Vision Research 122 (2016) 105–123

Contents lists available at ScienceDirect

Vision Research

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

Painfree and accurate Bayesian estimation of psychometric functions for (potentially) overdispersed data

Heiko H. Schütt [a,b,c,*], Stefan Harmeling [d], Jakob H. Macke [e,f,g], Felix A. Wichmann [a,g,h,*]

[a] Neural Information Processing Group, University of Tübingen, Tübingen, Germany
[b] Department of Psychology, Universität of Potsdam, Potsdam, Germany
[c] Graduate School for Neural and Behavioural Sciences IMPRS, Tübingen, Germany
[d] Institut für Informatik, Heinrich-Heine-Universität Düsseldorf, Germany
[e] Research Center Caesar, Bonn, Germany
[f] Max Planck Institute for Biological Cybernetics, Tübingen, Germany
[g] Bernstein Center for Computational Neuroscience, Tübingen, Germany
[h] Max Planck Institute for Intelligent Systems, Tübingen, Germany

ARTICLE INFO

Article history:
Received 28 June 2015
Received in revised form 22 February 2016
Accepted 27 February 2016

Keywords:
Psychometric function
Bayesian inference
Beta-binomial model
Overdispersion
Non-stationarity
Confidence intervals
Credible intervals
Psychophysical methods

ABSTRACT

The *psychometric function* describes how an experimental variable, such as stimulus strength, influences the behaviour of an observer. Estimation of psychometric functions from experimental data plays a central role in fields such as psychophysics, experimental psychology and in the behavioural neurosciences. Experimental data may exhibit substantial overdispersion, which may result from non-stationarity in the behaviour of observers. Here we extend the standard binomial model which is typically used for psychometric function estimation to a *beta-binomial* model. We show that the use of the beta-binomial model makes it possible to determine accurate credible intervals even in data which exhibit substantial overdispersion. This goes beyond classical measures for overdispersion—goodness-of-fit—which can detect overdispersion but provide no method to do correct inference for overdispersed data. We use Bayesian inference methods for estimating the posterior distribution of the parameters of the psychometric function. Unlike previous Bayesian psychometric inference methods our software implementation—*psignifit 4*—performs numerical integration of the posterior within automatically determined bounds. This avoids the use of Markov chain Monte Carlo (MCMC) methods typically requiring expert knowledge. Extensive numerical tests show the validity of the approach and we discuss implications of overdispersion for experimental design. A comprehensive MATLAB toolbox implementing the method is freely available; a python implementation providing the basic capabilities is also available.

1. Introduction

In psychophysics, experimental psychology and the behavioural neurosciences, researchers attempt to measure detection or discrimination behaviour as a function of stimulus level, i.e. some changeable aspect of the stimulus or experimental setup controlled by the researcher. The range of applications is vast, from simple detection of spots of lights or Gabor patches to categorical perception of faces in experimental psychology and from discrimination performance of a single neuron up to the behaving animal in neuroscience. After data collection, researchers frequently fit a psychometric function to their data—almost always an appropriately scaled cumulative probability density function—relating the independent variable on the abscissa to the observer's behaviour on the ordinate. Researchers then obtain the "threshold" and, sometimes, the slope from the estimated psychometric function. Detection or discrimination behaviour, or performance, is thus summarised using one or two values, namely the threshold and the slope.

Thus fitting the psychometric functions to experimental data is of central importance for many fields. Given this importance, much research was conducted to either investigate the efficiency and reliability of the data collection (e.g. Blackwell, 1952; Watson & Pelli, 1983; Green, 1990; Treutwein, 1995; García-Pérez, 1998; Kontsevich & Tyler, 1999; Jäkel & Wichmann, 2006; Shen & Richards, 2012) or how to obtain accurate estimates of the psychometric function parameters (e.g. O'Regan & Humbert, 1989; Treutwein & Strasburger, 1999; Wichmann & Hill, 2001a; Knoblauch & Maloney, 2012).

* Corresponding authors at: Neural Information Processing Group, University of Tübingen, Tübingen, Germany.
E-mail addresses: heiko.schuett@uni-tuebingen.de (H.H. Schütt), felix.wichmann@uni-tuebingen.de (F.A. Wichmann).

http://dx.doi.org/10.1016/j.visres.2016.02.002
0042-6989/

However, unless one has collected infinitely many trials per psychometric function, the parameters of the psychometric function are not fully constrained by the data and there remains uncertainty regarding the estimated parameters. To be able to draw valid conclusions when comparing thresholds and slopes from different experimental conditions, it is essential that this uncertainty is quantified. Typically, the uncertainty is expressed in the form of confidence intervals around the point estimates. Unfortunately, a reliable and accurate characterisation of this uncertainty is more difficult to obtain than the estimates themselves, partly due to the small size of typical datasets collected during behavioural experiments.[1]

The bootstrap (Efron, 1979; Efron & Tibshirani, 1994) was the first numerical sampling method applied to psychophysical data in order to characterise the uncertainty of the point estimates, i.e. to obtain confidence intervals (Foster & Bischof, 1987; Maloney, 1990; Foster & Bischof, 1991, 1997; Wichmann & Hill, 2001b).[2] Hill (2002) showed, however, that bootstrapped confidence intervals in the context of psychometric function estimation can be too small, a result confirmed by both Kuss et al. (2005) and Fründ, Haenel, and Wichmann (2011).[3]

As an alternative to the bootstrap, Bayesian statistics,[4] is centred on the notion of how to quantify uncertainty, and thus Bayesian statistics, too, offers a suitable theoretical framework to analyse data obtained in psychophysics, experimental psychology and the behavioural neurosciences. Bayesian statistics is, furthermore, especially suited for the small datasets (sample sizes) typically gathered in behavioural experiments. Kuss et al. (2005) provide a detailed and tutorial-style introduction to Bayesian inference for psychometric functions, and show results from numerical simulations suggesting that credible intervals obtained from Bayesian inference are more accurate than those obtained using the bootstrap. Similar results were later obtained by Fründ et al. (2011).

Bayesian inference for psychometric functions cannot be performed analytically, and instead has to rely on numerical methods to obtain the posterior distribution of the parameters given the data. Both Kuss et al. (2005) and Fründ et al. (2011) use Markov chain Monte Carlo (MCMC) methods to generate samples from the posterior distribution over parameters. MCMC is a standard method in Bayesian inference in general, and, in principle, allows Bayesian inference to be performed on many statistical problems. Unfortunately MCMC requires considerable statistical expertise from the user to fine tune the proposal distribution and the sampling step size, and especially to detect when the sampling fails. MCMC methods thus rarely work "automatically" with no or little user intervention the way analytical methods and the bootstrap do. Kuss et al. aptly summarise the problem in their paper: *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.* Kuss et al. (2005, p. 491). For many researchers in psychophysics, experimental psychology and the behavioural neurosciences this difficulty precludes the use of the MCMC-based Bayesian methods introduced by Kuss et al. (2005) and Fründ et al. (2011), and they still have to rely on the easier to use, albeit less accurate, bootstrap-based methods, e.g. the psignifit 2.5 toolbox by Wichmann and Hill (2001a, 2001b).

Finally, there is one more hurdle for inferring the uncertainty about the psychometric function parameters: overdispersion. Overdispersion means that the variance of the measured data is larger than expected from the binomial model, which may happen due to fluctuations in attention, vigilance, criteria or unmodelled aspects of the stimulus. Consequently all estimates of the uncertainty based on the binomial model become too small, whether based on Bayesian or on frequentist statistics if the data are overdispersed. To prevent this, early approaches used goodness-of-fit measures like deviance to detect overdispersion but could only suggest to reject overdispersed datasets (Wichmann & Hill, 2001b). Later Fründ et al. (2011) presented a method to perform a post hoc corrections of error bars for overdispersed datasets. However there has been no method which directly incorporated overdisperion in psychometric function fitting, despite the fact that the beta-binomial model for overdispersed binomial data has been well established for many years (Williams, 1982; McCullagh & Nelder, 1989, chap. 4.5, exercise 4.17; also see Venables & Ripley, 2013, chap. 7.5).

1.1. Contributions of this paper

The contributions of the current paper are fourfold:

1. We extend psychometric function modelling from the standard binomial to a beta-binomial model to capture overdispersion. We show that this model not only allows statistical inference from overdispersed data from a beta-binomial observer, but yields reasonable results for other sources of overdispersion, e.g. stemming from several types of serial dependencies (Sections 3.1 and 3.2).
2. We show that fitting a beta-binomial model provides a way of detecting overdispersion consistent with goodness-of-fit measures. In contrast with these approaches which can merely reject overdispersed data, this method allows valid statistical inference even for overdispersed data.[5]
3. We introduce a pain-free method for Bayesian inference for psychometric functions. First, we compute the posterior distribution of the parameters using numerical integration without the need for MCMC sampling techniques and any user intervention. Second, we suggest default priors and parameters for the Bayesian inference which in our simulations and experience yield good results, again without user intervention.[6] Third, we provide an implementation of the method, *psignifit 4*, coded in pure MATLAB[7] without dependencies on external code (such as mex-files) or other toolboxes, which eases the installation for the user, and helps the platform-independence.

[1] Much of conventional statistics relies on the asymptotic behaviour of estimators and probability distributions, i.e. relies on—ultimately infinitely—large datasets. Wichmann and Hill (2001a, 2001b) showed that, for the typical size of psychophysical datasets, methods based on asymptotic theory are not always reliable.

[2] Note that the "confidence intervals" estimated in the frequentist statistical framework, e.g. via the bootstrap, are not the same as the "credible intervals" obtained from Bayesian statistics. For a discussion of this difference in the context of psychometric functions see Kuss, Jäkel, and Wichmann (2005), p. 480–481. We always calculate credible intervals in what follows. For readers unfamiliar with the distinction, Bayesian credible intervals are what most people intuit when they think about confidence intervals, whereas the frequentist confidence intervals do not provide this Hoekstra, Morey, Rouder, and Wagenmakers (2014).

[3] Knoblauch and Maloney (2012) provide a comprehensive and clearly presented different approach to psychometric function estimation using the well-established framework of generalised linear models (GLMs). Their GLM approach benefits from a broad array of existing tests, confidence intervals, and software implementations. However, fitting asymptotes requires an alternation between the fitting of the GLM and fitting the asymptotes and the methods available to calculate confidence intervals are either based on asymptotic distributions for the parameter estimates or on bootstrapping. Thus the GLM approach provides no alternative approach for uncertainty assessment and, thus, no principled treatment of overdispersion.

[4] Detailed treatments of Bayesian statistics be found in many available textbooks, for example in O'Hagan (1994), Gelman et al. (2013), Jaynes (2003), and Kruschke (2014).

[5] Assuming, of course, that the data are reasonably well modelled using a sigmoidal function.

[6] Expert users can, of course, override any of the default choices in the software implementation, see the third sense of being pain-free.

[7] Similarly the python implementation does not require the user to compile code or link-in compiled binaries; furthermore, it does not require the user to install "exotic" packages.

4. We show that the proposed method yields essentially unbiased point estimates of the width and the threshold of the psychometric function[8] as well as an accurate estimate of the residual uncertainty, i.e. credible intervals with appropriate coverage for both the standard binomial observer (supplementary material S.1–S.3) and overdispersed observers (Sections 3.1 and 3.2).

2. Models, priors, methods and notation

2.1. The standard binomial model for psychophysical data

The standard observer model for psychometric functions assumes that each trial is a Bernoulli-trial with the success probability given by the value of the psychometric function independent of all other trials. As a consequence, the number of success in a block of fixed stimulus level is binomially distributed. Note that we here refer to any collection of trials with a fixed stimulus level as a block, whether they were collected en bloc or not.

As it is common practice, we model the psychometric function ψ as a scaled sigmoid function S, which is a strictly monotonic function from the stimulus level x on the real line to the unit interval $[0,1]$ approaching 0 for small x and 1 for large x. To specify the shape of this sigmoid function we choose a family of sigmoid functions $S(x;m,w)$, with parameters m, the threshold, and w, the width. The threshold m is the level at which the *unscaled* sigmoid function has value 0.5, i.e. $S(x=m,m,w)=0.5$ and the width w is the difference between the levels at which the function reaches 0.05 and 0.95. These parameters are illustrated in Fig. 1A, and mathematical details are provided in Appendix A.

Equivalently the sigmoid S can be parametrised by a shift parameter α and parameter β that describes the slope. However, in this paper we follow Alcala-Quintana and Garcia-Perez (2004) and Kuss et al. (2005) who proposed the width-parametrization for the following reasons[9]: First, the threshold and the width have an easily accessible meaning and are measured in the same units—particularly the width is easier to interpret and understand than the slope parameter being directly, or even inversely, proportional to the slope of S. Second, the parameterisation in terms of threshold m and width w has the same meaning for all sigmoid families and thus allows easy comparison between data fitted with any sigmoid. Third, specifying an appropriate prior is much more difficult for the parameter β than for the width w—we refer the interested reader to Kuss et al. (2005) on pp. 484–485 and their Fig. 3. Fourth, because threshold m and width w have the same meaning independent of the choice of sigmoid S, we can choose a single set of priors independent of the choice of S.[10]

The toolbox provides several sigmoid functions. Additional functions can be easily added by the user.[11] Currently the following seven functions are implemented:

1. The *cumulative normal* function.
2. The *logistic* function.
3. The *cumulative Gumbel* distribution.
4. Its reversed form, the *reversed cumulative Gumbel*.
5. The *cumulative* t_1*-distribution* with one degree of freedom.
6. The *Weibull* function.
7. The *log-normal* function.

Numbers 1 and 2 are the most common psychometric functions and have similar, symmetric shapes; numbers 3 and 4 are asymmetric, i.e. approach one asymptote faster than the other; number 5 is a strongly heavy tailed distribution, i.e. approaches both asymptotes slower; numbers 6 and 7 are widely used for stimuli on a logarithmic scale.[12] These families are illustrated in Fig. 1B and their formulae are listed in Table A.1.

Using the chosen sigmoid family, the psychometric function ψ is defined with two additional parameters λ and γ for the upper and lower asymptote, scaling the sigmoid function:

$$\psi(x;m,w,\lambda,\gamma)=\gamma+(1-\lambda-\gamma)S(x;m,w) \tag{1}$$

In this parametrization λ denotes the probability of an incorrect answer at infinitely high stimulus levels; γ denotes the probability of a correct answer for infinitely low stimulus levels. These two parameters can be fixed or estimated from the data, depending on the experimental design (see Section 2.3 for details). For an ideal observer λ would always be zero. In practice however, λ should not be fixed at zero as Wichmann and Hill (2001a) showed that the estimates for width and threshold can be severely biased if the upper asymptote is not allowed to be lowered in response to stimulus-independent lapses (errors unrelated to the stimulus level).[13] This rescaling changes the percent correct at the threshold and the points which define the width, which thus do not have percent corrects of 5%, 50% and 95% necessarily.

2.2. The beta-binomial model for overdispersed data

The standard binomial observer model assumes that each trial is a Bernoulli-trial with a fixed success probability independent of all other trials. This assumption is an idealisation, however, and it is typically not exactly correct. It is well known that observers show fluctuations in performance due to fatigue or changes in their attentional state on longer time-scales, and that trials in psychophysics are not independent of each other on short time-scale either (for a recent overview and statistical analysis see Fründ, Wichmann, & Macke, 2014). Typically, fluctuations in performance and serial dependencies result in data with variances larger than the variance of the binomial distribution: Data are *overdispersed*. Thus the standard binomial model provides only a *lower bound* for the variance of the actual variance inherent in behavioural data. Hence the credible intervals derived from this model might be too narrow. Fründ et al. (2011) simulated two different overdispersed observers resulting from learning or performance fluctuations respectively, and showed confidence and credible intervals derived from the standard binomial model to be too small indeed. Analyses of real-world data confirmed both types of overdispersion to be present.[14]

[8] Meaning the bias is small compared to the credible intervals. Fully unbiased statistical estimators are rare and simulations cannot exclude small biases.

