Spatial four-alternative forced-choice method is the preferred psychophysical method for naïve observers

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

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

H. R. Blackwell (1952) investigated the influence of different psychophysical methods and procedures on detection thresholds. He found that the temporal two-interval forced-choice method (2-IFC) combined with feedback, blocked constant stimulus presentation with few different stimulus intensities, and highly trained observers resulted in the "best" threshold estimates. This recommendation is in current practice in many psychophysical laboratories and has entered the psychophysicists' "folk wisdom" of how to run proper psychophysical experiments. However, Blackwell's recommendations explicitly require experienced observers, whereas many psychophysical studies, particularly with children or within a clinical setting, are performed with naïve observers. In a series of psychophysical experiments, we find a striking and consistent discrepancy between naïve observers' behavior and that reported for experienced observers by Blackwell: Naïve observers show the "best" threshold estimates for the spatial four-alternative forced-choice method (4-AFC) and the worst for the commonly employed temporal 2-IFC. We repeated our study with a highly experienced psychophysical observer, and he replicated Blackwell's findings exactly, thus suggesting that it is indeed the difference in psychophysical experience that causes the discrepancy between our findings and those of Blackwell. In addition, we explore the efficiency of different methods and show 4-AFC to be more than 3.5 times more efficient than 2-IFC under realistic conditions. While we have found that 4-AFC consistently gives lower thresholds than 2-IFC in detection tasks, we have found the opposite for discrimination tasks. This discrepancy suggests that there are large extrasensory influences on thresholds—sensory memory for IFC methods and spatial attention for spatial forced-choice methods—that are critical but, alas, not part of theoretical approaches to psychophysics such as signal detection theory.

Keywords: psychophysics, psychophysical methods, temporal and spatial forced-choice procedures, efficiency, reliability, overdispersion, bias, sensory determinacy

Introduction

Virtually all psychophysical studies require the measurement of psychophysical thresholds, be it difference or absolute thresholds. A number of experimental methods exist, most prominently single-interval or yes–no methods, forced-choice methods, and the method of adjustment, to help obtain thresholds. In addition, at least for single-interval and forced-choice methods, there are at least three different experimental procedures for collecting data: adaptive procedures, constant stimulus with trial-by-trial randomization of signal intensities, and constant stimulus with blocked presentation of signal intensities (block design).

There is a comparatively large body of literature that explores and compares the *statistical* properties of different procedures (e.g., Garcia-Perez, 1998; Green, 1990; Kaernbach, 1991; Kontsevich & Tyler, 1999; Laming & Marsh, 1988; Leek, Hanna, & Marshall, 1992; Snoeren & Puts, 1997; Treutwein, 1995; Watson & Pelli, 1983) and threshold estimation once data collection is complete (e.g., Foster & Bischof, 1991; Kaernbach, 2001; Kuss, Jäkel, & Wichmann, 2005; Maloney, 1990; McKee, Klein, & Teller, 1985; Miller & Ulrich, 2001; Treutwein & Strasburger, 1999; Wichmann & Hill, 2001a, 2001b). In addition, signal detection theory (SDT; Green & Swets, 1988) offers a theoretical framework in which thresholds obtained using forced-choice methods can be converted to the equivalent single-interval thresholds and vice versa. SDT attempts to explain detection and discrimination performance in terms of sensory and decision processes: the (sensory) noise and signal-plus-noise distributions on the putative internal decision axis and the (decision) criterion adopted by the observer. We know, however, that this conception of detection and discrimination does not tell the whole story: Attention influences detection and discrimination, and at least under some circumstances, the influence manifests itself on sensory processing, for example, a sharpening of spatial frequency or orientation tuning (e.g., Itti, Koch, & Braun, 2000; Lee, Itti, Koch, & Braun, 1999). Thus, SDT's assumption of having a fixed sensory front end proved to be incorrect. This opens the possibility that different psychophysical methods require more or less attention and thus yield significantly different results despite the prediction of SDT that thresholds

doi: 10.1167/6.11.13 Received September 9, 2005; published November 10, 2006 ISSN 1534-7362 © ARVO

should be convertible using SDT formalisms. Furthermore, some methods or procedures may be easier to learn or may feel more "natural" to human observers, particularly naïve observers.

Nonetheless, comparatively few studies investigated such more *psychological* than statistical consequences of the different methods and procedures. The most thorough examination of different methods for visual threshold measurements was reported in a seminal paper by Blackwell (1952). Blackwell identified three criteria by which to judge psychophysical methods and procedures:

1. Sensory determinacy. Methods that give lower thresholds are to be preferred as higher threshold values may indicate that the method makes observers more prone to unwanted extrasensory influences.
2. Reliability. This refers to the extent to which threshold measurements vary over time under what seem to be identical experimental conditions.
3. Inferred validity. This refers to the extent to which variables that are thought to be irrelevant influence threshold measurements.

The variables Blackwell studied included the number, spacing, and order of stimuli; the motivational attitude the observers adopted; and the role of feedback. The methods he compared were all methods of constant stimuli and included a yes–no method as well as temporal and spatial two- and four-alternative forced-choice methods (2-AFC and 4-AFC, respectively). Blackwell did run a temporal four-interval forced-choice (4-IFC) condition and a spatial 4-AFC condition, but in the latter, the stimuli were presented comparatively far in the periphery, 7° N, E, S, or W of fixation, and the target stimulus was small (18.5 arcmin). Furthermore, Blackwell only analyzed the responses in the E quadrant; hence, his results for spatial 4-AFC may not be true in general. The omission of the method of limits or the method of adjustment, on the other hand, is not critical as these methods are known to be generally inferior (Higgins, Jaffe, Caruso, & de Monasterio, 1988; Vaegan & Halliday, 1982; Woods & Thomson, 1993) and are rarely used today. Blackwell tried to obtain low and reproducible thresholds—objectives that one would presume are universally endorsed by all psychophysicists. On the basis of these criteria and his extensive data, Blackwell made the following recommendations:

1. 2-AFC is to be preferred over yes–no tasks and 4-AFC.
2. Forced choice should involve temporal intervals rather than spatial locations.
3. Stimuli should be grouped into blocks of the same magnitude rather than being randomized; that is, a block design should be used.
4. Use as few stimuli, for example, signal intensities, as practicable.
5. Feedback should be provided.
6. Participants should have extensive experience in threshold measurements; that is, one should work with trained observers.

Many of Blackwell's recommendations—trained observers, temporal 2-AFC, feedback, blocked constant stimulus—are in current practice in many psychophysical laboratories and have entered the psychophysicists' "folk wisdom" of how to run proper psychophysical experiments.

However, Blackwell explicitly recommends working with experienced observers because the other recommendations depend on this. Whereas psychophysical experiments are frequently carried out with highly trained observers, there are reasons why one should work with naïve observers. When examining patients especially children, it may not be possible to achieve a high level of training, for example. The limited applicability of recommendations by psychophysicists has been noted in the clinical literature, for example, by Woods and Thomson (1993). Also, observers who are willing to observe for several hours per week over a long period are, unfortunately, rare. On the other hand, there is usually a big pool of students available who will participate in one short experiment for course credits or money.

From our own experience, naïve observers need considerable time to feel comfortable with temporal forced-choice tasks, whereas spatial forced-choice tasks seem to be more natural. In temporal forced-choice tasks, observers have to keep their sensory impression in mind while waiting for the second interval, and in addition, there is an arbitrary assignment of responses to intervals, which can cause problems. In a spatial forced-choice task, the alternatives are presented simultaneously. When the layout of the response keyboard matches that of the screen presentation or when one uses a touch screen, the relation between stimulus alternatives and response becomes very obvious. This opens up the possibility of giving up Blackwell's Recommendation 2 at least for naïve observers. Note that this could improve the *efficiency* of the method of constant stimuli. Showing two alternatives at two spatial locations is almost twice as fast as showing two alternatives one after the other (efficiency, however, appears not to have been a concern for psychophysicists in the early 1950s; this is one of the very few aspects of psychophysical methods and procedures Blackwell did not explore).

Furthermore, using more than two alternatives in a forced-choice task reduces the variance inherent in the responses of an observer. It has been observed in practice that a higher number of alternatives is statistically more efficient—and we will quantify this gain in efficiency under realistic conditions in the experiments below. Figure 1 shows psychometric functions for m-AFC, with m taken from $\{2, 4, 8\}$. Suppose that an experimenter already has a rough idea about where the psychometric function lies and now has to choose some stimuli for presentation. Usually, experimenters will try to distribute the stimuli evenly such that they cover the whole range, but other sampling schemes are possible and more efficient (Wichmann & Hill, 2001a). Then, the observer is presented with N trials of one of these stimuli and produces p correct answers. Assuming that the correct answers come from a binomial distribution with a fixed probability, the expected variance of the data (i.e., variance of the number of correct answers) is given by $Np(1 - p)$. In the right panel of Figure 1, we plot $p(1 - p)$ for different stimulus values, and it can be seen that for 2-AFC, there are large regions of stimulus space for which the expected variance is higher than the variance for m-AFC with $m > 2$. Indeed, unless the psychometric function was sampled very inefficiently using only positive stimulus values on the axes of Figure 1, the higher m is, the better the psychometric function estimation for a given number of trials is. In addition, for greater m, the point with the highest variance is shifted to the steeper part of the psychometric function where changes in the stimulus result in greater changes in the response probability. Thus, it is worth exploring how well observers do in spatial 4- or 8-AFC, which are conditions not (or not satisfactorily) explored by Blackwell, and it is worth seeing whether the increased efficiency is worth giving up Blackwell's Recommendations 1 and 2.

