributions for the no-conflict model. Although there is slightly more certainty about parameter p than there is about parameters q and c, the posterior distributions for all three parameters are relatively wide considering that they are based on data from as many as 562 participants.

The middle panel of Figure 14.6 shows the posterior distributions for the destructive-updating model. It is important to realize that when $d = 0$ (i.e., no destruction of the earlier memory) the destructive-updating model reduces to the no-conflict model. Compared to the no-conflict model, parameters p, q, and c show relatively little change. The posterior distribution for d is very wide, indicating considerable uncertainty about its true value. A frequentist point-estimate yields $\hat{d} = 0$ (Wagenaar & Boer, 1987; see also Table 14.1), but this does not convey the fact that this estimate is highly uncertain.

The lower panel of Figure 14.6 shows the posterior distributions for the coexistence model. When $s = 0$ (i.e., no suppression of the earlier memory), the coexistence model reduces to the no-conflict model. Compared to the no-conflict model and the destructive-updating model, parameters p, q, and c again show relatively little change. The posterior distribution for s is very wide, indicating considerable uncertainty about its true value.

The fact that the no-conflict model is nested under both the destructive-updating model and the no-

---

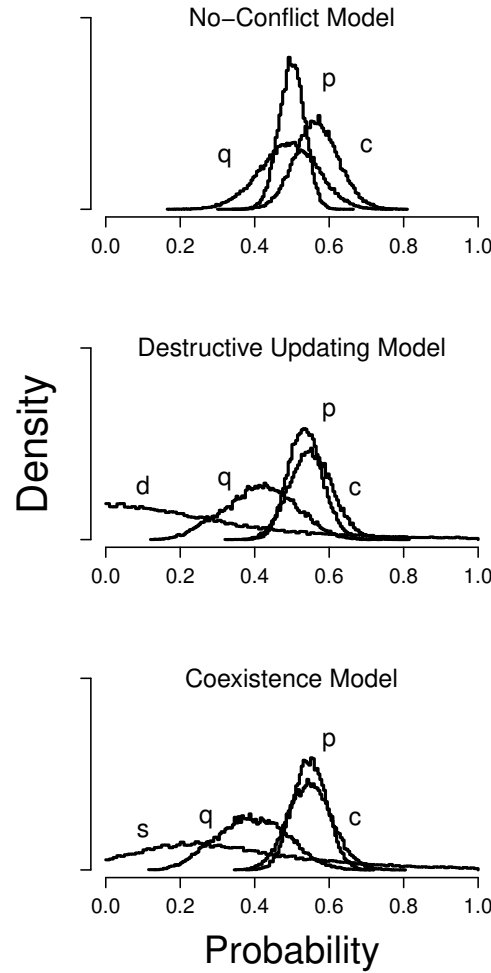[5]The second author used WinBUGS, the first and third authors used JAGS.

Figure 14.6: Posterior distributions for the parameters of the no-conflict MPT model, the destructive updating MPT model, and the coexistence MPT model, as applied to the data from Wagenaar and Boer (1987).

conflict model allows us to inspect the extra parameters d and s and conclude that we have not learned very much about their true values. This suggests that, despite having tested 562 participants, the data do not firmly support one model over the other. We will now see how Bayes factors can make this intuitive judgment more precise.

Importance sampling for the Bayes factor   We applied the Beta mixture importance sampling method to estimate marginal likelihoods for the three models under consideration. The results were confirmed by varying the mixture weight w, by independent implementations by the authors, and by comparison to the Savage-Dickey density ratio test presented later. Table 14.5 shows the results.

From the marginal likelihoods and the Jeffreys weights we can derive the Bayes factors for the pair-wise comparisons; the Bayes factor is 2.77 in favor of the no-conflict model over the destructive updating model, the Bayes factor is 1.39 in favor of the coexistence model over the no-conflict model, and the Bayes factor

Table 14.5: Bayesian evidence (i.e., the logarithm of the marginal likelihood), Jeffreys weights, and pairwise Bayes factors computed from the Jeffreys weights or through the Savage-Dickey density ratio, for the Wagenaar & Boer MPT models.

