Bayesian inference for psychometric functions

Malte Kuss Max Planck Institute for Biological Cybernetics, Tübingen, Germany

Frank Jäkel Max Planck Institute for Biological Cybernetics, Tübingen, Germany

Felix A. Wichmann Max Planck Institute for Biological Cybernetics, Tübingen, Germany

In psychophysical studies, the psychometric function is used to model the relation between physical stimulus intensity and the observer's ability to detect or discriminate between stimuli of different intensities. In this study, we propose the use of Bayesian inference to extract the information contained in experimental data to estimate the parameters of psychometric functions. Because Bayesian inference cannot be performed analytically, we describe how a Markov chain Monte Carlo method can be used to generate samples from the posterior distribution over parameters. These samples are used to estimate Bayesian confidence intervals and other characteristics of the posterior distribution. In addition, we discuss the parameterization of psychometric functions and the role of prior distributions in the analysis. The proposed approach is exemplified using artificially generated data and in a case study for real experimental data. Furthermore, we compare our approach with traditional methods based on maximum likelihood parameter estimation combined with bootstrap techniques for confidence interval estimation and find the Bayesian approach to be superior.

Keywords: psychometric function, Bayesian inference, Markov chain Monte Carlo, confidence intervals

Introduction

Psychophysics explores the connection between physical stimuli and subjective responses. The psychometric function relates the stimulus intensity ("physics") on the abscissa to the observer's response ("psychology") on the ordinate and is the central function in the analysis of data obtained from psychophysical studies. This is true not only in classical psychophysical settings in experimental psychology but also equally true in clinical or developmental studies where the datasets are typically even smaller, and thus proper statistical procedures are even more important. It is also true in awake-behaving neurophysiology studies where the datasets may be larger but the problem of stimulus-independent errors or "lapses" (Wichmann & Hill, 2001a) may be more pronounced.

Given that psychophysical experiments tend to be time consuming and tiring for the observers, many methods have been developed to estimate only a single point of the psychometric function, typically a point in the interval between 50% and 90% correct performance termed the *threshold*. These so-called *adaptive methods* vary the stimulus strength based on previous responses of the observer; adaptive methods can be divided into nonparametric (Garcia-Perez, 1998; Rose, Teller, & Rendleman, 1970; Taylor, 1971; Wetherill & Levitt, 1965) and parametric (Alcalá-Quintana & Garcia-Pérez, 2004; King-Smith & Rose, 1997; Kontsevich & Tyler, 1999; Madigan & Williams, 1987; Pelli, 1987; Pentland, 1980; Snoeren & Puts, 1997; Watson & Pelli, 1983; Watt & Andrews, 1981), the latter including some methods that are explicitly Bayesian (Alcalá-Quintana & Garcia-Pérez, 2004; Kontsevich & Tyler, 1999; Watson & Pelli, 1983). For a review of some of the most common adaptive methods, see Treutwein (1995).

However, in many cases it is important not only to know a single point of the psychometric function but also to estimate it in its entirety. Differences between experimental conditions may not lead to different threshold values but the slope of the psychometric functions could have changed significantly (Green & Swets, 1966; Wichmann, 1999). In principle, all trials taken during a run of an adaptive method could be used to estimate the complete psychometric function, but this is not recommended because the sampling, optimized to estimate only a single point, is sub-optimal for complete function estimation (Kaernbach, 2001).

There exists a fairly comprehensive literature on estimating the psychometric function (Klein, 2001; O'Regan & Humbert, 1989; Treutwein & Strasburger, 1999), with some studies additionally covering sampling issues and goodness-of-fit (Wichmann & Hill, 2001a, 2001b) or nonparametric estimation methods (Miller & Ulrich, 2001). Comparatively few studies, however, have investigated how to obtain reliable confidence intervals for the estimated parameters of psychometric functions (Finney, 1971; Foster & Bischof, 1991, 1997; Maloney, 1990; McKee, Klein, & Teller, 1985; Wichmann & Hill, 2001a, 2001b). There appears to exist a general consensus, however, that bootstrap methods offer more reliable confidence intervals than methods based on asymptotic considerations due to small

doi:10.1167/5.5.8 Received January 10, 2005; published May 27, 2005 ISSN 1534-7362

datasets typical in psychophysical research (between 50 and 1000 trials per psychometric function). In this work, we present experiments suggesting that Bayesian inference methods are likely to lead to more accurate point estimates and confidence intervals than do bootstrap-based techniques.

The binomial mixture model

In this section, we formally derive a basic statistical model of the process that generates the data. The object of interest is a *parametric psychometric function* $F(x,\theta)$ parameterized by θ, which maps the stimulus intensity x to the [0,1] interval. This function is commonly chosen to have a sigmoidal form like cumulative density functions of various probability distributions. We will discuss several common choices below in Parameterization and prior distributions.

The psychometric function relates the observer's response to stimulus intensity. In an *n*AFC experimental setting, there is a *chance probability* π_c that the observer "guesses" the correct answer independent of the stimulus. This probability of making the correct guess is usually $\pi_c = 1/n$, where n is the number of possible choices (the n in *n*AFC). In a long sequence of experimental trials, the observer occasionally *lapses* (i.e., makes a random choice independent of the stimulus). In vision experiments an obvious example is blinking while the stimulus is presented. This probability of lapsing π_l is a nuisance parameter, but it is necessary to take its effect into account in statistical modeling as shown by Wichmann and Hill (2001a, 2001b).

We now have all quantities for a basic model to relate the psychometric function F to the probability of giving the correct answer in a single *n*AFC stimulus presentation. Given the stimulus intensity x, the event of correct discrimination is a Bernoulli variable with probability of success equal to

$$\Psi(x,\theta,\pi_c,\pi_l) = (1-\pi_l)\left[(1-\pi_c)F(x,\theta)+\pi_c\right]+\pi_c\pi_l\,, \quad (1)$$

where $F(x,\theta)$ characterizes the change of discriminability as a function of the stimulus intensity. The model comes in the form of a mixture of two Bernoulli distributions, which is again a Bernoulli distribution. With probability π_l the observer lapses and has chance π_c to guess the correct answer. With probability $(1 - \pi_l)$ the observer does not lapse and has a chance of $(1 - \pi_c)F(x,\theta)+\pi_c$, which is $F(x,\theta)$ scaled to the $[\pi_c,1]$ interval, to give the correct answer.

The psychophysical experiment can be seen as a sequence of such Bernoulli trials. Often only a small number $\{x_1,\ldots,x_k\}$ of distinct stimulus intensities are used in an experiment, which allows a more compact representation. By aggregating the trials for identical stimulus intensities, we compress the data to a set of triples $D = \{(x,N,n)_i | i = 1,\ldots,k\}$ such that at contrast x_i we conducted N_i trials and observed n_i correct responses. Because n_i is a sum of Bernoulli variables, it has a binomial distribution

$$\begin{aligned} p(D \mid \theta,\pi_l,\pi_c) &= \prod_{i=1}^{k} p(n_i \mid N_i,x_i,\theta,\pi_l,\pi_c) \\ &= \prod_{i=1}^{k} \text{Binomial}(n_i \mid N_i,\Psi(x_i,\theta,\pi_l,\pi_c)), \end{aligned} \quad (2)$$

where Ψ is given by Equation 1. Equation 2 describes the assumed generative model of the data (i.e., the *sampling distribution*). Furthermore, read as a function of θ and π_l for observed D, we refer to it as the *likelihood* of the binomial mixture model.

What we have described thus far is the standard binomial mixture model for parametric psychometric functions as assumed in virtually every study on psychometric function fitting (e.g., Klein, 2001; Maloney, 1990; Treutwein & Strasburger, 1999)—except for the addition of the nuisance parameter π_l (Wichmann & Hill, 2001a). Furthermore, this model is easy to analyze and efficient to implement. Nevertheless, in data analysis one should always be aware of the model's assumptions, and the conclusions drawn from an analysis should obviously not be trusted more than the assumptions they are based on, whether we apply Bayesian inference or any other statistical method. For example, the assumption that for a given stimulus intensity the Bernoulli trials all have the same probability of success ignores adaptation processes, learning, and other forms of nonstationarity. For well-trained psychophysical observers this assumption is justified (Blackwell, 1952), but for naïve observers it may not always hold true, and the residuals of the fit have to be examined in detail (Wichmann & Hill, 2001a). A second potential worry concerns the choice or assumption of a particular parametric form of F. Typically, there are few a priori reasons to favor one sigmoidal function over another. In practice, however, the estimates of threshold and slope of the psychometric function rarely differ significantly from one F to another (Wichmann & Hill, 2001b).

Bayesian inference for psychometric functions

Here we describe how the data collected in psychophysical experiments can be used to do Bayesian *inference* about the parameters θ of a psychometric function $F(x,\theta)$ and the lapse probability π_l.

First we give a general description of how inference is performed in the Bayesian framework, using a simplified notation. Starting point is a model of the process of how the data that we can observe is generated. Let $p(D|\phi)$ be a statistical description of this model where D denotes observable data and ϕ are model parameters. In a nutshell, the problem is that the *true* generating parameter ϕ^* is hidden, but by observing data we can reduce our uncertainty about its value. In the Bayesian framework, probability distributions over parameter values are used to describe beliefs and uncertainties about the parameter value in the data-generating process.