Adaptive procedures are frequently assumed and sometimes argued to be *statistically* more efficient than the method of constant stimuli (e.g., Watson & Fitzhugh, 1990—this issue is not uncontroversial, see Hill, 2001, pp. 225–228). However, adaptive procedures violate Recommendations 3 and 4 by Blackwell. On every trial, or on nearly every trial depending on the adaptive procedure, a new stimulus is presented to the observers, and this may prevent them from learning to improve their performance for this stimulus. Furthermore, the frequent change of stimuli may make it hard for (naïve) observers to concentrate on particular features of the stimulus in question. Finally, some adaptive procedures are very sensitive to serial dependencies in the participant's responses, which mislead the procedure (Burns & Corpus, 2004; Friedman, Carterette, Nakatani, & Ahumada, 1968; Lages & Treisman, 1998; Treisman & Williams, 1984). Thus, it is not surprising that in real psychophysical experiments, the reliability of adaptive procedures is lower than that of constant stimuli methods (Woods & Thomson, 1993). Whereas one may be willing to sacrifice reliability for speed in certain settings, reliability is usually more important for basic scientific questions. Similarly, unless one is investigating a phenomenological aspect of perception, AFC tasks are preferred to yes–no tasks, as shown by Blackwell and others (e.g., Derrington & Henning, 1981). Hence, in this article, we will focus on forced-choice tasks and the method of constant stimuli, but we will explore whether we can improve the efficiency of m-AFC by using more alternatives—m taken from $\{2, 4, 8\}$—without sacrificing reliability, inferred validity, and sensory determinacy (low threshold values).

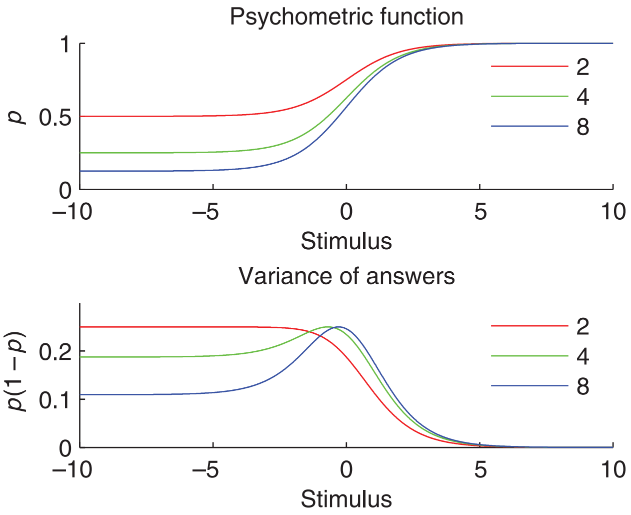

Figure 1. A psychometric function for different chance levels. We consider 2-, 4-, or 8-AFC and assume that the shape of the psychometric function does not change. If one were to sample N points from one location on one of these psychometric functions, the number of correct answers is assumed to be binomially distributed with a probability of success p. In this case, the variance of the number of correct answers would be $Np(1 - p)$. Hence, in the lower panel, we show how the variance in the data changes with the stimulus. For 2-AFC, there are large regions in stimulus space where the variance in the data is very high.

However, increasing m in an m-AFC paradigm may introduce new problems that possibly outweigh this advantage. For 2-AFC, response biases are usually thought to be low for trained observers, and they can be corrected. This is not necessarily so when $m > 2$, which may pose an even more serious problem for naïve observers (Green & Swets, 1988). Any assessment of a psychophysical method should thus include an estimation of response biases, an issue that was not yet well appreciated in Blackwell's days.

Summary and outlook

In this study, we assessed the thresholds (sensory determinacy), reliability, efficiency, and bias of different m-AFC tasks in naïve and experienced observers. We followed Blackwell's Recommendations 3, 4, and 5, namely, block design with as few signal intensities as practicable, combined with feedback. The number of signal levels, the amount of randomization, and whether or not feedback was provided were the variables that contributed to inferred validity in Blackwell's original study. We kept all those at the values Blackwell found to be optimal, as we see no reason to change those or explore them again. Our main aim is to explore the consequences of giving up Recommendation 6, experienced observers, on Recommendations 1 and 2, the number of response alternative in forced-choice methods and whether we should use temporal or spatial intervals.

As contrast sensitivity functions (CSFs) are well understood and have clinical relevance, they have frequently been used to assess psychophysical methods (Higgins et al., 1988; Long & Tuck, 1988; Woods & Thomson, 1993), and therefore, we will also use them. Contrast sensitivity measurements are based on a detection task. As the purpose of this study was to give general recommendations for psychophysical studies, one potential concern is that our recommendations may be specific to detection tasks and might not hold true for, say, discrimination tasks. Blackwell (1952) only considered detection tasks, as some properties of different psychophysical methods are probably independent of the sensory aspects of a task, for example, statistical efficiency, reliability, and response biases. These properties can be studied in a detection task and are extremely likely to generalize to other tasks. Ideally, psychophysical methods only measure the sensory aspects of a task. In both detection and discrimination tasks, one is interested in the sensory limits to perception. However, different methods have different extrasensory components that potentially "contaminate" the purely sensory aspects that one seeks to measure. For example, if two stimuli that have to be compared are presented in successive temporal intervals, it will be unavoidable for the threshold measurements to have a memory component. If two stimuli are presented simultaneously at different positions on the screen, it will be unavoidable for the threshold measurements to have a spatial attention component. As the resources that are necessary to perform a task are highly dependent on the task, we decided to study not only a detection task but also a discrimination task.

This article has two main sections (other than the Introduction and the Conclusions section). First, we examine *detection* in a standard CSF measurement. Our main finding is that spatial 4-AFC is the most reliable and most efficient method. The other section is concerned with *discrimination* in a sinusoidal contrast discrimination task. Here, we concentrate on the extrasensory components that influence the results for a spatial 4-AFC task, namely, spatial attention. For discrimination, 4-AFC is still the most efficient method; however, the thresholds are no longer the lowest—but they are consistent across experimental variations.

Detection

We measured CSFs with five different forced-choice procedures. In all cases, the participants had to perform a detection task. They had to indicate in which of the alternative positions or intervals they were able to detect a low-contrast sine grating. We used spatial AFC tasks with two, four, and eight alternatives. These were compared with temporal IFC tasks. In the spatial AFC tasks, it is not possible to have all alternatives foveated, but the alternative positions were very close to the fovea. To assess possible effects of retinal eccentricity, we compared the spatial AFC tasks to temporal 2-IFC with stimuli in the fovea and also with stimuli in the near periphery, at the same eccentricity as that of the spatial AFC tasks. In short, we refer to the spatial tasks as 2-AFC, 4-AFC, and 8-AFC and we refer to the temporal tasks as 2-IFCf (fovea) and 2-IFCp (periphery).

Methods

Stimuli were presented on a Clinton Monoray CRT; the monitor was controlled by a Cambridge Research Systems VSG 2/5 graphics controller driving the monitor at 150 Hz noninterlaced with a spatial resolution of 848 × 636 pixels. The display system was linearized using a Cambridge Research Systems OptiCAL photometer. Background luminance was measured to be 50 cd/m^2; none of the detection targets presented changed the mean luminance of the display. Pixels on the monitor were carefully adjusted to be square with 0.39-mm sides. Observers sat in a dimly lit experimental cubicle that was an arm's length away from the screen (38 cm) with their heads on a chin rest; observers viewed the screen binocularly. The experiment was controlled by a special-purpose software using the MATLAB (MathWorks, Inc.) toolbox provided by Cambridge Research Systems. Stimuli (targets) were horizontally oriented sine wave gratings at five different spatial frequencies: 0.5, 1.1, 2.1, 4.3, and 8.5 cpd, corresponding to wavelengths of 32, 16, 8, 4, and 2 pixels on the screen. All stimuli were bitmaps with a size of 99 × 99 pixels (5.9°); they were spatially vignetted using a modified Hanning window with a central circular patch of full contrast of radius 25 pixels (diameter, 3°), and beyond this radius, the stimulus contrast was ramped down to zero with a cosine at a radius of 25 to 50 pixels (ring of diameter, 3–5.9°). The (spatial) AFC methods presented the alternatives simultaneously on the screen. In the 8-AFC task, the possible locations on the screen were determined by the cells of a regular 3 × 3 grid; the central cell of the 3 × 3 grid was not used. The eight possible locations for the center of the stimulus were (−50,−50), (−50,0), (−50,50), (0,−50), (0,50), (50,−50), (50,0), and (50,50) pixels from the center of the screen. For the 4-AFC task, only the corners of this square were used. For the 2-AFC task, only the locations left and right of the center were used. This means that the stimulus appeared only at 2.9° eccentricity in the 2-AFC task, only at 4.2° in the 4-AFC task, but at both eccentricities in the 8-AFC task. A pilot study (data not shown) indicated that this difference in eccentricity did not have a large and systematic influence on the detection thresholds. This is because the stimuli were large (5.9°/99 × 99 pixels) compared with the differences in center offsets. The observers' responses during *m*-AFC were collected using a touch screen. The touch screen (IntelliTouch, ELO TouchSystems, with a 1,200 × 1,000 pixel resolution) was mounted as close as possible in front of the monitor using a frame that was built for this purpose. Pilot data (not shown) indicated that the mapping between the monitor coordinates and the touch screen coordinates was a simple affine transformation. The observers' responses were calibrated to precision, which is in the range of a few millimeters, by means of a least squares fit to 18 well-defined calibration targets displayed on the monitor prior to each experimental session. Pilot data indicated further that the variability in an observer's pointing movements to well-defined targets on the monitor had a standard deviation of 10 pixels (4 mm). For the 2-AFC and 4-AFC tasks, the response cells were 100 × 100 pixels in size, and responses could, thus, always be assigned to cells unambiguously. In the 8-AFC task, however, response cells were only 50 × 50 pixels in size. In our pilot study, we found very occasional misassignments for response cells of this size when observers were instructed to respond as fast as possible. In the experiments reported in this article, however, we instructed all observers to point as accurately as possible without undue time pressure; thus, we do not expect our data to be contaminated by a significant number of misassignments. Each trial for all conditions started with a fixation cross that was displayed at the center of the screen.