[9] If, for whatever reason, the researcher prefers α and β, it is easy to convert one set of parameters to the other, see Appendix A.1. *Psignifit 4* can convert m and w to α and β for any of the supplied sigmoid families.

[10] The equivalent prior on the individual standard parametrisation differs between different sigmoids and can be complicated depending on the mapping between parameters.

[11] *Psignifit 4* is explicitly coded to be extendible; if a user requires a particular sigmoid not supplied by default, they can add it easily, and make this extension accessible to the scientific community via github as explained on https://github.com/wichmann-lab/psignifit.

[12] In extremely rare cases no sigmoid may be a good model for the psychometric function, e.g. for the unusually shaped psychometric functions reported by Henning, Millar, and Hill (2000) exhibiting a plateau region. In such cases a non-parametric fit may be the better choice as advocated by Żychaluk and Foster (2009), who propose a locally weighted linear fit with a link function to map the prediction to the $[0,1]$ interval

[13] Note that the interpretation of a scaled sigmoid function is mathematically equivalent to a binomial mixture model: In the binomial mixture model the proportion correct answers of the observer results from two independent Bernoulli processes, one of (stimulus independent) guessing, and one being governed by the experimental design and the sensitivity of the observer; proof of the equivalence is provided in A.2.

[14] Fründ et al. (2011) cite preliminary conference reports; the experiments are now published in Maertens and Wichmann (2013) and Schönfelder and Wichmann (2013).

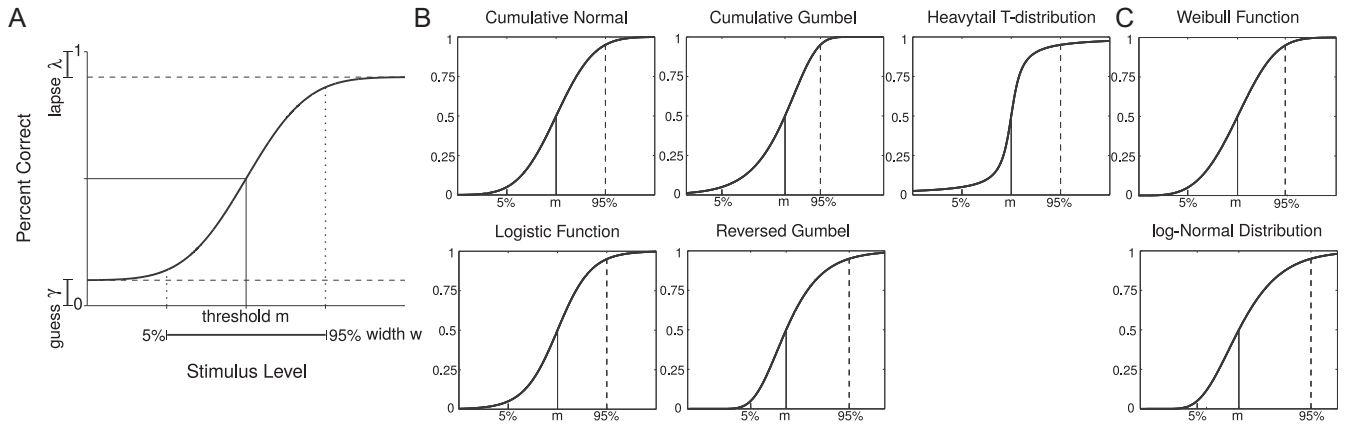

Fig. 1. A. Definition of the parameters: threshold m, the stimulus level at which the unscaled psychometric function reaches 0.5; width w, the difference between the stimulus levels for which the unscaled function reaches 0.05 and 0.95, respectively; the lapse rate λ, the difference between the upper asymptote and 1; the guess rate γ, the difference between the lower asymptote and 0.0. B. The implemented psychometric function families on a linear stimulus level: The cumulative normal, the logistic, the Gumbel and reversed Gumbel (both asymmetric) and a T_1-distribution with $df = 1$ (heavy tail distribution). C. The two implemented psychometric function families for logarithmic scales, for which we define the width and the threshold on the natural logscale: The Weibull and the log-normal.

Many different goodness-of-fit measures exist to measure the amount of overdispersion in data, typically termed lack of fit in this context. The most classical and perhaps most well-known measure is the overdispersion factor i.e. how many times larger the variance of the observations around the psychometric function is compared to the predicted (binomial) variance:

$$\frac{1}{N_b}\sum_{i=1}^{N_b}\frac{\left(\frac{y_i}{n_i}-\psi(x_i)\right)^2}{\frac{1}{n_i}\psi(x_i)(1-\psi(x_i))} \tag{2}$$

where N_b is the number of blocks, x_i is the stimulus level for block i and y_i and n_i are the number of correct trials and the total number of trials respectively.

This lack-of-fit measure converges to 1.0 for binomial data and many large blocks. Some variants of this formula exist depending on the variance normalisation and the weighting of blocks (see Collett (2002) for other goodness of fit measures and their discussion).

However such measures are not recommended for binomial data, as the distribution of binomial data becomes highly asymmetric for expected values near 0 and 1; furthermore the above estimates are biased for small sample sizes (Collett, 2002). Instead, for binomial data, *deviance D* is the recommended goodness-of-fit measure, and it is defined as:

$$D = -2[\log(P(\text{data}|M_{\text{Max}})) - \log(P(\text{data}|\theta))], \tag{3}$$

i.e. as -2 times the difference of the model log-likelihood ($\log(P(\text{data}|\theta))$) to the log-likelihood of the model M_{Max} ($\log(P(\text{data}|M_{\text{Max}}))$) with a free parameter for each block (i.e. fitting the data perfectly). This measure of goodness-of-fit for psychometric functions was explored in detail by Wichmann and Hill (2001a). Deviance converges to 1.0 for a large block from the binomial model and thus to the number of blocks for the whole function.

Fründ et al. (2011) suggested to apply a post hoc fix to improve the coverage of the confidence or credible intervals if data are overdispersed: they increased the size of the estimated intervals using a post hoc correction factor. The correction factor is based on an estimate of the residual dispersion around the most probable psychometric function from a standard binomial fit. This model does not, however, provide principled methods for full Bayesian inference in a beta-binomial model.

Here, we provide methods for full Bayesian inference in a beta-binomial model (Prentice, 1986) in which overdispersion is treated as an additional parameter. The beta-binomial model assumes that the success-probability per block at a constant intensity x is itself a beta-distributed random variable with mean $\psi(x)$. Thus the success-probability is not fixed at $\psi(x)$ as in the standard binomial model but is drawn randomly once for each block. The variance of the success-probability is scaled by a new scale parameter η—ranging from 0 and 1—such that the variance of the success probability equals $\eta^2\psi(x)(1-\psi(x))$.

Thus for the beta-binomial model with scale parameter η the mean percent correct remains $\psi(x)$ as in the standard binomial model, but the variance of percent correct becomes $\left(\eta^2+\frac{1-\eta^2}{N}\right)\psi(x)(1-\psi(x))$ for a block of N trials. For $\eta = 0$ the variance reaches its lower limit given by $\frac{1}{N}\psi(x)(1-\psi(x))$, i.e. the variance of the standard binomial observer. For $\eta = 1$ the variance reaches its maximum given by $\psi(x)(1-\psi(x))$, i.e. independent of N; in the most extreme case increasing the number of trials N does not lead to any reduction in the uncertainty about the location of the mean.

Consequently the variance of percent correct according to the beta-binomial model is $1+(N-1)\eta^2$ times the one of the binomial distribution. Thus the beta-binomial distribution is overdispersed by this factor and measures of overdispersion yield approximately this value.[15]

Thus for $\eta = 0$ the beta-binomial model reduces to the standard binomial model; but for values of $0 < \eta \leqslant 1$ the beta-binomial model becomes progressively more overdispersed. Note that the factor of overdispersion depends on the number of trials per block, as η scales the standard deviation of a distribution which is drawn from once per block. Thus any interpretation of η depends on the number of trials per block; an issue we return to in the discussion in Section 4.2.

The likelihood for the beta-binomial model can still be calculated directly for each observation given the (now) five parameters: $\boldsymbol{\theta} = (m, w, \lambda, \gamma, \eta)$ (see A.1)—thus, from a computational point of view, it results in a simple and straightforward extension of standard binomial model.

Given a formula for the likelihood we can also compute the deviance of the model and the deviance distribution according to the beta-binomial model with different η. In agreement with the variance measures the deviance depends on the number of trials measured in a block. For the binomial model it converges to 1 with

[15] Measures of overdispersion only approximate this value due to the necessarily approximate numerical estimation from data and uncertainty about the true $\psi(x)$.

growing block size. For the beta-binomial model it diverges at different speeds depending on η (see Fig. 2F).

To provide some intuition about the beta-binomial model, we show example data and how the variability and expected deviance changes for different beta-binomial observers (Fig. 2). We show three observers for illustration: The standard binomial observer (black; panel A), a moderately overdispersed beta-binomial observer ($\eta = 0.2$; dark blue; panel B) and a strongly overdispersed beta-binomial observer ($\eta = 0.5$; light blue; panel C). For all three observers a sample dataset with 500 trials in 20 blocks of 25 trials each was generated from the same generating psychometric function ($m = w = 1.0$ and $\lambda = \gamma = 0.02$). The generating psychometric function is shown as the dashed line, the Maximum A Posteriori (MAP) estimates of *psignifit 4* as solid lines in their respective colour. The increasing variance in the data from panel A–B and C is visible as the increased number of data points at "larger" distances from the generating psychometric function.

If a binomial observer has a probability $\overline{p}$ of being correct, then its success probability p is $\overline{p}$ in every block. In the beta-binomial model, however, this is a random variable and only the mean corresponds to $\overline{p}$. On a block-by-block basis p—the success probability—is drawn from a beta-distribution. We plot the probability density functions (PDFs) of these beta-distributions for four proportions correct ($\overline{p} = 0.2$, 0.5, 0.7 and 0.9) and the two beta-binomial observers with $\eta = 0.2$ and 0.5 as introduced in the previous paragraph in panel D and mark the mean $\overline{p}$ with red circles. For the binomial observer this graph would be a delta function at $\overline{p}$ (not shown).

To illustrate how the block variability depends on η and on the number of measured trials we plot the standard deviation of proportion correct in a block of trials against the number of trials measured, normalised to the variance of a single Bernoulli trial, for a number of η's (Panel E). For the standard binomial observer ($\eta = 0$, black curve) the standard deviation decreases proportional to the square-root of N. For beta-binomial observers, however, increasing the number of trials N results in a smaller reduction of the remaining uncertainty and especially the standard deviation does not converge to 0 but to η times the standard deviation of a single trial such that the factor to the binomial model grows for larger blocks.

Finally, we show the expected deviance of data generated by the three model observers against a binomial model. The deviance is shown as a function of the number of trials, and for four different (mean) proportions correct ($\overline{p} = 0.2$, 0.5, 0.7 and 0.9). The dashed line at 1.0 marks the asymptotic value for binomial data, the dotted line a typical cut off of 2.0 for what is often informally regarded as a still "well behaved" dataset. First of all, note that for a single trial the beta-binomial model degenerates to the binomial model (all curves start exactly the same points). Second, note that deviance for binomial data asymptotically converges to 1.0 (by design), but that convergence is a function of $\overline{p}$. For the beta-binomial observers deviance increases with increasing number of trials per block. As standard goodness-of-fit measures are based on detecting overly large deviance this graph illustrates their sensitivity to detect overdispersion in (beta-binomial) data: Stronger overdispersion increases deviance more making it detectable with smaller blocks.

The beta-binomial model implies a specific form of overdispersion, namely that the underlying performance level is constant within each block of data and changes randomly from block to block. For data obtained in a traditional blocked design—i.e. stimuli are grouped into blocks of the same magnitude rather than being randomised, sometimes termed *blocked constant stimulus*—this is a reasonable assumption. If the data were not collected in blocks, however, the situation is not as straightforward. The good news is that we show that the beta-binomial model can still be used to model overdispersion under conditions of stimulus randomisation; we discuss the issue in detail in Section 4.2.

2.3. Prior distributions for the parameters

We use Bayesian inference to obtain parameter estimates, and Bayesian inference requires a prior distribution $P(\boldsymbol{\theta})$ over the five parameters $\boldsymbol{\theta} = (m, w, \lambda, \gamma, \eta)$. The prior distribution should ideally reflect the scientists' degree of belief about the true parameters of the model. The default priors we provide and test in the simulations reported in this article were chosen pragmatically: they yield good inference results without the need of user intervention.[16] For psychometric functions it is reasonable to assume the priors of the five parameters to be independent of each other, i.e. we can factorise $P(\boldsymbol{\theta})$ and define the prior as the product of the marginal priors for each parameter: $P(\boldsymbol{\theta}) = P(m)P(w)P(\lambda)P(\gamma)P(\eta)$.

The numerical values for threshold m and width w can be any real or positive number respectively, depending on the type of experiment and parametrisation of the independent variable chosen by the scientist. Thus there is no universal prior for threshold m and width w that could ever be derived from statistical considerations alone. However, a non-informative (or "flat") prior would not be a good choice either, because the likelihood has an infinite integral over the real line. Thus maximum likelihood estimates of the parameters may diverge for small datasets and a flat prior typically does not capture the scientist's belief about the true parameters: In foveal contrast detection, e.g., one rarely beliefs that contrast thresholds are below 0.001% contrast or above 20% for motivated adult subjects and spatial frequencies around 2–4 c/deg. Hence choosing a prior assigning equal probability to the threshold being at 0.001%, 1.0% and 20% would not be reasonable—at least outside an ophthalmologist's laboratory.[17] To overcome this problem, we derive the priors $P(m)$ and $P(w)$ automatically from the x-levels of the data set to be analysed:

- For the threshold we choose a uniform prior over the range of the data with a cosine fall off to 0 over half the range of the data. This prior expresses the belief that the threshold is anywhere in the range of the tested stimulus levels with equal probability and may be up to 50% of the spread of the data outside the range with decreasing probability (see Fig. 3, first column, top and bottom rows).
- For the width we choose a uniform prior between two times the minimal distance of two tested stimulus levels and the range of the stimulus levels with cosine fall offs to 0 at the minimal difference of two stimulus levels and at 3 times the range of the tested stimulus levels (see Fig. 3, second column, top and bottom rows).

For λ and γ we consider three types of scenarios:

- *N-alternative forced choice* (N-AFC) experiments, when the lower asymptote is known and fixed at $\gamma = 1/N$ and the upper asymptote is free.
- *Equal asymptote* experiments in which the asymptotes are free, but assumed to be equal ($\lambda = \gamma$).
- *YesNo* experiments with two independent free asymptotes.

Whenever the asymptotes are free we chose a beta-distribution with parameters $(1, 10)$ as a prior for both $P(\lambda)$ and $P(\gamma)$. This is a

[16] Clearly, our default priors are not the only possible or accurate ones, nor do we claim that they reflect the belief of every user; users of the software can use other priors they deem appropriate for their problem.

[17] In the section "Prior distributions" by Kuss et al. (2005, p. 483–484), the authors present another argument against the use of flat priors for parameters: The difference between flat priors in model space versus flat priors on the parameters of a model.