The *prior* distribution $p(\phi)$ represents beliefs about the value of the true parameter ϕ^* previous to an inference step. By inference we refer to the process of integrating the information contained in observed data D and the prior $p(\phi)$ into a *posterior* distribution $p(\phi|D)$. The posterior is obtained according to Bayes' rule:

$$p(\phi \mid D) = \frac{p(D \mid \phi)p(\phi)}{p(D)}. \quad (3)$$

This can be understood as a weighting in which prior beliefs about ϕ^* are weighted proportionally to their compatibility with the observed data. The weighting is given by the likelihood function, which is $p(D|\phi)$ as a function of ϕ for given D. Prior and posterior are probability distributions describing two states relative to an inference step and correspond to potentially different beliefs about the value of the parameter that generated the data. For details, the reader is referred to O'Hagan (1994) and Jaynes (2003), to mention only two textbooks on Bayesian statistics.

We now describe how this framework can be applied to infer something about the parameters of psychometric functions. In the following, we assume the data are generated according to the binomial mixture model for some specific parametric type of $F(x,\theta)$. In psychophysical studies data $D = \{(x,N,n)_i \mid i = 1,\ldots,k\}$ are collected to learn about θ and π_l. Again Bayes' rule describes how the observed data consistently reduce the uncertainty about the underlying value of θ and π_l. Formally, the posterior is obtained according to Bayes' rule

$$p(\theta,\pi_l \mid D,\pi_c) = \frac{p(D \mid \theta,\pi_l,\pi_c)p(\theta)p(\pi_l)}{\int p(D \mid \theta,\pi_l,\pi_c)p(\theta)p(\pi_l)d\theta d\pi_l}, \quad (4)$$

where $p(\theta)$ and $p(\pi_l)$ are prior distributions, $p(D|\theta,\pi_l,\pi_c)$ acts as the likelihood, and $p(\theta,\pi_l|D,\pi_c)$ is the posterior. The posterior distribution summarizes all information contained in the observations and the prior about θ and π_l. Unfortunately, solving the integral in the denominator appears to be analytically intractable such that the posterior cannot be computed in closed form. Even if we could compute the posterior, the distribution would be of a nonstandard type, and we would be unable to work with it analytically. We therefore have to use approximative techniques to describe the information presented by the posterior.

Point estimates and confidence intervals

The most simple approximation to the information represented by the posterior distribution is to state a single point estimate of the true parameter values. In the Bayesian framework, choosing a point estimate is considered a *decision* problem in which the decision maker minimizes an expected risk, where the expectation is taken with respect to the posterior distribution (Jaynes, 2003). The risk function characterizes the *loss* associated with a discrepancy between the point estimate and the unknown true parameter value. For example, the expected absolute error is minimized by the median of the posterior distribution (MED). Likewise, minimizing the expected squared error leads to choosing the mean of the posterior (MEAN).

The mode of the posterior, which will be referred to as the maximum a posteriori (MAP) estimate, is obtained for a loss function, which is zero if the estimate and the true value match exactly and 1 otherwise. In the binomial mixture model, assuming the psychometric function is differentiable, gradient-based methods can be used to find the MAP point estimate

$$(\theta,\pi_l)^{MAP} = \underset{\theta,\pi_l}{\operatorname{argmax}}\ p(D \mid \theta,\pi_l,\pi_c)p(\theta)p(\pi_l). \quad (5)$$

If the prior $p(\theta)p(\pi_l)$ is taken to be constant (i.e., a flat prior), the maximum likelihood (ML) estimator is derived as a special case.

Another simple approximation technique is Laplace's method, by which the posterior is approximated by a Gaussian distribution that is found by a second-order Taylor expansion around the mode. This method is applicable in the proposed setting but the approximation might be poor. An obvious drawback is that the approximation is symmetric and fitted locally around the mode but can be poor in approximating the tails of the posterior. We therefore do not consider this method in the following.

The posterior distribution represents the remaining uncertainty after having seen the data. An obvious problem is that any notion of uncertainty or confidence is lost when only point estimates are stated. A convenient way of expressing how narrowly a parameter is determined is to state confidence regions. Given a confidence level $\gamma \in (0,1)$, the notion of a confidence region is conceptually different in frequentist and Bayesian statistics (DeGroot & Schervish, 2002).

In the Bayesian framework, it is valid to define a γ confidence region simply as a region in which the true parameter values are believed to lie with probability γ. This can be stated because the parameters are random variables, and we can express our degree of belief for any statement regarding the parameters by evaluating the statement under the posterior distribution.

In frequentist statistics confidence regions are constructed and interpreted differently. In this setting, the region itself is a random variable that contains the true parameter value with probability γ. This means that if the experiment is repeated infinitely many times, the proportion of computed regions containing the true value would be γ. For a particular data set, it is not possible to state a probability assignment that the true parameter lies in a computed confidence interval.

In case the distribution of an estimator cannot be computed analytically, a common strategy in frequentist statistics is to compute approximate confidence intervals using bootstrap methods (Efron & Tibshirani, 1993). The basic idea is to repeatedly generate artificial data sets. For each artificial data set, the parameters are re-estimated and

the variability of these estimates is used to estimate confidence intervals. In parametric bootstrap methods, artificial data sets are generated from the model using maximum likelihood estimates of the parameters. Wichmann and Hill (2001b) describe parametric bootstrap techniques to estimate confidence intervals for parameters of psychometric functions. The accuracy of confidence intervals obtained by this method crucially depends on the accuracy of the maximum likelihood estimate.

From a Bayesian point of view, the posterior represents the uncertainty about the true parameter values and should therefore be used to make confidence statements. Approximations using sampling methods are also common in Bayesian statistics in situations in which the posterior cannot be computed analytically. The difference is that in the Bayesian framework, samples are generated from the posterior over parameters. This can be implemented using Markov chain Monte Carlo techniques.

Approximate inference by Markov chain Monte Carlo sampling

In this section, we describe the basic idea of using Markov chain Monte Carlo (MCMC) methods for approximate Bayesian inference. For more technical introductions, the reader is referred to MacKay (1999, 2003), while more comprehensive reviews can be found in Neal (1993) and Gilks, Richardson, and Spiegelhalter (1996).

Recall the simplified notation introduced in the previous section. Assume some data D have been observed and we want to compute the posterior according to Bayes' rule Equation 3. A common situation is that we can evaluate the likelihood $p(D|\phi)$ and the prior $p(\phi)$ for every possible value of ϕ, but we cannot compute or work with the posterior analytically. MCMC methods sidestep this problem by generating samples from the posterior $p(\phi|D)$ using only evaluations of the *unnormalized* posterior $q(\phi|D) = p(D|\phi)p(\phi)$. The idea behind this is that the samples characterize the posterior distribution sufficiently well. In particular, statistics of the samples can be used to approximate properties of the posterior distribution. For example, the mean of the samples is an approximation to the mean of the posterior distribution.

To generate a sample from the posterior, a random sequence of parameter values $\phi_1, \phi_2, \ldots, \phi_n$ is generated such that the distribution of ϕ_n asymptotically becomes identical to the posterior as the length of the sequence n increases. In the MCMC terminology, the sequence is called a *chain* and each element is referred to as a *state*. In practice the chain is generated for a finite length n and the state ϕ_n is interpreted as a sample of the posterior. The procedure is repeated until enough samples are obtained such that the characteristics of the posterior distribution can be well approximated by statistics of the generated samples.

For this mechanism to work, the sequence has to be constructed in a particular way following the Metropolis-Hastings rule, which determines how consecutive states in a chain are found. Assume ϕ_t is the current state. To find a valid consecutive state ϕ_{t+1}, a candidate value ϕ' is proposed from a *proposal distribution* $p(\phi'|\phi_t)$, for example, a Gaussian distribution centered at ϕ_t. The decision whether ϕ' is accepted as consecutive state depends on the ratio

$$\pi = \frac{q(\phi' \mid D)}{q(\phi_t \mid D)} \frac{p(\phi_t \mid \phi')}{p(\phi' \mid \phi_t)}. \quad (6)$$

According to the Metropolis-Hastings rule, the proposal ϕ' is accepted if $\pi > 1$; otherwise, the probability of acceptance is equal to π. The first fraction in the definition of π captures whether the proposed state is in a region of higher posterior density than the current state. The second term captures the reversibility of the proposed state transition in case the proposal distribution is asymmetric (for the Gaussian example mentioned above this term is equal to 1). Because ϕ_{t+1} depends only on ϕ_t and not on the history of states, the resulting chain is called a Markov chain. See Figure 1 for an example.

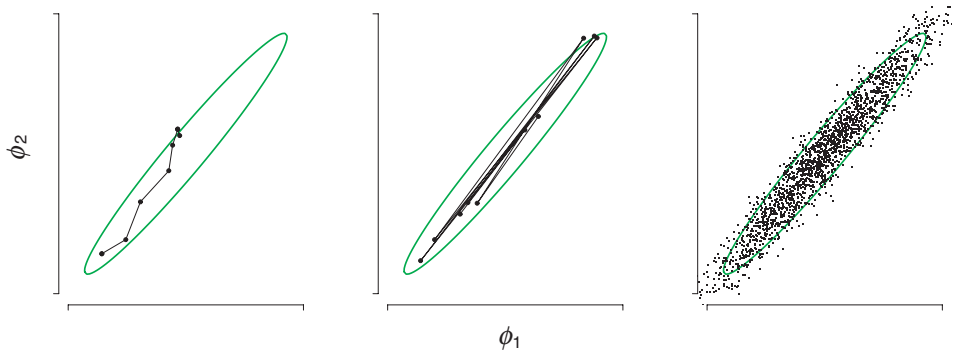

Figure 1. Illustration of Markov chain Monte Carlo (MCMC) sampling. The ellipse represents a contour of a posterior $p(\phi|D)$ we want to sample (i.e., that we want to approximate). The left panel shows a chain (after 10 proposed states) generated by Metropolis-Hastings sampling using local moves [the proposal distribution is Gaussian $p(\phi'|\phi_t) = N(\phi'|\phi', \sigma^2 I)$]. States are depicted by points and consecutive states are connected by lines. The central panel shows a chain (after 10 proposed states) obtained using hybrid MCMC sampling. Note that the states appear to be less dependent, whereas the number of accepted states is larger. The right panel shows 2000 samples generated using hybrid MCMC.

