Bernstein Center for
Computational Neuroscience
Berlin

Cognitive Science for Machine Learning 2

Empirical Methods: Psychophysical Methods, Signal Detection Theory & Response Times

Felix A. Wichmann

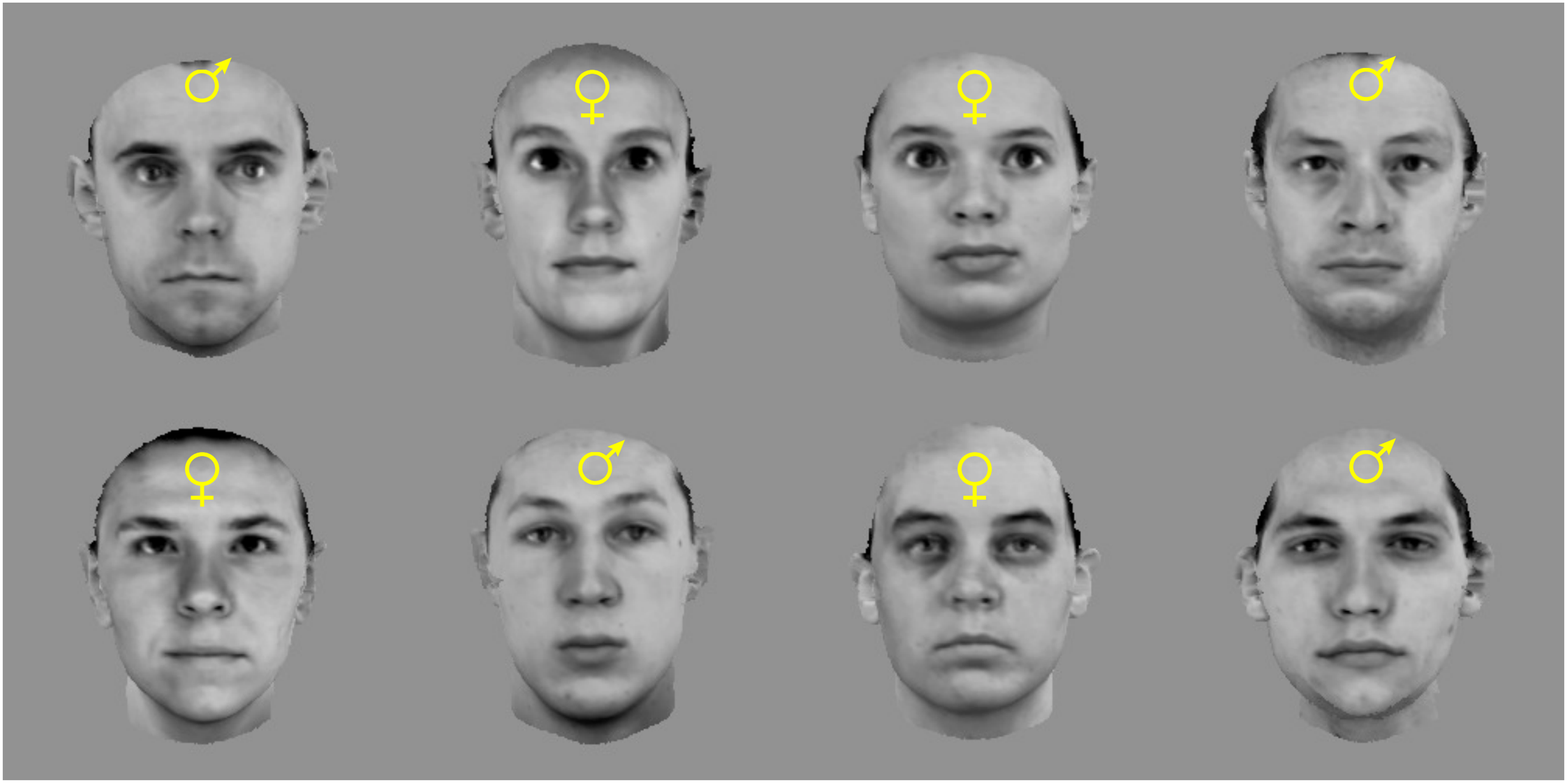
Modelling of Cognitive Processes Group
Bernstein Center for Computational Neuroscience
and
Technische Universität Berlin

