Inference for psychometric functions in the presence of nonstationary behavior

Ingo Fründ

N. Valentin Haenel

Felix A. Wichmann

Measuring sensitivity is at the heart of psychophysics. Often, sensitivity is derived from estimates of the psychometric function. This function relates response probability to stimulus intensity. In estimating these response probabilities, most studies assume stationary observers: Responses are expected to be dependent only on the intensity of a presented stimulus and not on other factors such as stimulus sequence, duration of the experiment, or the responses on previous trials. Unfortunately, a number of factors such as learning, fatigue, or fluctuations in attention and motivation will typically result in violations of this assumption. The severity of these violations is yet unknown. We use Monte Carlo simulations to show that violations of these assumptions can result in underestimation of confidence intervals for parameters of the psychometric function. Even worse, collecting more trials does not eliminate this misestimation of confidence intervals. We present a simple adjustment of the confidence intervals that corrects for the underestimation almost independently of the number of trials and the particular type of violation.

Keywords: inference, psychometric function, nonstationary behavior


Introduction

Sensitivity measurements form the basis of psychophysical research. Often, sensitivity is inferred from estimates of the psychometric function, the function that relates an observer’s performance to the intensity of a presented stimulus. In most cases, the psychometric function increases monotonically with stimulus intensity. Absolute sensitivity is inversely related to the horizontal shift of the psychometric function along the stimulus intensity axis. To refer to the “horizontal shift,” physicists often use the word “threshold,” being aware of the fact that theories that assume a hard threshold cannot explain decision behavior in psychophysical tasks (Swets, 1961). The psychometric function also provides information about sensitivity to differences or relative sensitivity: If the psychometric function is very steep, this indicates that an observer can discriminate small stimulus differences; if the psychometric function is very shallow, the observer can only discriminate relatively coarse differences. Different procedures have been proposed to determine thresholds. A number of authors have proposed adaptive procedures (Alcalá-Quintana & García-Pérez, 2005; Cornsweet, 1962; Dixon & Mood, 1948; García-Pérez & Alcalá-Quintana, 2007; Kontsevich & Tyler, 1999; Levitt, 1970; Watson & Pelli, 1983); others have used the method of constant stimuli (Treutwein & Strasburger, 1999; Ulrich & Miller, 2004; Zychaluk & Foster, 2009). One approach that has proven particularly successful consists of (a) recording a block of trials for each stimulus intensity of interest (Blackwell, 1952) and (b) fitting a parametric model to the response counts in each block (Kuss, Jäkel, & Wichmann, 2005; Wichmann & Hill, 2001a).

Fitting such a model makes two critical assumptions (Treutwein & Strasburger, 1999, p. 102):

1. It is assumed that responses depend on stimulus intensity only. Factors such as learning, fatigue, and fluctuations of attention are neglected.
2. It is assumed that responses follow a binomial distribution, where the success probability is given by the psychometric function model. This implies that the single trials are assumed to be independent of each other. Thus, any tendency of an observer to adjust his or her decision criterion after a trial is neglected (Lages & Treisman, 1998; Treisman & Williams, 1984). Although such criterion
effects might only be expected for single-interval designs (yes–no experiments are the most common example, but the argument also holds for single-interval identification tasks), recent evidence suggests that forced-choice tasks may also show bias (Yeshurun, Carrasco, & Maloney, 2008). In addition, Jäkel and Wichmann (2006) noted several inconsistencies in forced-choice tasks, which cast doubt on the assumption that forced-choice tasks provide a criterion-free measure of performance. It is clear that both assumptions are most likely violated: Observers learn during experiments, they experience fatigue, and they might even adapt their response strategies. In statistical terms, these and other violations can be summarized by the term “non-stationarity”: the distribution of the observers’ responses changes over the run of the experiment. Unfortunately, we neither know whether nonstationarity has an impact on psychometric function estimates at all nor how severe such an effect would be. It might well be that point estimates of parameters of the psychometric function are still valid, i.e., that a point estimate represents the average performance over time or an upper asymptotic performance level. Many nonstationary effects are likely, however, to lead to increased variance of the data. If this increased variability of the observers’ responses is not accounted for, researchers might draw the wrong conclusions from their data: The estimated confidence intervals around the point estimates are too small, and researchers judge differences between experimental conditions in relation to these confidence intervals. Thus, spurious differences might seem significant although they are, in fact, nothing more than random variations.

We performed a series of simulations to investigate the effect of nonstationarity on psychometric function estimates. We first report the results of these simulations. Second, we study a number of diagnostic tools to detect nonstationarity in psychophysical response data. Finally, we suggest a procedure for correcting the estimated confidence intervals for nonstationarity.

Simulated observers

We simulated three different observers. The first observer, termed binomial observer, was in perfect accordance with the typical assumptions about a stationary psychophysical observer. The other two observers violated common assumptions about psychophysical observers, displaying different forms of nonstationarity.

Binomial observer

For the binomial observer, responses were sampled from a stable, nonchanging binomial distribution with success probabilities given by the psychometric function. Thus, success probabilities of this observer were only dependent on the stimulus intensity and all responses were independent. This observer represents the optimal case for the estimation of a psychometric function. In addition, it is the model that is typically assumed in methods that are designed to obtain estimates of the psychometric function. There are methods available to analyze responses from this observer (Kuss et al., 2005; Wichmann & Hill, 2001a, 2001b). It has been shown that for binomial observers psychometric functions can be fitted successfully and that error estimates are reasonably valid (Hill, 2001). More details about this observer can be found in Appendix A.1. The psychometric function for this observer is shown in Figure 1a.

Figure 1. Psychometric functions for simulated observers. (a) Binomial observer: Response probabilities were always taken from the same psychometric function. (b) Learning observer: The psychometric function becomes steeper and thresholds decrease during the experiment. The plot shows psychometric functions separated by 50 trials. (c) Betabinomial observer: Response probabilities are selected from a Beta distribution centered on the psychometric function. The shading indicates the probability to select the respective response probability. The solid line marks the mode of the Beta distribution, which is equal to the psychometric function of the binomial observer.
Learning observer

A common source of nonstationarity is learning. In particular, untrained observers often learn in the beginning of an experiment (Jaekel & Wichmann, 2006). We simulated learning by continuously changing the parameters of the psychometric function: Over the time course of the experiment, the psychometric function became steeper while thresholds slowly decreased. Learning approached an asymptote. After 184 trials, parameters remained within 1% around their final level. Learning was independent of the encountered stimuli and of the performed responses. Learning parameters were set such that the observer initially performed worse than the binomial observer but performed better than the binomial observer close to the end of the experiment. During the experiment, the threshold decreased to 70% of its initial level; the width of the nondecision interval w (see Figure 2) decreased to 48% of its initial level. Averaged over the largest experiment, the parameters of the learning observer’s psychometric function were the same as for the binomial observer. More details about this observer can be found in Appendix A.2. Sample psychometric functions for this observer are shown in Figure 1b.