	Bayesian evidence	Jeffreys weight	Bayes factor (Savage-Dickey) over NCM	over DUM	over CXM
No-conflict model (NCM)	−30.55	0.36	1	2.77 (2.81)	0.72 (0.80)
Destructive updating model (DUM)	−31.57	0.13	0.36 (0.36)	1	0.26 (0.28*)
Coexistence model (CXM)	−30.22	0.51	1.39 (1.25)	3.86 (3.51*)	1

* Derived through transitivity: $2.81 \times 1/0.80 = 3.51$.

is 3.86 in favor of the coexistence model over the destructive updating model. The first two of these Bayes factors are anecdotal or "not worth more than a bare mention" (Jeffreys, 1961), and the third one just makes the criterion for "moderate" evidence, although any enthusiasm about this level of evidence should be tempered by the realization that Jeffreys himself described a Bayes factor as high as 5.33 as "odds that would interest a gambler, but would be hardly worth more than a passing mention in a scientific paper" (Jeffreys, 1961, pp. 256-257). In other words, the Bayes factors are consistent with the intuitive visual assessment of the posterior distributions: the data do not allow us to draw strong conclusions.

We should stress that Bayes factors apply to a comparison of any two models, regardless of whether or not they are structurally related or *nested* (i.e., where one model is a special, simplified version of a larger, encompassing model). As is true for the information criteria and minimum description length methods, Bayes factors can be used to compare structurally very different models, such as for example REM (Shiffrin & Steyvers, 1997) versus ACT-R (Anderson et al., 2004), or the diffusion model (Ratcliff, 1978) versus the linear ballistic accumulator model (Brown & Heathcote, 2008). In other words, Bayes factors can be applied to nested and non-nested models alike. For the models under consideration, however, there exists a nested structure that allows one to obtain the Bayes factor through a computational shortcut.

The Savage-Dickey approximation to the Bayes factor for comparing nested models Consider first the comparison between the no-conflict model $\mathcal{M}_{\text{NCM}}$ and the destructive updating model $\mathcal{M}_{\text{DUM}}$. As shown above, we can obtain the Bayes factor for $\mathcal{M}_{\text{NCM}}$ versus $\mathcal{M}_{\text{DUM}}$ by computing the marginal likelihoods using importance sampling. However, because the models are nested we can also obtain the Bayes factor by considering only $\mathcal{M}_{\text{DUM}}$, and dividing the posterior ordinate at $d = 0$ by the prior ordinate at $d = 0$. This surprising result was first published by Dickey and Lientz (1970), who attributed it to Leonard J. "Jimmie" Savage. The result is now generally known as the *Savage-Dickey density ratio* (e.g., Dickey, 1971; for extensions and generalizations see Chen, 2005; Verdinelli & Wasserman, 1995; Wetzels, Grasman, & Wagenmakers, 2010; for an introduction for psychologists see Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010; a short mathematical proof is presented in O'Hagan & Forster, 2004, pp. 174-177).[6] Thus, we can exploit the fact that $\mathcal{M}_{\text{NCM}}$ is nested in $\mathcal{M}_{\text{DUM}}$ and use the Savage-Dickey density ratio to obtain the Bayes factor:

$$BF_{\text{NCM,DUM}} = \frac{\boldsymbol{m}(y \mid \mathcal{M}_{\text{NCM}})}{\boldsymbol{m}(y \mid \mathcal{M}_{\text{DUM}})} = \frac{\boldsymbol{p}(d = 0 \mid y, \mathcal{M}_{\text{DUM}})}{\boldsymbol{p}(d = 0 \mid \mathcal{M}_{\text{DUM}})}. \tag{14.11}$$

The Savage-Dickey density ratio test is visualized in Figure 14.7; the posterior ordinate at $d = 0$ is higher than the prior ordinate at $d = 0$, indicating that the data have increased the plausibility that d equals 0. This means that the data support $\mathcal{M}_{\text{NCM}}$ over $\mathcal{M}_{\text{DUM}}$. The prior ordinate equals 1, and hence $BF_{\text{NCM,DUM}}$ simply equals the posterior ordinate at $d = 0$. A nonparametric density estimator