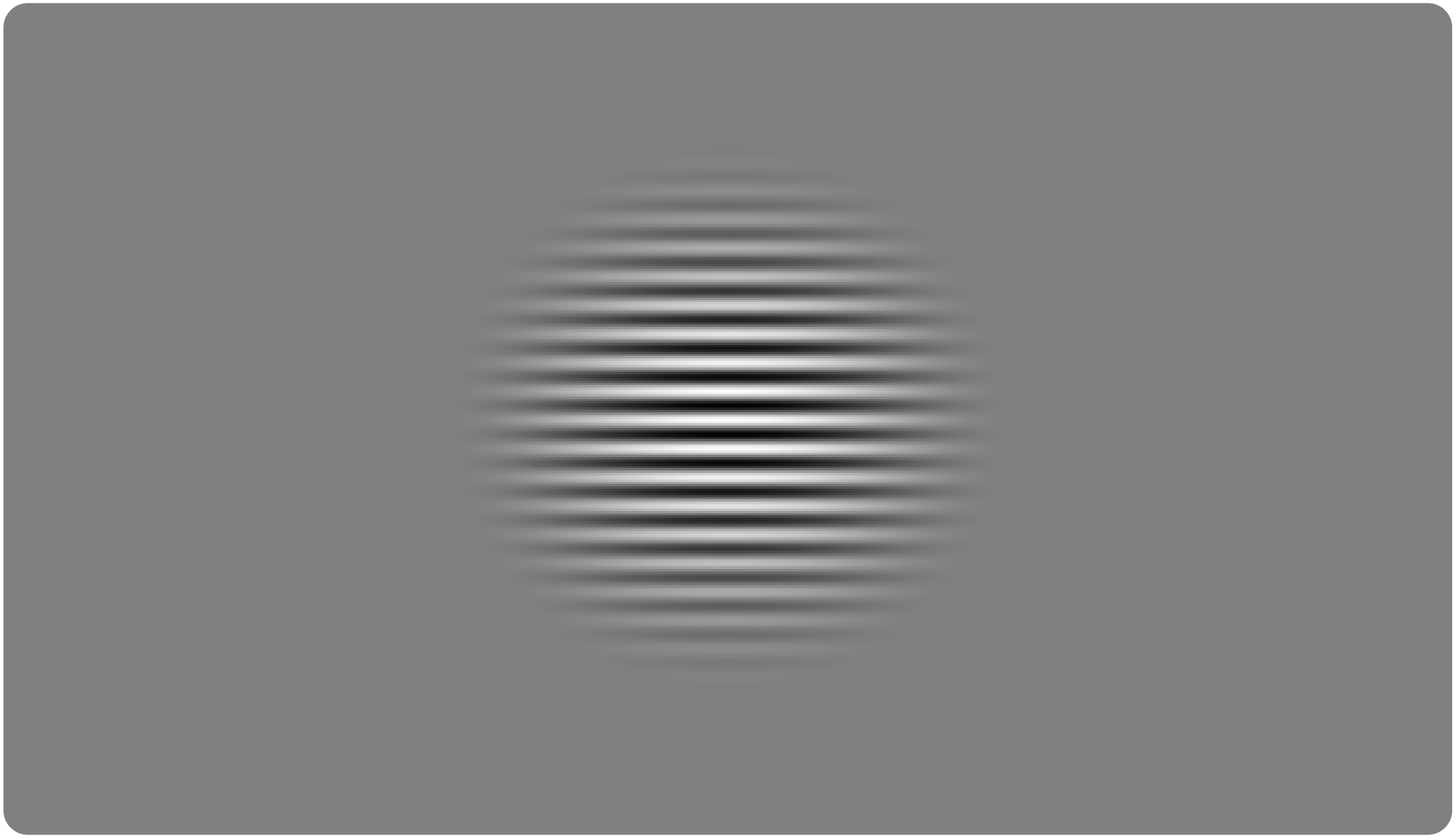
felix.wichmann@tu-berlin.de

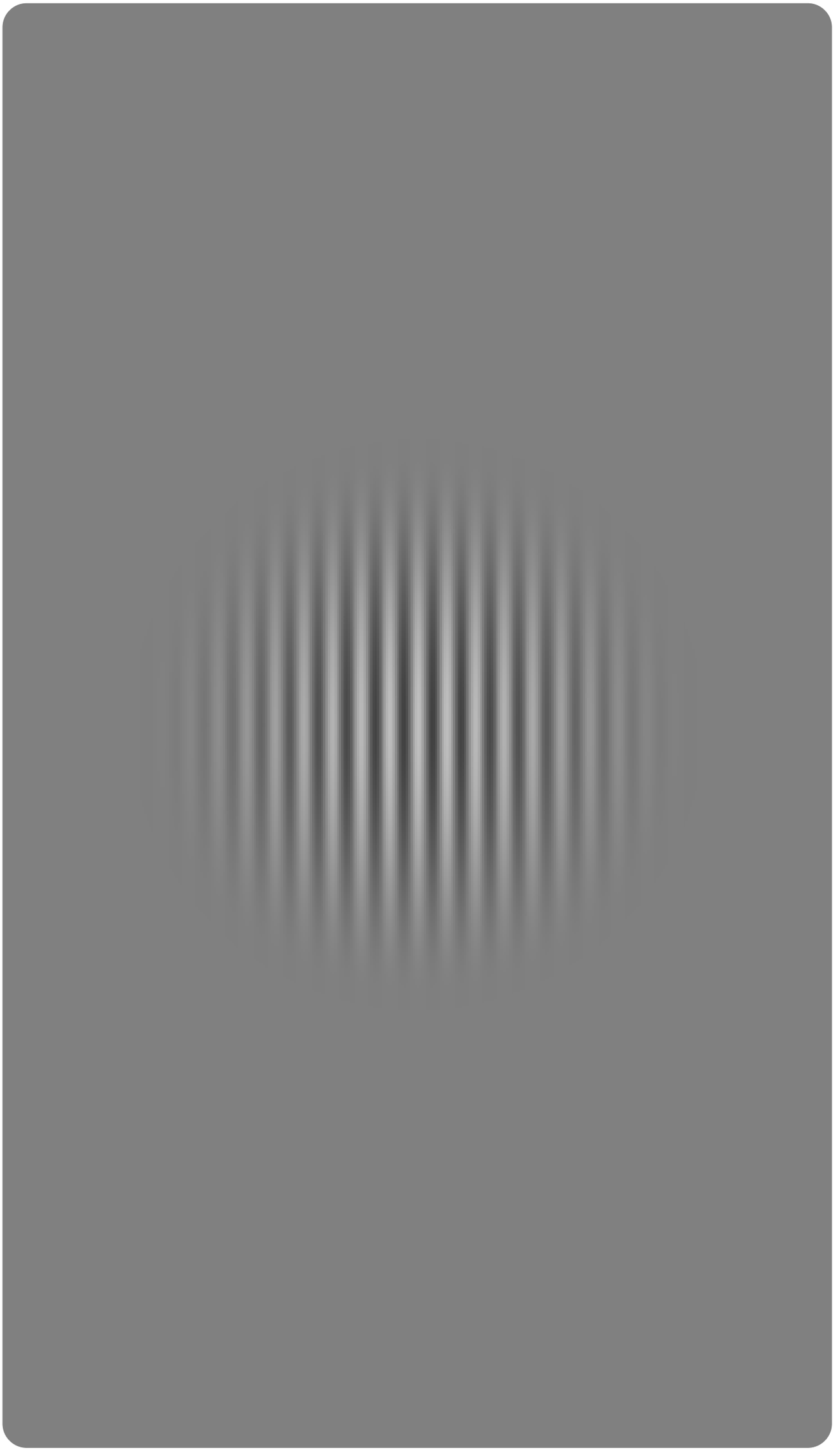
Bundesministerium
für Bildung
und Forschung

The Experimental Method. But psychology is passing into a less simple phase. Within a few years what one may call a microscopic psychology has arisen in Germany, carried on by experimental methods, asking of course every moment for introspective data, but eliminating their uncertainty by operating on a large scale and taking statistical means. This method taxes patience to the utmost, and could hardly have arisen in a country whose natives could be *bored*. Such Germans as Weber, Fechner, Vierordt, and Wundt obviously cannot ; and their success has brought into the field an array of younger experimental psychologists, bent on studying the *elements* of the mental life, dissecting them out from the gross results in which they are embedded, and as far as possible reducing them to quantitative scales. The simple and open method of attack having done what it can, the method of patience, starving out, and harassing to death is tried ; the Mind must submit to a regular *siege*, in which minute advantages gained night and day by the forces that hem her in must sum themselves up at last into her overthrow. There is little of the grand style about these new prism, pendulum, and chronograph-philosophers. They mean business, not chivalry. What generous divination, and that superiority in virtue which was thought by Cicero to give a man the best insight into nature, have failed to do, their spying and scraping, their deadly tenacity and almost diabolic cunning, will doubtless some day bring about.

William James (1890), *The Principles of Psychology,* ch. 7.







Psychophysical Methods

Traditional Psychophysical Methods

Method of Limits:
Experimenter shows a descending (ascending) sequence of stimuli (e.g. tone with less and less (more and more) energy, until the subject cannot experience it anymore (starts to experience it).

Method of Adjustment:
Similar to method of limits only the subject herself controls the intensity until she cannot experience the stimulus anymore.

Sometimes a difference is made between analogue and discrete presentations—do not take this too seriously unless the step size in the discrete presentation mode is too coarse.

Modern Psychophysical Methods

Method of Constant Stimulus:

Stimuli are presented at a small number of K (4 to 10) fixed intensity levels only. Typically $n_i = 5$, 10 or up to 50 repetitions at exactly the same signal intensity ("blocked constant stimulus") before a different intensity is selected. Alternatively, stimuli all all K intensity levels are intermixed.

Adaptive Procedures:

Latest development; here an algorithm selects the next presentation level based on the intensity of the stimulus and the response history of the subject. At least several dozen variants both non-parametric ("up-down methods") as well as parametric (typically Bayesian) ones.

PERGAMON

Vision Research 38 (1998) 1861–1881

Vision Research

Forced-choice staircases with fixed step sizes: asymptotic and small-sample properties

Miguel A. García-Pérez *

Departamento de Metodología, Facultad de Psicología, Universidad Complutense, Campus de Somosaguas, 28223 Madrid, Spain

Received 14 June 1996; received in revised form 11 February 1997; accepted 16 September 1997

Abstract

Visual detection and discrimination thresholds are often measured using adaptive staircases, and most studies use transformed (or weighted) up/down methods with fixed step sizes—in the spirit of Wetherill and Levitt (Br J Mathemat Statist Psychol 1965;18:1–10) or Kaernbach (Percept Psychophys 1991;49:227–229)—instead of changing step size at each trial in accordance with best-placement rules—in the spirit of Watson and Pelli (Percept Psychophys 1983;47:87–91). It is generally assumed that a fixed-step-size (FSS) staircase converges on the stimulus level at which a correct response occurs with the probabilities derived by Wetherill and Levitt or Kaernbach, but this has never been proved rigorously. This work used simulation techniques to determine the asymptotic and small-sample convergence of FSS staircases as a function of such parameters as the up/down rule, the size of the steps up or down, the starting stimulus level, or the spread of the psychometric function. The results showed that the asymptotic convergence of FSS staircases depends much more on the sizes of the steps than it does on the up/down rule. Yet, if the size Δ^+ of a step up differs from the size Δ^- of a step down in a way that the ratio Δ^-/Δ^+ is constant at a specific value that changes with up/down rule, then convergence percent-correct is unaffected by the absolute sizes of the steps. For use with the popular one-, two-, three- and four-down/one-up rules, these ratios must respectively be set at 0.2845, 0.5488, 0.7393 and 0.8415, rendering staircases that converge on the 77.85%-, 80.35%-, 83.15%- and 85.84%-correct points. Wetherill and Levitt's transformed up/down rules—which require $\Delta^-/\Delta^+ = 1$—and the general version of Kaernbach's weighted up/down rule—which allows any Δ^-/Δ^+ ratio—fail to reach their presumed targets. The small-sample study showed that, even with the optimal settings, short FSS staircases (up to 20 reversals in length) are subject to some bias, and their precision is less than reasonable, but their characteristics improve when the size Δ^+ of a step up is larger than half the spread of the psychometric function. Practical recommendations are given for the design of efficient and trustworthy FSS staircases.

and furthermore ...

Statistically, adaptive procedures are more efficient than the method of constant stimulus (i.e. fewer trials for same confidence interval around threshold)—however, they are more prone to be influenced by lapses, serial dependencies etc. all known to afflict living subjects ... .

SDT—certainly in single-interval designs one has to place the criterion optimally for best performance. If signal strength varies from trial-to-trial as in adaptive procedures this may be difficult. (Even worse if several staircases are interleaved!)

Standard Three-Step Routine in Modelling Data

1. Parameter estimation.

2. Goodness-of-fit assessment.

3. Assessing the variability of fitted parameters.

Definitions and Notation

K denotes the number of blocks of trials,
N the total number of trials, thus $N = \sum_{i=1}^{K} n_i$.
$\mathbf{x}$ is the vector $(x_1, \ldots, x_K)$ containing the signal levels (independent variable).
$\mathbf{y}$ is the vector $(y_1, \ldots, y_K)$ containing the corresponding *proportion* of correct responses.
$y_i n_i$ thus represents the *number* of correct responses at signal level x_i.

The psychometric function $\Psi(x)$ relates an observer's performance to an independent variable, usually some physical quantity of a stimulus in a psychophysical task. As a general form we define

$$\Psi(x|\alpha, \beta, \gamma, \lambda) = \gamma + (1 - \gamma - \lambda)F(x; \alpha, \beta).$$

The shape is determined by the parameter vector $\theta = (\alpha, \beta, \gamma, \lambda)$ and the choice of F, typically a two-parameter sigmoidal function such as the Weibull, logistic or cumulative Gaussian distribution.