Betabinomial observer

A second source of nonstationarity are fluctuations in the observer’s attention and motivation or changes in the stimulus features used by the observer to complete the task. It is difficult to find a simple mechanistic and psychologically plausible model for these fluctuations. We decided to model fluctuations in the observer’s attention and motivation probabilistically: For each block of trials, we selected a response probability from a Beta distribution centered on the psychometric function. The response probability for each block could, in principle, be any value between 0 and 1, but values close to the psychometric function were more likely. Figure 1c illustrates this point: The shading of a point represents the probability of sampling the response probability on the y-axis given that a stimulus with the intensity on the x-axis is presented. The mode of the Beta distribution at each stimulus level was selected equal to the psychometric function of the binomial observer. However, the standard deviation of the response counts was increased by approximately 15%. More details about this observer can be found in Appendix A.3.

The properties of the nonstationarities clearly depend on the parameters chosen for the learning and betabinomial observers. The parameters used for the simulations were not specifically adapted to experimental data. However, we will show later that we roughly reproduce experimentally observed nonstationarities.

Psychometric function model and fitting

The psychometric function was modeled as a binomial mixture model as suggested by Wichmann and Hill (2001a):

$$\Psi(x; \theta) = \gamma + (1 - \gamma - \lambda) F(x; \theta),$$

with a parameter vector $\theta = (m, w, \gamma, \lambda)$. The first two parameters $m$ and $w$ determine the shape of the psychometric function; $\gamma$ is the guessing rate and $\lambda$ is the lapse rate. The sigmoidal function $F: \mathbb{R} \to [0, 1]$ is required to be monotonically increasing. The meaning of these parameters is illustrated in Figure 2.

Model parameterization

Three different sigmoidal functions were used to analyze the simulated data: The logistic function, the Weibull cumulative distribution function, and the reversed Gumbel cumulative distribution function. These three sigmoidal functions were used to capture different symmetries of the psychometric function model: The logistic function is point symmetric with respect to its inflection point—the upper and lower asymptotes are approached equally fast. The Weibull cumulative distribution function is asymmetric—it approaches the lower asymptote faster than the upper asymptote. The reversed
Gumbel cumulative distribution function is also asymmetric. However, in this case, the upper asymptote is approached faster than the lower one. Depending on the sigmoidal function, the first two parameters, \( m \) and \( w \), have different interpretations.

The logistic function is given by

\[
F(x; \theta) = \left(1 + \exp\left(-\frac{z(a)}{w}(x - m)\right)\right)^{-1},
\]

with \( z(a) = 2\log(1/a - 1) \). In this case, \( m \) is the horizontal position of the sigmoid, the point at which the sigmoid is halfway between its lower asymptote and its upper asymptote. In many studies, \( m \) is taken to be the observers’ threshold estimate. The second parameter \( w \) is the width of the interval in which the sigmoid rises from \( a \) to \( 1 - a \). This is the width of the “interesting” region of the sigmoid, where changes in intensity result in large behavioral changes. The width is measured in units of stimulus intensity. We selected \( a = 0.1 \). The logistic function was used as the generating sigmoid for all simulated observers presented in this paper. In addition, it is the standard model for our analyses.

The Weibull function is given by

\[
F(x; \theta) = 1 - \exp\left(-\left(\frac{x}{m}\right)^w\right).
\]

Again, \( m \) is related to the horizontal position of the sigmoid, while \( w \) is inversely related to the width of the interval on which the function rises. However, this relation is less obvious in this case: The \( m \) parameter does not correspond to a pure horizontal shift, and the \( w \) parameter changes the shape of the function instead of simply scaling the sigmoid along the \( x \)-axis.

Finally, the reversed Gumbel function is given by

\[
F(x; \theta) = \exp\left(-\exp\left(\frac{z(1 - a) - z(a)}{w}(x - m) + z(0.5)\right)\right),
\]

where \( z(a) = \log(-\log(a)) \). In this case, the interpretation of the parameters \( m \) and \( w \) is again precisely the same as for the logistic function: \( m \) is the horizontal position of the sigmoid and \( w \) is the width of the interval in which the sigmoid rises from \( a \) to \( 1 - a \).

Fitting procedures

Two different fitting procedures were used: (a) A maximum likelihood procedure combined with bootstrap sampling to determine confidence intervals as suggested by Maloney (1990) as well as Wichmann and Hill (2001a, 2001b) and (b) a Bayesian inference based on samples from the posterior distribution as suggested by Kuss et al. (2005). In performing Bayesian inference, we assumed prior distributions for the model parameters \( m \sim N(0, 100) \), \( w \sim \text{Gamma}(1.01, 2000) \), \( \lambda \sim \text{Beta}(2, 50) \) (\( \gamma \sim \text{Beta}(1, 10) \)). Both fitting procedures can be used to obtain estimates of the psychometric function as well as credible intervals (i.e., confidence intervals for bootstrap sampling and posterior intervals for Bayesian inference). Details about the fitting procedures can be found in Appendix B.

Coverage

One primary interest of the current study was how well credible intervals for parameters of the psychometric function match the desired 95%. This can be quantified in terms of coverage: The coverage of a credible interval is the probability that the credible interval contains the true value. To estimate this probability, we simulated each combination of observer and fitting procedure 1000 times. From these 1000 repetitions, we calculated the fraction of simulations for which the generating parameters (of the observer) were actually within the estimated credible intervals. Thus, all reported coverages refer to fractions of 1000 repetitions in which the estimated credible intervals contained the generating parameters of the observer. This was repeated for different numbers of blocks and different block sizes. In total, we determined \( 3 \times 2 \times 3 \times 3 \) (observers: binomial, betabinomial, learning) \( \times 3 \) (psychometric function models: logistic, Weibull, reversed Gumbel) \( \times 2 \) (tasks: yes–no or 2AFC) \( \times 3 \) (inference paradigms: parametric bootstrap, nonparametric bootstrap, Bayesian inference) \( \times 1000 \) (repetitions) = 52,000 credible interval estimates for each combination of block size (10, 20, or 40 trials per block) and number of blocks (4, 6, 8, 12, or 24 blocks), totaling 780,000 estimated credible intervals with 40 to 960 trials per psychometric function.

Coverage for nonstationary observers

Figure 3 shows the coverage of credible interval estimates from different types of observers in a simulated two-alternative forced-choice (2AFC) task. The pattern is similar for all three types of inference: Coverage of Bayesian posterior intervals for a binomial observer approaches 95% at the latest after 100 trials but remains too low for both types of nonstationary observers. Bootstrap confidence intervals generally have coverages below 95%; however, the underestimation is clearly less severe for the binomial observer. This underestimation of bootstrap confidence intervals has already been reported by...
Wichmann and Hill (2001b) and Hill (2001). We will return to this issue below. The fact that coverage remained too low for the nonstationary observers means that credible intervals are seriously underestimated for these observers: instead of 95%, the credible intervals include the generating parameters only in about 60%. If we were to compare psychometric functions from two different conditions based on these underestimated confidence intervals, we would “find significant differences” between the experimental conditions in about 40% of the cases where, in fact, no difference exists! In order to get an intuition for the severity of this effect, we derived a scaling factor for the credible intervals. If the simulated parameters followed a normal distribution, the scaling factor would be the factor by which the credible intervals would have to be scaled to achieve 95% coverage is marked on the right axis (in deriving this scale, the parameter distribution was assumed to be normal, which is generally not true).