In an *m*-AFC task, the centers of the *m* different alternatives where a target could appear were marked with single pixels. One of the *m* alternatives was picked randomly and independently on every trial with equal probability. After 100 ms, a beep indicated the start of the stimulus presentation. The fixation cross and the marks for the alternatives disappeared, and 200 ms later, the target sine wave grating was presented. The temporal characteristics of the target presentation followed a modified Hanning window (100 ms fading-in using a cosine ramp, 100 ms nominal contrast presentation, 100 ms fading-out using a cosine ramp). Another beep indicated the end of the presentation; fixation cross and marks reappeared, and observers touched the mark on the screen where they believed the target to have been.

There were two variants of the (temporal) 2-IFC task: one with the stimulus presented in the fovea (2-IFCf) and one more peripheral (2-IFCp) at 2.9° eccentricity. In the foveal condition, the stimuli were presented at the position of the fixation cross; in the peripheral condition, there was also a little mark that was 50 pixels (2.9° eccentricity) above the fixation cross to indicate where the center of the stimulus would appear in one of the two intervals. The interval containing the target sine wave grating was chosen randomly and independently on every trial with equal probability. The temporal characteristics of the target presentation were the same as described above for the *m*-AFC task (nominal presentation time, 300 ms); the interstimulus interval had a length of 700 ms, and the beginning of both observation intervals was marked by a beep. A third beep prompted the observer to respond using a button box.

In all conditions, observers of the *m*-AFC, as well as the IFC variants, received auditory feedback as to whether their response had been correct. Altogether, there were 25 experimental conditions per observer: five spatial frequencies (0.5, 1.1, 2.1, 4.3, and 8.5 cpd) and five methods (2-AFC, 4-AFC, 8-AFC, 2-IFCp, and 2-IFCf). For each of the conditions, the psychometric function relating the probability of a correct response to contrast was obtained. We obtained two psychometric functions with eight stimuli and 400 trials each on two different days for each observer and condition. This allows us to assess the stability of the psychometric function over time. Four naïve observers (K.P., F.E., R.Z., and D.C.; two female, two male; mean age, 25 years) and one highly experienced psychophysicist who has performed at least 1 million 2-IFC trials during his distinguished career (G.B.H.) took part in this study. In total, we thus conducted 5 × 5 × 5 × 800 = 100,000 detection trials (the total number was not exactly 100,000: Participant D.C. performed only 400 trials instead of 800 trials for 2 of his 25 conditions and other participants did more than 800 trials for some conditions. The total number of trials that we analyzed is 107,850. If we were to include the 5 trials at the beginning of each block that were discarded for the analysis, we then would have 118,635 trials altogether).

All observers had normal or corrected-to-normal vision. The naïve observers were paid for their participation. Observers came to the laboratory for at least 10 sessions; each session took 2–2.5 hr. At this time, we collected between 30 and 40 blocks with 55 trials each using the method of constant stimuli. We followed Blackwell (1952) and presented only one stimulus intensity per block (block design). We considered the first 5 trials of each block as practice and discarded them. The naïve observers had very little or no experience of taking part in psychophysical experiments. For data analysis, we used psignifit (Wichmann & Hill, 2001a, 2001b), a MATLAB toolbox for psychometric function fitting. In addition to the thresholds and slopes of a fitted psychometric function, this toolbox also calculates confidence intervals for these values by means of parametric bootstrapping. For one of the naïve observers (R.Z.), we ran an additional session using the different *m*-AFC tasks and controlled eye movements with an eye tracker (Eyelink II, SR Research). This was done to check whether it was possible for the naïve observers to keep their fixation on the fixation cross during the various spatial *m*-AFC conditions. Fixation stability of R.Z. was very good: In the extremely rare case of R.Z. initiating a saccade, it would only start after the stimulus had already disappeared.

Results

Sensory determinacy

For each observer, we combined the data for each condition and fitted a psychometric function using maximum likelihood as implemented in psignifit (Wichmann & Hill, 2001a, 2001b). The performance of the participant for a stimulus s

$$\Psi(s) = \frac{1}{m} + \left(1 - \frac{1}{m} - \lambda\right) F(s;\alpha,\beta) \quad (1)$$

depends on chance performance $1/m$, a lapse parameter λ, and a function F with two parameters α and β that control the threshold and the slope of the psychometric function. Here, we have always chosen F to be a Weibull function. Figure 2 shows the functions for one observer at one spatial frequency measured with the five different tasks.

We define the 50% threshold to be the contrast at a performance level that lies halfway between chance level (given by $1/m$) and the maximum performance of the observer (determined by his or her lapse rate), that is, the stimulus s for which $F(s;\alpha,\beta) = 0.5$. Contrast sensitivity is defined as the reciprocal value of this 50% threshold.

All CSFs (for all five observers and all five methods) can be seen in Figure 3. The vertical lines are 95% confidence intervals that were found by bootstrapping the fitted psychometric function. All tasks recover the characteristic shape of a CSF, and differences between the tasks are small compared with the effect of varying spatial frequency. The experienced observer (G.B.H.) is older than the other observers and has a lower sensitivity, especially at higher spatial frequencies (Higgins et al., 1988). For each observer and spatial frequency, we calculated the minimum, the maximum, and the mean sensitivity over the different tasks. These quantities are shown in Figure 3 as dotted lines. The four inexperienced observers (K.P., F.E., D.C., and R.Z.) consistently show higher sensitivities than the mean for 4-AFC, whereas 2-IFCf results in sensitivities lower than the mean. The highly trained observer (G.B.H.) with years of experience in 2-IFC paradigms shows the opposite pattern and exactly replicated the findings of Blackwell (1952). To emphasize this point, we calculated for each method the differences from the mean sensitivity (central dotted line in Figure 3). We call this quantity the sensitivity difference. The sensitivity differences for all naïve observers (mean over spatial frequencies) can be seen in the left panel of Figure 4. The right panel shows the same for the highly trained observer.

All naïve observers show the lowest sensitivities for the 2-IFC variants whereas the experienced and highly trained observer showed the highest sensitivity for 2-IFCf. For the naïve observers, thresholds are consistently lowest for 4-AFC. An odd finding is that the naïve observers seem to perform better in the peripheral 2-IFCp task than in the foveal 2-IFCf task, but note that this difference is much smaller than the difference of 2-IFCf to 4-AFC.

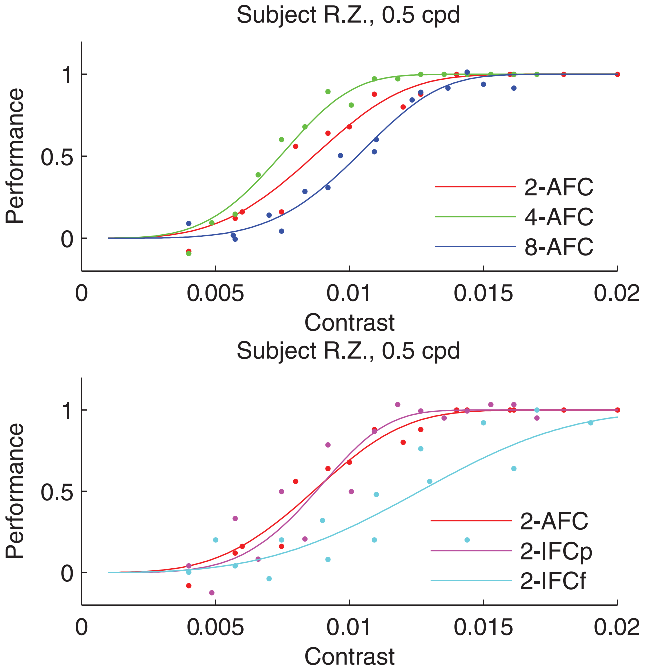

Figure 2. Psychometric functions Ψ were fitted (maximum likelihood) to the data from observer R.Z. at one spatial frequency (0.5 cpd). The figure shows the best fitting functions F ranging from 0 to 1 to facilitate comparison between conditions. The upper panel shows the AFC tasks, and the lower panel shows the tasks with two alternatives or two intervals (the 2-AFC condition is shown in both panels).