---

[6]Note that the Savage-Dickey density ratio requires that when $d = 0$ the prior for the common parameters p, c, and q is the same for $\mathcal{M}_{\text{DUM}}$ and $\mathcal{M}_{\text{NCM}}$. That is, $\boldsymbol{p}(p, c, q \mid d = 0, \mathcal{M}_{\text{DUM}}) = \boldsymbol{p}(p, c, q \mid \mathcal{M}_{\text{NCM}})$.
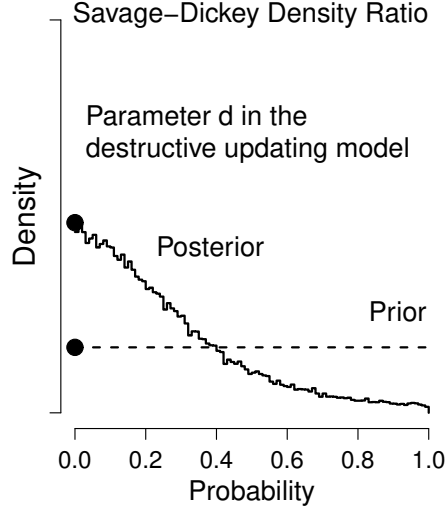
Figure 14.7: Illustration of the Savage-Dickey density ratio test. The dashed and solid lines show the prior and the posterior distribution for parameter d in the destructive updating model. The black dots indicate the height of the prior and the posterior distributions at $d = 0$.

(Stone, Hansen, Kooperberg, & Truong, 1997) that respects the bound at 0 yields an estimate of 2.81. This estimate is close to 2.77, the estimate from the importance sampling approach.

The Savage-Dickey density ratio test can be applied similarly to the comparison between the no-conflict model $\mathcal{M}_{\text{NCM}}$ versus the coexistence model $\mathcal{M}_{\text{CXM}}$, where the critical test is at $s = 0$. Here the posterior ordinate is estimated to be 0.80, and hence the Bayes factor for $\mathcal{M}_{\text{CXM}}$ over $\mathcal{M}_{\text{NCM}}$ equals $1/0.80 = 1.25$, close to the Bayes factor obtained through importance sampling, $BF_{\text{CXM,NCM}} = 1.39$.

With these two Bayes factors in hand, we can immediately derive the Bayes factor for the comparison between the destructive updating model $\mathcal{M}_{\text{DUM}}$ versus the coexistence model $\mathcal{M}_{\text{CXM}}$ through transitivity, that is, $BF_{\text{CXM,DUM}} = BF_{\text{NCM,DUM}} \times BF_{\text{CXM,NCM}}$. Alternatively, we can also obtain $BF_{\text{CXM,DUM}}$ by directly comparing the posterior density for $d = 0$ against that for $s = 0$:

$$
\begin{aligned}
BF_{\text{CXM,DUM}} &= BF_{\text{NCM,DUM}} \times BF_{\text{CXM,NCM}} \\
&= \frac{\boldsymbol{p}(d = 0 \mid y, \mathcal{M}_{\text{DUM}})}{\boldsymbol{p}(d = 0 \mid \mathcal{M}_{\text{DUM}})} \times \frac{\boldsymbol{p}(s = 0 \mid \mathcal{M}_{\text{CXM}})}{\boldsymbol{p}(s = 0 \mid y, \mathcal{M}_{\text{CXM}})} \\
&= \frac{\boldsymbol{p}(d = 0 \mid y, \mathcal{M}_{\text{DUM}})}{\boldsymbol{p}(s = 0 \mid y, \mathcal{M}_{\text{CXM}})},
\end{aligned}
\tag{14.12}
$$

where the second step is allowed because we have assigned uniform priors to both d and s, so that $\boldsymbol{p}(d = 0 \mid \mathcal{M}_{\text{DUM}}) = \boldsymbol{p}(s = 0 \mid \mathcal{M}_{\text{CXM}})$. Hence, the Savage-Dickey estimate for the Bayes factor between the two non-nested models $\mathcal{M}_{\text{DUM}}$ and $\mathcal{M}_{\text{CXM}}$ equals the ratio of the posterior ordinates at $d = 0$ and $s = 0$, resulting in the estimate $BF_{\text{CXM,DUM}} = 3.51$, close to the importance sampling result of 3.86.