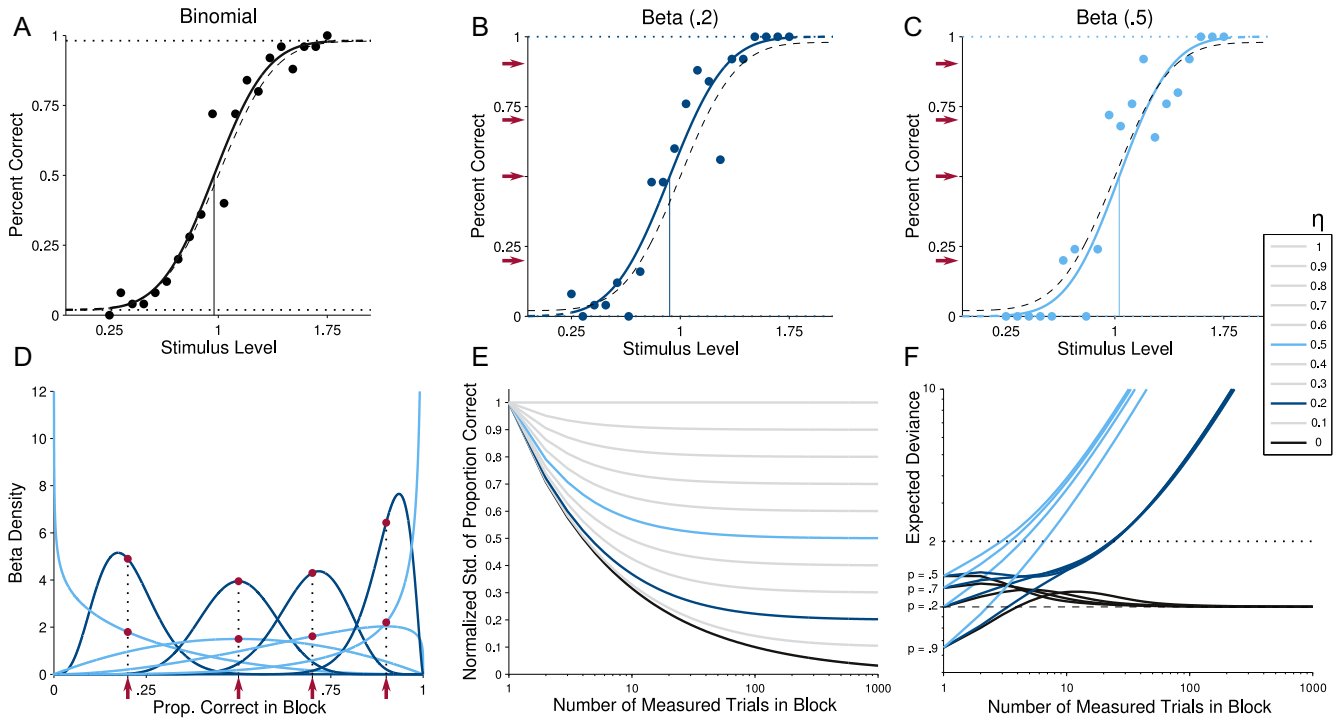

Fig. 2. Illustrations of the beta-binomial model we use to model overdispersion. A–C: example datasets for a binomial and two beta-binomial observers with $\eta = 0.2$ and $\eta = 0.5$ scaling the overdispersion, 500 simulated trials over 20 linearly spaced stimulus levels each, $\lambda = \gamma = 0.02$ and threshold = width = 1.0. The dashed line shows the generating function. D: beta-distribution densities for different means marked by the red arrows and standard deviations corresponding to $\eta = 0.2$ or 0.5. E: standard deviation of the proportion correct in a block of trials against the number of trials measured, normalised to the variance of a single Bernoulli trial. F: expected deviance of binomial data (black; panel A), weakly beta-binomial data (dark blue; panel B) and strongly beta-binomial data (light blue, panel C) against a binomial model; the deviance is shown as a function of the number of trials, and for four different (mean) proportions correct (p = 0.2, 0.5, 0.7 and 0.9). The dashed line at 1 marks the asymptotic value for binomial data, the dotted line a typical cut off around the value 2 for what is often regarded as a still "well behaved" dataset.

prior on the unit interval with a finite peak at 0 and a mean of $\frac{1}{11} \approx 9.09\%$. This prior has the same precision as the posterior after 9 correct trials which test only λ or γ: The most probable lapse or guessing rate is 0, i.e. the PDF peaks at zero and higher rates are progressively less probable. This choice of priors assumes that lapses (and guesses in YesNo) are rare. This prior is shown in Fig. 3, third column, top and bottom rows.

We chose this prior because it yielded relatively unbiased results for threshold and width for a large range of simulated lapse rates. There are other priors with similarly good inference properties, but with a peak away from zero. We chose the prior with the peak at zero because in the absence of evidence for lapsing it induces a (small) bias towards larger widths and smaller lapse rates—we mainly work with highly trained psychophysical observers. In different contexts and for different users a prior with a peak away from zero may well be preferable. The supplementary material S.4 contains a detailed evaluation of the influence of different prior distributions on inference.

For η we again chose to use a beta-distribution with parameters $(1, k)$ and thus a finite peak at 0, expressing the belief that, in the absence of data suggesting otherwise, a standard binomial observer with $\eta = 0$ to be most likely. As there are little data available on the overdispersion observed in typical experiments, we had little intuition about the prior for η. Thus we simulated three different observers: A standard binomial observer ($\eta = 0$), a moderately overdispersed beta-binomial observer ($\eta = 0.2$) and a strongly overdispersed beta-binomial observer ($\eta = 0.5$) (see Fig. 2). After observing the range of coverages we obtained for different strengths of the prior we chose $k = 10$ as default value. A beta-distribution with parameters $(1, 10)$ represents a prior which did not yield dramatically too small—nor dramatically inflated—coverages of the credible intervals for any of the simulated observers (details are presented in S.5, Fig. S9).

Note that all priors are derived solely from the tested stimulus levels ("x-values"), chosen by the experimenter, not on the data from the observer. The heuristics for threshold and width are based on the assumption of an experimenter choosing the stimulus levels reasonably in the range where the behaviour changes. For other sampling schemes—adaptive procedures, or situations in which observers never reach performance levels above 60% in YesNo or 80% in 2-AFC—this heuristic will be sub-optimal. In such cases the user of *psignifit 4* is required to supply a manually chosen x-range, to replace the default stimulus range (see Section S.3).

2.4. Bayesian inference

To estimate the parameters from the data we use Bayesian inference (see Gelman et al., 2013; Kruschke, 2014, for an introduction). Given the beta-binomial model described in Section 2.2, we can calculate the likelihood $L(\boldsymbol{\theta}|\text{data}) = P(data|\boldsymbol{\theta})$ (A.1). Together with the prior $P(\boldsymbol{\theta})$ as described in Section 2.3, we can compute the five-dimensional posterior $P(\boldsymbol{\theta}|\text{data})$ over all parameters $\boldsymbol{\theta}$ as:

$$P(\boldsymbol{\theta}|\text{data}) = \frac{P(\boldsymbol{\theta})\,L(\boldsymbol{\theta}|\text{data})}{\int_\Omega P(\boldsymbol{\theta})\,L(\boldsymbol{\theta}|\text{data})\,d\boldsymbol{\theta}} \tag{4}$$

The main calculation step is to calculate the integral over all possible parameters in the denominator. To calculate this integral we first span a sparser grid over the 5 dimensional space and evaluate the likelihood and the prior on all grid points. Then we omit parts of the grid which contain less than $e^{-7} \approx \frac{1}{1000}$ of the marginal posterior to focus the grid on the parts that contain probability mass.

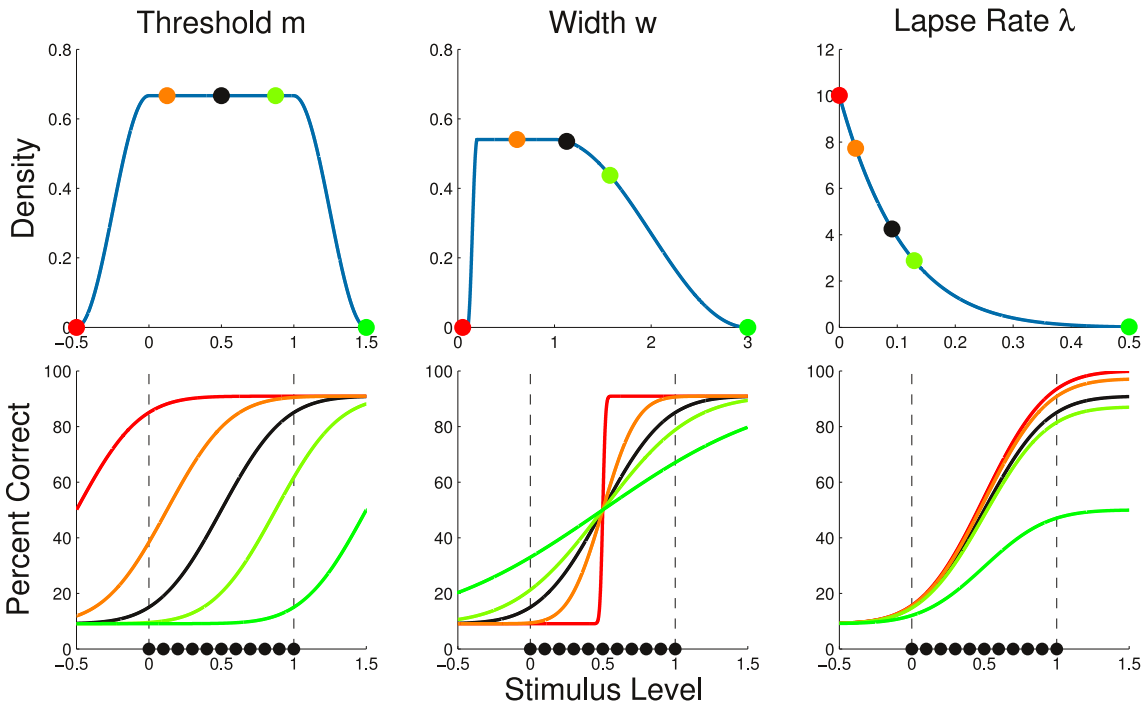

Fig. 3. Illustration of the priors on the different parameters for measurements at 11 equally spaced x-levels from 0.0 to 1.0. The solid blue curve in the panels in the top row shows the heuristically derived prior densities from the range and spacing of the x-levels for m and w in the first and second column, respectively. In the third column we show the default prior for λ (and γ if it is free), a beta-distribution with parameters $(1, 10)$. The coloured psychometric functions correspond to the 0%, 25%, 75% and 100% quantiles of the prior. All thresholds (first column) and widths (second column) between the orange and bright green curves receive a non-zero prior probability. All other parameters in the respective panels were set to the mean prior values corresponding to the black plots. The priors shift and scale with the stimulus range as described in the main text.

Subsequently we run a second evaluation of the non-normalised posterior on the refined, denser grid. For this final grid we use 40 gridpoints along the threshold and width and 20 for λ, γ and η. The chosen numbers of 40 and 20 are based on simulations exploring the change in the size of the 95% credible intervals as a function of the number of gridpoints, ensuring an error below 5% on the size of the credible intervals (see Appendix B.2, Fig. B.13). From the evaluations of the unscaled posterior on the refined grid we perform numerical integration to obtain the normalisation constant and proceed to normalise the posterior (see Appendix B.1 for details).

The posterior distribution represents the complete result of the Bayesian analysis. From the posterior distribution we can compute a maximum a posteriori (MAP) estimate by optimisation and (grid evaluations of) posterior marginal distributions for each parameter separately, as well as for parameter combinations. These marginal posterior distributions can be used to obtain credible intervals and posterior correlations of parameters. From the marginal distributions we can also compute a mean and a median estimate. Throughout this paper we use the MAP estimate as the point estimate, as we found that the mean and median estimates for low numbers of trials per psychometric function ($N < 200 - 300$) are often strongly biased for the width, as the default prior influences the final estimate too strongly (supplementary material: S.6).

An example result for the Bayesian analysis is displayed in Fig. 4, emphasising that the result of the analysis is not only a single parameter value but a distribution over the possible parameter values. Note that for clarity we restrict the presentation to the threshold and width parameters, but the posterior extends over the other three parameters as well.

3. Evaluations of the method and software

To test the statistical method and software implementation we performed extensive numerical simulations. Each simulation proceeded as follows. First, we generated datasets from one of many different simulated observers. Second, we ran *psignifit 4* on the datasets—using the default parameters and priors. Third, we evaluated the accuracy (bias) of the point estimates for threshold and slope as well as the coverage of the credible intervals. The coverage of the credible interval assesses how often they contain the parameter used to generate the data, i.e. a 95%-credible interval should contain the generating parameter 950 times in a simulation using 1000 repetitions. We report all results throughout the paper for the 95% confidence level. However we confirmed that all conclusions drawn from the 95% data remain true for the 68% confidence level, i.e. roughly ± 1 standard deviation of a Gaussian.

First we confirmed that the method, implementation, and default priors work well and as expected with binomial data using various sampling schemes, data collected using adaptive methods (Wetherill & Levitt, 1965; Levitt, 1971; Watson & Pelli, 1983; Pelli, 1987), and for data with lapse rates as high as 10%. Details and simulation results can be found in the supplementary material S.1–S.4.

3.1. Psychometric function estimation with overdispersed, beta-binomial, data

The purpose of the simulations reported in this section is a test of the software implementation and default priors for the beta-binomial model. To this end we simulate data from (stationary) beta-binomial observers, and then apply Bayesian inference using the exact model from which the data were generated.