Reliability

Reliability is commonly assessed in a test–retest design. We measured each psychometric function on two different days with each method. Figure 5 shows the logarithm of the threshold values for the first and second measurements of the psychometric function. The correlation coefficient for the test and the retest thresholds is generally very high (>0.9) for all methods even for the naïve observers. For 4-AFC and 8-AFC, it is exceptionally high (around 0.99). For the naïve observers, the second block of trials results in lower thresholds than expected by chance alone (2-AFC: 14/20, 4-AFC: 14/19, 8-AFC: 14/20, 2-IFCp: 16/19, 2-IFCf: 13/20; one observer did not participate in a second block of trials for two of the methods). This is an indication that the naïve observers were still learning. This is both undesirable and unavoidable if one is working with naïve observers, children, or patients. The absolute size of the learning effect was small, however, as can be seen in Figure 5, where the points all cluster around the positive diagonal. Wichmann and Hill (2001a) describe a statistical test to detect (severe) data contamination by learning; the learning effects of our observers

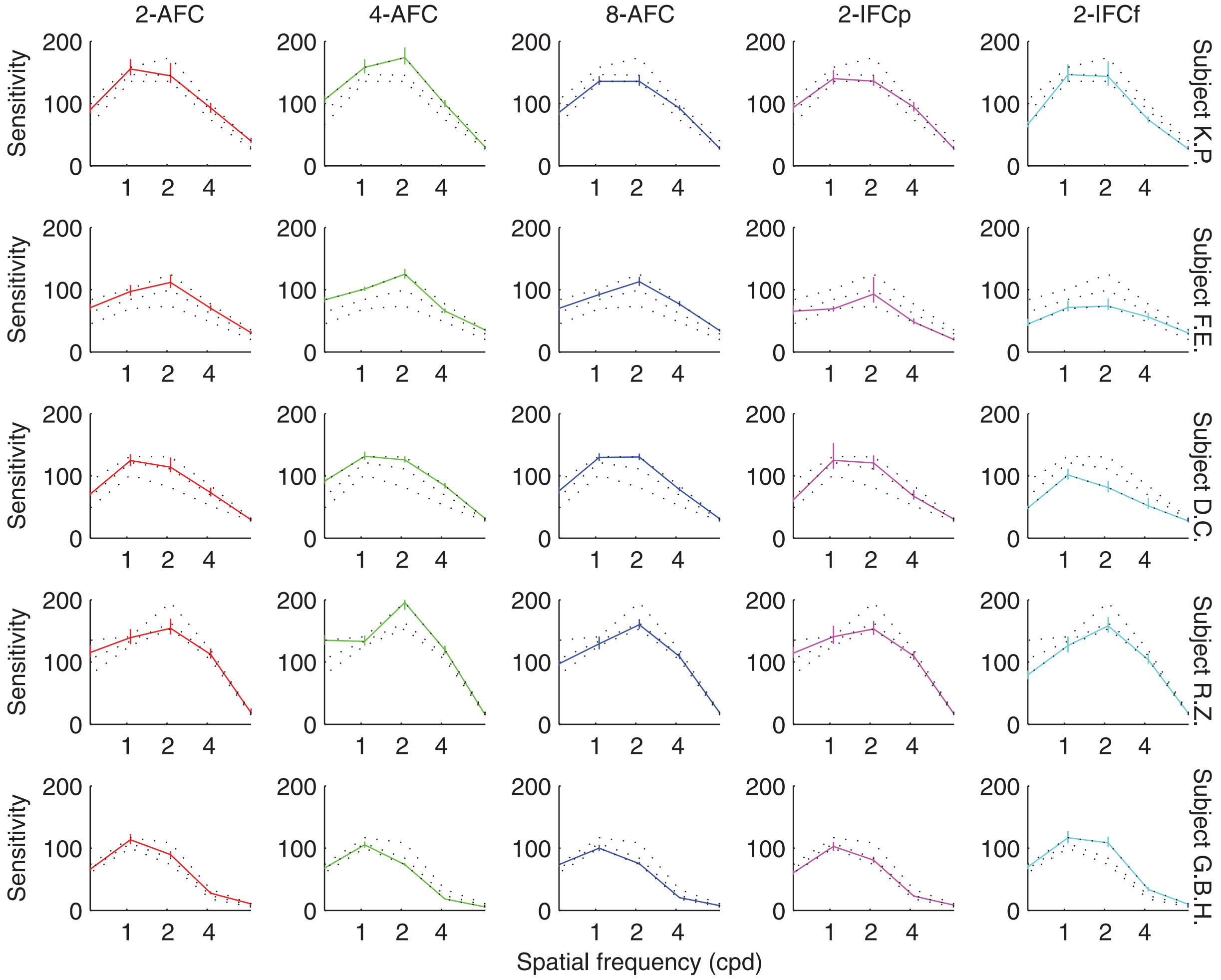

Figure 3. CSFs for all observers determined with the five different tasks. Vertical lines depict bootstrapped 95% confidence intervals. The dotted lines show the minimum, mean, and maximum sensitivity that the observer reached for each spatial frequency.

were not large enough to be noticed in this test. For the experienced observer (G.B.H.), all test–retest correlations are extremely high (>0.99) for all methods, and he shows remarkable stationarity in his responses (48% of the measured thresholds were lower in the first block of trials, whereas 52% were lower in the second block of trials).

In addition to the test–retest correlation, we examined the goodness of fit for the measured psychometric functions. The model for the psychometric function assumes a stationary psychometric function with a binomial distribution of correct responses. If the psychometric function is not stationary for a participant but is subject to random fluctuations or learning effects, this will result in overdispersion; that is, the observed variability of the responses is higher than expected from the binomial model (Collett, 1991; Prentice, 1986; Williams, 1982). If there is noise in the data that cannot be accounted for by the model, one will find bad model fits. Hence, by assessing the goodness of fit, we might find reason to doubt the assumed stationarity of the psychometric function, which, in turn, might indicate low reliability of a psychophysical method. As confidence intervals for the parameters of the psychometric function are obtained by using the model assumptions, a violation of the stationarity assumption can lead to overconfidence in the obtained parameter estimates. Therefore, model fits should always be accompanied by an examination of goodness of fit.

To assess goodness of fit, we calculated the deviance of the data from the fit. Deviance is a common summary statistic (Wichmann & Hill, 2001a), not unlike a χ^2 statistic, but more generally appropriate for maximum-likelihood fitting in contexts other than the least squares setting. Given a fit, Monte Carlo simulations can determine the distribution of this summary statistic against which the calculated value can be compared to judge significance.

We observed an unexpectedly high number of unlikely deviance values indicating bad model fits for our naïve observers. We analyzed each experimental method (2-, 4-, and 8-AFC as well as the 2-IFC variants) separately to see whether some result in worse fits than others. Recall that each fitted psychometric function comes from two runs on two different days with eight blocks each (each block is 50 trials). To minimize the effect of learning found

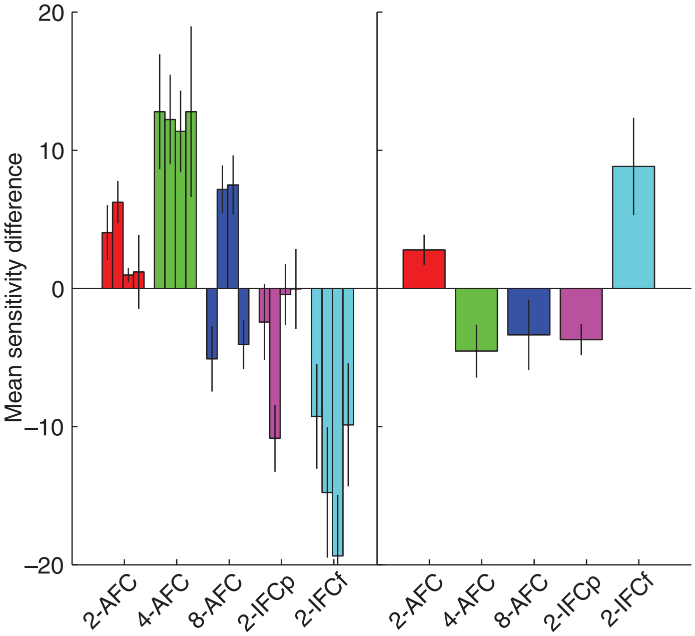

Figure 4. Difference of sensitivity from mean sensitivity for the five different methods averaged over all spatial frequencies. The left panel depicts the naïve observers; the right panel depicts the experienced observer. Vertical lines are standard errors.

previously—which will result in overdispersion—we only analyzed the data from the second day. For all five observers in all conditions, that is, for each of the 25 combinations of the five spatial frequencies and the five methods, we have a fit and, therefore, a deviance value. This deviance value can be compared with the distribution of deviances that would be observed were the fit is the true psychometric function. For the four naïve observers, 5 of the 20 (25%) psychometric function fits for 2-AFC could be rejected at the 5% significance level. For 4-AFC and 8-AFC, 15% and 30% of the fits could be rejected, respectively. For 2-IFCp and 2-IFCf, the values are 20% and 45%, respectively. For the experienced observer, on the other hand, only 2 of 25 fits (8%) could be rejected at the 5% level—we would have expected 1 or 2 by chance alone; thus, this is consistent with a stationary binomial observer.

Obviously, a misspecification of the parametric form of the psychometric function could also lead to bad model fits. However, for the Weibull function that we used in all cases, there were no obvious systematic deviations from the model as assessed by inspecting the residuals of the fit (Wichmann & Hill, 2001a). It is thus likely that the bad model fits for the naïve observers are due to random nonstationarity in their psychometric functions. The bad model fits should therefore be a warning signal that there are noise components in the data that the model does not capture. This, in turn, could lead to overconfidence in the estimated parameters of the psychometric function. However, comparing the confidence intervals for the threshold after 400 trials with the differences we found in the test–retest situation, we see that they are of the same magnitude. Therefore, the violations of the model assumption of binomially distributed data do not seem too grave. This is only true, however, as long as one is only interested in threshold measurements and not in the slope of the psychometric function. For comparisons of slopes using naïve observers, one needs to obtain more conservative confidence intervals by taking the overdispersion into account. Luckily, several ways to deal with overdispersion have been suggested in the literature (Collett, 1991; Prentice, 1986; Williams, 1982).

For the naïve observers, we found the most stable stationary psychometric functions for 4-AFC—both the lowest number of goodness-of-fit rejections and a very high test–retest reliability. Recall that it was for 4-AFC, too, that we found the lowest thresholds (highest sensory determinacy in Blackwell's terminology). The common 2-IFCf, on the other hand, has the worst fits, the highest thresholds, and the smallest test–retest correlation for naïve observers. The experienced observer, by contrast, behaved just as desired, that is, as a stationary binomial observer, seemingly unfazed by whatever method we put in front of him.