Comparison of Model Comparisons

We have now implemented and performed a variety of model comparison methods for the three competing MPT models introduced by Wagenaar and Boer (1987): we computed and interpreted the Akaike information criteria (AIC), Bayesian information criteria (BIC), the Fisher information approximation of the minimum description length principle (FIA), and two computational implementations of the Bayes factor (BF).

The general tenor across most of the model comparison exercises has been that the data do not convincingly support one particular model. However, the destructive updating model is consistently ranked the worst of the set. Looking at the parameter estimates, it is not difficult to see why this is so: the d parameter of the destructive updating model (i.e., the probability of destroying memory through updating) is estimated at 0, thereby reducing the destructive updating model to the no-conflict model, and yielding an identical fit to the data (as can be seen in the likelihood column of Table 14.2). Since the no-conflict model counts as a special case of the destructive updating model, it is by necessity less complex from a model selection point of view—the d parameter is an unnecessary entity, the inclusion of which is not warranted by the data. This is reflected in the inferior performance of the destructive updating model according to all measures of generalizability.

Note that the BF judges the support for NCM to be 'anecdotal' even though NCM and DUM provide similar fit and have a clear difference in complexity—one might expect the principle of parsimony to tell us that, given the equal fit and clear complexity difference, there is massive evidence for the simpler model, and the BF appears to fail to implement Occams razor here. The lack of clear support of the NCM over the DUM is explained by the considerable uncertainty regarding the value of the parameter d: even though the posterior mode is at $d = 0$, much posterior variability is visible in the middle panel of Fig. 14.6. With more data and a posterior for d that is more peaked near 0, the evidence in favor of the simpler model would increase.

The difference between the no-conflict model and the coexistence model is less clear-cut. Following AIC, the two models are virtually indistinguishable—compared to the coexistence model, the no-conflict model sacrifices one unit of log-likelihood for two units of complexity (one parameter). As a result, both models perform equally well under the AIC measure. Under the BIC measure, however, the penalty for the number of free parameters is more substantial, and here the no-conflict model trades a unit of log likelihood for $\log(N) = 6.33$ units of complexity, outdistancing both the destructive updating model and the coexistence model. The BIC is the exception in clearly preferring the no-conflict model over the coexistence model. The MDL, like the AIC, would have us hedge on the discriminability of the no-conflict model and the coexistence model.

The BF, finally, allows us to express evidence for the models using standard probability theory. Between any two models, the BF tells us how much the balance of evidence has shifted due to the data. Using two methods of computing the BF, we determined that the odds of the coexistence model over the destructive updating model almost quadrupled ($BF_{\mathrm{CXM,DUM}} \approx 3.86$), but the odds of the coexistence model over the no-conflict model barely shifted at all ($BF_{\mathrm{CXM,NCM}} \approx 1.39$). Finally, we can use the same principles of probability to compute Jeffreys weights, which express, for each model under consideration, the probability that it is true, assuming prior indifference. Furthermore, we can easily recompute the probabilities in case we wish to express a prior preference between the candidate models (for example, we might use the prior to express a preference for sparsity, as was originally proposed by Jeffreys, 1961).

Concluding Comments

Model comparison methods need to implement the principle of parsimony: goodness-of-fit has to be discounted to the extent that it was accomplished by a model that is overly complex. Many methods of model comparison exist (Myung et al., 2000; Wagenmakers & Waldorp, 2006), and our selective review fo-
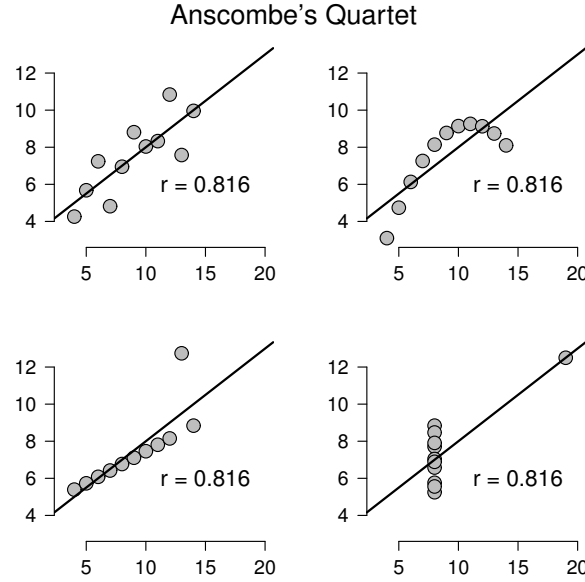
Figure 14.8: Anscombe's Quartet is a set of four bivariate data sets whose basic descriptive statistics are approximately identical. In all cases, mean of X is 9, variance of X is 11, mean of Y is 7.5, variance of Y is 4.1, and the best fitting linear regression line is $y_i^{est} = 3 + 0.5x_i$, which explains $R^2 = 66.6\%$ of the variance in Y. However, in two of the four cases, the linear regression is clearly a poor account of the data. The relative measure of model fit (R^2) gives no indication of this radical difference between the data sets, and an absolute measure of fit (even one as rudimentary as a visual inspection of the regression line) is required. (Figure downloaded from Flickr, courtesy of Eric-Jan Wagenmakers.)