The simulated observers had a threshold and width of 1.0, i.e. $m = w = 1.0$, and for every simulated dataset the lapse rate was drawn from a uniform distribution between 0 and 0.1, i.e. $\lambda \sim \mathcal{U}(0, 0.1)$. In case of simulated YesNo experiments the lower asymptote was drawn from a uniform distribution between 0 and 0.1 as well, i.e. $\gamma \sim \mathcal{U}(0, 0.1)$. Both λ and γ thus had considerable variability; furthermore, the distribution of both parameters

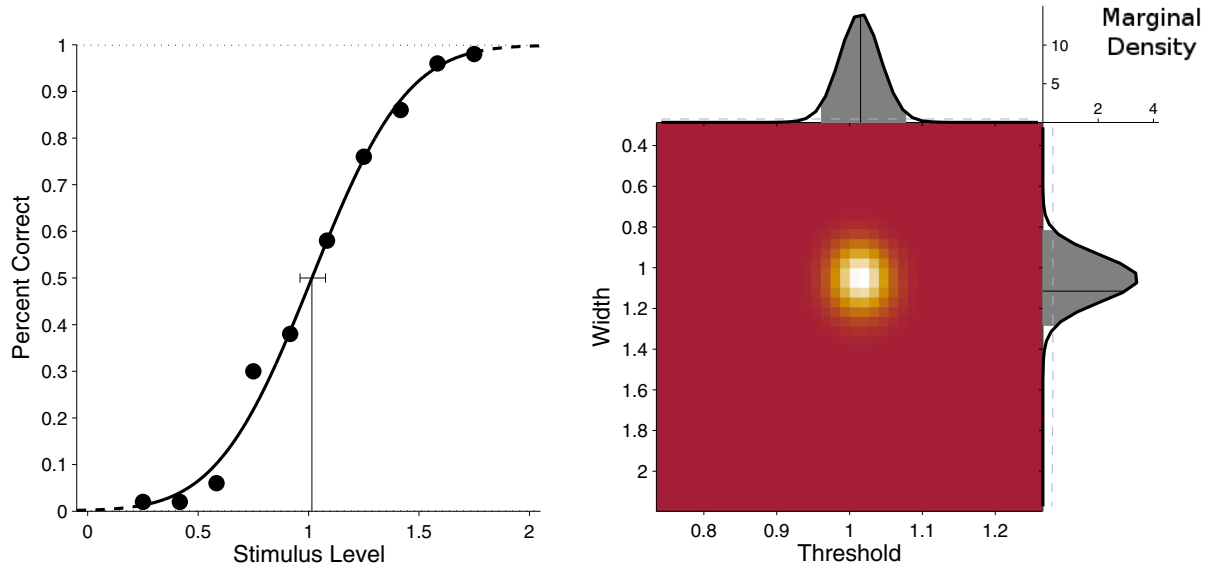

Fig. 4. Example results of the Bayesian analysis for a psychometric function measured with 500 trials. The data with the Maximum A Posteriori (MAP) estimate of the function shape are plotted on the left. On the right the marginal posterior for the threshold and the width is displayed. Attached to it are marginal distributions for the single parameters. The prior is also displayed as the dashed grey line; the grey shade filling most of the marginal distributions corresponds to the extracted 95% credible intervals.

does not match the used priors: The default priors $P(\lambda)$ and $P(\gamma)$ for λ and γ are beta-distributions with parameters $(1, 10)$ (see Section 2.3; Fig. 3). Both the large variability as well as the generat ing-distribution-to-prior-mismatch were chosen to ensure realistic and taxing tests for the methods and software.[18]

We simulated:

- Two degrees of overdispersion: We used the values $\eta = 0.2$ to simulate a moderately overdispersed observer, and $\eta = 0.5$ for a strongly overdispersed observer. The two values correspond to those shown in Fig. 2.
- Three types of experiments:
  1. 2-AFC with three estimated parameters (threshold m, width w, lapse rate λ);
  2. YesNo with four estimated parameters (threshold m, width w, lapse rate λ, lower asymptote γ);
  3. YesNo with equal asymptotes, i.e. three estimated parameters (threshold m, width w, lapse rate equal to the lower asymptote: $\lambda = \gamma$).
- Two different sigmoidal functions S: The cumulative Gaussian and the Weibull.
- Constant stimulus design with three sampling schemes: Five blocks, ten blocks and the number of blocks equal to the number of trials (see Fig. 5). Note that the priors $P(m)$ and $P(w)$ were automatically obtained from the x-levels of the data, as described in Section 2.3.
- Five different number of trials per psychometric function: 50, 100, 200, 400, and 800.
- One thousand repetitions for each of the combinations above, to be able to assess coverage.

Thus we simulated a $2 \times 3 \times 2 \times 3 \times 5 \times 1000$ design with 180,000 datasets and a total of 55.8×10^6 trials.

The sampling schemes we used are meant to represent typical sampling schemes. They all cover the range of the psychometric function well, but they are not optimal. Shen and Richards (2012) and Kontsevich and Tyler (1999) discuss optimal sampling points for the parameters of the psychometric function and discuss how to sample them adaptively during data acquisition.

The key results of these simulations for the moderately beta-binomial observer with $\eta = 0.2$ are shown in Fig. 6. Threshold estimates are unbiased (panel A), and width estimates are unbiased, too (panel B). The coverages of the credible intervals for threshold and width are reasonable and in the range of 92–97%, see panels C and D.

Fig. 7 shows the data for the strongly beta-binomial observer with $\eta = 0.5$. Note that such a strong overdispersion is more severe than exhibited by any observer in the substantial body of real data we analysed, or the simulated serially-dependent observers (see Sections 3.2 and 3.3 and Fig.'s &9–11C). Nonetheless, even for this severely overdispersed observer—and the default prior for the strength of the beta-parameter η—we estimate threshold and width accurately (panels A and B). The coverage of the credible intervals is too small, however, in the range around 80% for the threshold and 85% for the width for data from blocked sampling schemes.

In summary, this set of simulations shows that the method works accurately and with good credible interval coverage for the problem of estimating psychometric functions for *realistically* overdispersed data stemming from a moderately beta-binomial observer ($\eta = 0.2$). For strongly overdispersed data the implemented default prior precludes fitting appropriately large values for the beta parameter η. As a result the credible interval coverage falls short of the desired 95% by up to 10–15% in the worst cases of our simulations.[19]

3.2. Robustness against other sources of overdispersion

The purpose of the simulations reported in this section is to test whether the beta-binomial model can adequately cope with other

[18] Wichmann and Hill (2001a, 2001b), for example, only used six different values for λ, the maximum being 0.05—exactly the mean of the λ's we used, and only half the maximum value of 0.1 during our simulations. In addition, the distribution and range of their λ's corresponded to their uniform prior.

[19] If an experimenter is faced with data with extreme overdispersion, they can always change the prior parameters of *psignifit 4* to accommodate such datasets— for example to *beta*$(1, 5)$. For an uniform prior—as an extreme case—the inference becomes slightly conservative even for the strongly overdispersed observer, and strongly conservative for binomial observers (supplementary material S.5).

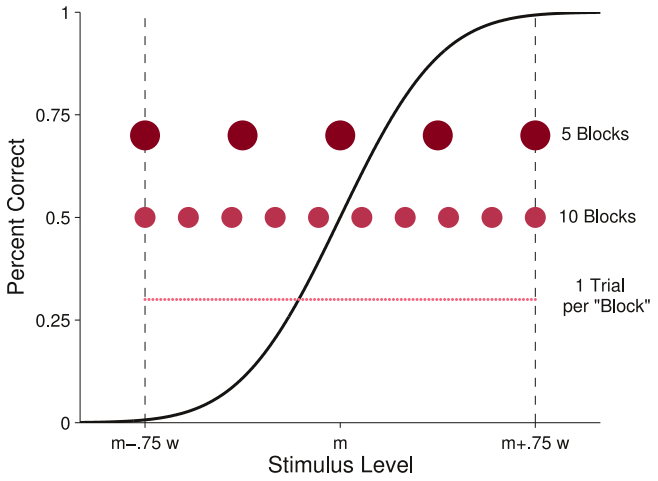

Fig. 5. Constant stimulus designs used for our simulations. For each sampling scheme we placed blocks of equal number of trials uniformly distributed between $m - 0.75w$ and $m + 0.75w$. We simulated three placements, one with five blocks, one with 10 blocks, and one with as many blocks as trials (shown here with 100 trials).

forms of overdispersion. To this end we modelled three observers exhibiting serial-dependencies between their responses. The existence of serial-dependencies between successive responses is well known at least since the early 1950s (Senders & Sowards, 1952; Verplanck, Collier, & Cotton, 1952; Howarth & Bulmer, 1956; Green, 1964; for a recent review and statistical assessment see Fründ et al., 2014).

- A *fluctuating* observer, whose threshold m varied over time according to an autoregression model from trial i to trial $i+1 : m_{i+1} = 1 + 0.95(m_i - 1) + 0.05N(0,1)$,[20] which results in slow fluctuations of the threshold around 1.0 roughly ranging from 0.6 to 1.4. This could be caused by drifts in attention, concentration or sleepiness.
- A *sequential dependent* observer, whose threshold was 0.2 lower after correct trials and 0.2 higher after incorrect trials, i.e. whose performance depends on the previous trial.[21]
- An *interval biased* observer, who prefers the interval which was the correct one in the previous trial, simulated by a shift in threshold of 0.2 up when the target was in the preferred interval, and 0.2 down if it was not.

The first two observers have a higher variance in their responses at a given x-level, while the third creates no dependence between the correctness of trials and thus for fitting is mostly equivalent to a binomial observer with a wider psychometric function.

For the sampling schemes we again chose the two well-behaved sampling schemes with 5 or 10 blocks shown in Fig. 5. In addition we simulated data by either blocking or mixing the trials of different stimulus levels in random order, as the observers simulated here behave differently depending on the previous trial—there is no difference between blocked and intermixed trials for binomial or beta-binomial observers investigated in the previous three sections (see the discussion of blocking of trials either during the experiment or during the analysis in Section 4.2). Thus we simulated a $2 \times 3 \times 4 \times 3 \times 5 \times 1000$ design with $360,000$ datasets and a total of 11.16×10^7 trials.

[20] $N(0,1)$ indicating a standard, normally distributed random variable with zero mean and unit variance.

[21] This introduces positive correlations for the correctness of trials and thus overdispersion (Prentice, 1986).

The key results for all three observers are shown in Fig. 8. We normalised the width estimates and credible intervals to the value obtained by fitting a single function to the data from all 1000 datasets, which always yielded well constrained fits. This was necessary as the threshold fluctuations on average lead to a wider psychometric function, which we define here as the "correct" function.[22] Threshold estimates are again unbiased (panel A), and width estimates are essentially unbiased, i.e. no more biased than for the binomial and beta-binomial observers (panel B).[23] The coverages of the credible intervals for the threshold are reasonable in the range of 75–98% (panel C) although the fluctuating observer is covered worse than the other observers. Note the interactions between the number of trials and blocks for the sequential dependent observers; Note in addition that the interval sequential dependent observer shows no overdispersion at all, although its width is inflated. Finally the coverage for the width is reasonable for all observers, although there is a slight decline for the blocked sampling schemes with 10 blocks to approximately 90% coverage for the sequential dependent and the fluctuating observer (Panel D).

To see how the beta-binomial model compares to the standard binomial analysis we fitted the same data setting $\eta = 0$, i.e. performing a pure binomial analysis —as previous psychometric function estimation methods would have done. In all cases and for all overdispersed observers—regardless of the cause of the overdispersion—the beta-binomial analysis produces superior results, especially for larger trial numbers, when the binomial analysis yields progressively worse credible intervals (see Fig.'s 9 and 10).

In summary, this set of simulations shows that the method works accurately and with reasonable credible interval coverage not only for the beta-binomial type of overdispersion, but also for at least three types of sequentially dependent observers. Especially coverage is improved compared to the standard binomial analysis, using the *default settings* of *psignifit 4*.

3.3. Real data: inflated variance versus goodness-of-fit

The purpose of the tests reported in this section are twofold: First, we compare the fitted psychometric functions of *psignifit 4* on real world data from contrast detection and discrimination experiments (Wichmann, 1999) to those obtained using the "old" psignfit 2.5 toolbox (Wichmann & Hill, 2001a). Second, we compare fitted estimates of *overdispersion*, η, in *psignifit 4* to *deviance* in the "old" psignfit 2.5 toolbox. We show that the beta-binomial model increases credible intervals mostly when deviance signals a bad fit, i.e. the beta-binomial model provides a viable inference method.

We used 540 datasets from contrast detection and discrimination experiments originally from Wichmann's DPhil thesis; some were published later (Wichmann, 1999; Bird, Henning, & Wichmann, 2002; Henning, Bird, & Wichmann, 2002); the experiments used different presentation times, spatial frequencies, observers and numbers of trials (100–1550 trials per function, usually in blocks of 50 trials with constant stimulus level, mean number of trials: 504). For 51 of these psignifit 2.5 crashed[24], while *psignifit 4* fitted all datasets using the default settings; thus we were left with 489 datasets for which both toolboxes succeeded in fitting the data.

[22] The "correct" functions had similar thresholds as the ones we used for generation for the fluctuating observer (Range: $[0.9763, 1.0126]$) and the interval biased observer (Range: $[0.9598, 1.0176]$). For the sequential dependent observer some functions for the 2AFC case yielded lower thresholds extending the range to lower values ($[0.8708, 1.0176]$). For the width the fluctuating observer yielded higher values (Range: $[1.0287, 1.2609]$), the sequential dependent observer yielded both higher and lower values (Range: $[0.7086, 1.4211]$), and the interval dependent observer yielded only higher values again (Range: $[1.1789, 1.4210]$).

[23] Most likely caused by the lapse rate prior as discussed in detail in S4.

[24] Crashes were caused by divergence of the estimates for the original data or the datasets for bootstrapping.

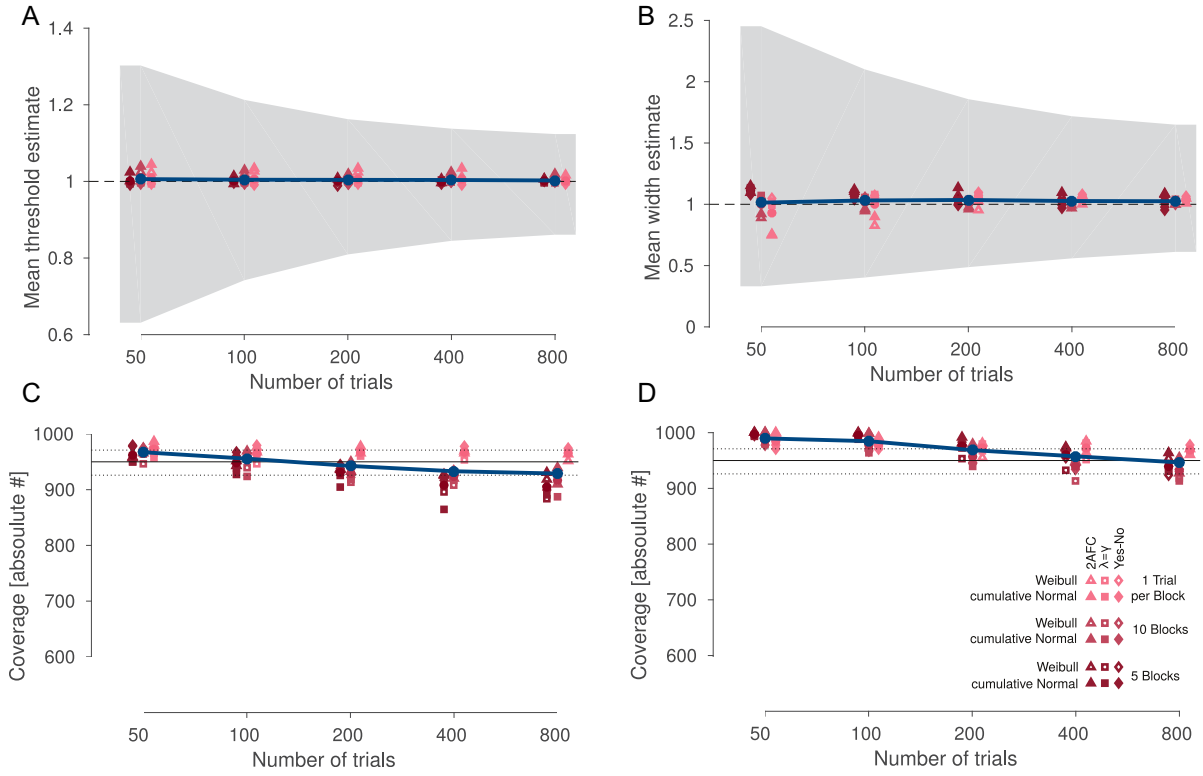

Fig. 6. Simulation results for the moderately overdispersed observer with $\eta = 0.2$ for linearly spaced constant stimulus designs. A: average MAP estimates against the number of trials. The thick line marks the grand average, the coloured symbols individual conditions computed from 1000 simulations each. The dashed line marks the true parameter value of 1.0 and the grey shade the average 95% credible interval over all conditions. B: MAP estimates of the width, plotting conventions as in A. C: coverage of the 95% credible intervals for the threshold. The continuous black line marks the nominal value of 950 of the 1000 simulations, the dotted lines mark the interval [926,971], which would contain the measured coverage in 99.9% of cases if the true one was exactly 95%. D: coverage of the credible intervals for the width. A perfectly unbiased estimate would lie at exactly 1 in A and B and perfect credible interval size would produce a coverage of 950 in all conditions in C and D.