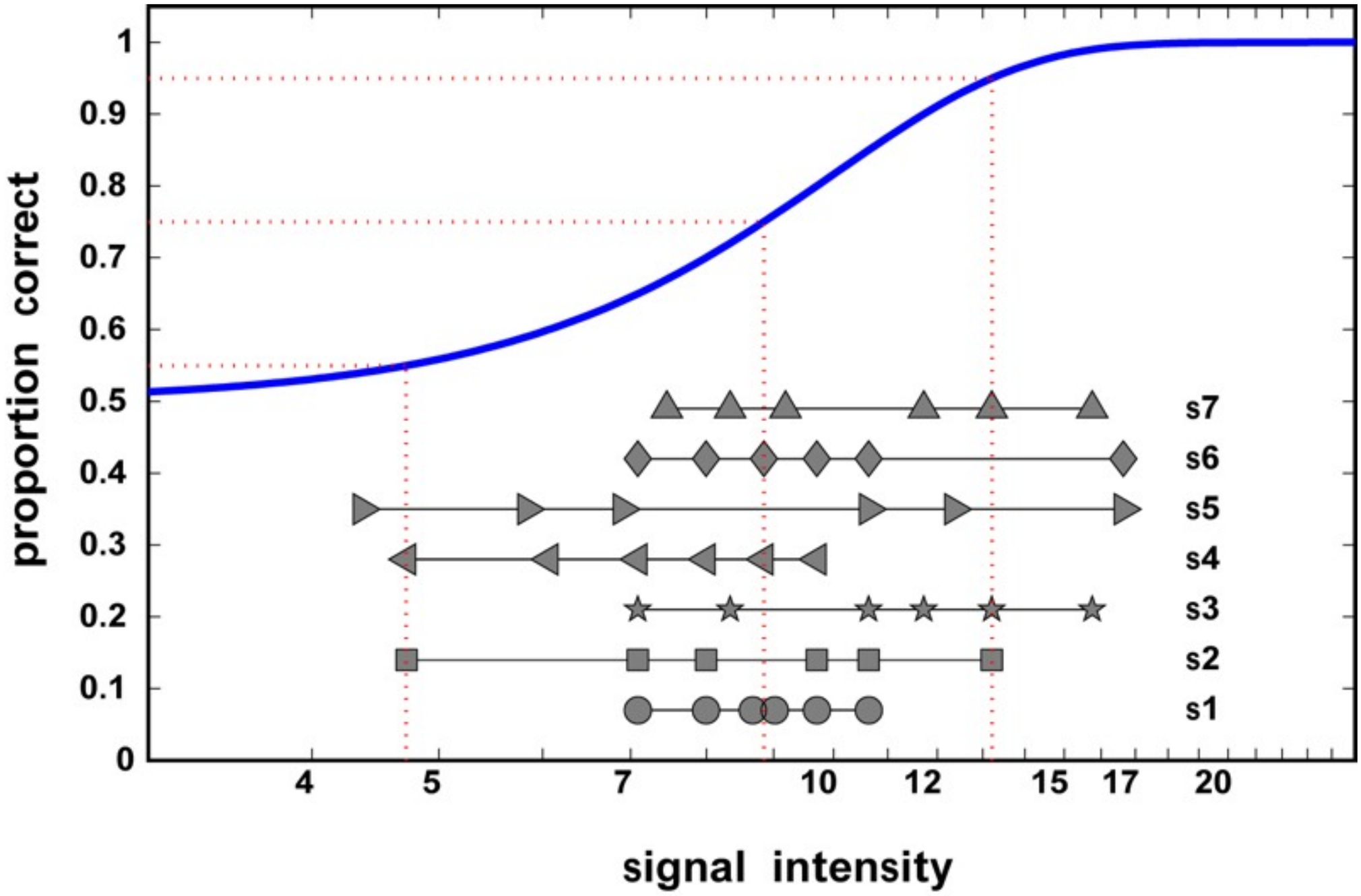
Definitions and Notation

$$\text{Weibull: } f(x) = 1 - \mathrm{e}^{-(\frac{x}{\alpha})^{\beta}}$$

$$\text{Logistic: } f(x) = \frac{1}{1 + \mathrm{e}^{\frac{\alpha - x}{\beta}}}$$

$$\text{Cumulative Gaussian: } f(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\beta^2}} \mathrm{e}^{-\frac{1}{2}(\frac{y-\alpha}{\beta})^2} dy$$

Definitions and Notation



Parameter Estimation

We assume that psychophysical trials are Bernoulli processes—the outcome (detection, no detection, or whatever it may be) is either 0 or 1—correct or incorrect—and there is an underlying probability p to get the corrrect answer ($0 \leq p \leq 1$).
For n trials at the same stimulus intensity x the number of correct answers, z, then is binomially distributed:

$$f(z) = \frac{n!}{z!(n-z)!} p^z (1-p)^{n-z} = \binom{n}{z} p^z (1-p)^{n-z}$$

If we substitute the proportion correct, $y = \frac{z}{n}$, we obtain:

$$f(y|p,n) = \binom{n}{ny} p^{ny} (1-p)^{n(1-y)}$$

Repeated in words: f(y) provides the probability of obtaining the proportion y correct responses *given* n trials and a success (detection, discrimination ...) probability p.

Parameter Estimation

In a psychophysical setting we typically have a vector $\mathbf{y} = (y_1, \ldots, y_K)$ of proportions of correct responses resulting from K blocks of trials with $N = \sum_{i=1}^{K} n_i$ trials—note that the n_i need not be the same for all i.
As we assume not only all trials to be independent sets of Bernoulli trials but the blocks of trials to be independent, too, the joint likelihood is the product of the individual likelihoods:

$$L(\mathbf{p}|\mathbf{y}, \mathbf{n}) = \prod_{i=1}^{K} \binom{n_i}{n_i y_i} p_i^{n_i y_i} (1 - p_i)^{n_i (1 - y_i)}$$

The maximum-likelihood estimate $\hat{\mathbf{p}}$ is again simply the vector $\mathbf{y} = (y_1, \ldots, y_K)$ of proportions of correct responses.

Parameter Estimation

Frequently, $\mathbf{y} = (y_1, \ldots, y_K)$ will not be monotonically increasing with increasing x. Thus the "raw" maximum-likelihood estimate of psychophysically observed data $\mathbf{y}$, $\hat{\mathbf{p}} = \mathbf{y}$ is not satisfactory. We want a psychometric function to be a *monotonically* increasing (decreasing) function of signal intensity.

So, effectively, we *regularise* our maximum-likelihood fit by forcing it to be monotonic. One way of doing this is to link the individual p_i's using a monotonic function:

$$p_i = \Psi(x_i|\alpha, \beta, \gamma, \lambda) = \gamma + (1 - \gamma - \lambda)F(x_i|\alpha, \beta).$$

with the parameter vector $\theta = (\alpha, \beta, \gamma, \lambda)$.
Likelihood, finally, becomes

$$L(\theta|\mathbf{x}, \mathbf{y}, \mathbf{n}, F) = \prod_{i=1}^{K} \binom{n_i}{n_i y_i} (\Psi(x_i|\theta))^{n_i y_i} (1 - \Psi(x_i|\theta))^{n_i(1-y_i)}$$

and only depends on θ (and the monotonic function F we choose).

Parameter Estimation

For reasons of convenience—small numbers, cumbersome derivatives—typically the log-likelihood is maximised. (Any monotone transform of likelihood returns the same $\hat{\theta}$ but the log has a number of nice properties, see below.)

Log-Likelihood, $l(\theta)$, finally, is

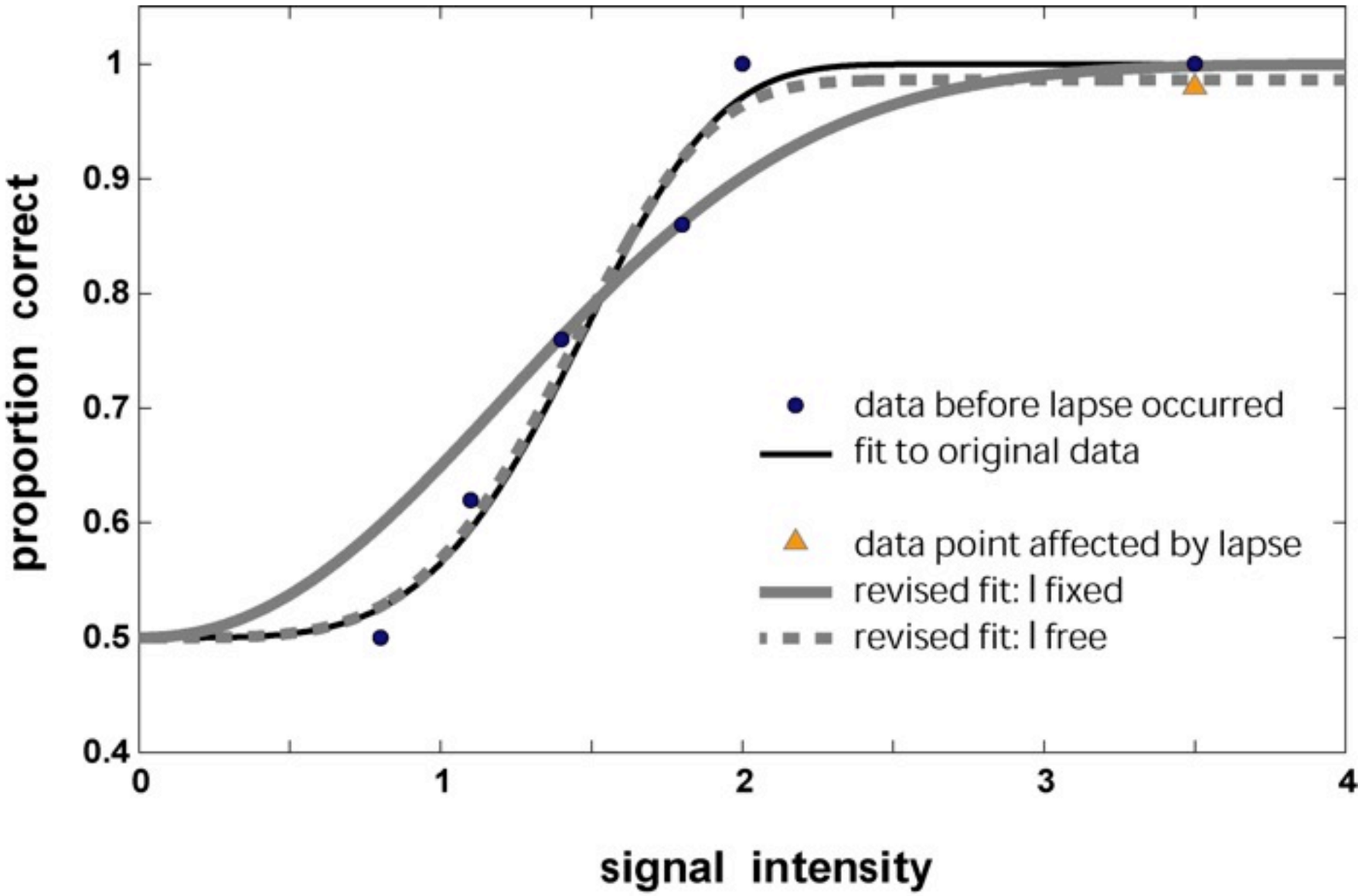
$$l(\theta|\mathbf{x},\mathbf{y},\mathbf{n}) = \sum_{i=1}^{K} \log \binom{n_i}{y_i n_i} + y_i n_i \log \Psi(x_i;\theta) + n_i(1-y_i)\log(1-\Psi(x_i;\theta)).$$