The computational efficiency of MCMC sampling depends on how the consecutive state is proposed. While simulating the Markov chain, states that occur close-by in the chain are dependent through the proposal distribution. Refinements of this scheme are directed toward improved proposal distributions such that this dependence is reduced. This has the effect of reducing the length of the sequence n after which the state can be considered an approximately independent sample of the posterior.

In the following we use *hybrid* Monte Carlo sampling, which is also known as Hamiltonian sampling, as described by Neal (1993) and MacKay (2003). New states are proposed using a procedure that can be understood as a discrete simulation of Hamiltonian dynamics. The sampling scheme requires additional parameters to be set, namely the number of steps (so-called *leapfrog* steps) and the step sizes used in the discrete simulation.

The main idea of this work is to use hybrid MCMC sampling to generate samples from the posterior (Equation 4) over the parameters of psychometric functions. Once we are convinced that the MCMC samples we have generated are representative of the posterior, they can be used to estimate certain characteristics of the posterior distribution. The empirical mean of the samples can be used as an estimate of the expectation of the posterior distribution (MEAN). Likewise, the sample median is an approximation to the median of the posterior distribution (MED). The empirical quantiles of the samples can be interpreted as estimates of the quantiles of the posterior distribution. We refer to the interval between the $(1 - \gamma)/2$ and $(1 + \gamma)/2$ empirical quantiles of the samples as an *approximate Bayesian* γ *confidence interval*.

Before we present examples of this approach, the following section describes parameterizations of psychometric functions and the role of prior distributions in the analysis.

Parameterization and prior distributions

In psychophysical practice the experimenter has certain beliefs about the mechanism of interest, otherwise the experiment could not be designed. Expressing prior beliefs and parameterization of the model go hand in hand. It is therefore advantageous to parameterize the model close to the way the scientist thinks about the mechanism it describes. In the following section, we describe a convenient parameterization of psychometric functions, before we discuss various forms of prior distributions on their parameters.

Parameterization of the psychometric function

Let $F(x,\theta)$ be the psychometric function and F^{-1} its inverse. In the analysis of psychophysical data, a common interest is to locate the *threshold* $m = F^{-1}(0.5)$ and a range for which the detectability varies with the stimulus intensity. A common way of characterizing the sensitivity of an observer is the (inverse) slope of the psychometric function at the threshold location. Another way of describing the range of interest is the *width* w defined as $w = F^{-1}(1-\alpha) - F^{-1}(\alpha)$. This is the length of the interval between $F^{-1}(\alpha)$, the stimulus intensity at which $F(x,\theta) = \alpha$, and $F^{-1}(1 - \alpha)$ for some small α. As default we use $\alpha = 0.1$, such that w is the range in which F ranges from 0.1 to 0.9. (see Figure 2a). The parameterization of the psychometric function in terms of threshold and width has been proposed by Alcalá-Quintana and Garcia-Pérez (2004). An advantage of this parameterization is that w comes in the scale of the stimulus itself, whereas the value of the slope is usually difficult to interpret. Furthermore, the width is more general than the slope in the sense that it can be used in

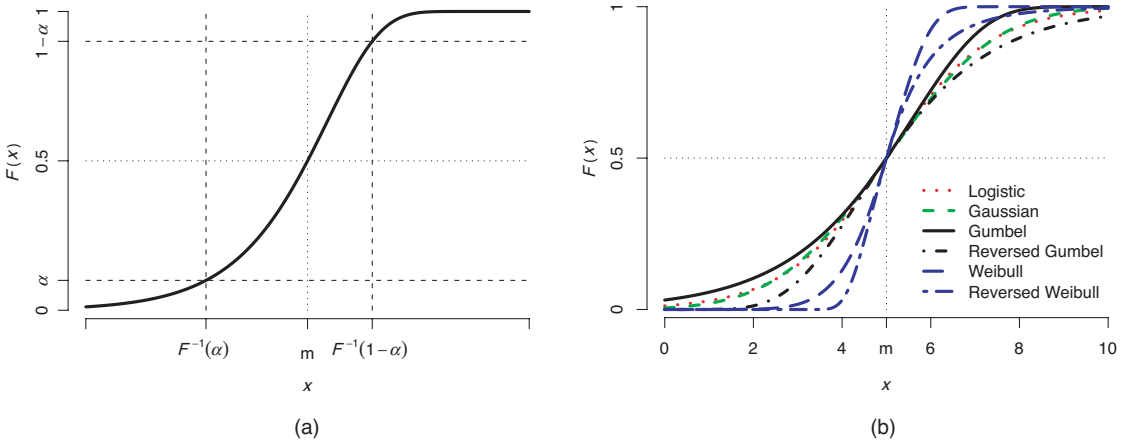

Figure 2. Types and parameterization of psychometric functions. (a) illustrates the parameterization of psychometric functions in terms of threshold location m and width w between $F^{-1}(\alpha)$ and $F^{-1}(1-\alpha)$. The example shows a Gumbel function for $\alpha = 0.1$. (b) exemplifies different types of psychometric function. The logistic, Gaussian, and Gumbel functions are shown for $m = 5$ and $w = 5$. The Weibull functions are plotted for $m = 5$ and $s = 0.5$.

models for which the slope at a particular point does not determine the entire psychometric function.

We now show how various common functions used to model F can be parameterized such that $\theta = [m,w]$. Note that the approach described in this study is not limited to this particular parameterization of F. Many of the functions used to model F also appear in statistical generalized linear models (GLMs) in which they are called response functions (McCullagh & Nelder, 1989).

The logistic function, which is called *logit* response function in GLMs, can be parameterized as

$$F_{\text{logistic}}(x,\theta) = \left(1+\exp\left(-\frac{z(\alpha)}{w}(x-m)\right)\right)^{-1}, \tag{7}$$

where $z(\alpha) = 2\ln(1/\alpha-1)$. The function is point symmetric around the threshold. If w is positive the functions have positive slope and negative slope if w is negative.

The cumulative density function (cdf) of the normal distribution Φ, the *probit* response, can be parameterized as

$$F_{\text{gauss}}(x,\theta) = \Phi\left(x \mid m, \frac{w}{z(\alpha)}\right), \tag{8}$$

where $z(\alpha) = \Phi^{-1}(1-\alpha)-\Phi^{-1}(\alpha)$ using the quantile function Φ^{-1} (inverse of cdf) of the standard normal distribution. The resulting functions often appear very similar to the logistic (see Figure 2b).

The Gumbel function can be derived from the cdf of the Gumbel distribution and is known in GLMs as the *log-log* model. We use the parameterization

$$F_{\text{gumbel}}(x,\theta) = 1-\exp\left(-\exp\left(\frac{z(\alpha)-z(1-\alpha)}{w}(x-m)+z(0.5)\right)\right), \tag{9}$$

where $z(\alpha) = \ln(-\ln(\alpha))$. The Gumbel function is asymmetric. For small x the function is similar to the logistic function but approaches 1 faster as the stimulus intensity gets larger. The asymmetry can be reversed, and we obtain the reversed Gumbel function

$$F_{\text{rgumbel}}(x,\theta) = \exp\left(-\exp\left(\frac{z(1-\alpha)-z(\alpha)}{w}(x-m)+z(0.5)\right)\right), \tag{10}$$

where again $z(\alpha) = \ln(-\ln(\alpha))$.

Another frequently found functional form is related to the cdf of the Weibull distribution, which is also asymmetric. Using the Weibull function is equivalent to using a Gumbel function for log-transformed stimulus intensities. Unfortunately, the Weibull function cannot be parameterized in terms of a width parameter w. Instead, we parameterize the function by threshold location m and slope at threshold $s = \frac{\partial F}{\partial x}\Big|_m$. So we use the following parametric form for the Weibull function

$$F_{\text{weibull}}(x,\theta) = 1-\exp\left(-\exp\left(\frac{2sm}{\ln(2)}\left(\ln(x)-\ln(m)\right)+\ln(\ln(2))\right)\right) \tag{11}$$

and

$$F_{\text{rweibull}}(x,\theta) = \exp\left(-\exp\left(-\frac{2sm}{\ln(2)}\left(\ln(x)-\ln(m)\right)+\ln(\ln(2))\right)\right)$$

for the reversed Weibull function. Both Weibull functions are defined for $x > 0$ and tend to 0 as $x \to 0$, which makes them conceptually appealing in many psychophysical settings.

Prior distributions

Ideally, a prior distribution describes the scientist's degree of belief for all hypotheses about the true model parameters. For continuous parameters, one could "draw" a curve over the parameter space representing the shape of the prior. The line would be at zero for parameter values that are believed to be absolutely impossible and otherwise proportional to the degree of belief in the hypothesis that the pen is over the true value. Using a prior from a parametric family of distributions can be seen as a convenient approximation to this "drawn prior" because it reduces the prior to a parametric form with a few parameters. In practice, a simple technique to find a parametric representation of prior beliefs is to plot probability density functions from a convenient family of distributions. Varying the parameters one can often find a function that is close to the drawn prior. One should also sample from the prior and inspect whether the corresponding model is consistent with prior beliefs.