Efficiency

Our measure of efficiency is the time needed to collect sufficient data per psychometric function given a desired precision target. The mean time for one psychometric function with 400 trials (8 blocks with 55 trials, where the first 5 trials of a block are discarded) was 17 min for the AFC methods and 28 min for IFC (excluding breaks between the blocks). AFC methods only need about 60% of the time of the IFC paradigms.

The second component of our efficiency measure—the time needed per psychometric function given a desired precision target—is the size of the bootstrapped confidence intervals. As shown in Figure 1, we expect a greater m to lead to smaller confidence intervals. In the following, we only consider the most frequently used summary statistic: the 50% threshold (as before, by 50% threshold, we refer to the stimulus intensity s such that $F(s;\alpha,\beta) = 0.5$ and *not* $\Psi(s;\alpha,\beta) = 0.5$). For each fit, we also calculated bootstrapped standard deviations for this quantity. However, some care has to be taken when one tries to compare these because for sinusoidal grating detection, the slope of the psychometric function is correlated with the threshold (for K.P., the correlation was 0.80; for F.E., it was 0.52; for R.Z., it was 0.83; for D.C., it was 0.64; and for G.B.H., it was 0.83), and therefore, higher thresholds imply larger confidence intervals. As we have found that IFC has higher thresholds (Figure 4), a direct comparison of the confidence intervals would thus not be fair. Instead, we calculate the ratio of the bootstrapped standard deviation to the threshold. We

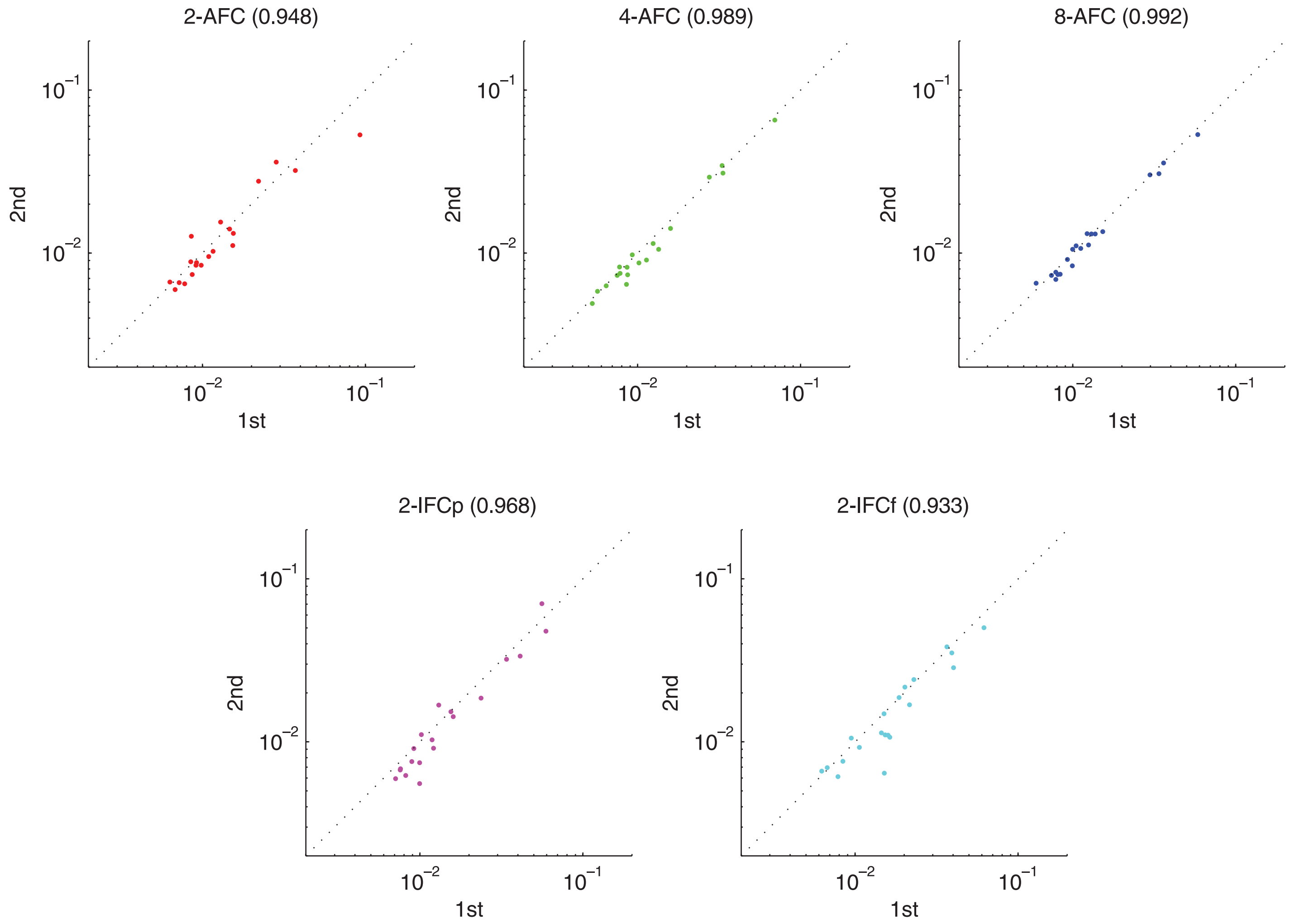

Figure 5. Test–retest results for the naïve observers in all five methods. Each psychometric function was measured in 2 days, and the logarithm of the thresholds was extracted. The correlation between the two measurements is given in parentheses.

call this measure threshold uncertainty. We determined threshold uncertainties in the two cases that we have already considered before: (a) the two runs from the two different days are pooled (which means 800 trials per fit) and (b) one psychometric function for each day fitted separately (only 400 trials per fit but twice the number of fits). Median values for both cases are given in Table 1. The median value for 4- and 8-AFC is smaller than that for 2-AFC and 2-IFC. The improvement is substantial as the median uncertainty for 4- and 8-AFC after 400 trials is already smaller than that for 2-AFC and 2-IFC after 800 trials. For the timings of our study, this means that 17 min of 4-AFC provide as much information about threshold as 56 min of 2-IFC—lower thresholds and higher reliability in less than a third of the time.

We explored this efficiency benefit in more detail by applying a nonparametric bootstrap technique. For each of the five observers, spatial frequencies, and methods—2-, 4-, and 8-AFC, 2-IFCf, and 2-IFCp—we have 800 trials per psychometric function in 16 blocks of 50 trials. We generated subsamples of these 800 trials with $N = \{50, 100, 150, \ldots, 800\}$ trials, to which we fitted psychometric functions and obtained confidence intervals as described above. The subsamples were taken randomly in proportion from each of the 16 blocks; for example, for a subsample of 100 trials, 12 of the blocks would be represented by 6 trials and 4 blocks would be represented by 7 trials. Thus, for each complete resampling run, we obtain 25 threshold uncertainties, one for each observer and spatial frequency. We repeated this procedure several times (the number of

	2-AFC	4-AFC	8-AFC	2-IFCp	2-IFCf
400 Trials (%)	5.1	3.7	3.5	5.0	6.1
800 Trials (%)	4.3	2.7	2.5	3.9	4.5
Ratio	1.19	1.37	1.40	1.28	1.35

Table 1. Efficiency of the different methods. For each fit, we obtained a standard deviation for the 50% threshold by means of parametric bootstrapping. Threshold uncertainty is the ratio of the standard deviation to the threshold. The median threshold uncertainty for all fits is given.

repetitions was different for each N, starting with 50 repetitions for a subsample of 50 trials and going down to 1 repetition for 800 trials), resulting in a distribution of threshold uncertainties for each N. Figure 6 shows the medians and 25–75% quantiles of the threshold uncertainty distributions thus obtained. The solid lines in Figure 6 plot the expected decrease in confidence interval widths based on the intervals for $N = 800$, scaled $\sim 1/\sqrt{N}$.

From Figure 6, it is apparent that, for example, for a threshold uncertainty of 5% under the realistic conditions of our study, one requires fewer than 250 trials for 4- and 8-AFC procedures but around 550 trials for the 2-IFC procedure, which is a 2.2-fold increase, consistent with estimates above. Combined with the 28/17 = 1.65 time factor, this means a more than 3.6-fold increase in overall efficiency for 4-AFC over 2-IFC. Note, finally, that most of the benefit comes from the increase from two to four alternatives; 8-AFC is not significantly more efficient than 4-AFC but shows considerably less sensory determinacy and reliability.

Overall, the threshold uncertainty values may appear small once the number of trials is larger than 300 or 400 per psychometric function. However, one should recall that the bootstrapped standard deviations are computed using the model assumptions, which we have found to be violated (i.e., data of naïve observers are overdispersed, showing extrabinomial variance—see the Reliability section). However, as explained above, we do not believe that these violations are grave, and therefore, the confidence intervals we report are at least in the right range.

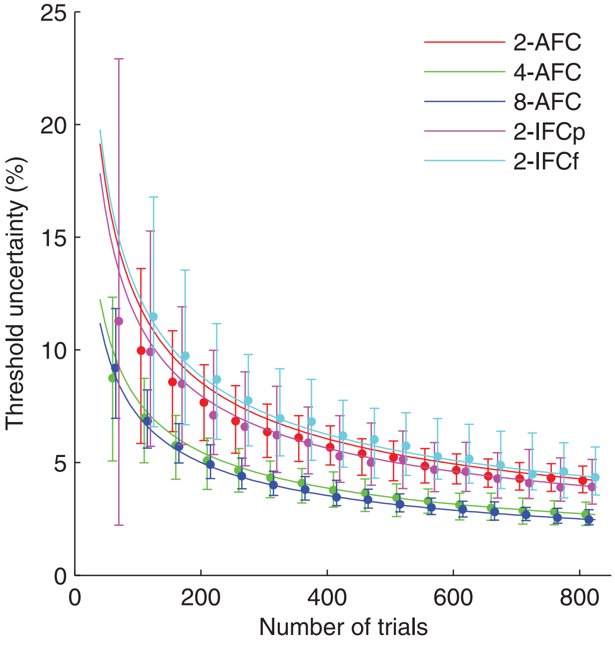

Figure 6. Threshold uncertainty for our five observers on the y-axis shown as a function of the number of trials on the x-axis; see text for details.