Parameter Estimation—Third Source of Error: Lapses

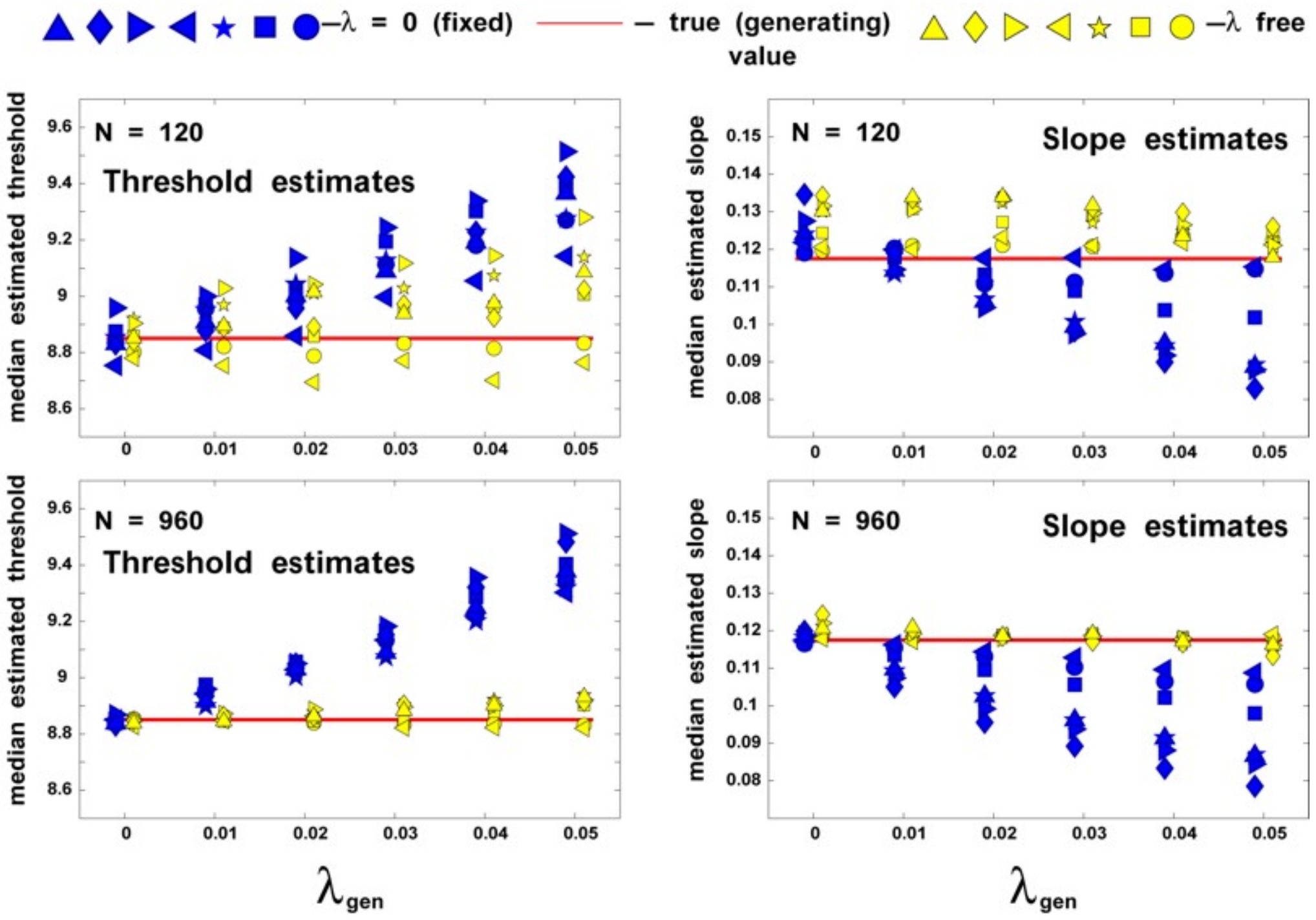
The upper bound of $\Psi(x;\theta)$, that is the performance for arbitrarily large stimulus signals, is given by $1-\lambda$. Typically the parameter λ is set to zero, and thus the range of the psychometric function is $[\gamma, 1]$.

However, human observers are seldom perfect—neglecting λ can lead to *major* mis-estimation of α and β, as shown in the following.

Parameter Estimation



Parameter Estimation



Statistical Considerations

An approximate solution to the exact problem is usually better than the exact solution to an approximate problem.

Better to have an approximate answer to the right question than a precise answer to the wrong question.

An appropriate answer to the right problem is worth a good deal more than an exact answer to an approximate problem.

Variations on the same theme all attributed to John W. Tukey (1915-2000).

Goodness-of-fit

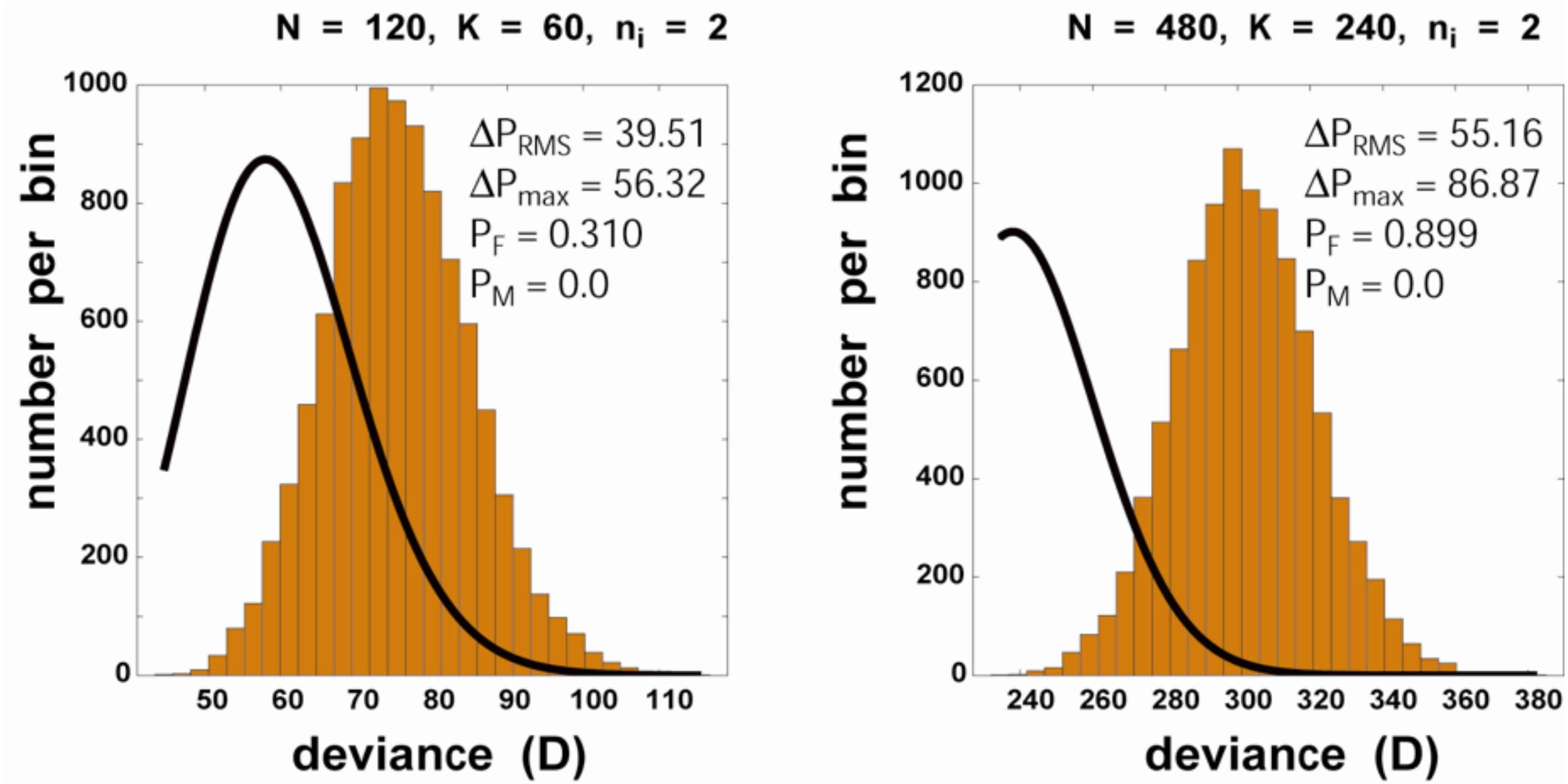
In maximum-likelihood parameter estimation the parameter vector $\hat{\theta}$ returned by the estimation routine is such that $L(\hat{\theta};\mathbf{y}) \geq L(\theta;\mathbf{y})$ for all θ. Thus, whatever error metric Z is used to assess goodness-of-fit, $Z(\hat{\theta};\mathbf{y}) \geq Z(\theta;\mathbf{y})$ should hold for all θ.

The log-likelihood ratio *deviance* is a monotonic transformation of likelihood and therefore fulfills this criterion. Deviance is defined as

$$D = 2\log\left(\frac{L(\theta_{max};\mathbf{y})}{L(\hat{\theta};\mathbf{y})}\right) = 2\left(l(\theta_{max};\mathbf{y}) - l(\hat{\theta};\mathbf{y})\right),$$

where θ_{max} denotes the parameter vector of the saturated model without residual error between empirical data and model prediction.
The reason why deviance is preferred in goodness-of-fit assessment over likelihood or log-likelihood directly is that, asymptotically, it is distributed as χ^2_K where K denotes the number of data-points (blocks of trials).