Often scientists unfamiliar with Bayesian data analysis feel that using informative priors—reflecting their understanding and uncertainties about the data-generating process—somehow "distorts" the inference process. Expressing prior beliefs is certainly nontrivial and requires great care. A common misconception is that using a flat, constant prior on model parameters is equivalent to expressing no prior information about the data-generating process—from a Bayesian point of view this is exactly what non-Bayesians do when they do not specify any prior explicitly. In fact, this prior describes the belief that *every* parameter value is equally likely to the scientist. However, this is typically *not* what the scientists intend: What they want is to allow every model or shape of the model to be equally likely, that is, they want a flat prior in model space. Typically, a flat prior on parameters is, unfortunately, not flat in model space. For psychometric functions this is illustrated in Figure 3. Here we show that using flat priors on the parameters θ

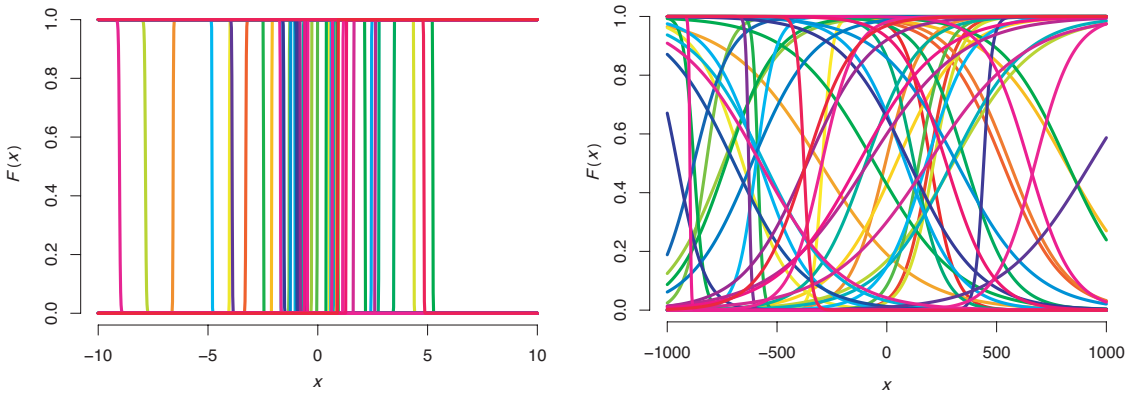

Figure 3. Simply changing the parameterization of the psychometric function makes flat priors favor steep psychometric functions–the prior is flat on parameters but not in function space. We approximate a flat prior on the elements of θ by uniform distributions on the interval [-1000,1000]. For (a) the logistic psychometric function was parameterized $F(x,\theta) = (1+\exp(-(\theta_1 x+\theta_2)))^{-1}$. We then sampled values of θ from the flat prior and plotted the corresponding psychometric function. For (b) the parameterization as shown in Equation 8 was used. Note the different x scales.

strongly favors very steep psychometric functions in Figure 3a, whereas if we simply re-parameterize our psychometric function, this tendency to favor steep psychometric functions disappears in Figure 3b.

Using a flat prior for the lapse rate π_l, a uniform distribution on [0,1], indeed represents maximal uncertainty. The hypotheses that every experimental observation was independent of the stimulus or that no lapse occurred are equally likely. That might reflect the uncertainties of a scientist under certain circumstances, but in general the notion of a lapse implies a rare event. Note that a flat prior on the lapse probability allows the model to explain all the data as a sequence of lapses, which intuitively minimizes the credibility of every observation. So if the scientist can safely assume that the lapse rate of an observer in a given task is small, the observations become more informative about the psychometric function and so its parameters can be better identified. On the other hand, excluding the potential existence of outliers forces the model to explain every single observation such that a single observation can become decisive. Note also that in the procedures described by Wichmann and Hill (2001a, 2001b) the parameter similar to the lapse rate is constrained during the ML optimization.

It can lend insight to examine how sensitively the posterior reacts to changes of the prior. The more data are available and the data are informative about the parameters, the less influential the prior will be. Comparing posteriors and priors can illustrate how informative the experimental data are about the parameters. When the data do not reduce the uncertainty about a certain parameter, then both distributions will be the same, expressing the beliefs are unchanged. The only warning is not to put zero prior probability on potential parameter values unless one knows that they are impossible (Cromwell's dictum).

We now describe some families of distributions that will be used as priors for the parameters of psychometric functions in the experiments described in the following sections. For details on the distributions, the reader is referred to any standard text book on statistics (e.g., DeGroot & Schervish, 2002).

The lapse parameter π_l takes values in the unit interval, and therefore the Beta distribution $p(\pi_l|\alpha,\beta) = \mathrm{Beta}(\pi_l|\alpha,\beta)$ is a convenient choice (see Figure 4a). For $\alpha = 1$ and $\beta = 1$ the uniform distribution on [0,1] is a particular case.

Considering priors for the elements of θ, parameterizing the psychometric function as described above is advantageous because the parameters have a more intuitive interpretation. For convenience, we specify the priors independently $p(\theta) = p(\theta_1)p(\theta_2)$. Especially for the location, parameter m, a normal (Gaussian) distribution, is a convenient choice if its value is unconstrained. By setting the standard deviation to increasingly large values, the prior becomes vaguer. For parameters that are known to be strictly positive, for example, the width w or the slope s, the gamma or the log-normal distribution can be used. If x is log-normal distributed, $\log(x)$ follows a normal distribution. See Figure 4b for examples of gamma and log-normal probability density functions. This section sketched only a small selection of possible densities one can use for specifying priors on the parameters of psychometric functions. If common distributions are not sufficient to model the prior, mixtures of distributions can also be used.

Experiments

Here we present and discuss simulations based on synthetic data and a case study in which we analyze real experimental data. Experiments in which the data are generated from the model can be useful for examining how well

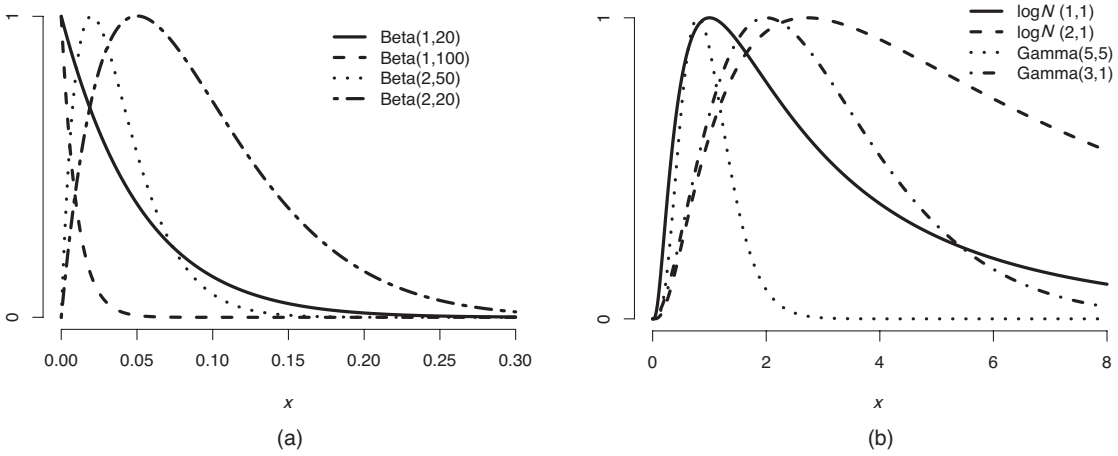

Figure 4. Illustration of the *form* of probability density functions of the beta, gamma, and log-normal distribution for different parameter values. Note that the probability density functions are scaled to the unit interval.

the true parameters can be identified, depending on the properties of the data. We do not aim at providing an exhaustive set of experiments for all possible data situations. Instead, the focus will be on understanding the advantages and difficulties of the proposed method.

Synthetic data

For illustration purposes, a data set from the binomial mixture model (Equation 2) is generated, where F is a Gumbel function with threshold location $m = 5$, width $w = 3$ (for $\alpha = 0.1$), and lapse probability $\pi_l = 0.05$. For $k = 6$ stimulus intensities x_i, corresponding to the F values equal to 0.1, 0.3, 0.6, 0.74, 0.84, and 0.94, we generate $N_i = 60$ samples respectively, which sums to 360 Bernoulli trials in total.

How to choose a prior in artificial experiments is a problematic issue. In the following examples the prior should be accepted as a toy-prior for demonstration purposes. For the lapse probability we use a Beta(2,50) prior (see Figure 4a). On the threshold location we put a wide normal prior with mean $\mu = 2$ and *SD* $\sigma = 10$, which expresses very little information about m. On the width we put a log-normal prior distribution $\log N(1,1)$ (see Figure 4b). Using hybrid MCMC sampling we simulate a Markov chain of 2000 samples from the posterior with 100 leapfrog steps and step sizes (0.5,0.1,0.2), which were chosen to obtain an acceptance rate of approximately 80% and very little autocorrelation between samples.