cused on methods that are popular, easy-to-compute approximations (i.e., AIC and BIC) and methods that are difficult-to-compute "ideal" solutions (i.e., minimum description length and Bayes factors). We applied these model comparison methods to the scenario of three competing MPT models introduced by Wagenaar and Boer (1987). Despite collecting data from 562 participants, the model comparison methods indicate that the data are somewhat ambiguous; at any rate, the data do not support the destructive updating model. This echoes the conclusions drawn by Wagenaar and Boer (1987).

It is important to note that the model comparison methods discusses in this chapter can be applied regardless of whether the models are nested. This is not just a practical nicety; it also means that the methods are based on principles that transcend the details of a specific model implementation. In our opinion, a method of inference that is necessarily limited to the comparison of nested models is incomplete at best and misleading at worst. It is also important to realize that model comparison methods are *relative* indices of model adequacy; when, say, the Bayes factor expresses an extreme preference for model A over model B, this does not mean that model A fits the data at all well. Figure 14.8 shows a classic but dramatic example of the inadequacy of simple measures of relative model fit.

Because it would be a mistake to base inference on a model that fails to describe the data, a complete inference methodology features both relative and absolute indices of model adequacy. For the MPT models under consideration here, Wagenaar and Boer (1987) reported that the no-conflict model provided "an almost

perfect fit" to the data.[7]

The example MPT scenario considered here was relatively straightforward. More complicated MPT models contain order-restrictions, feature individual differences embedded in a hierarchical framework (Klauer, 2010; Matzke et al., in press), or contain a mixture-model representation with different latent classes of participants (for application to other models see Scheibehenne, Rieskamp, & Wagenmakers, 2013; Frühwirth–Schnatter, 2006). In theory, it is relatively easy to derive Bayes factors for these more complicated models. In practice, however, Bayes factors for complicated models may require the use of numerical techniques more involved than importance sampling. Nevertheless, for standard MPT models the Beta mixture importance sampler appears to be a convenient and reliable tool to obtain Bayes factors. We hope that this methodology will facilitate the principled comparison of MPT models in future applications.

Acknowledgements

Glossary

Akaike's information criterion (AIC) A quantity that expresses the generalizability of a model, based on the likelihood of the data under the model and the number of free parameters in the model.

Akaike weights A quantity that conveys the relative preference among a set of candidate models, using AIC as a measure of generalizability.

Anscombe's quartet A set of four bivariate data sets whose statistical properties are virtually indistinguishable until they are displayed graphically, and a canonical example of the importance of data visualization.

Bayes factor (BF) A quantity that conveys the degree to which the observed data sway our beliefs towards one or the other model. Under a-priori indifference between two models $\mathcal{M}_1$ and $\mathcal{M}_2$, the BF expresses the a-posteriori relative probability of the two.

Bayesian information criterion (BIC) A quantity that expresses the generalizability of a model, based on the likelihood of the data under the model, the number of free parameters in the model, and the amount of data.

Fisher information approximation (FIA) One of several approximations used to compute the MDL.

Goodness-of-fit A quantity that expresses how well a model is able to account for a given set of observations.

Importance sampling A numerical algorithm to efficiently draw samples from a distribution by factoring it into an easy-to-compute function over an easy-to-sample from density.