The key results of these simulations are displayed in Fig. 11. We plot the point estimates from *psignifit 4* against those of psignifit 2.5 for threshold and width. These show that the methods are reasonably consistent in their parameters estimates (Panel A & B); the red symbols mark datasets for which the toolboxes deviated by more than 20% from each other and we inspected those $N = 63$ datasets by hand, see below. In panel C we plot the η parameter of the estimated overdispersion against the deviance from the binomial model as estimated by psignifit 2.5 (see explanation in panel F of Fig. 2 for the exact relation of these two measures). First, panel C shows that considerable over-dispersion exist even in simple contrast detection and discrimination experiments with experienced observers. Second, for all 293 statistically "innocuous" datasets for which psignifit 2.5 estimated a p-value below 0.975 in the deviance test, *psignifit 4* barely inflates variances, i.e. $\eta < .00014$. Thus the fitting of a beta-binomial model with an appropriate prior does not lead to a dramatic increase of credible intervals for real world data, mirroring our simulation results (c.f. panels A in Fig.'s 9 and 10). Third, 85 poorly fit datasets, with average deviance values of 1.259 are rejected by psignifit 2.5. The practical usefulness of this rejection is questionable, however, and we discuss this point below in Section 4.3. *Psignifit 4*, on the other hand, inflates the variance and, very likely, returns correct credible intervals.

As mentioned above, there were $N = 63$ datasets for which the toolboxes deviated by more than 20% from each other, and they are shown as the red symbols in panels A and B of Fig. 11. Inspection of all the individual datasets and their fits revealed that, first, *psignifit 4*'s estimates are preferable (see below), and, second, that the differences are explicable in terms of the different priors used in the two toolboxes. We observed three different scenarios:

1. 45 datasets displayed a lapse rate substantially larger than the 5% (default) limit on the lapse rate of psignifit 2.5. On these datasets psignifit 2.5 fitted a lapse rate of 5% and was thus forced to fit a larger width than *psignifit 4*. This explains all datasets for which *psignifit 4* fitted smaller widths.
2. 9 datasets were sampled extremely sparsely (2–4 blocks in total). For these datasets psignifit 2.5 frequently fitted a step function between two blocks. The prior on the width w in *psignifit 4*, on the other hand, enforces a width of at least the difference of the nearest blocks. This explains some cases for which *psignifit 4* fitted larger widths.
3. 9 shallow psychometric functions without data at high x-levels were fit with 5% lapse rate by psignifit 2.5, and with 0% in *psignifit 4*; this clearly reflects their priors on λ in the absence of evidence—uniform between 0% and 5% in psignifit 2.5, a beta-function $(1, 10)$ peaking at zero in psignifit 4—which explains the remaining cases for which *psignifit 4* fitted larger widths.

Differences in threshold occurred only together with differences in width estimation; as the different estimates of the width and the lapse rate also affect the threshold estimates, the three factors above are sufficient to explain all differences between the two methods and their implementations.

In summary, this set of tests shows the method using the *default settings* to work accurately for estimating psychometric functions from real world data, containing a substantial number of rather poorly sampled psychometric functions. We showed that, first, *psignifit 4* did not fail to fit a single of the 540 datasets, whereas psignifit 2.5 failed on 51 of them (9.1% failure rate). Second, if *psignifit 4* and psignifit 2.5 differed in their fitted threshold or width,

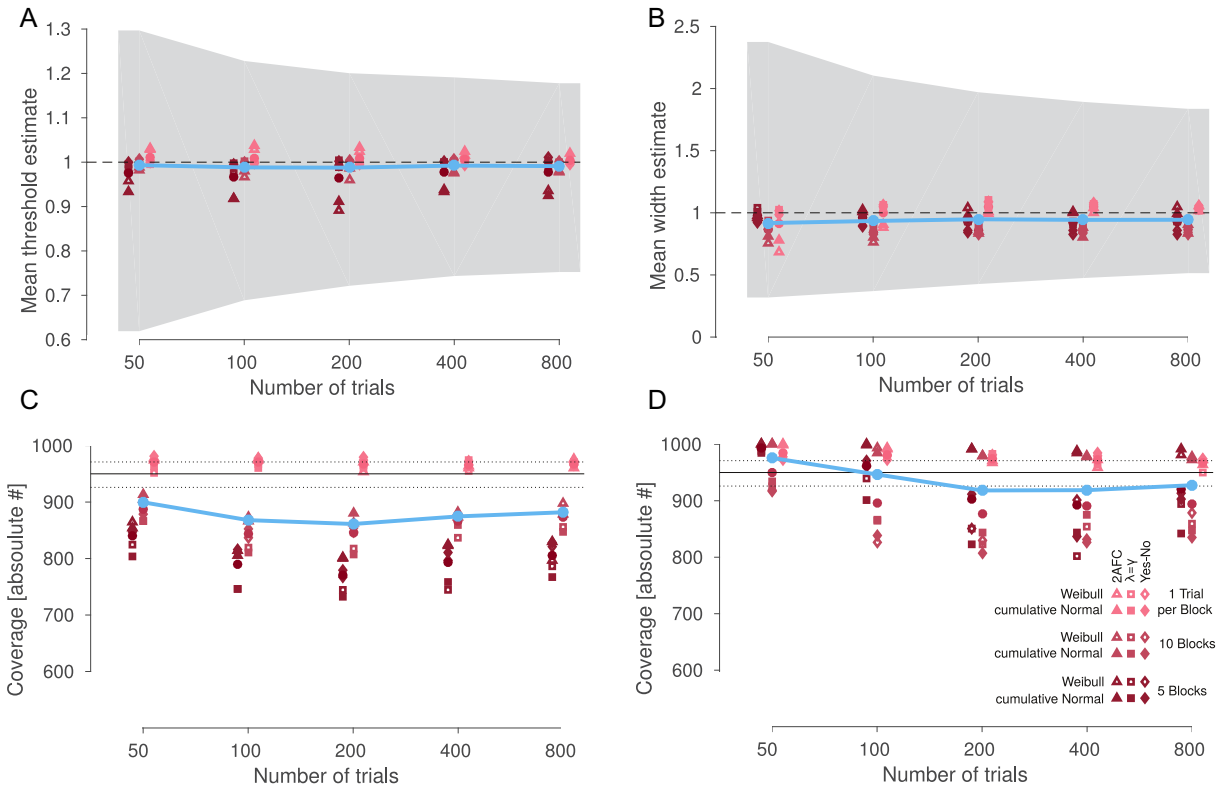

Fig. 7. Simulation results for the stronger beta-binomial observer ($\eta = 0.5$). Plotted are simulation results for well-sampled, linearly spaced constant stimulus designs with plotting conventions as in Fig. 6. Even for severely overdispersed data and the default prior for the strength of the beta-parameter η, we estimate threshold and width accurately, although the coverage of the credible intervals is too small (range 75–95%).

we preferred the results returned by *psignifit 4*. Third, for statistically innocuous datasets *psignifit 4* did not unnecessarily inflate the variance controlled by the η parameter of the beta-binomial model. Fourth, for datasets lacking in goodness-of-fit—large absolute average deviance—*psignifit 4* increased η in line with the increasing deviance.

3.4. Blocking of trials

For the statistical analysis of the binomial model blocking of trials has no influence on the results, as the likelihood does not change depending on how or whether data are blocked. However, the beta-binomial model assumes that the percent correct is drawn new once per block. Hence it matters for the beta-binomial model which trials form a block, or are grouped together as a block. In addition, the size of blocks puts an upper bound on how large the overdispersion can be in any model with independent blocks—independent blocks of single trials cannot be overdispersed and all models reduce to the binomial model.

To assess how the inference depends on how data are blocked we performed another set of simulations. Most parameters we kept as in the other simulations, e.g. we simulated cumulative Normal and Weibull functions, for the three types of experiments and two levels of overdispersion, moderate and very strong $\eta = 0.2$ and 0.5. We simulated 4 trials at 12 stimulus levels (48 trails in total) as well as 4, 8, 16 and 32 at 25 stimulus levels each (resulting in 100, 200, 400 and 800 trials in total). However, during the simulations we simulated 2 blocks of equal even number of trials at each stimulus level—each with a newly drawn percent correct according to a beta-binomial model. Thus the "true" underlying block sizes were half those given above, i.e. 2×2 trials at 12 levels; 2×2, 2×4, 2×8 and 2×16 trials at 25 stimulus levels. Then we performed Bayesian inference three times: Either we used the correct number of blocks per stimulus level (2), or we fused all data at one stimulus level into one block, or we split every block in half yielding 4 blocks per stimulus level. We refer to the conditions as *correct pooling* (as in the data generation), *more pooling* because one block in the analysis is based on two simulated blocks, and *less pooling* because the analysis used twice as many blocks of data (with half the number of trials per block) as during data generation.

The coverage results for these simulations are displayed in Fig. 12. The point estimates were as good as in the other simulations, and were essentially independent of how the data were blocked.

In summary, these simulations for randomly fluctuating (beta-binomial) observers indicate the following: When the correct number of trials or more were pooled per block, the results are correct for all conditions, i.e. too much pooling is not detrimental. If more blocks were formed—and thus fewer trials pooled per block—the coverage became too small for larger datasets and the strongly overdispersed observer, i.e. too little pooling is problematic.

4. Discussion

4.1. Extension to the beta-binomial model

We extend psychometric function modelling from the standard binomial to a beta-binomial model to capture overdispersion. We show that this model works well both for binomial data (see sections in supplementary material, S.1–S.3) as well as for overdis-

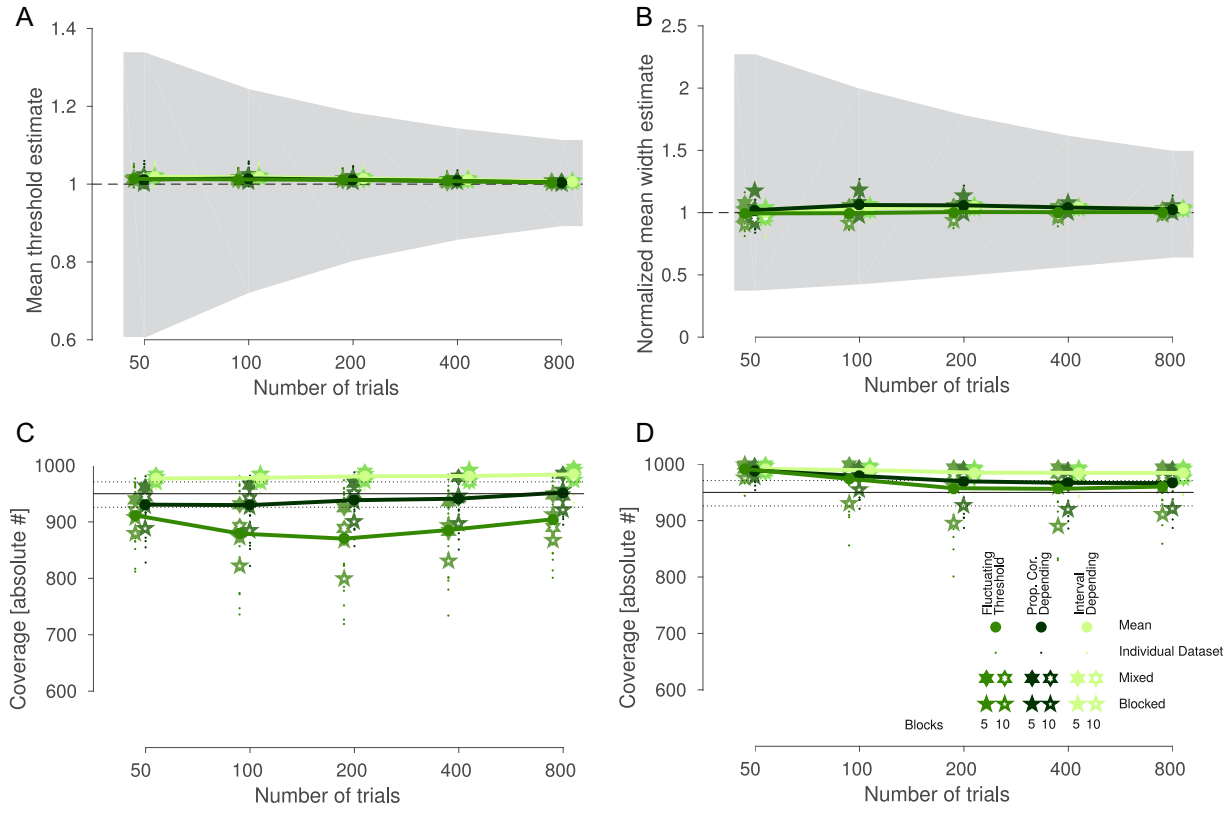

Fig. 8. Results from three non binomial observers differentiated by colour. One observer whose threshold fluctuates over time, one whose probability correct depends on the outcome of the previous trial and one whose bias depends on the previous trial. Plotting conventions as in Fig. 6, but with different symbol shapes for the different sampling schemes and whether stimuli were blocked or not.

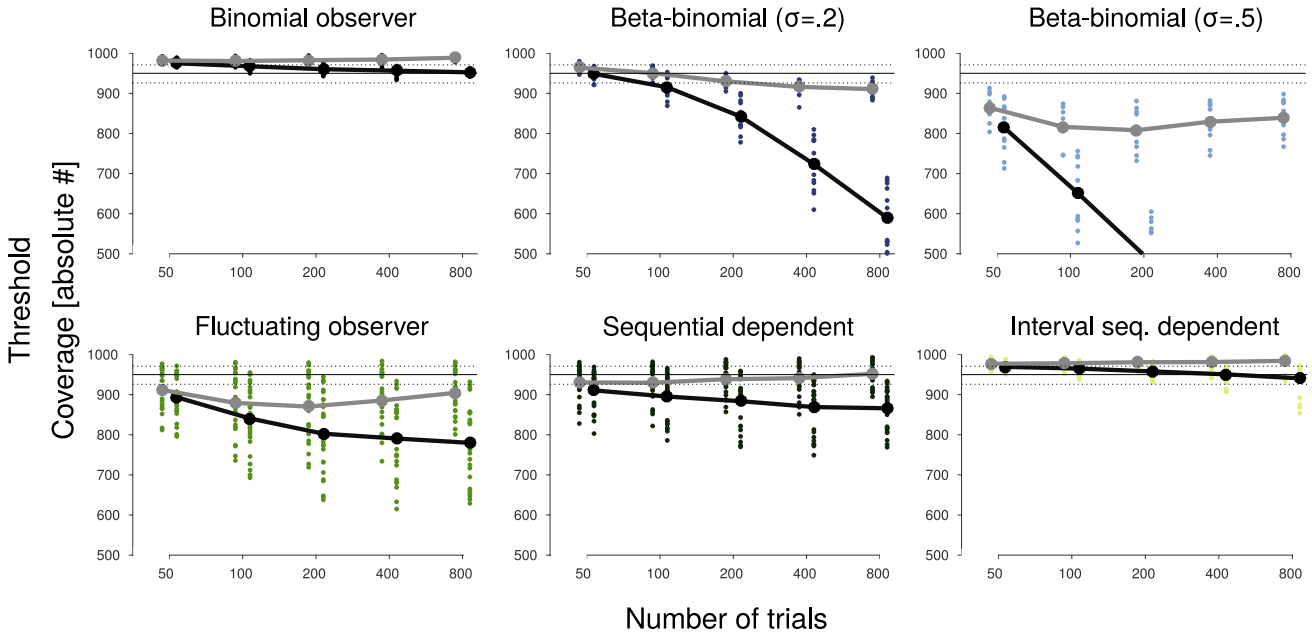

Fig. 9. Beta-binomial coverage (grey) and standard binomial coverage (black) for the threshold for six different simulated observers (with single datasets colour coded as in the previous Fig.'s 6–8); see text for details. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

persed data from a beta-binomial observer (Fig.'s S1, 6 and 7), for which the standard binomial analysis fails (Fig.'s 9 and 10).