Goodness-of-fit



p_{gen} uniformly distributed on $[.52, \ldots, .85]$

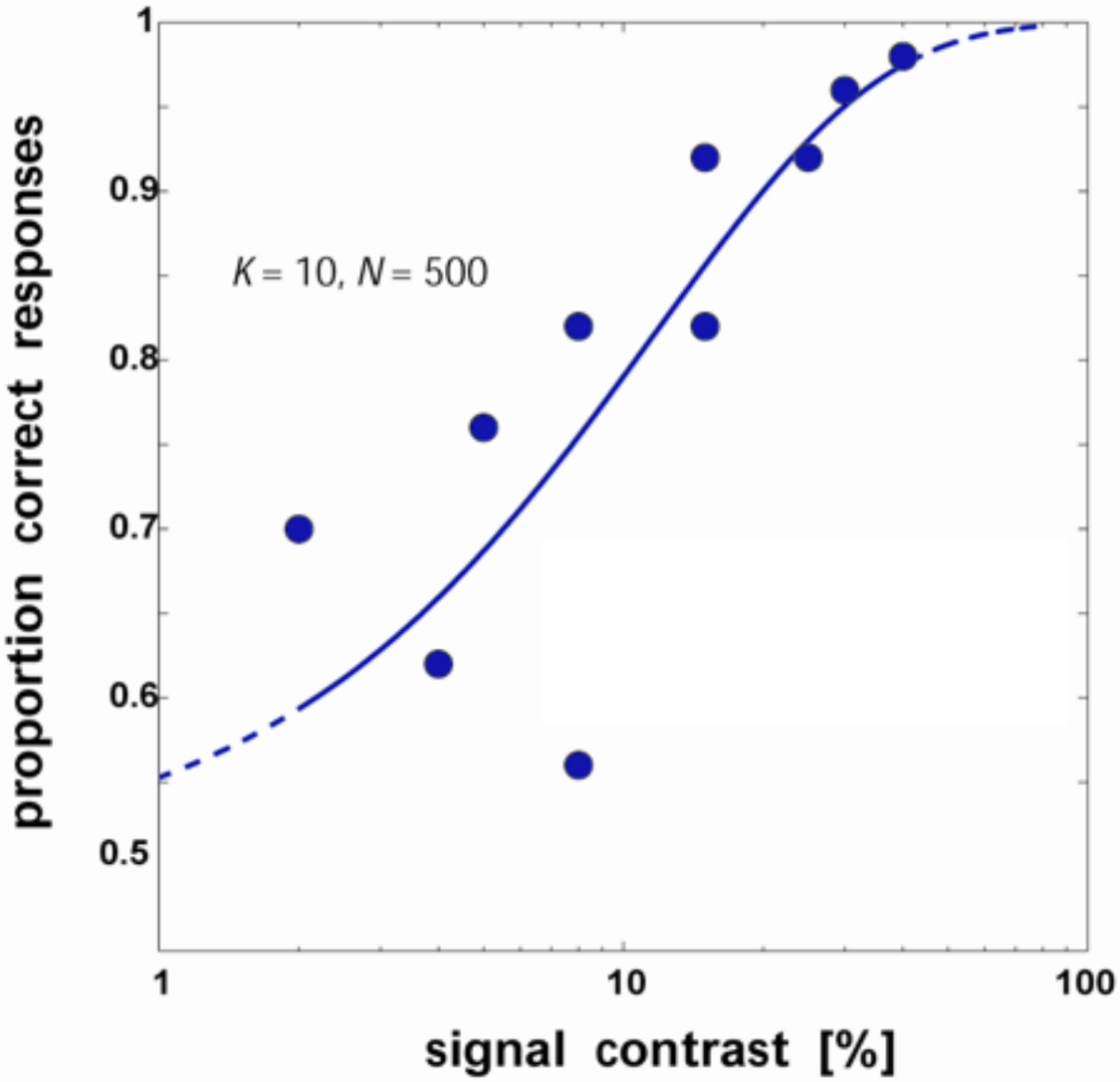
Goodness-of-fit

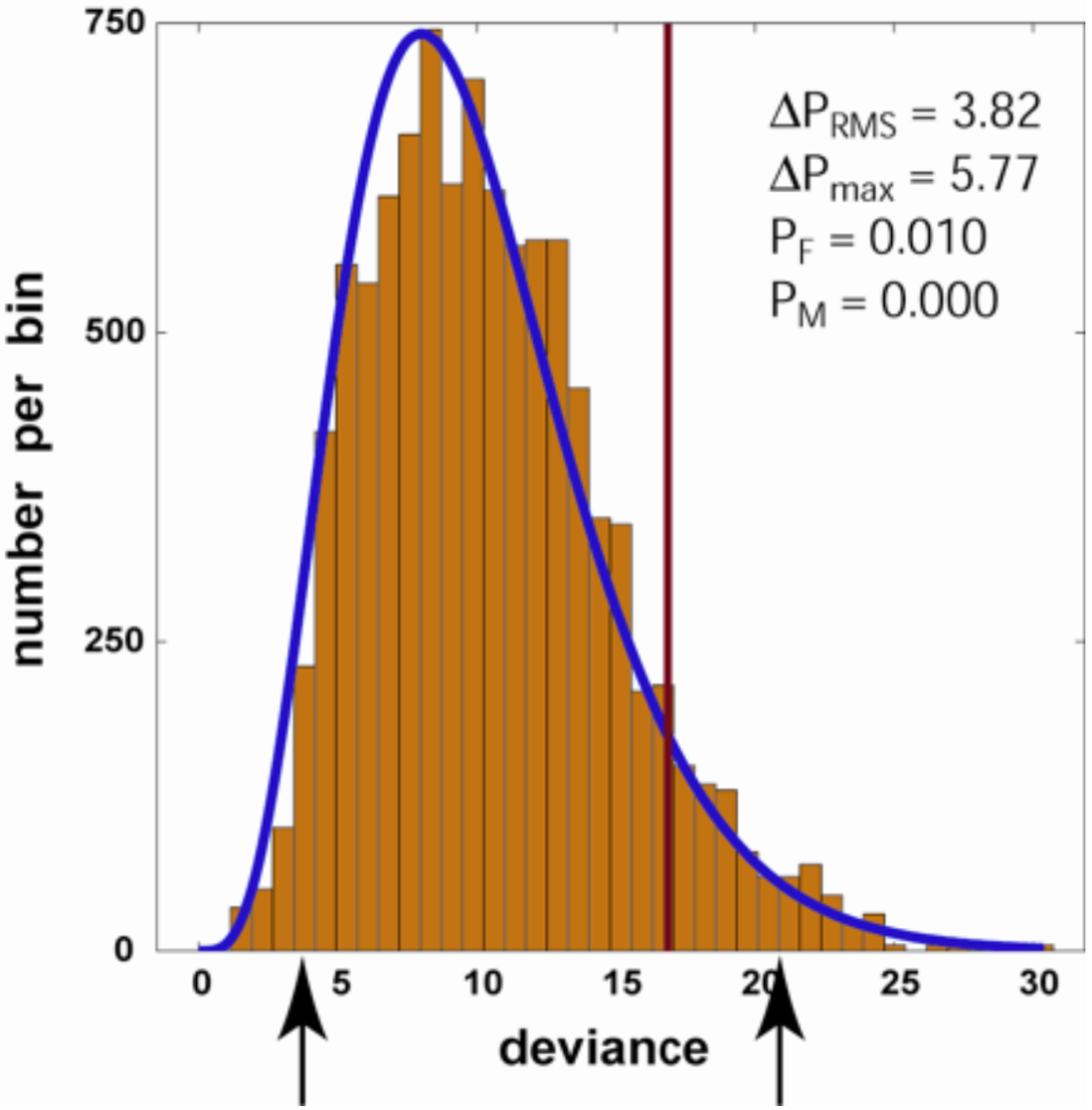
Frequently it is helpful to examine the residuals directly, rather than just a summary statistic such as deviance. Each deviance residual d_i is defined as the square root of the deviance value for datapoint i in isolation, signed according to the direction of the arithmetic residual $y_i - p_i$. For binomial data in the case of psychometric function fitting this results in

$$d_i = \text{sgn}(y_i - p_i)\sqrt{2\left(y_i n_i \log\left(\frac{y_i}{p_i}\right) + n_i(1-y_i)\log\left(\frac{1-y_i}{1-p_i}\right)\right)}$$