Figure 3 illustrates another important point about the detection of nonstationarity. With larger blocks, coverage is worse than with smaller blocks. In experiments with 40 trials per block, coverages go to levels as low as 50% (scaling factors as high as 2.5, betabinomial observer, diamond symbols). In contrast with the same total number of trials, smaller blocks had considerably higher coverage (up to 20% more coverage for the betabinomial observer with 10 trials per block, circles in Figure 3). These effects could also be observed for the learning observer. However, they were less severe here (60% coverage with 40 trials...
per block, improvement with 10 trials per block between 10% and 15%).

Another observation in Figure 3 is that confidence intervals derived from the two bootstrap procedures have coverages below 95% even for a binomial observer. This is particularly true for the threshold \( m \). Wichmann and Hill (2001b) deal with this problem by exploring the likelihood function in a neighborhood of the maximum likelihood estimate and extending their confidence limits accordingly. Here, we see that an alternative solution might be estimation of Bayesian posterior intervals: Bayesian posterior intervals have the desired coverage of 95% already after 80 trials (40 trials for threshold only). Although bootstrap confidence intervals for the width \( w \) of the interval on which the psychometric function rises have coverages close to 95%, the underestimation of the confidence intervals is also obvious in this case. The nonparametric bootstrap procedure does not assume any dependencies between different blocks in the resampling process. Thus, one would expect the nonparametric bootstrap to be more robust with respect to violations of the binomiality assumption that goes into the fitting process. However, this is not the case. The two bootstrap procedures do not differ much with respect to their coverages.

We also determined credible intervals for simulated yes–no tasks (data not shown). In these cases, the psychometric function model had an additional free parameter, the guessing rate \( \gamma \). Coverages for the simulated yes–no task showed a similar pattern as for the 2AFC task: Coverage for the binomial observer was close to 95% in the Bayesian setting and approached 95% with increasing number of trials for the two bootstrap procedures; coverages for the two nonstationary observers were clearly too low. Although coverages for the two bootstrap procedures approached 95%, it remained between 90% and 95% for above 400 trials (200 trials if only width is considered) and did not reach 95% for any number of simulated trials. Importantly, coverage determined for the learning observer was no better than for a betabinomial observer. This highlights the fact that there is no ordering between the different observers: The learning observer does not generally result in better credible intervals than the betabinomial observer. This clearly depends on the parameters used for our simulations of the simulated observers. The choice of parameters is justified in detail in the Discussion section. In experiments with 40 trials per block, coverages as low as 55% (scaling factor nearly 3) were observed in the worst case.

The data reported so far refer to the unrealistic scenario that we know the analytical form of the psychometric function in advance. In this case, estimation of the psychometric function consists of estimating a number of unknown parameters. However, in a real experiment, we will not know the analytical form of the psychometric function. Thus, in most realistic cases, the model that we fit to the data will not match the process that generated the data. To simulate this situation, we created a mismatch between the sigmoidal function \( F \) (see Equation 1) for generating and analyzing the data. The generating sigmoidal function was, as always, the logistic. The analyzing sigmoidal functions were the Weibull cumulative distribution function (see Equation 3) and the Gumbel on reversed \( x \)-axis (see Equation 4). If the sigmoidal function used to analyze the data differed from the sigmoidal function used to generate the data, coverage was slightly worse for low numbers of trials (10% less coverage with less than 100 trials per psychometric function). All other observations were still valid: coverage was around 70–90% for nonstationary observers, it was worse with smaller blocks, and Bayesian inference resulted in more realistic credible intervals than the two bootstrap procedures. Results of all other analyses reported here were virtually the same. Therefore, we only report results for those cases in which the analytical form of the generating and analyzing psychometric functions was the same.

### Diagnostics for nonstationarity

In the previous section, we demonstrated that nonstationarity can result in serious underestimation of credible intervals. It is, thus, important to be able to detect data sets that display nonstationarity in order to treat these data sets with special caution. Therefore, we derived a number of tests for nonstationary behavior. Two different diagnostics for nonstationarity have been suggested by Wichmann and Hill (2001a). These are easily applicable in a maximum likelihood setting. However, these diagnostics do not directly map to a Bayesian inference framework. We will first review the procedures suggested by Wichmann and Hill (2001a) and then discuss possible extensions of these ideas to a Bayesian inference framework. In a third step, we will present an alternative approach to diagnosing nonstationarity that is based on the estimation of influential observations.

### Nonstationarity in a maximum likelihood setting

As a very global measure of the goodness of fit for a psychometric function model, Wichmann and Hill (2001a) suggest to simulate data from the fitted model using the assumptions associated with a binomial observer and re-fit the model to these simulated data again. For each of these fits, the deviance is calculated. Deviance is a generalization of the sum-of-squares error metric that also applies to binary/binomial data (e.g., Dobson & Barnett, 2008). We derived deviance from 2000 simulated data sets from each fitted psychometric function and compared the deviance of the fit to the original data set with these.
simulated data sets. Whenever the deviance of the fit to the original data set exceeds the 95th percentile of the deviances from the fits to the simulated data sets, this suggests larger variance than would be expected for a binomial observer. We quantified the power of this test by analyzing how well the test discriminated between the binomial and betabinomial observers described above.

As a more specific test for the goodness of fit for a psychometric function model, Wichmann and Hill (2001a) suggest to decompose the deviance into deviance residuals. Deviance residuals are a generalization of standard residuals that also applies to binary/binomial data (Dobson & Barnett, 2008). If the deviance residuals varied systematically during the experiment, this would constitute good evidence for nonstationarities such as learning or fatigue that introduce long-lasting trends in the data. It is clear that this test will be insensitive to the increased variance that comes with the kind of nonstationarities that we simulated with the betabinomial observer model. To quantify these long-lasting trends, we calculated the correlation between deviance residuals and the sequence in which the stimulus blocks were presented for each data set that was simulated from the fitted psychometric function model. Note that the simulated data sets were generated with the assumption that the observer was perfectly stationary. Thus, the correlations derived on the simulated data sets describe the distribution of correlations that would be expected if the observer was perfectly stationary. If the correlation between deviance residuals and recording sequence in the original data set differs significantly from the correlations that were derived from the simulated data sets, this is evidence for learning or fatigue. We quantified the power of this test by analyzing how well the test discriminated between a binomial and a learning observer.

Nonstationarity in a Bayesian setting

We will now generalize the nonstationarity diagnostics from the previous section to a Bayesian setting. A main difference between the maximum likelihood setting and the Bayesian setting is that in Bayesian statistics both the data and the model parameters are described in a probabilistic manner. Thus, the fit of a model to a data set consists of writing down the posterior distribution of parameters given the observed data set. This has one particular implication for generalization of the nonstationarity diagnostics from the previous section to a Bayesian setting: In a Bayesian setting, simulating data from a fitted model means sampling from the joint distribution of data and model parameters. This technique is sometimes termed “posterior predictive simulation” (Gelman & Meng, 1996).