The beta-binomial model models a specific kind of overdispersion. We have shown that fitting the beta-binomial model yields an accurate statistical description of the amount of overdispersion, but the fitted model does not provide any information about the cause of the overdispersion. What makes the beta-binomial useful in practice is that correcting the statistical inference under overdispersion seems to work with the beta-binomial model relatively independent of the cause for the overdispersion. The beta-

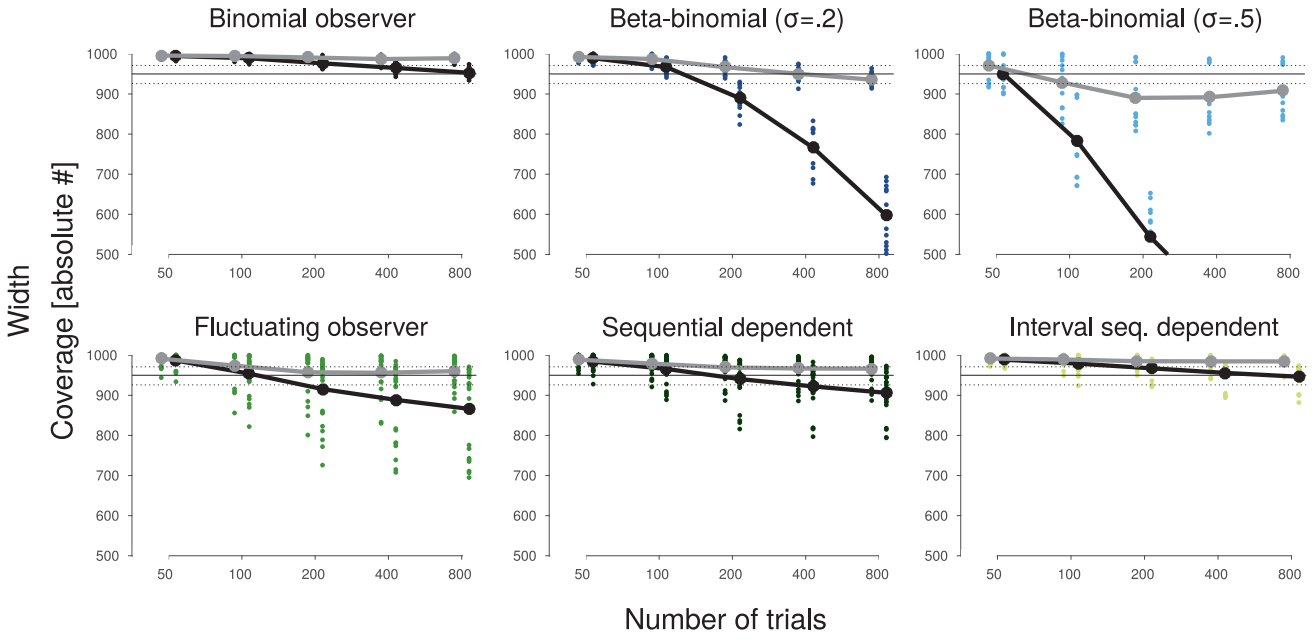

Fig. 10. Same as Fig. 9 but for the width.

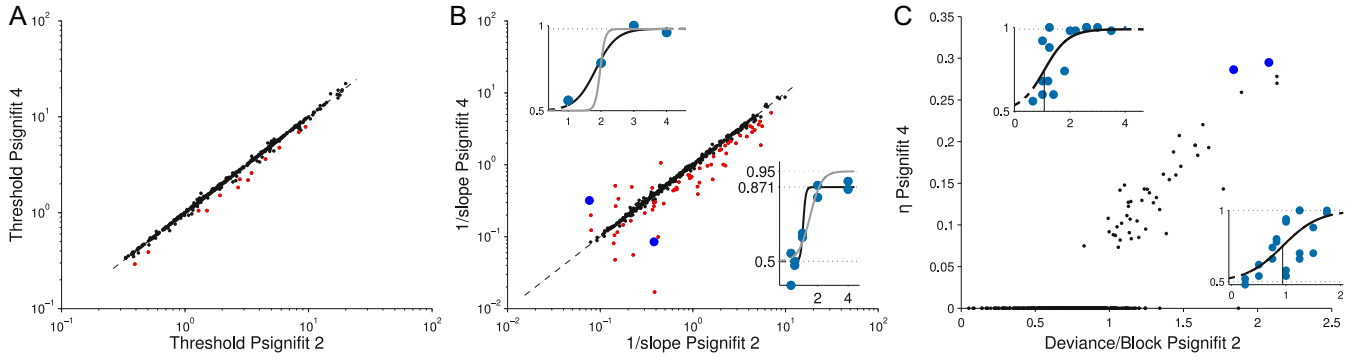

Fig. 11. The point estimates of psignifit 2.5 and 4 are largely consistent on 489 real data psychometric functions for contrast detection and discrimination. A, B: log–log scatterplots of threshold and of 1./slope at 50% correct respectively. Fits marked in red deviate more than 20% from each other. These were investigated by hand and results are described in the main text. The insets show two psychometric functions for which psignifit 2 and 4 disagree, marked with blue dots as examples for the situations described in the text. The black line is the fit by psignifit 4, the grey line the fit of psignifit 2. C: scatterplot of average absolute deviance residual per block from psignfit 2.5 vs. fitted η value from psignifit 4. A deviance of 1.0 per block is expected for a binomial observer. Larger values indicate overdispersion. Note that few datasets have a η of around 0.2; none have values larger than 0.35, supporting our argument made above that a η of 0.5 denotes an extreme form of overdispersion. The insets show the two most overdispersed psychometric functions, again marked with blue dots in the scatter plot.

binomial model yields reasonable results for other sources of overdispersion stemming from several types of serial dependencies (Fig. 8), for which the standard binomial model fails (Fig. 9 and 10).

If one desires to study the overdispersion itself, different analyses are necessary to differentiate different causes for overdispersion. For example, fits to different parts of the data can be calculated to detect learning or fatigue or the dependence of trials can be studied as demonstrated by Schönfelder and Wichmann (2013) or Fründ et al. (2014).

4.2. Blocking

Two issues need to be considered concerning the blocking of trials: Whether to block them during the statistical analysis, and whether to block them during experiments.

4.2.1. Blocking during the statistical analysis

Our simulations of randomly fluctuating observers indicate that pooling blocks which, in reality, were independent is not problematic and does not degrade inference, whereas not pooling enough results in credible intervals with too little coverage (see 3.4). Thus from a practical point we advise to pool trials at a given stimulus level and otherwise identical conditions during psychometric function estimation unless performance might have changed over the experiment systematically.

However, one should never pool data automatically, without careful inspection of the data: Our simulations were performed using a *randomly* fluctuating (beta-binomial) observers. That is, the *time-averaged* success probability of such an observer is constant, but at any point in time the success probability fluctuates up and down. If data exhibit *systematic* overdispersion, resulting from learning for example, one should *not* pool trials which might have different success probabilities. If blocks from a learning observer are pooled, the different success probabilities of the subsets are not represented any more: Pooling results in a single large block of average performance instead of a block with low performance at the beginning, and a block with higher performance at the end due to learning. Consequently one cannot notice that there

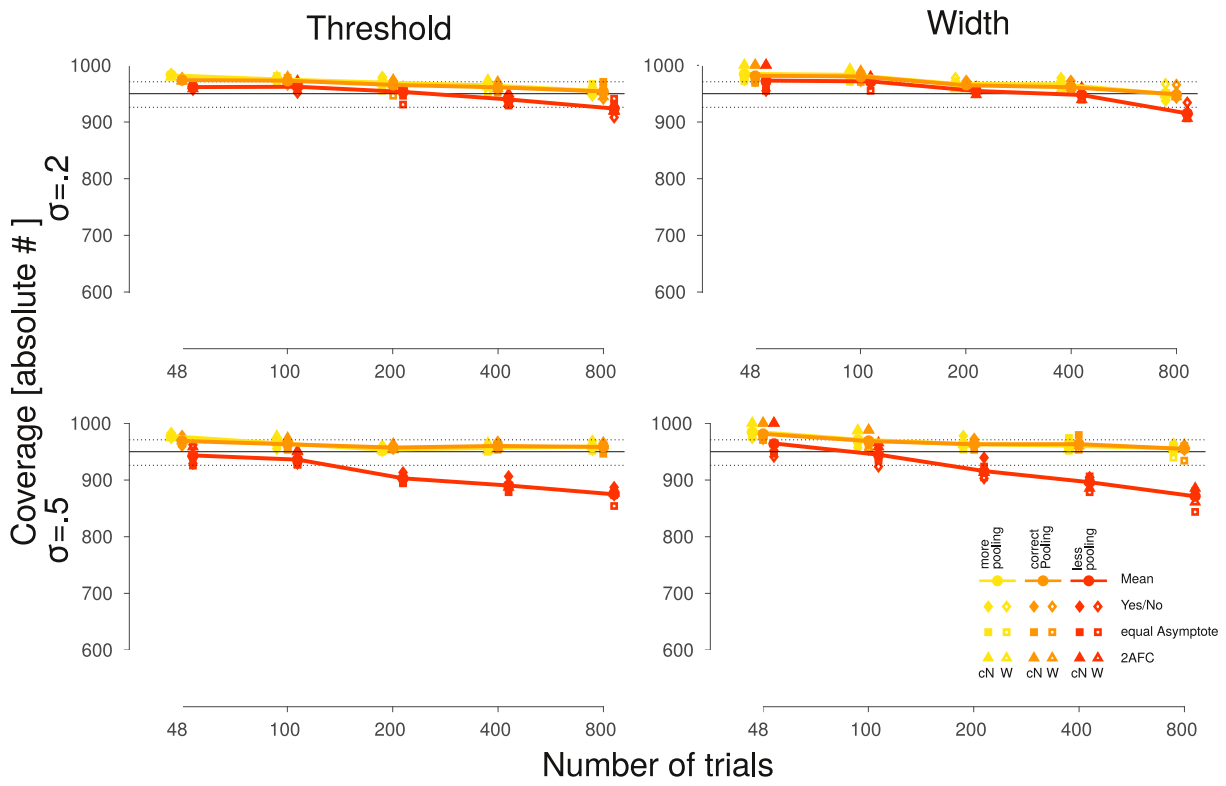

Fig. 12. Results for simulations when the pooling differs from the simulated blocks of the beta-binomial model. Only the coverage results are shown, as the average point estimate is largely independent of the pooling. Correct pooling means that each block in the simulation was one block in the fit. For less pooling we split each simulated block in two for the fit. For more pooling we pooled two simulated blocks of the same stimulus level into one block. Plotting conventions as in Fig. 6C and D.

were systematic performance differences. This problem occurs only if the deviations from the mean psychometric function cancel reliably due to the systematicity of the source of overdispersion. In such cases η would erroneously be estimated as too close to zero (no or little overdispersion when in fact there was a lot due to learning).

Thus if one suspects or knows that observers improve during the experiment, or for other reasons expects *systematic* performance variations—e.g. when collecting data for natural images and one has reason to believe that some images make the task much harder than others—it is advisable to keep the data for potentially different situations separate, and not to pool excessively.

This scenario did not occur in our simulations, as we always assumed that the means of the blocks *randomly* fluctuate around the psychometric function. Then the pooled blocks are still overdispersed and the inference remains correct.

4.2.2. Blocking during experiments

For data collection classical psychophysics recommended to block trials of equal signal strength, i.e. to use the method of blocked constant stimulus with as few different stimulus levels as possible (Blackwell, 1952; Jäkel & Wichmann, 2006). Such an experimental design yields lower and more stable and repeatable thresholds, at least in highly experienced observers. Contrary to these recommendations the typical advice in psychological methods classes is to "randomise presentation order as much as possible." This is suggested to minimise the influence of the many uncontrolled parameters that may influence an observer, and thus to reduce overdispersion.

In accordance with the general psychology recommendation we find statistical inference to be more stable for mixed presentations for the realistic simulated nonstationarities we explored in Section 3.2. Both the binomial and the beta-binomial analyses yield better coverage for the mixed presentations than for the blocked *presentations* (see Fig. 8).[25]

4.2.3. Discussion on blocking

Our results confirm the idea that mixing stimulus levels may reduce overdispersion caused by sequential dependencies (Fig. 8). At the same time, our results suggest it is beneficial to pool trials into one block if possible—unless one suspects data to be affected by learning; furthermore, for η to be estimated from the data one requires approximately ten or more blocks, providing a limit on the amount of pooling—for the estimation of the psychometric function. The latter is not justified by a better correspondence of model and reality, but for purely pragmatic reasons: We find that blocking too many trials (within reasonable limits, see comment above) into one block degrades inference much less than blocking too few (Fig. 12). For experimental design we thus recommend: First, and in accordance with Blackwell, to measure more trials per level rather than more levels, as small blocks weaken the inference about overdispersion.[26] Second, an unlike Blackwell, to randomise the order of trials to minimise the unwanted overdispersion resulting from extraneous influences. Third, to pool data at a given level from a single session into one block during psychometric function estimation.

[25] Note that data are still overdispersed and the coverage of the credible intervals is reduced, i.e. mixing trials of different stimulus levels reduces, but does not overcome the overdispersion problem.

[26] Especially adaptive procedures violate this in their current form, as they often sample only single trials at a given stimulus level. Maybe enforcing them to choose from a (much) smaller set of possible stimulus levels might be beneficial, i.e. may represent a suitable compromise between their efficiency to find the "threshold" during the experiment, and the ability to estimate overdispersion from the collected data after the experiment.

4.3. Going beyond goodness-of-fit

In earlier publications on psychometric functions it was emphasised to check the goodness-of-fit, that is, how well the model explains the data (Wichmann & Hill, 2001a). The aim of these checks is to detect when a model does not fit the data. In general this is obviously sound advice: If a model does not fit the data, it is very often of little use. However, in the context of psychometric function fitting this recommendation may be too strict, or at least of little practical use: Often there is no alternative to fitting a psychometric function, and the goodness-of-fit rejection leaves researchers without a viable option to proceed.

With the beta-binomial model, however, researchers can draw valid conclusions from overdispersed data, providing them with a viable option. The overdispersion parameter of the beta-binomial model measures a very similar property as classical measures of goodness-of-fit, namely the additional variance around the function (Section 2.2; c.f. Fig. 11). But in the case of the beta-binomial model the overdispersion parameter can be used to increase the uncertainty until the data are consistent with the fitted model, instead of rejecting the model.

4.4. Pain-free implementation

We provide an implementation of the method in pure MATLAB[27], which is largely automated, designed to be easy to use and reasonably fast on modern laptop or desktop computers ($\approx$ 1 s for 4 parameters and $\approx$ 20 s for 5). This implementation is discussed in B.1 and its numerical stability in B.2. It uses numerical integration and thus does not need MCMC sampling techniques; furthermore, the default priors and parameter settings are general enough to cover at least the broad range of simulations presented in this paper (with the exception of the prior if data come from adaptive procedures, as discussed in Section S.3).