Jeffreys weights A quantity that conveys the relative preference among a set of candidate models, using BF as a measure of generalizability.

---

[7]We confirmed the high quality of fit in a Bayesian framework using posterior predictives (Gelman & Hill, 2007), results not reported here.

Likelihood principle A principle of modeling and statistics which states that all information about a certain parameter that is obtainable from an experiment is contained in the likelihood function of that parameter for the given data. Many common statistical procedures, such as hypothesis testing with p-values, violate this principle.

Minimum description length (MDL) A quantity that expresses the generalizability of a model, based on the extent to which the model allows the observed data to be compressed.

Monte Carlo sampling A general class of numerical algorithms used to characterize (i.e., compute descriptive measures) an arbitrary distribution by drawing large numbers of random samples from it.

Nested models Model $\mathcal{M}_1$ is nested in Model $\mathcal{M}_2$ if there exists a special case of $\mathcal{M}_2$ that is equivalent to $\mathcal{M}_1$.

Overfitting A pitfall of modeling whereby the proposed model is too complex and begins to account for irrelevant particulars (i.e., random noise) of a specific data set, causing the model to poorly generalize to other data sets.

Parsimony A strategy against overfitting, and a fundamental principle of model selection: all other things being equal, simpler models should be preferred over complex ones; or: greater model complexity must be bought with greater explanatory power. Often referred to as *Occam's razor*.

Rissanen weights A quantity that conveys the relative preference among a set of candidate models, using FIA as a measure of generalizability.

Savage-Dickey density ratio An efficient method for computing a Bayes factor between nested models.

Schwartz weights A quantity that conveys the relative preference among a set of candidate models, using BIC as a measure of generalizability.

References

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

Akaike, H. (1974a). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Akaike, H. (1974b). On the likelihood of a time series model. *The Statistician*, *27*, 217–235.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*, 1036–1060.

Andrieu, C., De Freitas, N., Doucet, A., & Jordan, M. I. (2003). An introduction to MCMC for machine learning. *Machine Learning*, *50*, 5–43.

Ardia, D., Baştürk, N., Hoogerheide, L., & van Dijk, H. K. (2012). A comparative study of Monte Carlo methods for efficient evaluation of marginal likelihood. *Computational Statistics and Data Analysis*, *56*, 3398–3414.

Batchelder, W. H., & Riefer, D. M. (1980). Separation of storage and retrieval factors in free recall of clusterable pairs. *Psychological Review*, *87*, 375–397.

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review*, *6*, 57–86.

Batchelder, W. H., & Riefer, D. M. (2007). Using multinomial processing tree models to measure cognitive deficits in clinical populations. In R. W. J. Neufeld (Ed.), *Advances in clinical cognitive science: Formal modeling of processes and symptoms* (pp. 19–50). Washington, DC: American Psychological Association.

Berger, J. O. (2006). Bayes factors. In S. Kotz, N. Balakrishnan, C. Read, B. Vidakovic, & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences, vol. 1 (2nd ed.)* (pp. 378–386). Hoboken, NJ: Wiley.

Berger, J. O., & Berry, D. A. (1988). Statistical analysis and the illusion of objectivity. *American Scientist*, *76*, 159–165.

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

Berger, J. O., & Wolpert, R. L. (1988). *The likelihood principle (2nd ed.)*. Hayward (CA): Institute of Mathematical Statistics.

Bernardo, J. M., & Smith, A. F. M. (1994). *Bayesian theory*. New York: Wiley.

Brown, S. D., & Heathcote, A. (2005). Practice increases the efficiency of evidence accumulation in perceptual choice. *Journal of Experimental Psychology: Human Perception and Performance*, *31*, 289–298.

Brown, S. D., & Heathcote, A. J. (2008). The simplest complete model of choice reaction time: Linear ballistic accumulation. *Cognitive Psychology*, *57*, 153–178.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information–theoretic approach (2nd ed.)*. New York: Springer Verlag.

Busemeyer, J. R., & Diederich, A. (2010). *Cognitive modeling*. Thousand Oaks, CA: Sage.

Chechile, R. A. (1973). *The relative storage and retrieval losses in short–term memory as a function of the similarity and amount of information processing in the interpolated task*. Unpublished doctoral dissertation, University of Pittsburgh.