The tests from the previous section were extended to posterior predictive simulation. To this, the deviance of every sampled parameter vector from the posterior distribution was calculated with respect to two data sets: the original data set and a data set from the joint distribution of data and model parameters. From these, a Bayesian \( p \)-value was calculated, which is the fraction of simulation runs for which the deviance on the simulated data set was larger than the deviance on the original data set (Gelman & Meng, 1996). If the Bayesian \( p \)-value is close to 0 or 1, the data set is unlikely to result from the fitted joint distribution with respect to its deviance. Because the joint distribution of data and model parameters was fitted based on the assumption of a binomial observer, this represents evidence for nonstationarity. We quantified the power of this test by analyzing how well the test discriminated between a binomial and a betabinomial observer.

Posterior predictive simulations were also used to derive a Bayesian \( p \)-value for the correlations between deviance residuals and recording sequence.

Influential observations

Another perspective on nonstationary behavior can be gained by analyzing the influence of single data blocks on the fitted psychometric function. If a data block has an unduly large influence on the fitted psychometric function, this can have two reasons. Either the psychometric function has been sampled largely suboptimally and is determined only by a small fraction of the data points or the high influence of single data blocks is due to the fact that one data block represents a completely different psychological process than the others. An example for the second situation might be a data block that was recorded at the beginning of learning and, thus, represents unlearned behavior, while all other data blocks refer to more or less learned behavior. An observer that switches strategies from block to block might also result in single blocks with performances that are largely different from the remaining data set. Under some conditions, such an untypical block can have a large—and unwanted—influence on the fitted psychometric function.

Wichmann and Hill (2001a) propose to determine influential data blocks by a jackknife procedure: To detect an undue influence of data block \( i \) on the estimated model parameters, we perform two fits. The first fit is performed on the complete data set, while the second fit is performed on a modified data set in which the \( i \)th data block has been omitted. If the fitted model parameters differ between the complete data set and the modified data set, this is evidence that the \( i \)th data block has an undue effect on the estimated model parameters. We can extend this strategy to obtain a continuous measure of influence. To this end, we scaled the difference between the complete data fit and the modified data fit such that a difference of 1 means “the modified data fit is exactly on the 95%
confidence limit of the complete data fit.” This scaling has been applied to each parameter in isolation. Subsequently, the influence of a single data block was the maximum influence it had on any parameter in isolation.

There is no straightforward generalization of this idea to a Bayesian setting. Because a fitted model in a Bayesian sense is a complete probability distribution, we have to compare the complete data posterior distribution to the modified data posterior distribution. To compare these two distributions, we used a common measure from information theory: The Kullback–Leibler divergence of the complete data posterior distribution under the modified data posterior distribution. If both distributions are equal, the Kullback–Leibler divergence will be zero and it will be larger as they grow different. However, Kullback–Leibler divergence is not symmetric: For two distributions $p$ and $q$, the Kullback–Leibler divergence of $p$ under $q$ is, in general, not equal to the Kullback–Leibler divergence of $q$ under $p$. We quantified the influence of block $i$ by the Kullback–Leibler divergence of the complete data posterior distribution under the modified data posterior distribution. Thus, high influence is assigned to blocks for which the modified data posterior distribution has a lot of mass in areas in which the complete data posterior distribution does not have mass. Such blocks either result in a shift of the posterior distribution or make the posterior distribution considerably sharper. A way how Kullback–Leibler divergence can be derived from samples from the posterior is outlined in Appendix C.

Detection of nonstationarity

Figure 4 shows the performance of the nonstationarity tests outlined in the previous section. After approximately

![Figure 4](https://jov.arvojournals.org/pdfaccess.ashx?url=/data/journals/jov/932793/)
300 trials, the deviance test based on maximum likelihood simulation (Figure 4a) has about 80% power to discriminate between a binomial and a betabinomial observer. The test based on posterior predictive simulation (Figure 4b) has slightly less power—80% power require between 300 and 500 trials. Note that the power of the posterior predictive simulation was evaluated on the basis of the Bayesian $p$-value. This value does not necessarily lead to the predefined $\alpha$ level of 5%. In fact, it does not in this case. The true $\alpha$ level for decisions based on posterior predictive simulations of the deviance is at about 2.5% for trial numbers larger than 200. This implies that the test could be more sensitive if a less conservative criterion on the Bayesian $p$-value was chosen. The Bayesian $p$-value incorporates the probabilities that the observed data are unexpectedly good or that the observed data are unexpectedly bad. The observed 2.5% true $\alpha$ level matches well with the fact that typically the fit of the observed data was worse than that for the simulated data and, thus, the test was, in fact, a one-sided test.

We also tested how well a learning observer could be detected on the basis of correlations between block index and deviance residuals. In the maximum likelihood setting (Figure 4c), the power of the test increased with the number of trials but did not reach values larger than 50% in any of the simulated experimental designs. Similar to the results observed for coverage, designs with small blocks yielded higher power for the detection of a learning observer: With equal total numbers of trials per psychometric function, simulated experiments with 10 trials per block (circles) resulted in considerably higher power than simulated experiments with 40 trials per block (diamonds). When the correlation between block index and deviance residuals was tested using posterior predictive simulations (Figure 4d), power increased slightly slower than in the maximum likelihood setting. Power of more than 50% was not reached for psychometric functions in any simulated experimental design, even for $N = 960$ trials. Similar to the results observed in the maximum likelihood setting, designs with small blocks yielded higher power for the detection of a learning observer. Thus, the tests for nonstationarity gave very similar results no matter whether they were generated based on maximum likelihood simulation or posterior predictive simulation.

These results were very similar, if data were recorded in a yes–no task instead of a 2AFC task. Discriminative power to detect the learning observer from simulations of the correlation between block index and deviance residuals was higher (above 80% for experiments with 10 trials per block after a total of 240 trials, and above 75% in experiments with 20 trials per block after a total of 480 trials, with 40 trials per block, power was about 45% only for the highest number of trials, 960). This is in accordance with the stronger underestimation of credible intervals for a learning observer that was observed in the yes–no task.