Bias

Researchers are often worried about possible contamination of their threshold measurements by response bias. The use of forced-choice methods avoids the worrisome bias stemming from criterion placement or shifts in single-interval (yes–no) methods, but it may suffer from a different type of bias: temporal or position bias. An observer may have a preference (bias) for a certain interval in time (IFC) or a position in space (AFC). Some observers may be biased to see stimuli on the left half of the screen, whereas other observers may be biased to see stimuli on the right. Bias may not only be perceptual; however, right-handed observers may, for example, be biased to press the right button on the response keyboard. We will assume that this type of bias for a certain position in time or space is constant irrespective of the performance level but that it can, of course, be overcome given sufficiently strong sensory information. In theory, such a bias can be corrected for using methods from SDT. However, standard tools from SDT give little guidance in the case considered here. First of all, SDT usually deals with the case where the bias is varied but the performance level is fixed. The objective of the researcher is to find a good measure for the sensitivity and treat the bias as a nuisance parameter. This is in contrast to measurements of psychometric functions where the sensitivity is varied and the bias is assumed to be constant. There are several bias measures in the literature, but there have been few attempts to trace isobias curves to validate these measures (Dusoir, 1983, 1975). Secondly, SDT has seldom been applied to the m-AFC case where m is greater than 2. The reason for this is that the mathematics required for the generalization to the m-alternative case is “rather clumsy” (Luce, 1963) and the numerous assumptions necessary have much less empirical support than those required for yes–no or 2-AFC methods. Luce’s choice model, on the other hand, is much simpler and, in most cases, is a viable alternative to SDT. For yes–no, Luce’s choice model leads to ROC curves that are very similar to the Gaussian equal variance signal detection model. The few studies that compared the signal detection model to Luce’s choice model have found that the signal detection model fits the data slightly better but that Luce’s choice model is in any case a very good approximation (Luce, 1963, 1977; Treisman & Faulkner, 1985). However, for our purposes, Luce’s choice model has the advantage that it is straightforward to generalize it to m-AFC (Luce, 1963) and that the bias term is easy to interpret. Hence, we will use Luce’s choice model to separate sensitivity from response bias, be it temporal or position bias. In this model, the probability to respond with alternative i given that the stimulus s is presented at alternative j is given by

$$p(\text{resp. } i|s \text{ at alternative } j) = \frac{\eta_{s,i,j} b_i}{\sum_{k=1}^{m} \eta_{s,k,j} b_k}. \qquad (2)$$

The parameters b_i can be interpreted as bias terms. If their sum is normalized to 1, they give the a priori probability of the participant to respond with a certain alternative-irrespective of performance level. The $\eta_{s,i,j}$ model the sensitivity of the participant to stimulus s. If the sum of all $\eta_{s,k,j}$ over k is normalized to 1, we can interpret them as response probability for an unbiased observer. It is usually assumed that the probability that an unbiased observer correctly detects the stimulus does not depend on the alternative j at which it is presented; that is, the sensitivity is the same for all alternatives. If it is further assumed that, for an unbiased observer, the errors are spread evenly among all wrong alternatives, one parameter η_s is enough to model the sensitivity of the participant. In this case, the $\eta_{s,i,j}$ are chosen to be $\eta_{s,j,j} = \eta_s$ and $\eta_{s,i,j} = (1 - \eta_s)/(m - 1)$ for $i \neq j$. The model can be fitted by maximizing the likelihood of the data and optimizing over the m response bias terms and the sensitivity term for each block. This is what we have done to all participants and to all our methods. One example for observer D.C. is shown in Figure 7.

For three observers—including the experienced observer—we found a strong bias for the second interval for 2-IFC. The a priori probability of these observers to answer with the second interval was around 65%. The other two observers only had a small bias in 2-IFC (about 55%). All naïve participants were virtually unbiased in 2-AFC (less than 52% preference for one side). The left-handed experienced observer had an a priori probability of 56% for the left alternative. Note that for 2-AFC, we were using a touch screen and the participants were free to use whichever hand they preferred. 4-AFC and 8-AFC showed a less clear pattern, but for the right-handed naïve participants, there was a tendency to prefer the lower right—consistent with a bias for minimal arm and hand movements. For example, for the most biased participant, we found that, in 4-AFC, the top left, top right, lower left, and lower right had an a priori probability of 15%, 23%, 29%, and 33%, respectively.

With an estimate of the biases of our observers, it is now possible to correct for the bias. How many correct responses would the participants have had if they had been unbiased? Note that this means increasing the number of correct responses for unfavored responses but decreasing the number of correct responses for preferred responses—the overall number of correct responses in a block is only affected if there is a significant net gain. If we compare the observed number of correct responses in a block of 50 trials to the bias-corrected number of correct responses for this block, we find that the difference between the two is less than half a trial on average (2-AFC: 0.07 trials, 4-AFC: 0.2 trials, 8-AFC: 0.44 trials, 2-IFCf: 0.45 trials, 2IFCp: 0.44 trials). We also compared the thresholds that are obtained with bias-corrected blocks to the ones we found before: The improvement in the thresholds due to bias correction is much smaller than the size of the confidence intervals. Overall, we thus find observers to be more biased in 4-AFC than in 2-AFC, but we find them to be biased in 2-IFC too. Most important, however, we show that for all practical purposes, the influence of these biases on threshold estimation and confidence-interval width determination is all but negligible.

Discussion

We find a striking and consistent discrepancy between naïve observers' behavior and that reported for experienced observers by Blackwell (1952), which was replicated in

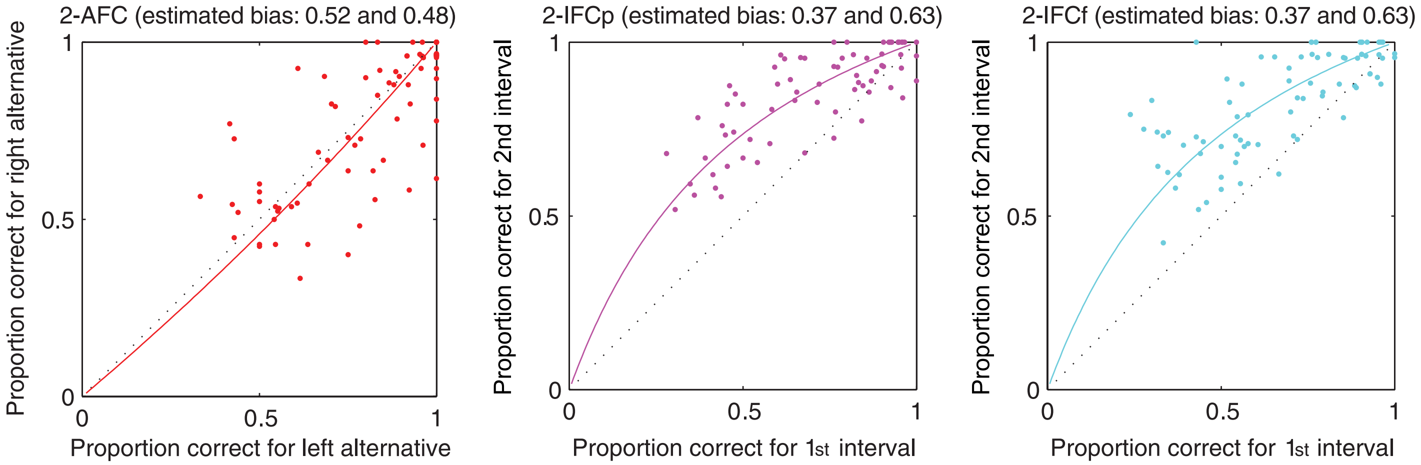

Figure 7. Response biases and isobias curves for participant D.C. for the three methods that involved only two possible responses. Each data point is a block of trials. The axes are the proportion of correct answers for a stimulus that is shown in the first or second interval or on the left or right side. The solid line is the maximum likelihood fit using Luce's choice model. For the IFC methods, there is a strong bias toward the second interval. In cases where the participants are unable to detect the stimulus, they give the correct answer in 63% of the trials if the stimulus is in the second interval but only in 37% if it is in the first interval.

our study by the one highly experienced observer we used. For the latter, 2-IFCf is the best psychophysical method in terms of sensory determinacy. 2-IFCf showed excellent reliability, too, although this was true for all the forced-choice methods for our experienced observer. (The biggest factor influencing reliability in Blackwell was yes–no versus forced choice; hence, we did not expect to find a large effect between the forced-choice variants for the experienced observer.) Naïve observers, on the other hand, performed worst during 2-IFC, in terms of both reliability and sensory determinacy, and they performed best for spatial 4-AFC. This is not just an overall group effect: Each of our four naïve observers showed the highest sensitivity and reliability during 4-AFC (see Figures 3 and 4). In addition, of all methods tested, 2-IFC showed the largest bias. For threshold estimation, the bias was shown not to be critical but there may be applications or circumstances under which this, too, may constitute an argument against the use of 2-IFC. Note that this large bias was consistently found for both our naïve and experienced observers.

Because our experienced observer replicated the findings of the Blackwell (1952) study, it is exceedingly unlikely that our equipment or laboratory routines mysteriously favored 4-AFC over 2-IFC. Rather, it appears as if during lifelong psychophysical training, performance for 2-IFC keeps improving, whereas performance for the other methods appears not to benefit as much from extensive practice—this was true back in 1952 and is still true today. This is, in its own right, reassuring. We can offer no explanation of why this should be or of what mechanisms may be responsible for this change in optimal psychophysical methods with increasing psychophysical training.