Furthermore, we compute the posterior sample MEAN, MAP, and ML point estimates, for which the corresponding $\Psi(x,\theta,\pi_l)$ are depicted in Figure 5a. By taking samples from the MCMC chain and plotting the corresponding $\Psi(x,\theta,\pi_l)$, we obtain Figure 5b. Each of the sampled functions represents a hypothesis about the underlying genera-

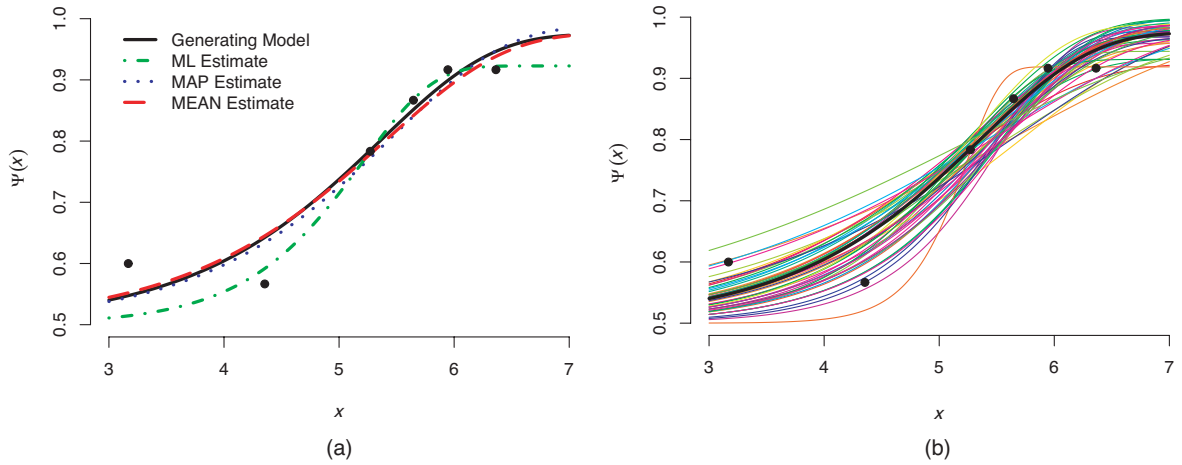

Figure 5. Synthetic data example. (a) shows the generated data set (dots), the $\Psi(x,\theta,\pi_l)$ that was used to generate the data, and three estimates thereof. The maximum likelihood estimate (ML) clearly overestimates π_l and infers a too small width w. The estimate that appears closest to the generating $\Psi(x,\theta,\pi_l)$ corresponds to the mean of MCMC samples and the maximum a posteriori (MAP) estimate. In (b) we plot a large number of hypotheses, each corresponding to an MCMC sample.

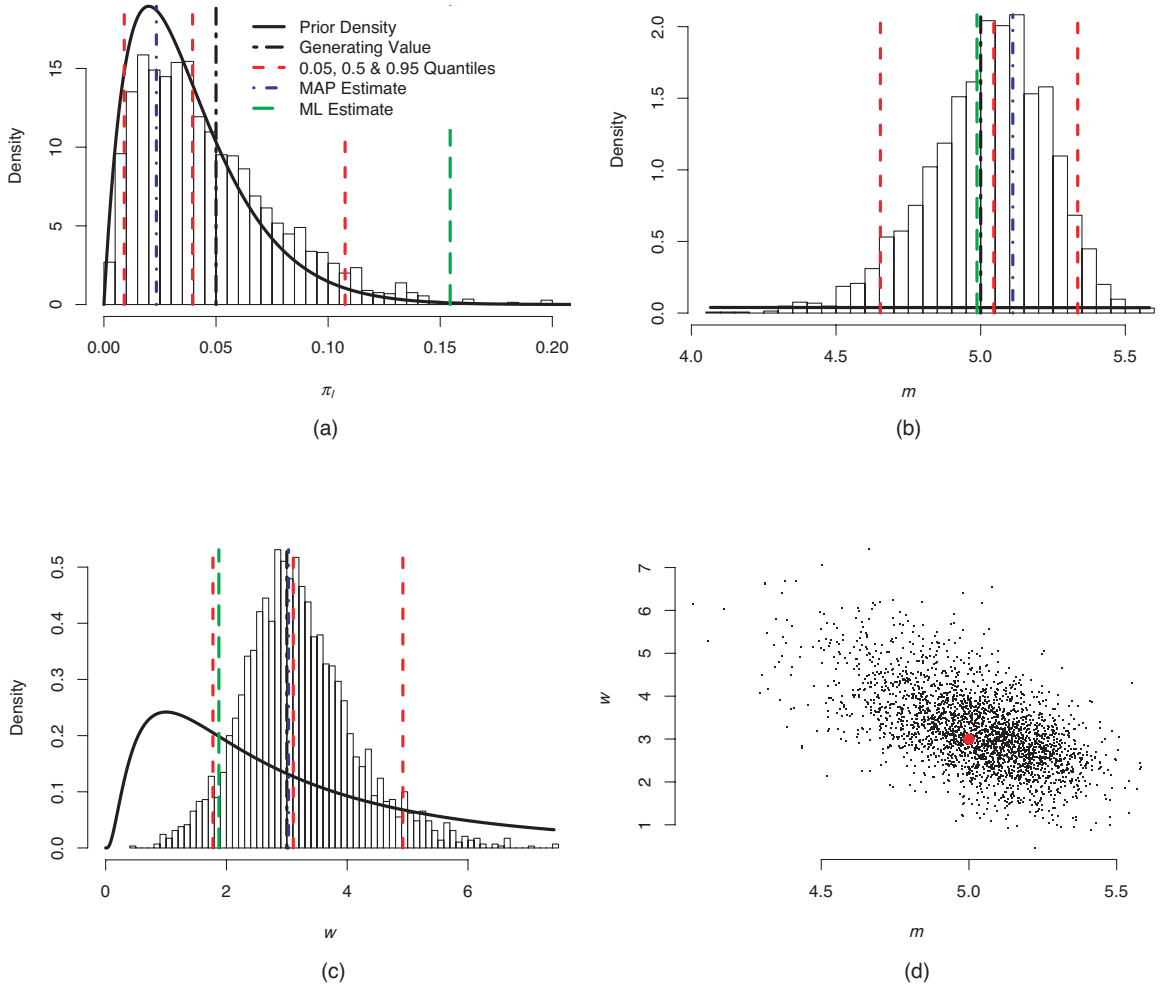

Figure 6. Synthetic data example. Plots of prior densities and histograms of posterior samples. Each plot (a–c) corresponds to one model parameter (π, m, and w) and shows the normalized histogram of MCMC samples of the posterior distribution in comparison to the prior density. Vertical lines mark the value that was used to generate the data, the ML, and MAP point estimates and the 0.05, 0.5, and 0.95 empirical quantiles of the MCMC samples. The interval between the 0.05 and 0.95 quantiles is the approximate Bayesian 90% confidence interval. (d) shows a scatter plot of w and m parts of the samples. Note the negative correlation, which corresponds to the necessity of steeper functions as the threshold location is moved to the right.

tive function valid under the posterior. The functions are relatively close for large values of x but show rather large differences for smaller stimulus intensities. This can be interpreted such that the experimental observations for low stimulus intensities do still support a rather wide range of hypotheses about the width of the psychometric function.

We can gain more insights into the posterior by inspecting the MCMC samples. To illustrate how much the data reduced the uncertainty about the parameters, we graphically compare priors and posteriors, of which the posterior is approximated by a normalized histogram of MCMC samples (see Figure 6).

For the lapse rate shown in Figure 6a, we observe that the posterior is very similar to the prior, which indicates that the data did not allow us to reduce our uncertainty about this parameter. In many experiments we observed that identifying the lapse rate is relatively difficult. Nevertheless, the posterior samples of m and w are generated, while the assumed π_l value varies according to the prior.

For the threshold location m, the samples shown in Figure 6b suggest that the threshold location is well inferred from the data. The prior, which was a wide normal, is approximately constant in the plotted region, and we observe that the data were very informative. The posterior samples of the width w illustrated in Figure 6c show that the data were informative about w but the remaining uncertainty is still relatively large. Note that the function samples given in Figure 5b already indicated that the posterior still supports a rather wide range of hypotheses on w.

We can use the empirical quantiles of the MCMC samples to estimate the quantiles of the posterior distribution. We take the range between the 0.05 and 0.95 empirical quantile as an approximation for the Bayesian 90% confidence interval.

To examine the accuracy of point estimates and the approximated Bayesian confidence regions, we conducted a large set of repeated experiments. We compare approximate Bayesian confidence intervals estimated from MCMC samples with confidence intervals obtained using a *parametric* bootstrap (Efron & Tibshirani, 1993). In parametric bootstrap methods $i = 1,...,B$, artificial or synthetic data sets D_i^* are generated from one's fitted model using the maximum likelihood estimates $\hat{\theta}$, $\hat{\pi}_l$, and, if appropriate, $\hat{\pi}_c$ of the parameters as generating parameters for the synthetic data sets. For each synthetic dataset D_i^*, we obtain another set of maximum likelihood estimates $[\hat{\theta}^*, \hat{\pi}_l^*, \hat{\pi}_c^*]_i$ and thus a bootstrap distribution of $i = 1,...,B$ values for each parameter. From these distributions we obtain *bias-corrected and accelerated* confidence intervals (Efron, 1987), currently considered state of the art in bootstrap techniques. (For the bootstrap experiments, we used the `psignifit` software implementation of the method sketched above and described in more detail by Wichmann and Hill, 2001b.)

In the experiments we varied the number of trials N and the lapse parameter π_l. The Gumbel function as described above with $m = 5$ and $w = 3$ and same sample locations were used. The data set size N takes the values 90, 360, and 900 and lapse rate is set to either 0.05 or 0.15. For each of the six conditions, we generated 1000 data sets.

Performing a large set of MCMC simulations is computationally demanding, and we cannot inspect each individual chain. We used one set of parameters for hybrid MCMC sampling for each of the six conditions and later removed those rare runs in which the acceptance rate was lower than 50%. As above we used priors $p(m) = N(2,10)$ and $p(w) = \log N(1,1)$.