As a complementary approach to the detection of nonstationarity, we attempted to detect influential observations as described in the previous section. If unusually many data blocks have a strong influence on the inference performed on a data set, this indicates a general tendency toward strong fluctuations in the data set (or very poor sampling of the psychometric function; this factor did not play a role in our simulations, however). If blocks close to the beginning and the end of an experiment have an unusually strong influence, this indicates a systematic trend in the data. Figure 5 shows distributions of influence measures for the three simulated observers and the three measures that were employed. It is clear that the influence of single data blocks is weak for a true binomial observer. In contrast, for the betabinomial observer, it is much more likely to have single data blocks influencing the inference performed. The positions of these influential data blocks do not vary systematically during the experiment. Finally, with a learning observer, data blocks close to the beginning of an experiment typically have a tendency to exert a slightly stronger influence on the inference performed on the data. Note that all these effects are particularly obvious for Bayesian inference, where influence is quantified in terms of the Kullback–Leibler divergence of the posterior distribution derived from the full data set under the posterior distribution derived from a modified data set in which a single data block has been removed. By comparing the first and second rows of Figure 5, we observe that the design has an impact on the distribution of influential observations. The data from the first and second rows of Figure 5 are from simulated experiments with the same total number of trials. These data differ in how the trials were distributed over experimental blocks: The data in the first row is from an experiment with few large blocks (6 blocks with 40 trials each), while the data in the second row is from an experiment with many smaller blocks (12 blocks with 20 trials each). We can see that the influence of individual blocks is typically much weaker if data were collected in a large number of small blocks than if data were collected in a small number of large blocks. In fact, with a large number of small blocks, the influence measure derived from the bootstrap procedures does not discriminate well between stationary and nonstationary observers. The above effects were particularly prominent with a learning observer in a yes–no task (not shown): The distribution of the Bayesian influence measure for the first trial obtained with small blocks was virtually nonoverlapping with the corresponding distribution for the binomial observer. The influence of single data blocks was increased for the betabinomial observer with respect to the binomial observer. However, the largest Bayesian influence observed for the first block
of the betabinomial observer was still below the median of the corresponding distribution for the learning observer.

**Correcting credible intervals**

In the Coverage for nonstationary observers section, we observed that credible intervals are seriously underestimated if psychometric functions are fitted to nonbinomial observers. Even worse, we observed that recording more trials does not solve this problem and that for “small” numbers of trials ($N < 300$) statistical tests to detect nonstationarities have insufficient power. Here, we present a procedure for correcting credible intervals that have been derived assuming a stationary binomial observer although the observer actually was not stationary. The basic idea is simple: nonstationarity always includes dependencies between trials. If we recorded $n$ trials in an experiment, these dependencies imply that the number of independent trials we recorded is less than $n$. Here, we show that based on the residuals of a psychometric function that was estimated assuming stationarity, it is possible to estimate the fraction of trials that were actually independent.

To do so, we assume that the decrease in the number of independent trials is equal for all blocks. Thus, we can say that the number $n_i$ of trials in the $i$th block is reduced to $\nu n_i$ trials, with some $\nu \in (0, 1]$. If $\nu = 1$, then the trials seem to be sufficiently independent such that no reduction in the trial number is necessary. To estimate $\nu$, we assume that the data vary around the estimated psychometric function according to a Beta distribution and choose $\nu$ to maximize the corresponding likelihood function. Details of this procedure are given in Appendix D. To illustrate the validity of $\nu$, we plotted the total effective number of trials $\nu n$ over the total number of recorded trials in Figure 6. For a binomial observer, the effective number of
trials \( \nu n \) is in close agreement with the number of recorded trials. Note, however, that the effective number of trials is slightly underestimated even for the binomial observer: Ideally, \( \nu \) should be one and all the data points for the binomial observers should be on the dotted diagonal. For the learning observer (light blue symbols) and for the betabinomial observer (red symbols), the plotted curves are clearly below the diagonal. This indicates that the effective number of trials is less than the number of recorded trials.

Once we have estimated \( \nu \), we can apply one of two different procedures to correct our inference for violations of the independence assumption:

1. We can scale the estimated credible intervals according to the formula that is known for the standard error of the mean:

\[
c \mapsto \frac{c - \hat{\theta}}{\sqrt{\nu}} + \hat{\theta}.
\]

That is, we scale the distance between a confidence limit \( c \) and the parameter estimate \( \hat{\theta} \) by a factor of \( 1/\sqrt{\nu} \). In a Bayesian posterior distribution, that is, the above transformation would be applied to each single sample \( c = \theta^{(j)} \). This correction is applied after the inference has already been performed. We will, therefore, refer to it as correction of inference.

2. Alternatively, we can scale the data used for the inference. To do so, we simply reduce the number of responses \( k_i \) and the number of trials \( n_i \) in the \( i \)th block by a factor \( \nu \):

\[
k_i \mapsto \lfloor \nu k_i \rfloor, \quad n_i \mapsto \lfloor \nu n_i \rfloor.
\]

Here, \( \lfloor x \rfloor \) denotes the largest integer \( n \) such that \( n \leq x \). This correction is applied to the data before inference has been performed. We will, therefore, refer to it as correction of data.

Both of these corrections have advantages and disadvantages: Correction of inference is very easily applied and can even be used in case previous (and possibly incomplete) data need to be reanalyzed. However, after correction of inference credible intervals might include unreasonable or even impossible values such as negative lapse rates or guessing rates larger than 1. Correction of data, on the other hand, requires the whole inference to be repeated with modified trial numbers. This results in some computational overhead. In addition, it is more difficult to perform this kind of correction on data that have been published in a paper. As an advantage, correction of data will always result in valid inference in terms of lapse rates and guessing rates. Another potential problem with correction of data is the fact that for very strong dependencies, \( \nu \) can be close to zero. In that case, we might end up with blocks that effectively contain no trials anymore. Even worse, we might even find that no data are left at all from our experiment. This did indeed happen in a number of our simulation runs (up to 10% for the betabinomial observer, about 3% for the learning observer, and about 1% or below for the binomial observer). In these cases, correction of data was not applicable.

To explore whether these correction strategies worked and, if so, which performed better, we estimated \( \nu \) from each simulated data set and corrected credible intervals accordingly. Figure 7 shows the results of the correction of inference. For bootstrap-based inference and with increasing trial number, the true coverage of 95% confidence intervals approaches the nominal coverage of 95%. However, serious underestimation of confidence intervals still occurs with less than approximately 500 trials per psychometric function—the problem, thus, remains severe for typical psychophysical experiments. Even worse, the underestimation of confidence intervals remains visible even for the highest trial number tested. The situation is different for Bayesian inference, however: For Bayesian inference, the true coverage is very close to 95% even for relatively small trial numbers. For the binomial observer, posterior intervals are slightly too broad resulting in coverages \( >95\% \), in accordance with the slight underestimation of the effective number of trials that was observed for the binomial observer in Figure 6.

Figure 8 shows the results of the correction of data. We can see that correction of data improves the inference in the desired way only for designs with large numbers of trials per block. For small numbers of trials per block, credible intervals for threshold had worse coverage after
Figure 7. Coverage of nonstationary observers after correction of inference. Details are as in Figure 3.

Figure 8. Coverage of nonstationary observers after correction of data. Details are as in Figure 3.
correction of data than they had without any correction. For width, credible intervals were slightly better than for thresholds. However, they were no better than without any correction. With small numbers of trials per block, correction of data can easily result in situations where a block reduces to only one or two valid trials. Under these conditions, our automatic inference procedure used for estimation failed altogether in many cases (convergence rates as low as 4.7% in the worst case, median: 47.1% for binomial observer, 24.4% for betabinomial observer, 25.4% for learning observer). We would like to emphasize that correction of data yields improved confidence intervals only in designs with large numbers of trials per block. Thus, correction of data is not a generally recommended way for correcting credible intervals for nonstationary behavior, at least not in the context of psychophysics with the comparatively small number of trials. In most realistic psychophysical scenarios, we thus recommend correction of inference.