Discrimination

A potential concern is that the advantage of 4-AFC over the other methods may be specific to threshold measurements in *detection* tasks and does not generalize to the arguably more common *discrimination* tasks. Thus, we also explored the influence of 2-IFC, 2-AFC, and 4-AFC on a contrast discrimination task (we are grateful to an anonymous reviewer for suggesting this to us).

The increased efficiency of 4-AFC is independent of the task: 4-AFC takes less time than 2-IFC and is statistically more efficient irrespective of whether a participant performs detection or discrimination. We also believe that response biases should be independent of the task. Therefore, we did not explore these factors for discrimination tasks.

Perhaps a surprising result of the experiments on detection was that, for naïve observers, thresholds are lowest for 4-AFC and highest for 2-IFC. We therefore examined the relationship between 2-IFC, 2-AFC, and 4-AFC in a contrast discrimination task with naïve observers. Introspectively, detection is preattentive: Participants monitor a uniform background and all they do is to note a change in their central visual field—this interpretation is at least consistent with our finding that focal attention as in 2-IFC does not yield lower thresholds than 4-AFC. For more complex discriminations, however, the thresholds may well be a function of attention and memory. To study possible attention effects, we considered different variants of a sinusoidal contrast discrimination task. In all tasks, participants were presented with sine gratings and were asked to choose the alternative with a higher contrast. In the "standard" discrimination task, the stimuli were circular patches with different contrasts. The circular patches were presented either consecutively (2-IFC) or simultaneously at different positions on the screen (2-AFC and 4-AFC).

We hypothesized that attention may be affected by the "objecthood" of the circular patches and therefore by the number of objects that have to share attentional resources. Thus, as a control, we included a discrimination task with only one object at a time: Instead of having four patches in the 4-AFC task, there was a large background with a sine pattern to which a contrast increment patch was added at one of four possible positions. This condition was intended as a hybrid between discrimination and detection: On the one hand, it is a discrimination task because participants have to respond to a contrast increment with a nonzero pedestal; on the other hand, this task is much more like detection because there is only one object against a patterned background—one location should "pop out" similar to a detection task. An alternative way in which attention may affect thresholds is by the size of the visual area that has to be monitored (size of the attentional "spotlight"). Thus, a second control condition manipulated the distance between the patches in 2- and 4-AFC.

Methods

The setup was the same as the one described above, except that we replaced the monochrome CRT (yellow phosphor) with a Sony F520 color CRT that was adjusted to have (almost) the same pixel size and linearized appropriately. The mean luminance of the Sony CRT was 48 cd/m^2, and the refresh rate was 140 Hz. The experimental procedures with all parameters were identical to those used for detection; the only change was in the stimuli.

For the standard discrimination task, the stimuli were circular patches of radius 40 pixels with an 8-pixel wavelength sine grating, corresponding to 2.1 cpd. This spatial frequency was chosen to be around the maximum of the CSF. The pedestal in the discrimination task had always a contrast of 0.1—nearly a log unit above the detection threshold of our participants and, thus, clearly above the pedestal dip, which is at a pedestal contrast approximately equal to the detection threshold (Bird, Henning, & Wichmann, 2002; Henning, Bird, & Wichmann, 2002; Wichmann, 1999). The contrast was actually generated by alternating frames with a contrast of 0.2 with mean luminance frames. The patch with higher contrast was generated by adding a sine patch that was ramped down with a modified Hanning window to the mean luminance of the interleaved frames. These stimuli were used in a 2-IFC, 2-AFC, and 4-AFC task, identical to the corresponding detection tasks.

For the hybrid task, the rationale was that it might be possible to have a discrimination task that feels more like a detection task. To this end, we had a background sine grating with a spatial frequency of 2.1 cpd, which covered the whole area where stimuli could occur in the 4-AFC, 2-AFC, and 2-IFC tasks. To avoid adaptation effects, the background was only present during the presentation intervals and not between trials or between intervals (when a fixation cross was shown on a mean luminance screen as in the detection task). The contrast of the sine wave background was 0.1, and again, this was achieved by interleaving a sine wave with a contrast of 0.2 with mean luminance frames. The contrast increment at the target positions was achieved by adding a sine patch (same as in standard discrimination) to the interleaved mean luminance frames.

Four new naïve observers (C.G.F., M.D.J., P.T., and V.K.M.) performed the standard discrimination task and the hybrid task. Psychometric functions for both tasks were obtained by means of 2-IFC, 2-AFC, and 4-AFC. At least eight blocks of constant stimuli with 55 trials for each of the six conditions were obtained for each participant.

In addition, one highly experienced (G.B.H.) and a moderately experienced (F.J.) observer performed the 4-AFC hybrid discrimination task with two different eccentricities of the targets. They performed the same task as the other participants (with a distance of 4.2° from the center), as well as an additional task, where the distance between the centers of the patches in 2- and 4-AFC was halved. This second condition leads to some overlap between the relatively large target regions (approximately 4° effective size).

Results

Figure 8 shows the thresholds for the four observers (C.G.F., M.D.J., P.T., and V.K.M.) who performed a standard discrimination task and the hybrid discrimination task described above. For the two tasks, thresholds were obtained by three different methods: 2-IFC, 2-AFC, and 4-AFC. The thresholds are, as in detection, given by the point halfway between chance and perfect performance (minus lapses) and were extracted from a maximum likelihood fit using a Weibull as the psychometric function. The independent variable is the contrast increment (in percentage) that was added to the pedestal with a contrast of 10%.

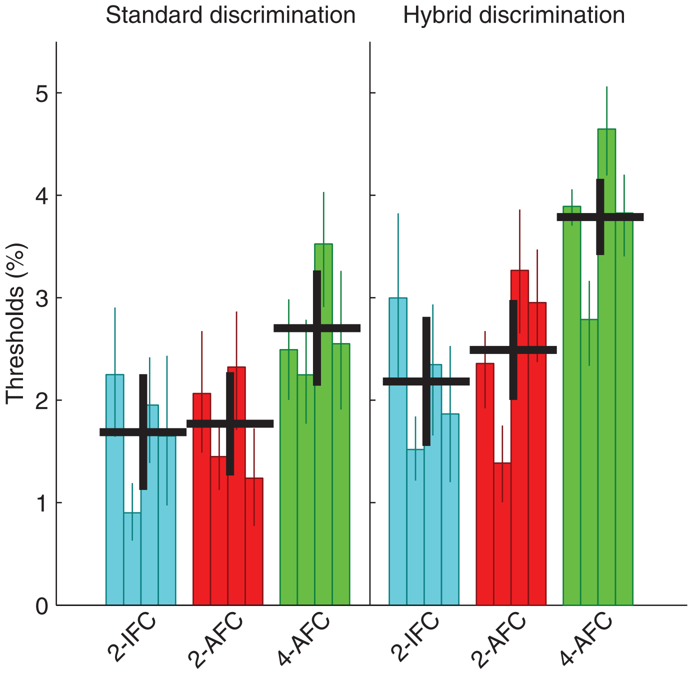

Figure 8. Thresholds for the standard and hybrid discrimination tasks: The thresholds are contrast increments (in percentage) from a 10% contrast pedestal. The first bar in each group corresponds to the first participant; the other bars correspond to the other participants. Vertical lines depict bootstrapped 95% confidence intervals. The solid black lines are the means over the participants with approximate 95% confidence intervals.

First, note that for all four observers (the four bars) and both tasks (the two panels), 4-AFC gives the highest thresholds. For most of the participants, the bootstrapped 95% confidence intervals for the threshold in 2-IFC and the threshold in 4-AFC do not overlap notably; thus, for discrimination, 4-AFC always gives the highest thresholds.

Second, the hybrid task almost always leads to higher thresholds than the standard discrimination task. The only exception to this is the threshold level of the second observer (M.D.J.) in the 2-AFC method. Eleven of 12 pairs (4 participants × 3 methods) show this pattern: a result that is extremely unlikely to occur by chance, strongly suggesting that "objecthood" is not a critical parameter that determines thresholds.

For the hybrid task, we also examined the dependence on the eccentricity of the targets. Two observers performed the same hybrid task that the other participants did, with an eccentricity of 4.2°, and an additional task with approximately half of this eccentricity. Observer F.J. had a threshold of 5.8 ± 0.53% (95% bootstrapped confidence interval) for the discrimination task with 4.2° eccentricity and a threshold of 3.81 ± 0.71% for the one with smaller eccentricity. For G.B.H., the thresholds were 5.75 ± 0.95% and 4.0 ± 0.52%, respectively, strongly suggesting that for attention-demanding discriminations, the size of the "attentional spotlight" is a crucial parameter.

Discussion

The ordering of the thresholds for the discrimination task is in stark contrast to the ordering for the detection task. In detection, 4-AFC consistently gave the lowest threshold, whereas in discrimination, 4-AFC leads to the highest thresholds. This dissociation is most likely due to attentional effects: Detection may be a preattentional task that requires only little attention, and therefore, increasing the number of possible spatial locations does not cost more attentional resources—at least up to 4-AFC because we found 8-AFC to be worse than 4-AFC. However, in sinusoidal contrast discrimination, observers may have to actively direct their attention to the possible spatial locations, and therefore, a greater number of possible spatial locations leads to a worse performance.

One could have expected that performance in the hybrid task is better because it is more similar to the simpler detection task. Instead of having to pay attention to four objects on the screen, there is just one object that "pops out" against a background. However, our results clearly show that the hybrid task is not easier than the standard discrimination task. Thus, the number of objects is not a decisive factor. Observers still have to monitor four spatial positions for the hybrid task in 4-AFC, and the objecthood of the targets seems irrelevant.