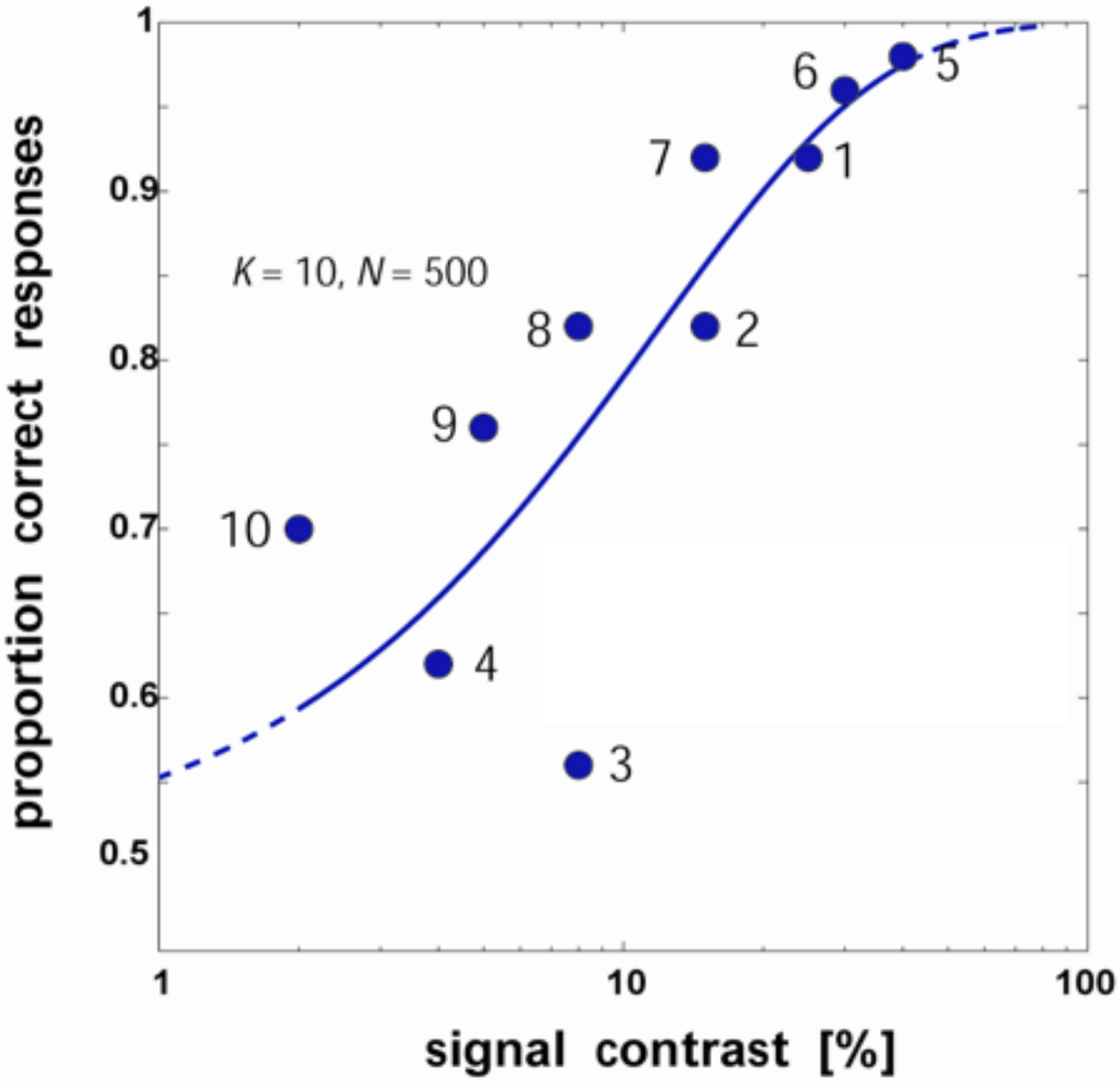
with $p_i = \Psi(x_i; \theta)$.