Similar results were observed if the data were collected in a simulated yes–no task instead of a 2AFC task. However, in this case, the correction was less effective, resulting in corrected coverages for the bootstrap procedures between 70% and 90% (for 4 blocks with 40 trials each, it was even as bad as 60%). The estimated fractions of effective trials were marginally lower in this case for all conditions.

**Discussion**

We performed a large number of simulations to investigate the effects of nonstationary behavior on estimation of psychometric functions. The effects of nonstationarity on the estimated credible intervals could not be neglected. When psychometric functions were estimated from nonstationary data, the credible intervals for threshold as well as width were much too small. Even worse, the typical strategy that is used to deal with corrupted data is to simply collect more data. This strategy does not work here: More bad data will not lead to good inference.

It is never advisable to detect nonstationary data by eye only. We, therefore, devised systematic tests to detect nonstationary data. These tests were based on simulations and tried to detect unreasonably large variances and systematic trends in the data. Both types of tests performed reasonably well. It is, thus, possible to detect data sets that are corrupted by nonstationarities such as learning or fluctuations in attention.

In a third step, we demonstrated that an analysis of the residuals of the fitted psychometric function can be used to correct the credible intervals. From this analysis, a correction factor that can be used to scale credible intervals or data was derived. This correction improved inference if credible intervals were scaled. Scaling the data did result in improvements only with relatively large blocks of trials.

**Realism of the simulated data**

The selection of the observers was relatively ad hoc: With different generating parameters, different results might have been obtained.

The estimated fractions of effective trial numbers can easily be compared to data from real psychophysical experiments. We analyzed data from 5 observers that detected luminance increments in different high-level contexts (three contexts, data kindly provided by Maertens & Wichmann, 2010). All observers showed at least some evidence for learning. The estimated $\nu$ was between 0.37 and 0.91 corresponding roughly to the quartile range of our simulated learning observer (quartile range of 0.37 to 1). Correction of inference would increase the credible intervals by 5% to 64%. We also analyzed data from 6 listeners from an auditory experiment (kindly provided by Schoenfelder & Wichmann, 2010). Listeners had to detect sine tones that were masked by band-limited noise. In total, each listener performed between 30,000 and 40,000 trials. Out of these trials, we analyzed the first 1600 trials and estimated $\nu$-values between 0.19 and 0.62. This coincided roughly with the quartile range of our simulated betabinomial observer, which was 0.17 to 0.77. Correction of inference would increase the credible intervals by 27% to 129%. Note, however, that in both of these studies, the data we analyzed were excluded from the final analysis of the data due to the observed nonlinearities. These data indicate that the simulated nonstationarities we used were at least roughly comparable to empirically observed nonstationarities. We also expect that stronger nonstationarities (i.e., lower $\nu$) will be observed for children, patients, or elder people.

Figure 4 indicates that only a part of the space has actually been explored with the current parameter settings. All simulated data for the betabinomial observer scatter around a well-defined curve that monotonically increases with trial number. However, not all types of nonstationarity have to be on this curve. For instance, an observer with weakly nonstationary behavior might have been tested in 500 trials. However, the nonstationarity might be weak enough to be detected by the deviance test only in 30% of the cases. The presented data do not provide much information as to how severe the misestimation of credible intervals is in such a case. We, therefore, repeated simulations for the betabinomial observer. In these simulations, the strength of the nonstationarity was varied to target-specific areas of the plot in Figure 4: These areas corresponded to rejection probabilities of approximately 30%, 60%, and 90% in order to cover the whole range of rejection probabilities. For a binomial observer, deviance asymptotically follows a $\chi^2$-distribution with the degrees of freedom equal to the number of data blocks minus the...
number of parameters. The $M$ parameter of the betabinomial observer was adapted to yield rejection rates of 30%, 60%, and 90% for this asymptotic distribution. Figure 9 illustrates the results of these additional simulations: The left panel shows the actual rejection probabilities that were observed in the simulations. The targeted values are not hit precisely, but the observed rejection probabilities now cover the whole range between 0 and 1 as intended. The middle panel presents uncorrected coverages. It is clear that nonstationarities that are easy to detect (90% desired rejection rates) result in very severe underestimation of credible intervals. Furthermore, nonstationarities that typically pass the deviance test (30% desired rejection rates; actually closer to 10%) still result in significant underestimation of credible intervals: The credible intervals had to be scaled by a factor of 1.2 to obtain the correct size. Thus, these data sets were very hard to tell apart from being generated by a stationary observer using diagnostic tests while still leading to 20% too small confidence intervals with only 85–90% coverage instead of the intended 95%. The right panel presents coverages after correction of inference. Similar to Figure 7, coverages are now very close to the desired 95%. This illustrates two points: First, nonstationarity may have an impact on credible intervals even though the nonstationarity is not detected in any formal test. Second, correction of inference also works for these weakly nonstationary data sets.

Alternatives to the proposed procedure

It is always statistically superior to use the correct model if it is known. This will result in improved estimation of unknown parameters and better coverage of credible intervals. However, for nonstationary observer behavior, there is no general and widely accepted model. The best one can hope for is, thus, to get as far as possible with the models we have and deal with the shortcomings of this model.

The strategy presented here differs from the one suggested by Wichmann and Hill (2001a, 2001b): They suggested to simply mark bad fits and repeat the experiment. Here, we took a completely different approach. Instead of repeating the experiment, we quantify the nonstationarity using a diagnostic measure and then inflate the credible intervals accordingly. Although this will result in larger credible intervals, it allows us to perform valid inference without losing data. This is particularly useful in studies with patients or children where it is difficult to record more data or in cases in which experimental time is very expensive such as in psychophysiological studies, e.g., functional magnetic resonance imaging (fMRI).

One might wonder why we need a correction for the credible intervals at all. The correction is strongly based on modeling the residuals of the fitted psychometric function by a Beta distribution. Thus, a Beta regression model (Cribari-Neto & Zeileis, 2010; Ferrari & Cribari-Neto, 2004) might provide the correct credible intervals in the first place. We believe that the approach at hand has the advantage that it makes a relatively clear distinction between those parameters that are relevant to obtain a correct point estimate and those that are relevant to obtain the correct interval estimates. We also believe that an advantage of the current approach is that it allows determination of the number of efficient trials even in published data. This does not imply that Beta regression will fail in these tasks.

We presented tests for stationarity of observers. Alternatively, one might think about comparing the fit of a binomial observer model with fits of observer models that violate the stationarity assumption. Although model selection methods have recently improved considerably
In this paper, we simulated behavior from models for nonstationarity. It should, however, be kept in mind that we do not argue that these models are correct nor that they are the only mechanisms that could generate nonstationary behavior. Thus, we do not select among different competing models because we do not know the competitors. The tests presented here do not require knowledge of alternative models. They can detect violations of the model’s assumptions without presenting an alternative. We also believe that in combination with the proposed correction method for credible intervals, these tests provide a reasonable way for dealing with nonstationary observers: The tests are simple and come with virtually no additional computational burden over the Monte Carlo techniques that are used to derive the credible intervals. The correction procedure is simple and intuitive. In addition, it provides a graded measure of the severity of the violations of the assumptions.