Despite the generality of the default priors there are some cases for which they should at least be carefully assessed, especially if the sampling scheme or the observers are not very typical.

The default priors for threshold and width assume that the sampling scheme covers the interval over which the psychometric function changes, i.e. contains the threshold and is nearly as wide as the changing part of the psychometric function. This is, e.g. frequently violated for adaptive sampling schemes which often cover only a narrow range of x-values. For such cases the priors can be adjusted easily in the software (see supplementary material S.3).

The default priors for the lapse and guessing rates were chosen for reasonably experienced psychophysical observers. Other observers in, e.g. clinical settings or animal research, might require different priors if they lapse much more. Also different tasks might require different assumptions about the guessing rate (see supplementary material S.4 for alternative suggestions for the lapse rate prior).

4.5. Correctness and robustness

In Sections 3.1, 3.2, 3.3 as well as in the appendices and the supplementary material we assessed the correctness and robustness of the method against a substantial number of simulated datasets. The method works well under ideal conditions (Fig. S1), but also for badly sampled data (Fig. S2), adaptively sampled data (Fig. S3) and different kinds of overdispersion (Fig.'s 6–8). Finally, we ensured that the method works robustly on real world data (Fig. 11). One aspect that allows the method to work well is the careful choice of priors for the parameters. This is especially true for the lapses, which we allowed to be as large as 10% during our simulations without resulting in poor fits (see supplementary material S.4).

Finally, we compared the results from different fitted sigmoids and show that the method provides a robust analysis even if the "wrong" sigmoid family is fitted (see supplementary material S.7)—this benefit stems from the common parameterisation of all sigmoidal families in terms of threshold and width.[28]

4.6. Higher confidence levels

We thoroughly tested and present our method for the 95% confidence level and checked the results for the 68% as well. Thus we are very sure about our results up to the 95% confidence level.

Testing coverage for higher confidence levels requires more simulations as the expected proportion of parameter estimates lying outside the credible interval shrinks. Thus we could not check coverage for higher confidence levels exhaustively.

However, we tested the 99% confidence level using only a small subset of all binomial simulations[29] with 10,000 repetitions each. For these we found the coverages to be between 99.28% and 100% for threshold and width indicating a slightly larger than nominal coverage than 99%—as for the 95% credible intervals. Furthermore the numerical approximations for 99% credible intervals still seem to be reliable. Thus 99% credible intervals appear to be functional using our approach, at least up to the degree we could test them.

If even higher confidence levels are required one should be careful though, as our numerical approximation might become inaccurate in the tails—a caveat applicable to many numerical methods. To emphasise this, our software issues a warning if users request confidence levels higher than 95%, reminding them that we did not test such confidence levels extensively. For these high confidence levels it might be necessary to change the settings for our numerical method, adding more gridpoints or reducing the threshold for parts of the parameter space to be included in the grid.

4.7. Bayesian inference

The estimation of psychometric functions particularly profits from the Bayesian perspective, for the following three reasons:

First, Bayesian inference provides a principled way of integrating out the uncertainty over parameters which are not well constrained by the data. Thus it works even in scenarios where only some, or even none, of the parameters are constrained by the data. For example, datasets with few blocks hardly constrain η, and datasets with little data in the asymptotic range hardly constrain λ. In such cases fitting a single value for these parameters, as frequentist statistics does, cannot take the uncertainty about other parameters into account. Thus fitting a single parameter value yields unstable estimates for all parameters. In contrast, Bayesian inference integrates the results for the different possible values for the unconstrained parameters yielding sensible estimates and credible intervals for the other parameters.

Secondly, we show that the method profits considerably from the use of carefully pre-determined default priors (see supplementary material S.4 and S.5). Importantly, however, the default priors do not unduly dominate the fits. For most real world datasets we fit very similar parameters values as the original psignifit 2.5 (Fig. 11),

[27] In addition we provide a basic version in python.

[28] The only exception is the heavy tailed cumulative T_1-distribution, which is steeper than the other functions for the same width parameter and thus produce width estimates which are consistently off by a factor; see supplementary material S.8 for details.

[29] We simulated 2AFC and equal asymptote psychometric functions with 50 and 800 trials, using either 5 or 10 linearly spaced blocks.

i.e. the Bayesian point estimates are consistent with earlier purely frequentist analyses; if both methods disagree, however, the new Bayesian method seems preferable.

Finally, Bayesian statistics were previously shown to produce credible intervals with better coverage than those obtained from the bootstrap (Kuss et al., 2005; Fründ et al., 2011). This means that for psychometric functions our Bayesian method even yields better intervals according to a frequentist evaluation criterion.

Apart from its accuracy and reasonable credible interval coverage for single psychometric functions, a further advantage of Bayesian inference is that it allows the posterior distribution to be used in further statistical analyses. We are aware that such inference is often not performed routinely, and most statistical inference in psychological and behavioural science is still done using hypothesis tests, although they are a weak solution (Wagenmakers, 2007; Nuzzo, 2014). But the posterior distributions our method and software calculate could be used by the user, for example, for Bayesian inference on a hierarchical model, predicting the parameters of the psychometric functions across conditions or observers.

4.8. Too large coverage for binomial observer

Our method is conservative in terms of the coverage of the credible intervals, when the data are not overdispersed, i.e. binomial. This is to be expected: For datasets of typical size it cannot be ruled out that there might have been some overdispersion, and the uncertainty about overdispersion leads to (slightly) larger credible intervals. If one is sure that there is definitely no overdispersion in the data, one should use the binomial model[30] yielding less conservative credible intervals. In general, however, yielding too high coverage is more tolerable than yielding too low coverage.

Acknowledgements

We would like to thank previous members of the Wichmann-lab who were involved in developing MCMC based methods of Bayesian inference for the psychometric function, most notably Ingo Fründ, Valentin Hänel, Frank Jäkel and Malte Kuss. Finally, we like to thank Frank Jäkel for fruitful discussion, and Nicole Eichert, Frank Jäkel, David Janssen, Britta Lewke, Lars Rothkegel, Tom Wallis, Uli Wannek, Bei Xiao as well as four anonymous reviewers for helpful comments on the manuscript. This work was supported by the German Research Foundation (DFG; Sachbeihilfe Wi 2103/4–1 to FAW) and the German Federal Ministry of Education and Research (BMBF) through the Bernstein Computational Neuroscience Program Tübingen (FKZ: 01GQ1002).

Appendix A. Mathematical details

A.1. Mathematical description of the model

We use the following model of psychometric functions: We start with a family of sigmoid functions in the stimulus level x, $S(x;m,w)$, with parameters m, the threshold and w the width, e.g. a strictly monotonically growing function $S:\mathbb{R}\to[0,1]$ with two parameters. For all families the parameters are chosen such that $m = S^{-1}(.5)$ and $w = S^{-1}(.95) - S^{-1}(.05)$.

We define the psychometric function ψ adding parameters λ and γ for the upper and lower asymptote:

$$\overline{p}_i = \psi(x;m,w,\lambda,\gamma) = \gamma + (1-\lambda-\gamma)S(x;m,w) \tag{A.1}$$

To generate block i of data with n_i trials at physical level x_i, first a success-probability for this block p_i is drawn from a beta-distribution with mean $\overline{p}_i = \psi(x_i;m,w,\lambda,\gamma)$ and variance $\eta^2\overline{p}_i(1-\overline{p}_i)$ using the fifth and last parameter $\eta\in[0,1]$ which scales the standard deviation of the beta-distribution between the most extreme possible values. Then, n_i Bernoulli-trials with success-probability p_i are drawn, to obtain the number of correct trials k_i, which is thus Binomial distributed with success-rate p_i. In formal short hand this is the following:

$$\begin{aligned} k_i &\sim \mathrm{Binom}(n_i,p_i)\\ p_i &\sim \mathrm{Beta}\left(\left(\frac{1}{\eta^2}-1\right)\overline{p}_i,\left(\frac{1}{\eta^2}-1\right)(1-\overline{p}_i)\right)\end{aligned} \tag{A.2}$$

Using this model we can compute the Likelihood of a parameter $\theta=(m,w,\lambda,\gamma,\eta)$ with $\eta' = \frac{1}{\eta^2}-1$ for N Blocks of data with k_i of n_i correct trials in Block i employing B, the beta function

$$\begin{aligned} L(\theta|data) &= \prod_{i=1}^{N}\binom{n_i}{k_i}\frac{B(k_i+\eta'\overline{p}_i,n_i-k_i+\eta'(1-\overline{p}_i))}{B(\eta'\overline{p}_i,\eta'(1-\overline{p}_i))}\\ &= \prod_{i=1}^{N}\binom{n_i}{k_i}\frac{\Gamma(k_i+\eta'\overline{p}_i)\Gamma(n_i-k_i+\eta'(1-\overline{p}_i))}{\Gamma(n_i+\eta')}\frac{\Gamma(\eta')}{\Gamma(\eta'\overline{p}_i)\Gamma(\eta'(1-\overline{p}_i))}\end{aligned} \tag{A.3}$$

And thus the log-likelihood l with $\gamma = \log\Gamma$ the log-Gamma function, by dropping the constant factor $\binom{n_i}{k_i}$ as all calculations require only a value proportional to the likelihood:

$$\begin{aligned} l(\theta|data) = &\sum_{i=1}^{N}\gamma(k_i+\eta'\overline{p}_i)+\gamma(n_i-k_i+\eta'(1-\overline{p}_i))\\ &-\gamma(n_i+\eta')+\gamma(\eta')-\gamma(\eta'\overline{p}_i)-\gamma(\eta'(1-\overline{p}_i))\end{aligned} \tag{A.4}$$

With these equations we perform standard Bayesian inference and calculate the posterior probability $P(\theta)$ by normalising the prior $p(\theta)$ times the likelihood to have integral 1, e.g. to be a probability:

$$P(\theta|data) = \frac{L(\theta|data)p(\theta)}{\int_\Omega p(\theta)L(\theta|data)d\theta} \tag{A.5}$$

Using these equations we can calculate marginal distributions for the parameters, a mean and a median estimate and also a MAP estimate by optimisation.

Table A.1
The formulae for the seven implemented sigmoid families $S(x;m,w)$. For notation we use x, the stimulus level, m and w, the threshold and the width, Φ and Φ^{-1}, the cumulative standard normal distribution and its inverse, T_1 and T_1^{-1}, the cumulative T-distribution with 1 degree of freedom and its inverse, and log and exp the natural logarithm and exponential functions.

Name	Formula
Cumulative Gauss	$S(x;m,w) = \Phi(C\frac{x-m}{w})$ $C = \Phi^{-1}(.95) - \Phi^{-1}(.05)$
Cumulative T_1-distribution	$S(x;m,w) = T_1(C\frac{x-m}{w})$ $C = T_1^{-1}(.95) - T_1^{-1}(.05)$
Logistic	$S(x;m,w) = \frac{1}{1+e^{-2\log(\frac{1}{.05}-1)\frac{x-m}{w}}}$
Gumbel	$S(x;m,w) = 1-\exp\left(\log(.5)e^{C\frac{x-m}{w}}\right)$ $C = \log(-\log(.05)) - \log(-\log(.95))$
Reversed Gumbel	$S(x;m,w) = \exp\left(\log(.5)e^{C\frac{x-m}{w}}\right)$ $C = \log(-\log(.95)) - \log(-\log(.05))$
Weibull	$S(x;m,w) = 1-\exp\left(\log(.5)e^{C\frac{(\log(x)-m)}{w}}\right)$ $C = \log(-\log(.05)) - \log(-\log(.95))$
Cumulative log-normal	$S(x;m,w) = \Phi\left(C\frac{\log(x)-m}{w}\right)$, $C = \Phi^{-1}(.95) - \Phi^{-1}(.05)$

[30] Our software allows users to fit the standard binomial model as well.

A.2. Binomial mixture model

In Section 2.1 and Eq. (1) we introduced the psychometric function as an appropriately scaled sigmoid function. This interpretation of a scaled sigmoid function is mathematically equivalent to a binomial mixture model (Kuss et al., 2005), as we show here for completeness.

In the binomial mixture model the proportion correct answers of the observer results from two independent Bernoulli processes: First, with a probability π_l the observer guesses independently of stimulus intensity, and has a probability π_c of guessing correctly. The probability π_c depends on the experimental design and is typically $\pi_c = 1/N$ where N denotes the number of response alternatives, i.e. $N = 2$ in 2AFC (For single-interval YesNo tasks, we expect π_c to be near 0). Second, with a probability $1 - \pi_l$ the observer attempts to solve the task the best she can; in this case the probability for a correct answer is $\psi(x; m, w, \pi_c) = \pi_c + (1 - \pi_c)S(x; m, w)$. Taken together these results in a Bernoulli variable with a probability of success as a function of the stimulus intensity x given by

$$\begin{aligned}\psi(x; m, w, \pi_l, \pi_c) &= (1 - \pi_l)[\pi_c + (1 - \pi_c)S(x; m, w)] + \pi_l\pi_c \\ &= (1 - \pi_l - \pi_c + \pi_l\pi_c)S(x; m, w) + (1 - \pi_l)\pi_c + \pi_l\pi_c \\ &= (1 - \pi_l - \pi_c + \pi_l\pi_c)S(x; m, w) + \pi_c \\ &= \pi_c + (1 - \pi_c - \pi_l(1 - \pi_c))S(x; m, w) \\ &= \gamma + (1 - \gamma - \lambda)S(x; m, w) \end{aligned} \quad (A.6)$$

If, in the second to last line, we substitute $\gamma = \pi_c$ and $\lambda = \pi_l(1 - \pi_c)$, we obtain the last line, which shows the equivalence of the binomial mixture model to the scaled asymptote formulation of equation No. A.1.

Appendix B. Computational details

The numerical method for finding the posterior works as follows: we first evaluate the likelihood and the prior on a grid over all parameters, compute the integral over the whole grid from these values and divide by this integral. In practice there are some choices to be made for this method though, for example how many grid points over which part of the parameter space to calculate.

After choosing these parameters of the method our final implementation still spends more than 90% of its running time evaluating the likelihood on the grid of possible parameter values. Thus decreasing the number of necessary gridpoints and optimising the evaluation of the likelihood was the main focus of our computational optimisations, which we describe and evaluate in this appendix.

B.1. Implementation details

First we need to choose the endpoints or borders of the grid. As an initial guess for the boundaries we can use the boundaries of the prior support (as we use priors with bounded support these boundaries are always finite). However, large parts of this area will have very low posterior probability (as they will be very unlikely under the likelihood model) and thus will contribute neither to the normalisation integral nor to the credible intervals. To avoid evaluation of the integral over areas with vanishingly low probability, we do a first approximation of the posterior with a sparse grid and remove parts with low posterior probability. For this calculation we use an initial grid with 20 grid points for threshold and width and 10 for each of the other parameters. We then find all gridpoints for which the area they represent has posterior marginal probability larger than $\frac{1}{1000}$ and choose the last gridpoint before these and the first one after them as new boundaries.

This refinement of the borders yields a marked improvement of the accuracy achievable with a given number of grid points per dimension. Preliminary simulations confirmed that this truncation of the integration region did not alter results noticeably.

Grid points are spaced linearly in parameter space between borders, and we were not able to identify non-linear spacings which consistently yielded superior results. For the Weibull and the cumulative normal this means that the tested thresholds and widths are placed logarithmically on the stimulus axis, as the parameters are defined on the logarithm of the stimulus.

As we use an orthogonal grid on the parameter space, many grid points share the same threshold and width. Thus evaluating the sigmoid function only once and reusing the value for the computation saves considerable computation time. Additionally the likelihood evaluation can be vectorized on a grid, which is very efficient in higher level programming languages and for parallelization. Together these improved the computation times by a factor of several hundred compared to separate computation for the different grid points.