Chechile, R. A., & Meyer, D. L. (1976). A Bayesian procedure for separately estimating storage and retrieval components of forgetting. *Journal of Mathematical Psychology*, *13*, 269–295.

Chen, M.-H. (2005). Computing marginal likelihoods from a single MCMC output. *Statistica Neerlandica*, *59*, 16–29.

Chen, M.-H., Shao, Q.-M., & Ibrahim, J. G. (2002). *Monte Carlo methods in Bayesian computation*. New York: Springer.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing account of the Stroop effect. *Psychological Review, 97*, 332–361.

D'Agostino, R. B., & Stephens, M. A. (1986). *Goodness-of-fit techniques.* Marcel Dekker Inc.

Dickey, J. M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Mathematical Statistics, 42*, 204–223.

Dickey, J. M., & Lientz, B. P. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *The Annals of Mathematical Statistics, 41*, 214–226.

Dutilh, G., Vandekerckhove, J., Tuerlinckx, F., & Wagenmakers, E.-J. (2009). A diffusion model decomposition of the practice effect. *Psychonomic Bulletin & Review, 16*, 1026–1036.

Erdfelder, E., Auer, T.-S., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models: A review of the literature. *Zeitschrift für Psychologie, 217*, 108–124.

Frühwirth–Schnatter, S. (2006). *Finite mixture and Markov switching models.* New York: Springer.

Gamerman, D., & Lopes, H. F. (2006). *Markov chain Monte Carlo: Stochastic simulation for Bayesian inference.* Boca Raton, FL: Chapman & Hall/CRC.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.* Cambridge: Cambridge University Press.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science, 7*, 457–472.

Gill, J. (2002). *Bayesian methods: A social and behavioral sciences approach.* Boca Raton (FL): CRC Press.

Good, I. J. (1985). Weight of evidence: A brief survey. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian statistics 2* (pp. 249–269). New York: Elsevier.

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

Grünwald, P. (2007). *The minimum description length principle.* Cambridge, MA: MIT Press.

Grünwald, P., Myung, I. J., & Pitt, M. A. (Eds.). (2005). *Advances in minimum description length: Theory and applications.* Cambridge, MA: MIT Press.

Hammersley, J. M., & Handscomb, D. C. (1964). *Monte Carlo methods.* London: Methuen.

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review, 7*, 185–207.

Heathcote, A., & Hayes, B. (2012). Diffusion versus linear ballistic accumulation: Different models for response time with different conclusions about psychological mechanisms? *Canadian Journal of Experimental Psychology, 66*, 125–136.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science, 14*, 382–417.

Jefferys, W. H., & Berger, J. O. (1992). Ockham's razor and Bayesian analysis. *American Scientist, 80*, 64–72.

Jeffreys, H. (1961). *Theory of probability* (3 ed.). Oxford, UK: Oxford University Press.

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

Klauer, K. C. (2010). Hierarchical multinomial processing tree models: A latent–trait approach. *Psychometrika, 75*, 70–98.

Klauer, K. C., & Kellen, D. (2011). The flexibility of models of recognition memory: An analysis by the minimum-description length principle. *Journal of Mathematical Psychology, 55*(6), 430–450.

Lee, M. D., & Wagenmakers, E.-J. (in press). *Bayesian modeling for cognitive science: A practical course.* Cambridge: Cambridge University Press.

Lewandowsky, S., & Farrell, S. (2010). *Computational modeling in cognition: Principles and practice.* Thousand Oaks, CA: Sage.

Lewis, S. M., & Raftery, A. E. (1997). Estimating Bayes factors via posterior simulation with the Laplace–Metropolis estimator. *Journal of the American Statistical Association*, *92*, 648–655.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of *g* priors for Bayesian variable selection. *Journal of the American Statistical Association*, *103*, 410–423.

Loftus, E. F., Miller, D. G., & Burns, H. J. (1978). Semantic integration of verbal information into a visual memory. *Journal of Experimental Psychology: Human Learning and Memory*, *4*, 19–31.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, *95*, 492–527.

Logan, G. D. (1992). Shapes of reaction–time distributions and shapes of learning curves: A test of the instance theory of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 883–914.