**Conclusion**

We performed large simulations to study the effects of nonstationary behavior on the estimation of psychometric functions. In particular, we demonstrated that credible intervals for parameters of the psychometric function are severely underestimated if the data are nonstationary. We presented diagnostic criteria for nonstationarity and suggested a simple correction formula to scale the confidence intervals to the correct size.

All methods discussed in this paper are available from http://psignifit.sourceforge.net.

**Appendix A**

**Simulated observers**

All simulated observers received stimuli at \( K \) different stimulus intensities \( x_1, \ldots, x_K \). Stimulus intensity \( x_i \) was given by

\[
x_i = F^{-1}(y_i) = m + \frac{w}{z(a)} \log \left( \frac{y_i}{1 - y_i} \right).
\]  

\( \gamma_i \)s were selected to linearly span the range from 0.01 to 0.99, independent of the value of \( K \).

**Binomial observer**

To simulate a binomial observer, response counts \( k_i \) were sampled from a binomial distribution \( k_i \sim \text{Binom}(n_i, \psi_i) \), where

\[
\psi_i := \Psi(x_i) = \gamma + (1 - \gamma - \lambda)F(x_i; m, w).
\]  

Here, \( \gamma \) is the guessing rate of the observer. For \( n \)-AFC tasks, \( \gamma \) was fixed to \( 1/n \). For a yes–no task, we set \( \gamma = 0.02 \). The lapse rate \( \lambda \) was always set to \( \lambda = 0.02 \). The sigmoid function \( F \) was a logistic function as parameterized in Kuss et al. (2005):

\[
F(x; m, w) = \left( 1 + \exp \left( -\frac{z(a)}{w} (x - m) \right) \right)^{-1}.
\]  

For the current simulations, we set the threshold \( m = 4 \) and the width of the rising interval of the psychometric function \( w = 2 \). The value \( z(a) = 2\log(1/(a - 1)) \) was set such that \( w \) is the width of the interval on which \( F \) rises from 0.1 to 0.9.

**Learning observer**

For the learning observer, the response probability on every trial was given by Equation A2. However, the parameters \( m \) and \( w \) were both functions of the trial number. Initially, we used \( m_0 = 4.7 \) and \( w_0 = 2.7 \). The parameters on trial \( t \) were then recursively defined by

\[
u_t = \frac{u_{t-1} - E_u}{\tau},
\]  

where \( u \) is either \( m \) or \( w \) and \( E_{m_0} = 3.3 \), \( E_{w_0} = 1.3 \), and \( \tau = 40 \). The observer improved to an asymptotic performance level that is given by \( m_\infty = E_m \), \( w_\infty = E_w \). The speed at which performance approaches the asymptotic level is quantified by \( \tau \): The higher the value of \( \tau \), the smaller is the improvement in performance on each trial and the slower the performance approaches the asymptotic level.

To analyze coverage for the learning observer, the generating parameters were averaged over time.

**Betabinomial observer**

To simulate the betabinomial observer, response counts \( k_i \) for block \( i \) were sampled from a binomial distribution
During the optimization, the lapse rate \( \lambda \) was determined by maximizing the log likelihood of the data.

To estimate confidence intervals, a bootstrap procedure (Wichmann & Hill, 2001b) was used. Two different sampling strategies were used to generate bootstrap samples (Efron & Tibshirani, 1993, chap. 6):

1. Parametric bootstrap refers to the case where new data sets were generated from the fitted psychometric function. That is, the response probability for block \( i \) is given by \( \Psi(x_i; \theta) \) and trials for that block are then simply generated from a binomial distribution.

2. Nonparametric bootstrap refers to the case where new data sets were generated from the relative response frequencies. That is, the response probability for block \( i \) is given by \( k_i/n \) and trials for that block are then generated from a binomial distribution. For each random data set, a new psychometric function was fitted. We generated 2000 bootstrap samples to obtain an approximation of the sampling distribution of the parameters. From these samples, 95% bias corrected and accelerated confidence intervals (Davison & Hinkley, 1997; Hill, 2001; Wichmann & Hill, 2001b) were derived.

**Appendix B**

**Fitting procedures**

**Bootstrap sampling**

A point estimate of the psychometric function was determined by maximizing the log likelihood of the data. During the optimization, the lapse rate \( \lambda \) was constrained to the interval [0, 0.1]. In yes–no tasks, the guessing rate \( \gamma \) was also constrained to this interval.

To estimate confidence intervals, a bootstrap procedure (Wichmann & Hill, 2001b) was used. Two different sampling strategies were used to generate bootstrap samples (Efron & Tibshirani, 1993, chap. 6):

1. Parametric bootstrap refers to the case where new data sets were generated from the fitted psychometric function. That is, the response probability for block \( i \) is given by \( \Psi(x_i; \theta) \) and trials for that block are then simply generated from a binomial distribution.

2. Nonparametric bootstrap refers to the case where new data sets were generated from the relative response frequencies. That is, the response probability for block \( i \) is given by \( k_i/n \) and trials for that block are then generated from a binomial distribution. For each random data set, a new psychometric function was fitted. We generated 2000 bootstrap samples to obtain an approximation of the sampling distribution of the parameters. From these samples, 95% bias corrected and accelerated confidence intervals (Davison & Hinkley, 1997; Hill, 2001; Wichmann & Hill, 2001b) were derived.

**Markov Chain Monte Carlo**

The posterior distribution of the parameters of the psychometric function given the data cannot be derived analytically. To approximate the posterior distribution, we used Markov Chain Monte Carlo sampling (Gilks, Richardson, & Spiegelhalter, 1996; Kuss et al., 2005). Briefly, Markov Chain Monte Carlo constructs a Markov Chain that converges to the desired posterior distribution. Thus, for an infinite number of samples, we can be sure that the samples are indeed from the posterior distribution. However, for any finite number of samples, this is not guaranteed. We employed a number of measures to ensure that only samples from chains that had converged to the posterior distribution were used.

Samples were generated using the Metropolis–Hastings algorithm (Hastings, 1970; Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953). This algorithm requires a number of parameters. These parameters were derived from a series of pilot samples: The first pilot sample was generated by starting the chain at the maximum of the posterior distribution and proposing new samples based on the Fisher information matrix. From this first sample, appropriate sampling parameters were derived using a method proposed by Raftery and Lewis (1996). Based on these sampling parameters, another pilot sample was generated, again starting at the maximum of the posterior distribution and proposing new samples based on sample variances of the previous pilot sample. These steps were repeated iteratively until the number of samples required to reach convergence as derived from the pilot sample in one iteration was smaller than or equal to the number of required samples derived from the pilot sample in the previous iteration.

Second, stationarity of individual chains was monitored by a procedure inspired by Geweke (1991). Each chain was split in 10 subsequent bins and the mean was calculated for each bin. If the mean in one bin differed from the overall mean of the chain by more than 2 standard deviations, the chain was rejected due to non-stationarity of the Markov Chain. Only chains that met this stationarity criterion were used for the analysis.