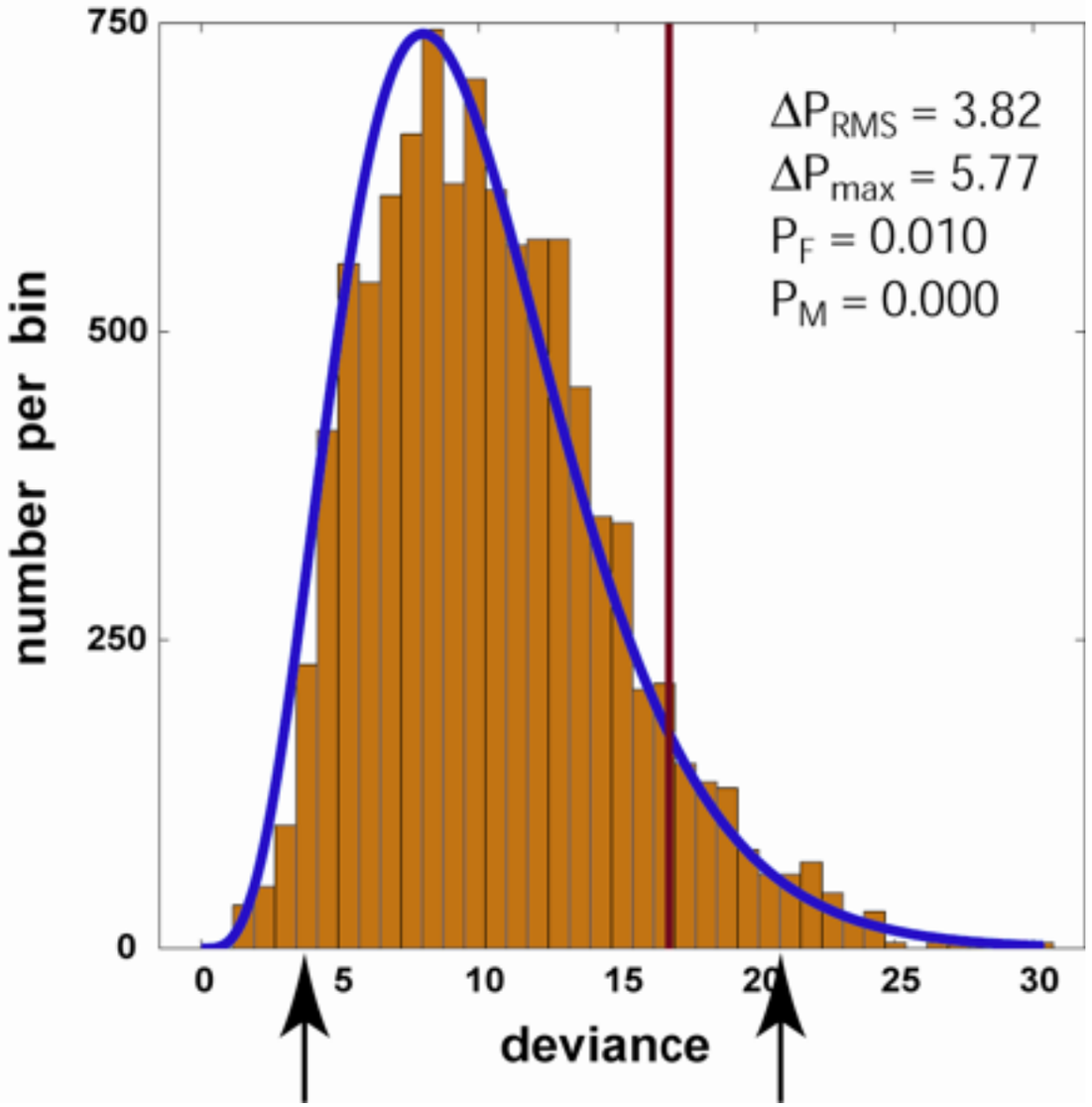
Goodness-of-fit



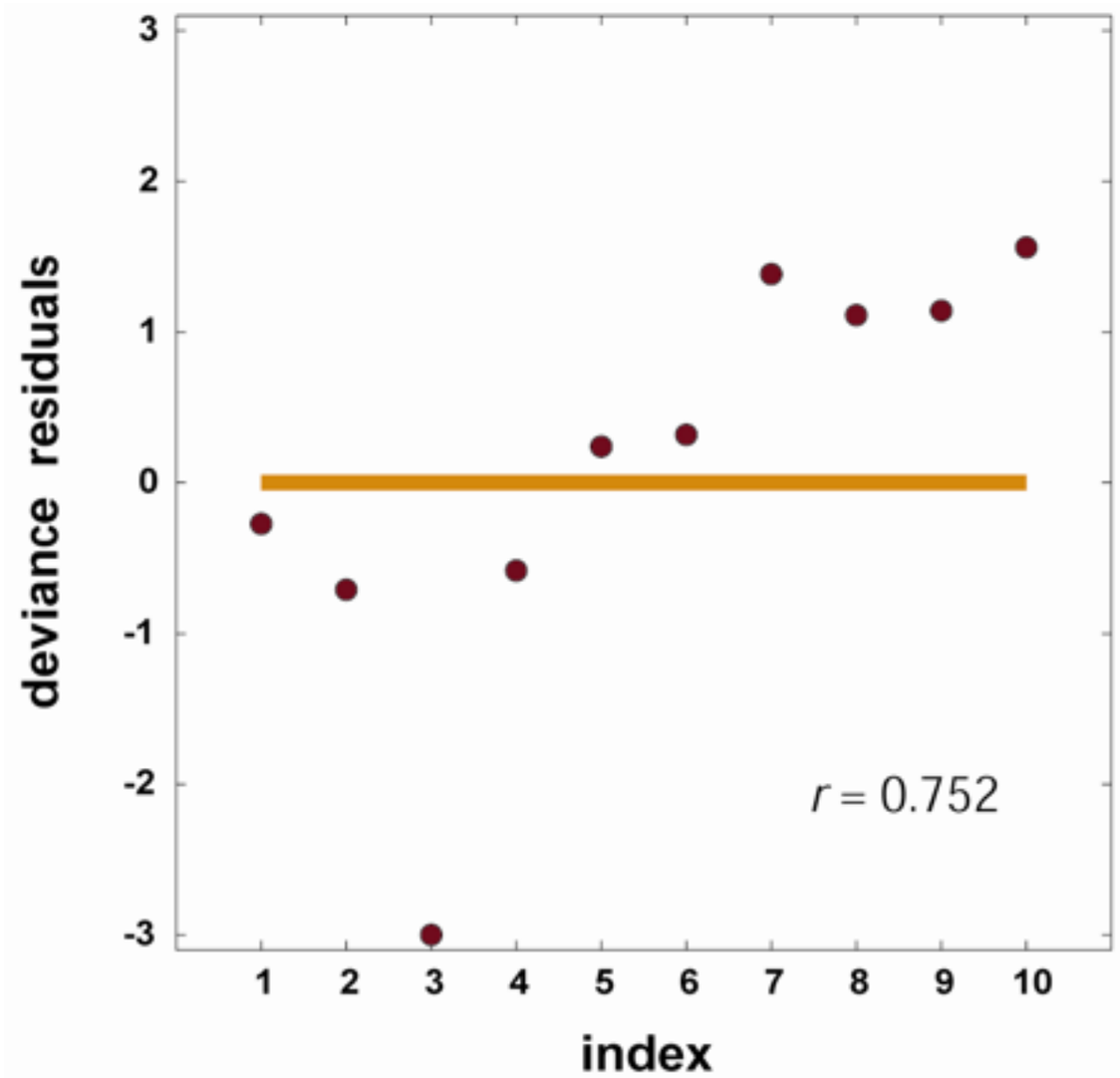


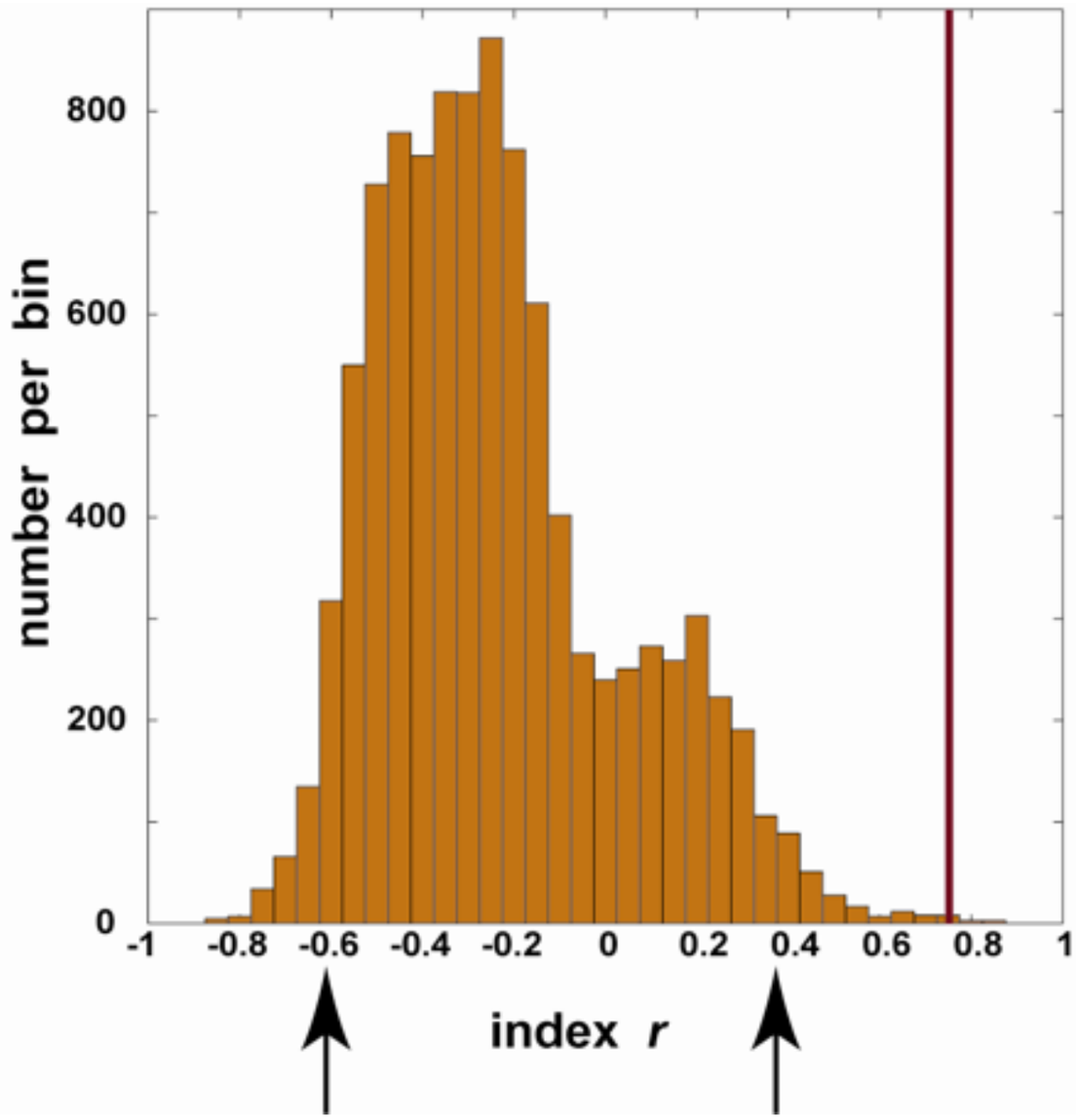
Goodness-of-fit





Goodness-of-fit





Estimates of Variability—The Way of Bayes ...

Journal of Vision (2005) 5, 478-492 http://journalofvision.org/5/5/8/ 478

Bayesian inference for psychometric functions

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

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

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

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

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

... or the Traditional Way

Perception & Psychophysics
2001, 63 (8), 1314-1329

The psychometric function: II. Bootstrap-based confidence intervals and sampling

Perception & Psychophysics
2001, 63 (8), 1293-1313

The psychometric function: I. Fitting, sampling, and goodness of fit

FELIX A. WICHMANN and N. JEREMY HILL
University of Oxford, Oxford, England

The psychometric function relates an observer's performance to an independent variable, usually some physical quantity of a stimulus in a psychophysical task. This paper, together with its companion paper (Wichmann & Hill, 2001), describes an integrated approach to (1) fitting psychometric functions, (2) assessing the goodness of fit, and (3) providing confidence intervals for the function's parameters and other estimates derived from them, for the purposes of hypothesis testing. The present paper deals with the first two topics, describing a constrained maximum-likelihood method of parameter estimation and developing several goodness-of-fit tests. Using Monte Carlo simulations, we deal with two specific difficulties that arise when fitting functions to psychophysical data. First, we note that human observers are prone to stimulus-independent errors (or *lapses*). We show that failure to account for this can lead to serious biases in estimates of the psychometric function's parameters and illustrate how the problem may be overcome. Second, we note that psychophysical data sets are usually rather small by the standards required by most of the commonly applied statistical tests. We demonstrate the potential errors of applying traditional χ^2 methods to psychophysical data and advocate use of Monte Carlo resampling techniques that

phys-
dness
hether
, how-
e only
in our
Monte
fitted

Literature on Psychometric Methods

Blackwell, H. R. (1946). Contrast threshold of the human eye. *Journal of the Optical Society of America*, *36*(11), 624-643.

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

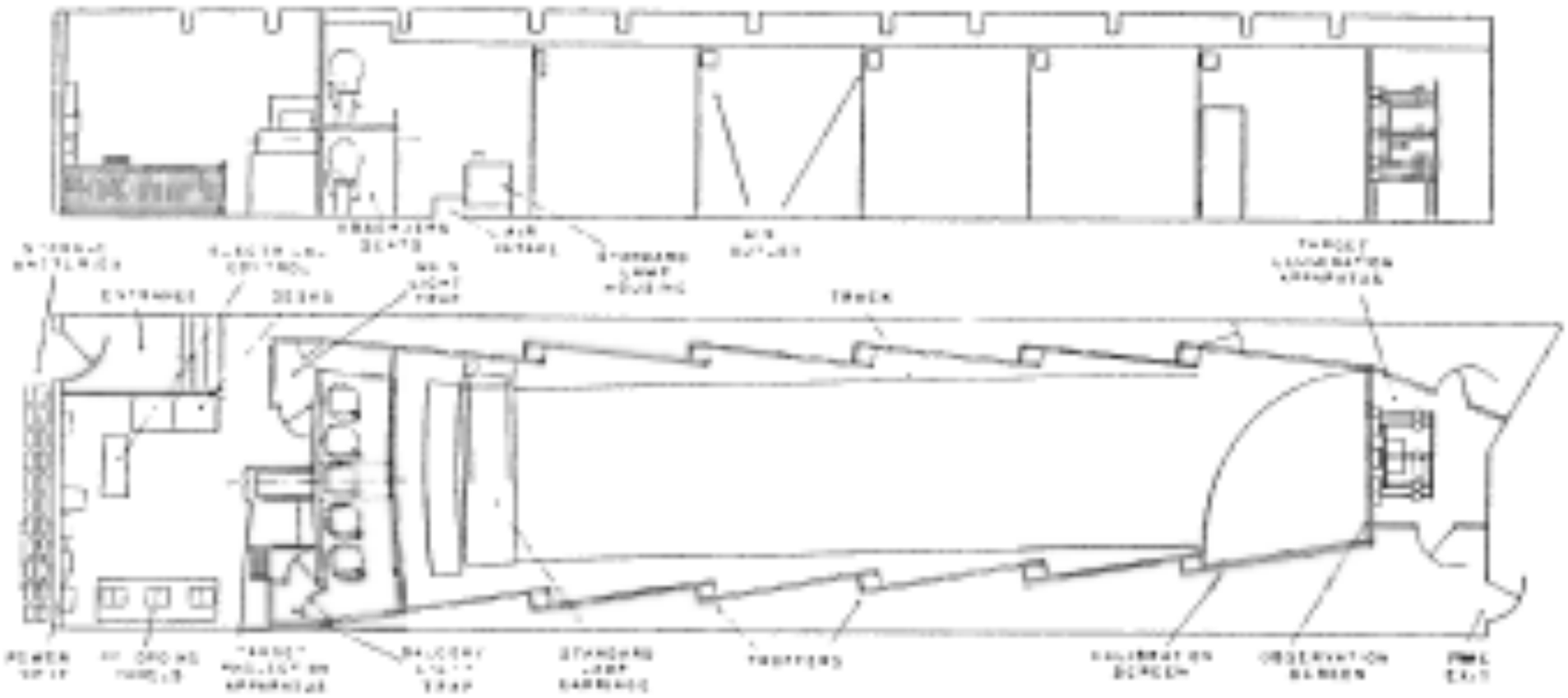


FIG. 2. Floor plan of laboratory. Dimensions of the plywood room (observation room) were: length, 64 feet; height, 10 feet, and width at the narrowest points, 10 feet.