For data sets generated with a lapse rate $\pi_l = 0.05$, we used a Beta(2,50) prior for π_l. To make similar information available to the bootstrap method, we constrained the π_l parameter to [0,0.1] during maximum likelihood estimation. For data sets generated with $\pi_l = 0.15$, we used Beta(2,20) and [0,0.25]. The use of different prior distributions is necessary because either method—Bayesian inference and bootstrapping—is highly sensitive to the prior information on π_l. This is not a problem, however, because it appears realistic to assume that an experimenter is to some degree aware of the lapse rate and asymptotic performance of a subject during experiments resulting in different prior beliefs.

First we examine the accuracy of several point estimates for m and w. We compare the MAP estimate, the MCMC sample mean (MEAN) and median (MED), the maximum likelihood (ML) estimate, and the constrained ML (CML) estimate computed by `psignifit`. Each line in the following table states the median of the absolute errors $|m-m^*|$ of these point estimates in 1000 repeated experiments for the different values of N and π_l.

N	π_l	MAP	MEAN	MED	ML	CML
90	0.05	0.301	0.316	**0.289**	0.349	0.336
90	0.15	0.370	0.426	**0.353**	0.418	0.401
360	0.05	0.147	0.141	**0.136**	0.166	0.165
360	0.15	0.232	**0.175**	0.179	0.241	0.230
900	0.05	0.102	**0.088**	0.090	0.109	0.110
900	0.15	0.183	**0.131**	0.139	0.166	0.155

Table 1. Median of absolute errors $|m-m^*|$ for threshold. π_l = lapse probability, MAP = maximum a posteriori estimate, MEAN and MED = mean and median of MCMC samples, ML = maximum likelihood estimate, CML = constraint maximum likelihood estimate.

For estimating m, the sample median MED consistently shows good accuracy. Note that the MED minimizes the expected absolute error so this result conforms to the theory.

N	π_l	MAP	MEAN	MED	ML	CML
90	0.05	**0.906**	1.331	1.024	1.446	1.363
90	0.15	**0.905**	2.113	1.128	1.717	1.655
360	0.05	0.479	0.517	**0.478**	0.629	0.635
360	0.15	**0.559**	0.616	0.574	0.828	0.826
900	0.05	**0.314**	0.334	0.320	0.400	0.401
900	0.15	0.464	0.431	**0.405**	0.502	0.495

Table 2. Median of absolute errors $|w-w^*|$ for width parameter.

The width w is best estimated by the MAP followed by the MED. The errors decrease with sample size and increase for the large lapse rate.

We now compare the reliability of bootstrapped and Bayesian confidence regions. We therefore compare the frequency at which the true generating value was included in the approximated 90% confidence interval. In theory this frequency should become exactly 90% for large numbers of repeated experiments. Larger values correspond to *over-conservative* statements, whereas smaller values indicate *over-confidence*. With the frequency we also report the median width of the computed confidence intervals.

		m		w	
N	π_l	Accuracy	Width	Accuracy	Width
90	0.05	0.783	1.422	0.903	9.049
90	0.15	0.707	1.558	0.911	18.258
360	0.05	0.863	0.757	0.883	3.184
360	0.15	0.795	0.915	0.865	4.148
900	0.05	0.848	0.488	0.859	1.934
900	0.15	0.818	0.682	0.867	2.573

Table 3. Accuracy and width of bootstrap confidence intervals.

		m		w	
N	π_l	Accuracy	Width	Accuracy	Width
90	0.05	0.911	1.765	0.933	8.103
90	0.15	0.922	2.340	0.981	11.882
360	0.05	0.918	0.750	0.931	2.959
360	0.15	0.926	0.884	0.937	3.745
900	0.05	0.919	0.457	0.916	1.694
900	0.15	0.901	0.585	0.916	2.209

Table 4. Accuracy and width of approximate Bayesian confidence intervals.

For the threshold location m, the approximated Bayesian confidence intervals exhibit accuracy close to the desired 90% for all six conditions. The bootstrap confidence intervals appear to be over-confident, especially for small data sets and high lapse rates. For small data sets $N = 90$, the width of the Bayesian confidence regions is larger, whereas for larger data set sizes, the Bayesian confidence intervals exhibit higher accuracy but smaller interval width.

For the w parameter, both the bootstrap and the Bayesian confidence regions were found to be relatively accurate. The Bayesian confidence regions tend to be conservative, especially for $N = 90$ and $\pi_l = 0.15$, whereas the median width over the confidence regions is consistently smaller.

In the presented set of synthetic experiments, the Bayesian MCMC sampling-based estimators were found to give more accurate point estimates and more accurate and tighter confidence regions.

A case study

Data for the case study are taken from a visual contrast-discrimination task published by Henning, Bird, and Wichmann (2002). Observers either performed sinusoidal contrast increment detection or detected a contrast increment applied to a pulse train grating; both were two-interval forced-choice tasks. For both conditions the contrast of the added signal was varied using the method of constant stimuli.

The aim of the experiment was to determine whether the two conditions yielded similar or different discrimination thresholds. Both stimuli—the sine wave and the pulse train—have the same fundamental frequency, but the pulse train has additional higher frequency components. Hence, one might expect that these facilitate discrimination and therefore the threshold for the pulse condition might be lower.

First we analyze the data from the sine wave condition. The data come from one of the observers and consist of 13 blocks with 50 trials each. Each block was measured at a different contrast between 0.5% and 7.5%. Using a Weibull function to explain the data is a common choice for contrast experiments. Instead of directly fitting a Weibull, we have found it more convenient to log-transform the contrast and use a Gumbel function. As pointed out before, these two possibilities are equivalent, but the Gumbel function allows a more intuitive parameterization in terms of the width.

Next, we have to specify our prior beliefs about the parameter values. For the lapse rate, a convenient choice for the prior is the beta distribution. We expect from our experience with observers that some of the trials are lapses. A reasonable choice that makes small lapse rates more likely than big lapse rates is $\alpha = 2$ and $\beta = 50$ (see Figure 4a). This prior also expresses our belief that observers are unlikely to perform errorlessly. The mean of the prior distribution is given by $\alpha/(\alpha + \beta) \approx 4\%$ and the mode is $(\alpha - 1)/(\alpha + \beta - 2) = 2\%$. Thinking about the stimulus, we can derive a conservative prior on the threshold location. At 100% contrast, a sine wave can clearly be seen, but at about 10%, the task becomes difficult. This is the range in which we would expect the threshold. At 1% the task seems almost impossible. A reasonable prior on log-contrast threshold therefore has a maximum at −1 (10%). We can take a Gaussian with this mean. Even though −2 corresponding to 1% and 0 corresponding to 100% seem to be unlikely threshold values, we do not want to rule out these hypotheses a priori. Therefore, *SD* 1 seems to be a conservative choice.

A width smaller than zero would correspond to a psychometric function for which performance increases with lower contrasts, so we constrained the prior to positive values. A width of 2 log-units is highly unlikely because it would mean that the psychometric function potentially ranges from 1% to 100% contrast. Therefore, a width between 0.5 and 1.5 log-units seems to be a reasonable range. For positive parameters, the gamma distribution is a common choice for the prior. By plotting a gamma distribution for $\alpha = 2$ and $\beta = 1.5$, we found it to be a good description of our beliefs. The mean is given by $\alpha/\beta = 1.33$ and the mode by $(\alpha - 1)/\beta = 0.66$. The standard deviation that is given by $\sqrt{\alpha/\beta} = 0.94$ is large enough to support values even bigger than 2.

Once we have specified the prior, we can sample from the posterior. For this we have to set values for the number and sizes of the leap-frog steps; this requires some "hands-on" experience with MCMC but is not particularly difficult. We come back to this issue at the end of the article.

Figure 7 presents histograms of the posterior samples generated by hybrid MCMC sampling. First, we inspect the posterior samples of the threshold parameter. The MAP, ML, and MEAN point estimate for the threshold all lie between −1.7 and −1.65 log-units (i.e., at a contrast between 2% and 2.2%). Furthermore, we compute the approximate Bayesian 90% confidence region from empirical quantiles of the samples. The Bayesian confidence interval ranges from −1.74 to −1.59 log-units or 1.8% and 2.6% contrast (the outermost dashed lines in Figure 7). Information like this is necessary if one wants to compare thresholds from different conditions.

When comparing priors and posteriors, we observe that the priors for the threshold and the width parameter were flat relative to the posterior distributions. The posterior of the lapse rate parameter is similar to the prior, which shows that the data did not allow a reduction in uncertainty about this parameter. Figure 7 also depicts the maximum likelihood (ML) and the maximum a posteriori (MAP) estimates. The difference between the two indicates how much the prior has influenced the MAP estimate. The difference for the lapse rate is substantial, which again emphasizes the importance of the prior on this parameter. The ML estimate suggests that far more than 10% of the observations were lapses, which also explains why the width is estimated to be relatively small.