However, the discrimination performance decreases with the eccentricity of the targets. Initially, one may be led to think that this has only to do with the peripheral drop-off in sensitivity. However, in our detection experiments, this effect was not decisive: 2-IFC and 2-AFC had higher thresholds than 4-AFC, which has the greatest eccentricity. Furthermore, in the discrimination experiments, we only used stimuli of rather low spatial frequency (2.1 cpd). At this frequency, the sensitivity does not fall off very rapidly, and our stimuli were presented not very far in the periphery (the centers were at most ±5°). Hence, the most likely explanation appears to be that it is the size of the spatial area that has to be monitored by attention that matters. If attentional resources have to be shared over a larger area, then performance decreases.

While this seems to suggest that spatial attention can be a crucial factor for the size of the thresholds, this does not necessarily constitute a strong argument against the use of 4-AFC as a psychophysical method for discrimination experiments. As long as the attentional load is constant over different conditions, all thresholds in all conditions will be higher compared with the 2-IFC method. Thresholds measured by 4-AFC certainly have an attentional component of unknown size, but thresholds measured by 2-IFC have a similar problem with sensory memory. A priori, it is not clear which of these two extrasensory factors is more worrisome for threshold measurements. We can only state that for naïve observers in detection tasks, 4-AFC leads to lower thresholds than 2-IFC, suggesting that sensory memory is a limiting factor in detection. As 2-IFC leads to lower thresholds than 4-AFC in a discrimination task, discrimination tasks seem to be more limited by attentional resources than by sensory memory. Unless one compares different psychophysical methods with different extrasensory influences on thresholds, it is not clear how big the extrasensory components are. In most experiments, however, one is not interested in the absolute limits of sensory perception. Instead, one tries to find differences among conditions. If the size of the extrasensory component is approximately constant over the conditions, the differences between the conditions will be preserved.

Conclusions

Undoubtedly, visual acuity is highest in the central fovea for all observers, but this advantage for 2-IFCf seems to be offset for naïve observers perhaps due to stronger sensory memory demands of temporal forced choice. This is highly speculative, but experienced 2-IFC observers may have developed more accurate or efficient sensory memory, enabling them to benefit from the higher visual acuity in the fovea. For naïve observers, however, sensory memory seems to be an important limiting factor for detection thresholds. In discrimination, however, we found higher thresholds using 4-AFC than 2-IFC. This could indicate that attention factors are more limiting in discrimination tasks than in detection tasks. If there is any truth to our speculations, this highlights two important psychological aspects of psychophysics that are not part of standard SDT: attention and sensory memory. Recently, the first network model combining dynamical processing of sensory stimuli with short-term memory has been developed by Machens, Romo, and Brody (2005). Perhaps, observers compare sensory or memory representations of the stimuli during temporal forced choice and not just the scalar activation values on the putative internal decision axis. Other aspects, for example, concerning bias, are also known to be missing from SDT (Friedman et al., 1968).

In addition, we find 4-AFC to be much more efficient than 2-IFC. The statistical advantage is 2.2-fold and the advantage in time is larger than 1.6-fold, combining into a 3.5-fold decrease in experimental time required to reach the same threshold uncertainty. Thus, it may well be beneficial to use 4-AFC in discrimination tasks—or with experienced observers in detection—despite the lower sensory determinacy. Clearly, not every experiment can be run as spatial 4-AFC, as stimuli may be too large or may really require the highest possible spatial resolution, but in many standard psychophysical experiments, 4-AFC could be used instead of 2-IFC. This is particularly the case in clinical settings or when one uses children as observers.

Finally, mastering the art of being a truly experienced psychophysical observer appears to be arduous and to take a lifetime: Each of our four naïve observers was still better at 4-AFC than at 2-IFC at the end of our study, that is, after 20,000 detection trials each, which is a long way to go before they can compete with G.B.H.

Acknowledgments

We would like to express our sincere gratitude to Bruce Henning for his willingness and patience to act as the experienced "reference observer" for our study while on sabbatical leave visiting F.A.W. at Bernhard Schölkopf's Department of Empirical Inference, MPI for Biological Cybernetics, in Tübingen, FRG. Furthermore, we are indebted to two anonymous reviewers for their insightful comments and in particular to one of them for suggesting to compare detection and discrimination performance. In addition, we would like to thank Jez Hill and Bruce Henning for helping us improve our manuscript.

Commercial relationships: none.

Corresponding author: Frank Jäkel.
Email: frank@tuebingen.mpg.de.
Address: Spemannstraße 38, 72076 Tübingen, Germany.

References

Bird, C. M., Henning, G. B., & Wichmann, F. A. (2002). Contrast discrimination with sinusoidal gratings of different spatial frequency. *Journal of the Optical Society of America A, Optics, Image Science, and Vision, 19,* 1267–1273. [PubMed]

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

Burns, B. D., & Corpus, B. (2004). Randomness and inductions from streaks: "Gambler's fallacy" versus "hot hand". *Psychonomic Bulletin & Review, 11,* 179–184. [PubMed]

Collett, D. (1991). *Modelling binary data*. Boca Raton, FL: Chapman & Hall/CRC.

Derrington, A. M., & Henning, G. B. (1981). Pattern discrimination with flickering stimuli. *Vision Research, 21,* 597–602. [PubMed]

Dusoir, A. E. (1975). Treatments of bias in detection and recognition models: A review. *Perception & Psychophysics, 17,* 167–178.

Dusoir, T. (1983). Isobias curves in some detection tasks. *Perception & Psychophysics, 33,* 403–412. [PubMed]

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.

Friedman, M. P., Carterette, E. C., Nakatani, L., & Ahumada, A. (1968). Comparison of some learning models for response bias in signal detection. *Perception & Psychophysics, 3,* 5–11.

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

Green, D. M. (1990). Stimulus selection in adaptive psychophysical procedures. *Journal of the Acoustical Society of America, 87,* 2662–2674. [PubMed]

Green, D. M., & Swets, J. A. (1988). *Signal detection and psychophysics* (Reprint ed.). Los Altos, CA: Peninsula Publishing.

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

Higgins, K. E., Jaffe, M. J., Caruso, R. C., & de Monasterio F. M. (1988). Spatial contrast sensitivity: Effects of age, test–retest, and psychophysical method. *Journal of the Optical Society of America A, Optics and Image Science, 5,* 2173–1280. [PubMed]

Hill, N. J. (2001). *Testing hypotheses about psychometric functions*. Unpublished doctoral dissertation, University of Oxford, Oxford, UK.

Itti, L., Koch, C., & Braun, J. (2000). Revisiting spatial vision: Toward a unifying model. *Journal of the Optical Society of America A, Optics, Image Science, and Vision, 17,* 1899–1917. [PubMed]

Kaernbach, C. (1991). Simple adaptive testing with the weighted up–down method. *Perception & Psychophysics, 49,* 227–229. [PubMed]

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

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

Kuss, M., Jäkel, F., & Wichmann, F. A. (2005). Bayesian inference for psychometric functions. *Journal of Vision, 5*(5), 478–492, http://journalofvision.org/5/5/8/, doi:10.1167/5.5.8. [PubMed] [Article]

Lages, M., & Treisman, M. (1998). Spatial frequency discrimination: Visual long-term memory or criterion setting? *Vision Research, 38,* 557–572. [PubMed]

Laming, D., & Marsh, D. (1988). Some performance tests of QUEST on measurements of vibrotactile thresholds. *Perception & Psychophysics, 44,* 99–107. [PubMed]

Lee, D. K., Itti, L., Koch, C., & Braun, J. (1999). Attention activates winner-take-all competition among visual filters. *Nature Neuroscience, 2,* 375–381. [PubMed] [Article]

Leek, M. R., Hanna, T. E., & Marshall, L. (1992). Estimation of psychometric functions from adaptive tracking procedures. *Perception & Psychophysics, 51,* 247–256. [PubMed]

Long, G. M., & Tuck, J. P. (1988). Reliabilities of alternate measures of contrast sensitivity functions. *American Journal of Optometry and Physiological Optics, 65,* 37–48. [PubMed]

Luce, R. D. (1963). Detection and recognition. In Luce, R. D., Bush, R. R., & Galanter, E., (Eds.), *Handbook of mathematical psychology* (pp. 103–189). New York: Wiley.

Luce, R. D. (1977). The choice axiom after twenty years. *Journal of Mathematical Psychology, 15,* 215–233.

Machens, C. K., Romo, R., & Brody, C. D. (2005). Flexible control of mutual inhibition: A neural model of two-interval discrimination. *Science, 307,* 1121–1124. [PubMed]

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

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

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

Prentice, R. L. (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,* 321–327.

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

Treisman, M., & Faulkner, A. (1985). On the choice between choice theory and signal-detection theory. *Quarterly Journal of Experimental Psychology Section A, Human Experimental Psychology, 37,* 387–405.

Treisman, M., & Williams, T. C. (1984). A theory of criterion setting with an application to sequential dependencies. *Psychological Review, 91,* 68–111.

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

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

Vaegan, & Halliday, B. L. (1982). A forced-choice test improves clinical contrast sensitivity testing. *British Journal of Ophthalmology, 66,* 477–491. [PubMed]

Watson, A. B., & Fitzhugh, A. (1990). The method of constant stimuli is inefficient. *Perception & Psychophysics, 47,* 87–91. [PubMed]

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

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

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

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

Williams, D. A. (1982). Extra-binomial variation in logistic linear models. *Applied Statistics, 31,* 144–148.

Woods, R. L., & Thomson, W. D. (1993). A comparison of psychometric methods for measuring the contrast sensitivity of experienced observers. *Clinical Vision Sciences, 8,* 401–415.