Logan, G. D. (2002). An instance theory of attention and memory. *Psychological Review*, *109*, 376–400.

Lunn, D., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. (2012). *The BUGS book: A practical introduction to Bayesian analysis.* Boca Raton (FL): Chapman & Hall/CRC.

Ly, A., Verhagen, J., & Wagenmakers, E.-J. (in preparation). A tutorial on the fisher information matrix.

MacKay, D. J. C. (2003). *Information theory, inference, and learning algorithms.* Cambridge: Cambridge University Press.

Matzke, D., Dolan, C. V., Batchelder, W. H., & Wagenmakers, E.-J. (in press). Bayesian estimation of multinomial processing tree models with heterogeneity in participants and items. *Psychometrika*.

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, *44*, 190–204.

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

Myung, I. J., Navarro, D. J., & Pitt, M. A. (2006). Model selection by normalized maximum likelihood. *Journal of Mathematical Psychology*, *50*, 167–179.

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

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

O'Hagan, A., & Forster, J. (2004). *Kendall's advanced theory of statistics vol. 2B: Bayesian inference (2nd ed.).* London: Arnold.

Pauler, D. K. (1998). The Schwarz criterion and related methods for normal linear models. *Biometrika*, *85*, 13–27.

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

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

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, & A. Zeileis (Eds.), *Proceedings of the 3rd international workshop on distributed statistical computing.* Vienna, Austria.

Raftery, A. E. (1995). Bayesian model selection in social research. 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.

Rickard, T. C. (1997). Bending the power law: A CMPL theory of strategy shifts and the automatization of cognitive skills. *Journal of Experimental Psychology: General*, *126*, 288–311.

Riefer, D. M., & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review*, *95*, 318–339.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, *14*, 445–471.

Rissanen, J. (1987). Stochastic complexity. *Journal of the Royal Statistical Society B*, *49*, 223–239.

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.

Rouder, J. N., Lu, J., Morey, R. D., Sun, D., & Speckman, P. L. (2008). A hierarchical process dissociation model. *Journal of Experimental Psychology: General, 137*, 370–389.

Scheibehenne, B., Rieskamp, J., & Wagenmakers, E.-J. (2013). Testing adaptive toolbox models: A Bayesian hierarchical approach. *Psychological Review, 120*, 39–64.

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

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review, 4*, 145–166.

Silver, N. (2012). *The signal and the noise: The art and science of prediction.* London: Allen Lane.

Singmann, H., & Kellen, D. (in press). MPTinR: Analysis of multinomial processing tree models with R. *Behavior Research Methods.*

Smith, J. B., & Batchelder, W. H. (2010). Beta–MPT: Multinomial processing tree models for addressing individual differences. *Journal of Mathematical Psychology, 54*, 167–183.

Stone, C. J., Hansen, M. H., Kooperberg, C., & Truong, Y. K. (1997). Polynomial splines and their tensor products in extended linear modeling (with discussion). *The Annals of Statistics, 25*, 1371–1470.

Townsend, J. T. (1975). The mind–body equation revisited. In C. Cheng (Ed.), *Philosophical aspects of the mind–body problem* (pp. 200–218). Honolulu, USA: Honolulu University Press.

Verdinelli, I., & Wasserman, L. (1995). Computing Bayes factors using a generalization of the Savage–Dickey density ratio. *Journal of the American Statistical Association, 90*, 614–618.

Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological methods, 17*, 228.

Wagenaar, W. A., & Boer, J. P. A. (1987). Misleading postevent information: Testing parameterized models of integration in memory. *Acta Psychologica, 66*, 291–306.

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

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

Wagenmakers, E.-J., & Waldorp, L. (2006). Model selection: Theoretical developments and applications [Special issue]. *Journal of Mathematical Psychology, 50*(2), –.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi. *Journal of Personality and Social Psychology, 100*, 426–432.

Wetzels, R., Grasman, R. P. P. P., & Wagenmakers, E.-J. (2010). An encompassing prior generalization of the Savage–Dickey density ratio test. *Computational Statistics & Data Analysis, 54*, 2094–2102.

Wu, H., Myung, J. I., & Batchelder, W. H. (2010). Minimum description length model selection of multinomial processing tree models. *Psychonomic Bulletin & Review, 17*, 275–286.