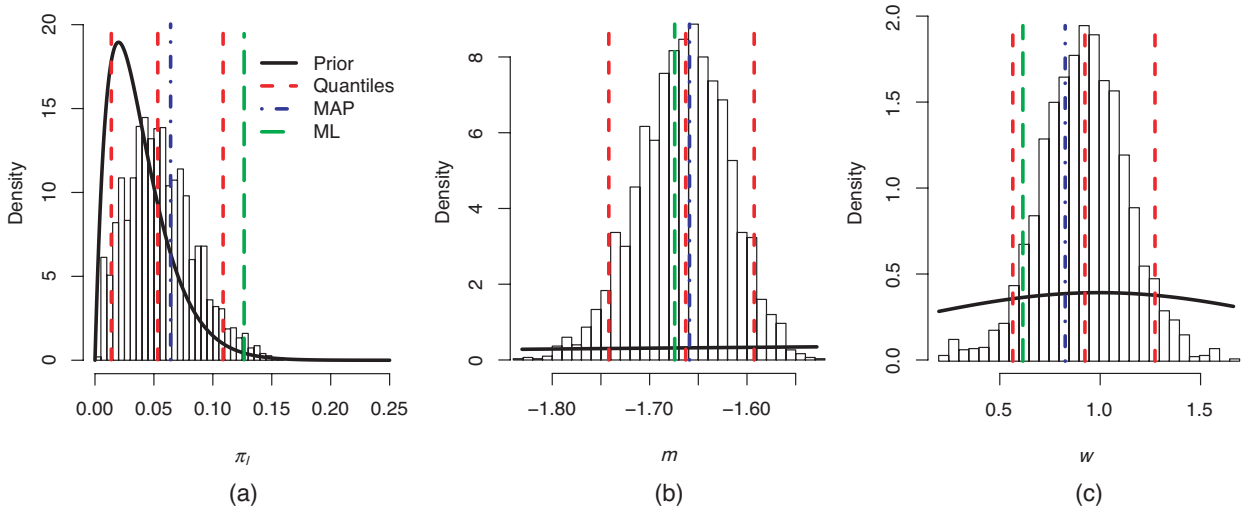

Figure 7. Sine condition. The estimated posterior distributions for the lapse parameter (a), the threshold (b), and the width (c) of the psychometric function. Vertical lines depict MAP estimates, ML estimates, and quantiles at 5%, 50%, and 95%. The solid black line shows the prior distribution. For the threshold and width, the prior is relatively flat compared to the posterior.

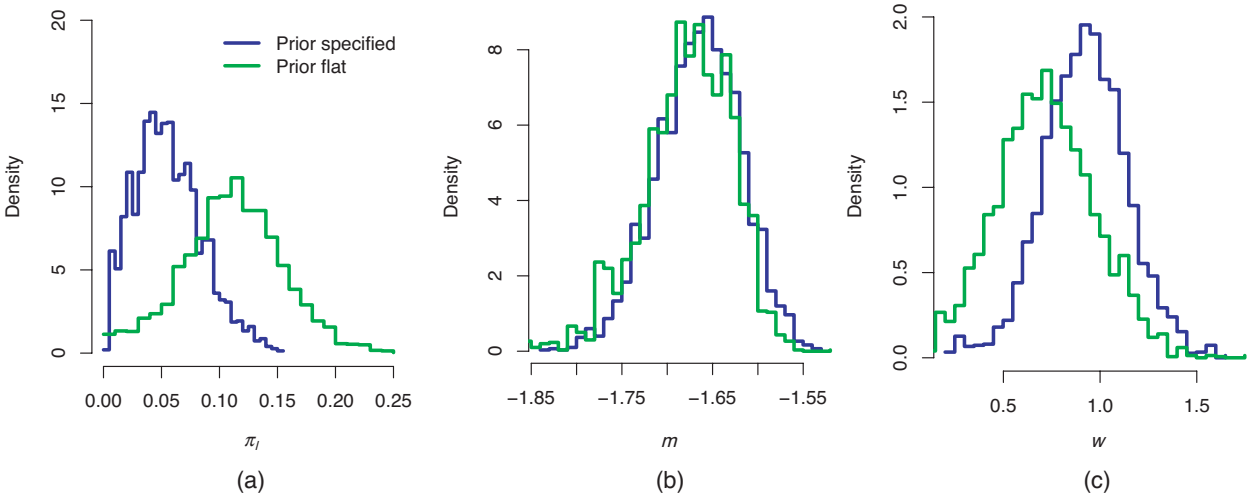

Figure 8. Sine condition. The estimated posterior distributions for the lapse parameter (a), the threshold (b), and the width (c) of the psychometric function. The posterior was computed with the same prior as in Figure 7 and for a flat prior for comparison. For the threshold parameter, there is hardly any difference between the two posteriors, which comes as no surprise because the Gaussian prior is almost flat in the relevant region. For the lapse parameter, the influence of the prior is substantial, as it is for the width.

An important question when using Bayesian methods in practice is the sensitivity of the results with respect to changes in the prior. For comparison, we repeat the sampling, this time using flat priors on all parameters. Note again that a flat prior is not uninformative. Figure 8 compares the posterior distributions that result from using either flat priors (on log-contrast) or the priors as specified above. It reveals that the choice of prior matters for the lapse parameter and the width, but in this case not so much for the threshold. Figure 9a shows the data and three point estimates of the psychometric function for the sine condition.

The second condition in the experiment was a pulse train instead of a sine wave discrimination task. The fundamental frequency of the pulse train was identical to the frequency of the sine wave. Because the pulse train has additional higher frequency components, one may expect that these facilitate discrimination. For the second condition, the data consist of 11 blocks with 50 trials each. The stimuli varied between 10% and 1% contrast. We used the same priors that we used for the first condition. The data can be seen in Figure 9b along with various estimates. The psychometric functions for the pulse and the sine condition look similar. Figure 10 compares the posteriors of both conditions, confirming that the approximated posterior distributions over the parameters are highly overlapping. In particular, the posterior distributions of the thresholds overlap. When comparing the psychometric functions, one

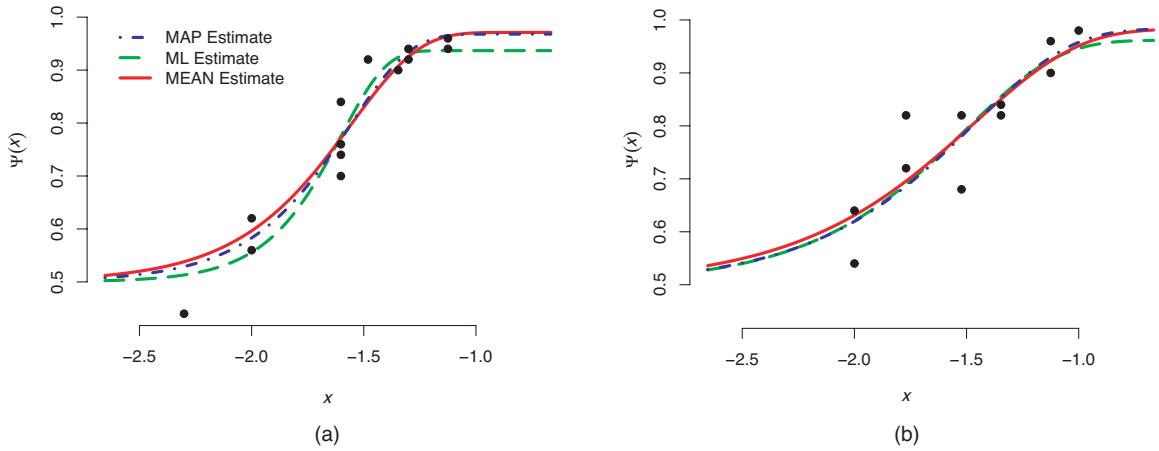

Figure 9. Point estimates for both conditions. (a) shows data from the sine condition and three estimates of the psychometric function (MAP, ML, and MEAN). Each data point represents 50 trials. In (b) we show the same for the pulse condition. For both conditions the three estimates are very similar.

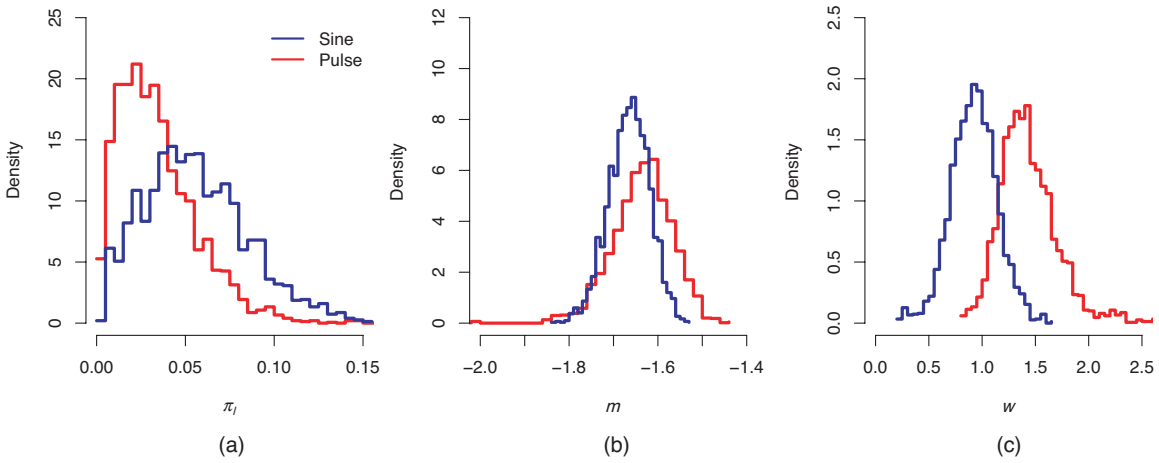

Figure 10. A comparison of the sine and the pulse condition. (a) shows the posterior distributions for the lapse parameter. In (b) the posteriors for the thresholds are similar. The uncertainty is a bit larger in the pulse condition (c). It cannot be concluded that the psychometric functions have a different width. The respective posteriors are not different enough. Usually, the lapse parameter and the width show a negative correlation. Hence, they need to be compared carefully. The smaller width of the sine condition goes along with a higher lapse rate.

might suspect a difference in the width. The posteriors over the width parameters are shown in the right panel of Figure 10. It has to be stressed that the lapse parameter and the width cannot be interpreted independently of each other, because a negative correlation between the two parameters can be seen in the MCMC samples. Intuitively, a higher lapse rate squeezes Ψ down, which is compensated by a smaller width.