As a third step, we sampled three chains to estimate the posterior distribution. The first chain started at the maximum of the posterior distribution; the two other chains were started at two randomly selected borders of the 95% posterior intervals of the posterior as estimated from the first chain. If the sampling algorithm really samples the posterior distribution, samples from these three chains should be statistically indistinguishable. Following a procedure proposed by Gelman (1996), we computed the ratio \( R \) of between-chain variance over within-chain variance. If this ratio was larger than 1.1, all three chains were rejected.

From chains that were considered as sampling the posterior distribution sufficiently closely, a point estimate
was derived as the average of the posterior samples (mean of the posterior estimate) and posterior intervals were taken as the 2.5th and the 97.5th percentile.

**Appendix C**

**Kullback–Leibler divergence from sampled data**

Consider two random variables with density functions \( f \) and \( g \). In our current application, we know \( f \) and \( g \) only up to an unknown scaling factor. That is, we could write

\[
f = \frac{1}{Z_f} \tilde{f},
\]

where we know \( \tilde{f} \), but we do not know \( Z_f \). The same holds for \( g \). The Kullback–Leibler divergence of \( f \) under \( g \) is given by

\[
D_{\text{KL}}(f\|g) = -\mathbb{E}_f \left( \log \frac{g}{f} \right),
\]

where \( \mathbb{E}_f \) denotes expectation under \( f \). The expectation on the right-hand side cannot be evaluated directly because we do not know the scaling factors. We write \( F(\theta) = -\log f(\theta), G(\theta) = -\log g(\theta) \) and can then use importance sampling to obtain (Bishop, 2006, p. 554)

\[
\frac{Z_g}{Z_f} \approx -\frac{1}{N} \sum_{i=1}^{N} \exp \left( -F(\theta^{(i)}) + G(\theta^{(i)}) \right),
\]

where \( \theta^{(i)}, \ell = 1, \ldots, N \) are samples from \( f \).

We can rewrite the right-hand side of Equation C2 as

\[
-\mathbb{E}_f \left( \log \frac{g}{f} \right) = -\mathbb{E}_f \left( \log \frac{\tilde{g}}{Z_g} + \frac{Z_f}{Z_f} \right)
\]

\[
= -\mathbb{E}_f \left( \log \frac{\tilde{g}}{\tilde{f}} \right) - \mathbb{E}_f \left( \log \frac{Z_f}{Z_g} \right)
\]

\[
= -\mathbb{E}_f \left( \log \frac{\exp(-G)}{\exp(-F)} \right) - \mathbb{E}_f \left( \log \frac{Z_g}{Z_f} \right)
\]

\[
= -\mathbb{E}_f (-G + F) + \log Z_f
\]

\[
\approx -\frac{1}{N} \sum_{i=1}^{N} \left( -G(\theta^{(i)}) + F(\theta^{(i)}) \right) + \log \left( \frac{1}{N} \sum_{i=1}^{N} \exp \left( -G(\theta^{(i)}) + F(\theta^{(i)}) \right) \right).
\]

Thus, it is sufficient to evaluate \( F(\theta^{(i)}) - G(\theta^{(i)}) \) with each sample \( \theta^{(i)} \) in order to approximate the Kullback–Leibler divergence of \( f \) under \( g \).

**Appendix D**

**Estimating the effective fraction of independent trials**

Assume that we fitted a psychometric function based on the assumption of binomial variance:

\[
k_i \sim \text{Binom}(n_i, \Psi(x_i|\bar{\theta})) =: \text{Binom}(n_i, \Psi_i). \quad (D1)
\]

Here, \( k_i \) is the number of correct responses that were obtained from the \( n_i \) trials presented at stimulus level \( x_i \), \( \theta \) is the fitted parameter vector. The index \( i \) goes from 1 to \( K \), where \( K \) is the number of blocks. We used the abbreviation \( \Psi_i = \Psi(x_i|\bar{\theta}) \) as in Equation A2.

Equation D1 can be modified to allow for larger variance, if we set

\[
p_i := \frac{k_i}{n_i} \sim \text{Beta}(\Psi_i n_i + 1, (1 - \Psi_i) n_i + 1). \quad (D2)
\]

If \( \nu = 1 \), Equations D1 and D2 are equivalent. However, we can now find a \( \nu^* \in (0, 1] \) such that the log likelihood

\[
\ell(\nu) = \sum_{i=1}^{K} \log(f(p_i; \Psi_i n_i + 1, (1 - \Psi_i) n_i + 1)), \quad (D3)
\]

is maximized. Here, \( \nu; (0, 1) \to \mathbb{R}_+ \) is the density of the Beta distribution. This \( \nu^* \) will then be the estimated fraction of effectively independent trials.

**Acknowledgments**

We want to thank Jakob Macke and the “Modelling of Cognitive Processes” group in Berlin for helpful comments on a previous version of this manuscript. This work was supported by the German Research Foundation (DFG Sachbeihilfe FR 2854/1-1 awarded to IF and FAW) and, in part, by the Bernstein Computational Neuroscience Program of the German Federal Ministry of Education and Research (Förderkennzeichen 01GQ0414).

Commercial relationships: none.

Corresponding author: Ingo Fründ.

Email: ingo.fruend@tu-berlin.de.
Address: Modellierung Kognitiver Prozesse, Technische Universität Berlin and Bernstein Center for Computational Neuroscience, Sekr 6-4 TU-Berlin Franklinstr. 28/29, Berlin, 10587, Germany.

Footnotes

1 It should be noted that this betabinomial observer is, in fact, stationary. However, it mimics the increased variance that would result from the above-mentioned fluctuations. This is why we will refer to the betabinomial observer as “nonstationary.”

2 Two of these sigmoidal functions are distribution functions of extreme value distributions. The Weibull function arises in contrast detection experiments if a high threshold model and a maximum decision rule are employed (Quick, 1974). It is often used as a purely descriptive model for the psychometric function in contrast detection experiments because it fits data very well due to its asymmetry; in addition, it is used as the underlying model of the QUEST adaptive procedure (Watson & Pelli, 1983). The Gumbel cumulative distribution function was chosen simply to have a function with the inverse symmetry. No theoretical interpretation was implied.

3 We term this function “reversed” because it corresponds to the cumulative distribution function of the Gumbel distribution on a reversed x-axis.

4 We use the term “credible intervals” here to refer to both frequentist confidence intervals as obtained from the bootstrap procedures as well as Bayesian posterior intervals. We also use the term “confidence intervals” or “posterior intervals” to refer to either the frequentist or the Bayesian version.

5 In fact, this definition is only true for Bayesian posterior intervals. The proper definition for confidence intervals is slightly more technical but is not the topic of the current investigation.

6 This might seem theoretically unsound. One should expect that at least a single trial at each stimulus level has been collected in any case. The slight underestimation of the effective number of trials might, however, result in data sets with effectively no trials remaining. In addition, if \( n_i \) is less than 1, \( [n_i] \) will be zero.

7 The deviance distribution in real experimental setups is often quite far from the asymptotic \( \chi^2 \)-distribution (Wichmann & Hill, 2001a).

References