Evaluation of the likelihood according to the beta-binomial model formulas becomes numerically unstable for small values of η, because the parameters of the beta-distribution diverge $\left(\lim_{\eta \to 0}\left(\frac{1}{\eta^2} - 1\right) = +\infty\right)$. As, for very small values of η the beta-binomial model is indistinguishable from a binomial model, we calculate the probability according to the binomial model for $\eta^2 < 10^{-9}$.

Once the likelihood is evaluated we use simple trapezoid addition to calculate the integral over the grid. While there are more sophisticated approaches for numerical integration like sparse grids (Gerstner & Griebel, 1998; Heiss & Winschel, 2008), we chose not to use such methods, because a full grid makes marginalizations and the visualisation of the actually integrated function easier.

To obtain the marginal distributions and point estimates, we calculate marginal distributions simply as a sum over the probability mass. Then we can compute the mean as the sum over the gridpoints weighted by their probability and the median by approximating the marginal distribution by the linear interpolation of the computed discrete values. For the MAP we start a Nelder-Mead simplex algorithm (Lagarias, Reeds, Wright, & Wright, 1998) as implemented in MATLAB (2012) as "fminsearch" from the maximum on the grid.

B.2. Tests of numerical approximation

The number of gridpoints on each dimension determines the trade-off between computation time and numerical accuracy. Here we chose the number of gridpoints based on the accuracy of the credible intervals obtained. We assessed this accuracy with two sets of simulations, first changing the number of grid points separately per parameter, then simultaneously for all parameters (Fig. B.13).

For each simulation, we first simulated 25 datasets for each experiment type, 50, 400 and 1000 trials, and the Weibull and the cumulative Gaussian. For the first simulation we fitted them changing the number of gridpoints on each dimension from 10 to 150 (100 for Yes No) in steps of 10 while keeping the others constant at 75 gridpoints. For the second simulation we kept the ratio of grid point numbers constant and changed the absolute number for threshold and width from 20 to 60 in steps of 2 resulting in steps of 1 from 10 to 30 for the other parameters. In total we simulated 450 datasets for each simulation and ran 48,750 + 9450 fits to choose the number of grid points.

The results of the simulations displayed in Fig. B.13 show that the credible interval estimates converge for increasing number of

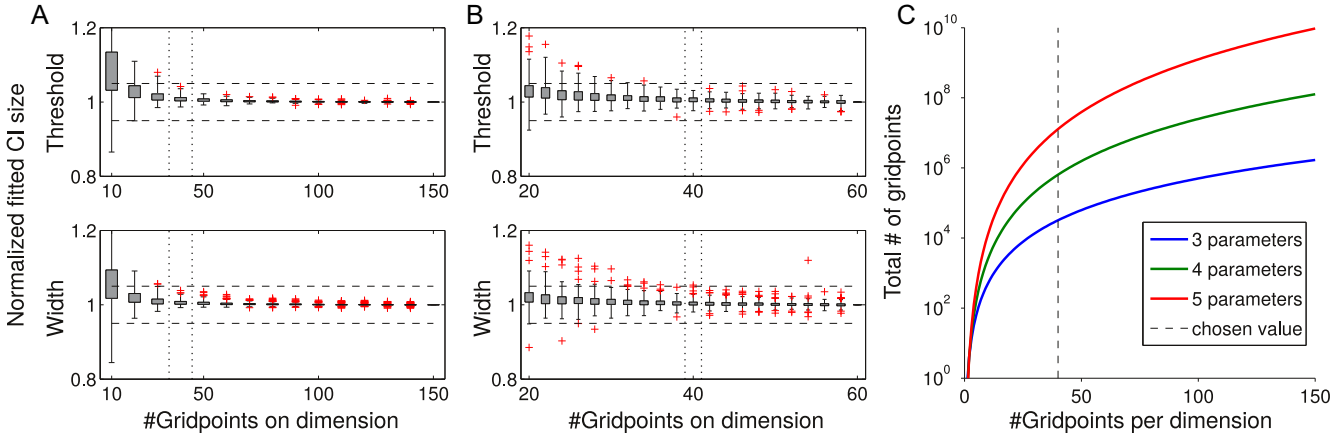

Fig. B.13. Tests of the numerical approximations used. A: accuracy of the credible interval estimates, changing sampling per dimension. Plotted is the size of the computed credible interval divided by the size computed with 150 gridpoints for 50 datasets × 3 sample sizes × 3 experiment types each, against the number of gridpoints used in the fit, changing the gridpoints of one dimension only. The dashed lines mark 5% error, which was the maximum we tolerated, the dotted lines mark the chosen value. B: as in A, but changing the number of gridpoints on all dimensions simultaneously using a fixed ratio between dimensions based on the result displayed in A. C: total number of gridpoints against the number of gridpoints on one dimension for fits with 3, 4 or 5 parameters. The number of gridpoints scales linearly with computation time. Roughly 10^7 datapoints take one second per datapoint to compute on an office computer.

grid points. We used the result with the maximum number of gridpoints as a proxi for the true value. Using this proxy, we find that the maximal errors in credible interval size were smaller than 5% whenever we used more than 40 grid points on threshold and width and 20 on the others, and we therefore chose these numbers for further simulations and as the default settings of the software.

We also analysed the numerical accuracy results for credible intervals with coverage from 15% to 95% levels in steps of 10% and at 99%. The results were very similar to the ones for 95% credible intervals we present here. Thus we can confirm the method to work for other credible interval sizes as well, but skip their detailed presentation here for brevity.

Using 40 grid points for threshold and width and 20 for the other 3 parameters, the calculation of the posterior takes roughly 1 s per data block on an ordinary office computer in the case of a full 5 parameter model and.05 or.0025 s for models with 4 or 3 parameters. (nAFC experiments with/without beta-binomial model), which is substantially faster than currently available MCMC based methods (Kuss et al., 2005). For standard data sizes the computations thus take only a few seconds even for the 5 parameter model and can be almost instantaneous for fewer free parameters.

To confirm that the numerical integration method and more standard MCMC methods yield similar results we ran an adaptive Metropolis Hastings algorithm to sample the posterior for each of the 18 combinations of the three sample sizes, experiment types and the two sigmoid shapes. For each psychometric function we ran 10 adaptive Metropolis Hastings chains (Haario, Saksman, & Tamminen, 2001), in which we adjusted the covariance matrix of a Gaussian proposal to be the covariance of the previous 90% of the samples after every 1000 samples, adapting the size by upscaling the proposal distribution by $\exp\left(\frac{15}{i+50}\right)$ after accepting the i-th proposal, and downscaling by $\exp\left(-\frac{5}{i+50}\right)$ after denying the ith proposal. Thus the acceptance probability converged to 25% (near the optimal value for Gaussians of 27.9/27.5% found by Gelman, Roberts, & Gilks, 1996). After adapting for 50 000 samples we left each chain running for another 50,000 samples with constant proposal distribution and kept only these second 50 000 trials as an approximation of the posterior. This MCMC method yielded and effective sample size of 9634–25,233 from the 500,000 samples per psychometric function.

The MCMC results confirmed the results we obtained from the numerical analysis, i.e. the confidence intervals from the MCMC analysis differed less than 5% from the numerical integration results in each of the 18 cases. We then sub-sampled the posterior samples independently to obtain an estimate how many samples one needs to estimate the credible intervals. Drawing 100 sets of samples for each size and psychometric function we found that roughly 5000 independently drawn samples were necessary to obtain estimates with less than 5% error in credible intervals size. For a Metropolis Hastings algorithm this corresponds to at least 100 000 samples, as Gelman et al., 1996 found that it can maximally reach an efficiency of 6.2%/7.6% even for a Gaussian distribution in 5/4 dimensions.

Drawing 100, 000 samples for a demo dataset took roughly 60 s on a usual desktop computer compared to 20 s for a 5 dimensional numerical integral and 1 s for a 4 dimensional numerical integral. Thus with our implementation the numerical integration is indeed considerably faster. However an implementation more optimised for single parameter values and a well tuned MCMC method could probably beat the numerical integration method for the 5 dimensional integral, while beating the method for the 4 or lower dimensional integrals would be challenging. Furthermore the numerical integration method saves the hassle of checking convergence and mixing of MCMC chains.

Appendix C. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.visres.2016.02.002.

References

Alcala-Quintana, R., & Garcia-Perez, M. A. (2004). The role of parametric assumptions in adaptive bayesian estimation. *Psychophysical Methods, 9*(2), 250–271.
Bird, C., Henning, G., & Wichmann, F. (2002). Contrast discrimination with sinusoidal gratings of different spatial frequency. *JOSA A, 19*(7), 1267–1273.
Blackwell, H. R. (1952). Studies of psychophysical methods for measuring visual thresholds. *JOSA, 42*(9), 606–614.
Collett, D. (2002). *Modelling binary data*. CRC Press.
Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics, 7*, 1–26.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap. Monographs on statistics and applied probability*. Chapman & Hall/CRC.
Foster, D. H., & Bischof, W. F. (1987). Bootstrap variance estimators for the parameters of small-sample sensory-performance functions. *Biological Cybernetics, 57*, 341–347.
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.
Fründ, I., Haenel, N. V., & Wichmann, F. A. (2011). Inference for psychometric functions in the presence of nonstationary behavior. *Journal of Vision, 11*(6), 1–19.
Fründ, I., Wichmann, F. A., & Macke, J. H. (2014). Quantifying the effect of intertrial dependence on perceptual decisions. *Journal of Vision, 14*(7), 1–16. 9.
García-Pérez, M. A. (1998). Forced-choice staircases with fixed step sizes: Asymptotic and small-sample properties. *Vision Research, 38*(12), 1861–1881.
Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Chapman & Hall/CRC Press.
Gelman, A., Roberts, G., & Gilks, W. (1996). *Efficient metropolis jumping rules. Bayesian statistics, 5*(599–608), 42.
Gerstner, T., & Griebel, M. (1998). Numerical integration using sparse grids. *Numerical Algorithms, 18*, 209–232.
Green, D. M. (1964). Consistency of auditory judgements. *Psychological Review, 71*(5), 392–407.
Green, D. M. (1990). Stimulus selection in adaptive psychophysical procedures. *Journal of the Acoustical Society of America, 87*(6), 2662–2674.
Haario, H., Saksman, E., & Tamminen, J. (2001). An adaptive metropolis algorithm. *Bernoulli*, 223–242.
Heiss, F., & Winschel, V. (2008). Likelihood approximation by numerical integration on sparse grids. *Journal of Econometrics, 144*, 62–80.
Henning, G., Bird, C., & Wichmann, F. (2002). Contrast discrimination with pulse trains in pink noise. *JOSA A, 19*(7), 1259–1266.
Henning, G. B., Millar, R. W., & Hill, N. J. (2000). Detection of incremental and decremental bars at different locations across Mach bands and related stimuli. *Journal of the Optical Society of America A, 17*(7), 1147–1159.
Hill, J. (2002). *Testing hypotheses about psychometric functions* (Ph.D. thesis).UK: St. Hugh's College, University of Oxford.
Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review, 21*(5), 1157–1164.
Howarth, C. I., & Bulmer, M. G. (1956). Non-random sequences in visual threshold experiments. *Quarterly Journal of Experimental Psychology, 8*, 163–171.
Jäkel, F., & Wichmann, F. A. (2006). Spatial four-alternative forced-choice method is the preferred psychophysical method for Naïve observers. *Journal of Vision, 6*(11), 1307–1322.
Jaynes, E. T. (2003). *Probability theory*. Cambridge: Cambridge University Press.
Knoblauch, K., & Maloney, L. T. (2012). *Modeling psychophysical data in R. Springer*.
Kontsevich, L. L., & Tyler, C. W. (1999). Bayesian adaptive estimation of psychometric slope and threshold. *Vision Research, 39*(16), 2729–2737.
Kruschke, J. (2014). *Doing Bayesian data analysis* (2nd ed.). A Tutorial with R, JAGS, and Stan. Academic Press.
Kuss, M., Jäkel, F., & Wichmann, F. A. (2005). Bayesian inference for psychometric functions. *Journal of Vision, 5*, 478–492.
Lagarias, J., Reeds, J. A., Wright, M. H., & Wright, P. E. (1998). Convergence properties of the Nelder–Mead simplex method in low dimensions. *SIAM Journal of Optimization, 9*(1), 112–147.
Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *Journal of the Acoustical Society of America, 49*(2), 467–477.
Maertens, M., & Wichmann, F. A. (2013). When luminance increment thresholds depend on apparent lightness. *Journal of Vision, 13*(6), 1–11. 21.
Maloney, L. T. (1990). Confidence intervals for the parameters of psychometric functions. *Perception & Psychophysics, 47*(2), 127–134.
MATLAB, 2012. 7.14.0.739 (R2012a). The MathWorks Inc., Natick, Massachusetts.
McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (Vol. 37). CRC Press.
Nuzzo, R. (2014). Statistical errors. *Nature, 506*(7487), 150–152.
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 and Psychophysics, 45*, 434–442.
Pelli, D. G. (1987). The ideal psychometric procedure. In: *Investigative ophthalmology & visual science. Vol. 28* (Suppl).
Prentice, R. (1986). Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors. *Journal of the American Statistical Association, 81*(394), 321–327.
Schönfelder, V. H., & Wichmann, F. A. (2013). Identification of stimulus cues in narrow-band tone-in-noise detection using sparse observer models. *Journal of the Acoustical Society of America, 134*(1), 447–463.
Senders, V., & Sowards, A. (1952). Analysis of response sequences in the setting of a psychophysical experiment. *The American Journal of Psychology, 65*(3), 358–374.
Shen, Y., & Richards, V. M. (2012). An updated maximum-likelihood procedure: Thresholds, slopes, and lapses of attention. *Journal of the Acoustical Society of America, 132*(2), 957–967.
Treutwein, B. (1995). Adaptive psychophysical procedures. *Vision Research, 35*(17), 2503–2522.
Treutwein, B., & Strasburger, H. (1999). Fitting the psychometric function. *Perception & Psychophysics, 61*(1), 87–106.
Venables, W. N., & Ripley, B. D. (2013). *Modern applied statistics with S-PLUS*. Springer Science & Business Media.
Verplanck, W. S., Collier, G. H., & Cotton, J. W. (1952). Non-independence of successive responses in measurements of the visual threshold. *Journal of Experimental Psychology, 44*(4), 273–282.
Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems OFP values. *Psychonomic Bulletin & Review, 14*(5), 779–804.
Watson, A. B., & Pelli, D. G. (1983). Quest: A bayesian adaptive psychometric method. *Perception & Psychophysics, 33*(2), 113–120.
Wetherill, G., & Levitt, H. (1965). Sequential estimation of points on a psychometric function. *British Journal of Mathematical and Statistical Psychology, 18*(1), 1–10.
Wichmann, F. A. (1999). *Some aspects of modelling human spatial vision: contrast discrimination* (Ph.D. thesis). University of Oxford.
Wichmann, F. A., & Hill, N. J. (2001a). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception and Psychophysics, 63*(8), 1293–1313.
Wichmann, F. A., & Hill, N. J. (2001b). The psychometric function: II. Bootstrap-based confidence intervals and sampling. *Perception and Psychophysics, 63*(8), 1314–1329.
Williams, D. A. (1982). Extra-binomial variation in logistic linear models. *Applied Statistics*, 144–148.
Żychaluk, K., & Foster, D. H. (2009). Model-free estimation of the psychometric function. *Attention, Perception, & Psychophysics, 71*(6), 1414–1425.