In conclusion, based on the data, we do not find evidence that pulse trains lead to a discrimination performance different from sine waves. The observation that there is no substantial difference between the conditions agrees with the conclusion that Henning et al. (2002) reached based on bootstrapping. As always, more data may unearth a small difference between the conditions, which, in the context of the study by Henning et al. (2002), would not affect the conclusions drawn. Note that this is true despite the substantial amount of data (>550 trials per psychometric function) collected. In addition, inspection of the posterior distributions gives an indication about how large differences between conditions would have to be to allow them to be statistically distinguished.

Conclusions

We presented a Bayesian approach to inference about the parameters of psychometric functions. Because computing the density of the posterior distribution is analytically intractable, we described how Markov chain Monte Carlo techniques can be used instead to generate samples from the posterior.

We illustrated that the proposed Bayesian method can produce more accurate point estimates and confidence intervals than the popular frequentist bootstrap technique. Although we cannot prove that this observation generalizes to all possible data situations, there is no reason to believe that the Bayesian approach should do worse on other datasets. Furthermore, the Bayesian approach exhibits several conceptual advantages. Yet another advantage is that by inspecting the MCMC samples and observing correlations and dependencies, we gain a deeper understanding of the process at hand.

We discussed the role of prior distributions in the analysis of experimental data and the difficulties of avoiding informative priors. We observed that the prior on the lapse parameter can be highly influential. For θ we found that even relatively small data sets are often informative enough to overrule the prior. However, a Bayesian analysis should always report prior and posterior distributions, and the latter should always be interpreted relative to the prior given the model.

A difficulty of the proposed method is that using Markov chain Monte Carlo methods is nontrivial and requires the Markov chains to be inspected and parameters to be set by the user. In practice, the parameters are found in a trial-and-error procedure. In a nutshell, the step-size parameter of the hybrid sampling scheme is adapted such that the acceptance rate is between 60% and 90% and the autocorrelation between samples is small. We usually fix the number of leapfrog-steps to 100. As a companion to this study, we released a software implementation package named `PsychoFun` for the (free) R environment for statistical computing. The `PsychoFun` package can be obtained from the authors' websites together with a technical report (Kuss, Jäkel, & Wichmann, 2005) describing details of the implementation and usage. The report also contains more guidance for setting the parameters of the hybrid Markov chain Monte Carlo sampling scheme and informally describes how to inspect the simulated chains. In addition, it contains the code for reproducing the experiments presented in this article.

Acknowledgments

We would like to thank Jeremy Hill and Carl Edward Rasmussen for helpful comments and discussion as well as two anonymous reviewers for helping us to improve our manuscript. MK was supported by the German Research Council (DFG) Grant RA 1030/1.

Commercial relationships: none.
Corresponding author: Malte Kuss.
Email: malte.kuss@tuebingen.mpg.de.
Address: Max Planck Institute for Biological Cybernetics, Tübingen, Germany.

References

Alcalá-Quintana, R., & Garcia-Pérez, M. A. (2004). The role of parametric assumptions in adaptive Bayesian estimation. *Psychological Methods, 9*(2), 250-271. [PubMed][Abstract]

Blackwell, H. R. (1952). Studies of psychophysical methods for measuring visual thresholds. *Journal of the Optical Society of America, 42*, 606-616. [PubMed]

DeGroot, M. H., & Schervish, M. J. (2002). *Probability and statistics* (3rd ed.). Boston: Addison-Wesley.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association, 82*, 171-200.

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

Finney, D. J. (1971). *Probit analysis* (3rd ed.). Cambridge: Cambridge University Press.

Foster, D. H., & Bischof, W. F. (1991). Thresholds from psychometric functions: Superiority of bootstrap to incremental and probit variance estimators. *Psychological Bulletin, 109*, 152-159.

Foster, D. H., & Bischof, W. F. (1997). Bootstrap estimates of the statistical accuracy of thresholds obtained from psychometric functions. *Spatial Vision, 11*(1), 135-139.

Garcia-Perez, M. A. (1998). Forced-choice staircases with fixed step sizes: Asymptotic and small-sample properties. *Vision Research, 38*(12), 1861-1881. [PubMed]

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (Eds.). (1996). *Markov chain Monte Carlo in practice*. London: Chapman & Hall.

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

Henning, G. B., Bird, C. M., & Wichmann, F. A. (2002). Contrast discrimination with pulse trains in pink noise. *Journal of the Optical Society of America A, 19*(7), 1259-1266. [PubMed]

Jaynes, E. T. (2003). *Probability theory*. Cambridge: Cambridge University Press.

Kaernbach, C. (2001). Slope bias of psychometric functions derived from adaptive data. *Perception & Psychophysics, 63*(8), 1389-1398. [PubMed][Article]

King-Smith, P. E., & Rose, D. (1997). Principles of an adaptive method for measuring the slope of the psychometric function. *Vision Research, 37*(12), 1595-1604. [PubMed]

Klein, S. A. (2001). Measuring, estimating, and understanding the psychometric function: A commentary. *Perception & Psychophysics, 63*(8), 1421-1455. [PubMed] [Article]

Kontsevich, L. L., & Tyler, C. W. (1999). Bayesian adaptive estimation of psychometric slope and threshold. *Vision Research, 39*(16), 2729-2737. [PubMed]

Kuss, M., Jäkel, F., & Wichmann, F. A. (2005). *Approximate Bayesian inference for psychometric functions using* MCMC *sampling (135)*. Tübingen, Germany: Max Plank Institute for Biological Cybernetics. [Report]

MacKay, D. J. C. (1999). Introduction to Monte Carlo methods. In M. I. Jordan (Ed.), *Learning in graphical models* (pp. 175-204). Cambridge, MA: MIT Press.

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

Madigan, R., & Williams, D. R. (1987). Maximum-likelihood psychometric procedures in two-alternative forced-choice: Evaluation and recommendations. *Perception & Psychophysics, 42*(3), 240-249. [PubMed]

Maloney, L. T. (1990). Confidence intervals for the parameters of psychometric functions. *Perception & Psychophysics, 47*, 127-134. [PubMed]

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). Boca Raton: Chapman & Hall.

McKee, S. P., Klein, S. A., & Teller, D. Y. (1985). Statistical properties of forced-choice psychometric functions: Implications of probit analysis. *Perception & Psychophysics, 37*(4), 286-298. [PubMed]

Miller, J., & Ulrich, R. (2001). On the analysis of psychometric functions: The Spearman-Kärber method. *Perception & Psychophysics, 63*(8), 1399-1420. [PubMed] [Article]

Neal, R. M. (1993). *Probabilistic inference using Markov chain Monte Carlo methods* (Technical Report CRG-TR-93-1). Department of Computer Science, University of Toronto. [Report]

O'Hagan, A. (1994). *Bayesian inference*. London: Arnold.

O'Regan, J. K., & Humbert, R. (1989). Estimating psychometric functions in forced-choice situations: Significant biases found in threshold and slope estimation when small samples are used. *Perception & Psychophysics, 45*, 434-442. [PubMed]

Pelli, D. G. (1987). The ideal psychometric procedure. *Investigative Ophthalmology and Visual Science* (Suppl.), *28*, 366.

Pentland, A. (1980). Maximum likelihood estimation: The best PEST. *Perception & Psychophysics, 28*(4), 377-379. [PubMed]

Rose, R. M., Teller, D. Y., & Rendleman, P. (1970). Statistical properties of staircase estimates. *Perception & Psychophysics, 8*(4), 199-204.

Snoeren, P. R., & Puts, M. J. H. (1997). Multiple parameter estimation in an adaptive psychometric method: MUEST, an extension of the QUEST method. *Journal of Mathematical Psychology, 41*(4), 431-439. [PubMed]

Taylor, M. M. (1971). On the efficiency of psychophysical measurement. *Journal of the Acoustical Society of America, 49*(2), 505-508. [PubMed]

Treutwein, B. (1995). Adaptive psychophysical procedures. *Vision Research, 35*(17), 2503-2522. [PubMed]

Treutwein, B., & Strasburger, H. (1999). Fitting the psychometric function. *Perception & Psychophysics, 61*(1), 87-106. [PubMed]

Watson, A. B., & Pelli, D. G. (1983). QUEST: A Bayesian adaptive psychometric method. *Perception & Psychophysics, 33*(2), 113-120. [PubMed]

Watt, R. J., & Andrews, D. P. (1981). APE: Adaptive probit estimation of psychometric functions. *Current Psychological Reviews, 1*(2), 205-213.

Wetherill, G. B., & Levitt, H. (1965). Sequential estimation of points on a psychometric function. *British Journal of Mathematical and Statistical Psychology, 18*(1), 1-10. [PubMed]

Wichmann, F. A. (1999). *Some aspects of modelling human spatial vision: Contrast discrimination*. Unpublished doctoral dissertation, Oxford University.

Wichmann, F. A., & Hill, N. J. (2001a). The psychometric function. I. Fitting, sampling and goodness-of-fit. *Perception & Psychophysics, 63*(8), 1293-1313. [PubMed][Article]

Wichmann, F. A., & Hill, N. J. (2001b). The psychometric function. II. Bootstrap-based confidence intervals and sampling. *Perception & Psychophysics, 63*(8), 1314-1329. [PubMed][Article]