FIG. 4. View of observers' stations.

FIG. 8. Permanent recording of experimental data

Literature on Psychometric Methods

Blackwell, H. R. (1946). Contrast threshold of the human eye. *Journal of the Optical Society of America*, *36*(11), 624-643.

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

Garcia-Perez, M. A. (1998). Forced-choice Staircases with Fixed Step Sizes: Asymptotic and Small-sample Properties. *Vision Research*, *38*(12), 1861-1881.

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

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

Kuss, M., Jäkel, F., & Wichmann, F. A. (2005). Bayesian inference for psychometric functions. *Journal of Vision*, 5, 478-492.

Jäkel, F., & Wichmann, F. A. (2006). Spatial four-alternative forced-choice method is the preferred psychophysical method for naive observers. *Journal of Vision*, 6, 1307-1322

Essentials of Psychometric Methods

- JND-style measurements most precise, reliable and immune to (unwanted) extraneous influences (mood, time of day, etc.); MCS (blocked) preferable to adaptive procedures.

- Rating and scaling experiments much less reliable and immune; important to provide an anchor as well as the extremes of the scale.

- Human observers and animals are prone to lapses: always fit a mixture model and explicitly estimate the upper (lower) asymptote(s).

- Goodness-of-fit and estimates of variability (CI's) require Monte Carlo methods as datasets are typically too small for asymptotic theory to apply.

- CI's are likely too small—binomial variability is only the lower bound: data are overdispersed due to learning, fatigue, cue-shifting, serial dependencies ... .

- As thresholds and their CI's are used to test hypotheses:

Many more experimental differences in psychophysics are known to be true than are!

Signal Detection Theory

> Diagnostic problems abound for individuals, organizations, and society. The stakes are high, often life and death. Such problems are prominent in the fields of health care, public safety, business, environment, justice, education, manufacturing, information processing, the military, and government.

Particular diagnostic questions are raised repetitively, each time calling for a positive or negative decision about the pres- ence of a given condition or the occurrence (often in the future) of a given event. Consider the following illustrations: Is a cancer present? Will this individual commit violence? Are there explosives in this luggage? Is this aircraft fit to fly? Will the stock market advance today? Is this assembly-line item flawed? Will an impending storm strike? Is there oil in the ground here? Is there an unsafe radiation level in my house? Is this person lying? Is this person using drugs? Will this applicant succeed? Will this book have the information I need? Is that plane intending to attack this ship? Is this applicant legally disabled? Does this tax return justify an audit? Each time such a question is raised, the available evidence is assessed by a person or a device or a combination of the two, and a choice is then made between the two alternatives, yes or no. The evidence may be a x-ray, a score on a psychiatric test, a chemical analysis, and so on.

In considering just yes–no alternatives, such diagnoses do not exhaust the types of diagnostic questions that exist.

Swets, J.A., Dawes, R.M. & Monahan (2000), Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest,* 1(1), 1-26.

Binary World

	“World”	
	“present” (signal)	“absent” (noise)
“Decision”		
“present” (Yes)	**Hit**	**False Alarm**
“absent” (No)	**Miss**	**Correct Rejection**

Decision Theory Applied to Perception & Cognition

- Internal noise: Assumption that even in the absence of any external stimulus there is (variable) internal activity in the nervous system.

- Decision-axis: For binary problems there is always a one-dimensional sufficient statistic independent of the dimensionality of the observation space.

Because we have assumed that the decision rule must say either H_1 or H_0, we can view it as a rule for dividing the total observation space Z into two parts, Z_0 and Z_1, as shown in Fig. 2.4. Whenever an observation falls in Z_0 we say H_0, and whenever an observation falls in Z_1 we say H_1.

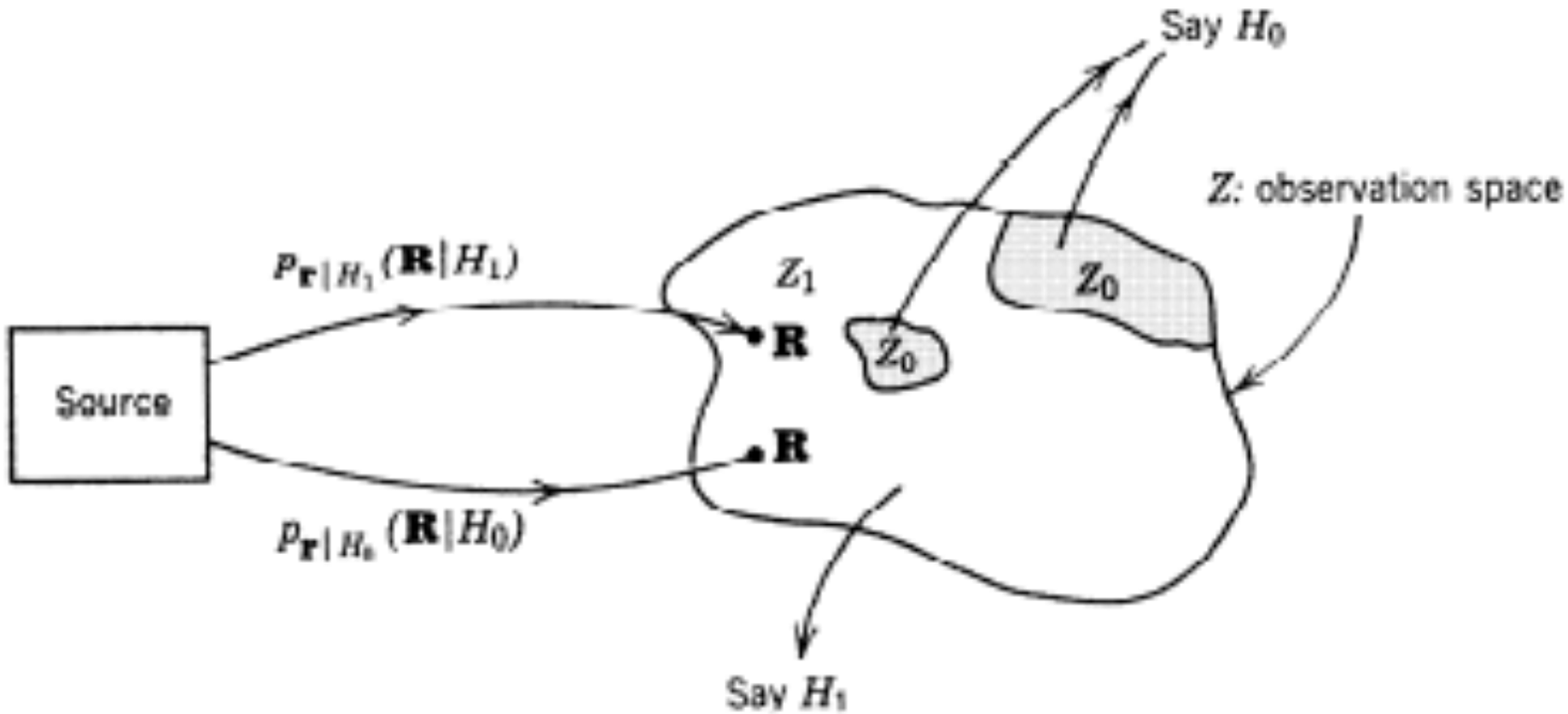


Fig. 2.4 Decision regions.

Van Trees, H. L. (1968). *Detection, Estimation, and Modulation Theory.* New York: John Wiley & Sons, p. 24.

We can now write the expression for the risk in terms of the transition probabilities and the decision regions:

$$\begin{aligned}\mathcal{R} = {} & C_{00}P_0 \int_{Z_0} p_{\mathbf{r}|H_0}(\mathbf{R}|H_0)\, d\mathbf{R} \\ & + C_{10}P_0 \int_{Z_1} p_{\mathbf{r}|H_0}(\mathbf{R}|H_0)\, d\mathbf{R} \\ & + C_{11}P_1 \int_{Z_1} p_{\mathbf{r}|H_1}(\mathbf{R}|H_1)\, d\mathbf{R} \\ & + C_{01}P_1 \int_{Z_0} p_{\mathbf{r}|H_1}(\mathbf{R}|H_1)\, d\mathbf{R}. \end{aligned} \tag{5}$$

For an N-dimensional observation space the integrals in (5) are N-fold integrals.

Van Trees, H. L. (1968). *Detection, Estimation, and Modulation Theory.* New York: John Wiley & Sons, p. 25.

Alternately, we may write

$$\frac{p_{\mathbf{r}|H_1}(\mathbf{R}|H_1)}{p_{\mathbf{r}|H_0}(\mathbf{R}|H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{P_0(C_{10} - C_{00})}{P_1(C_{01} - C_{11})}. \tag{12}$$

The quantity on the left is called the *likelihood ratio* and denoted by $\Lambda(\mathbf{R})$

$$\boxed{\Lambda(\mathbf{R}) \triangleq \frac{p_{\mathbf{r}|H_1}(\mathbf{R}|H_1)}{p_{\mathbf{r}|H_0}(\mathbf{R}|H_0)}.} \tag{13}$$

Because it is the ratio of two functions of a random variable, it is a random variable. We see that regardless of the dimensionality of $\mathbf{R}$, $\Lambda(\mathbf{R})$ is a one-dimensional variable.

The quantity on the right of (12) is the threshold of the test and is denoted by η:

$$\eta \triangleq \frac{P_0(C_{10} - C_{00})}{P_1(C_{01} - C_{11})}. \tag{14}$$

Thus Bayes criterion leads us to a *likelihood ratio test* (LRT)

$$\Lambda(\mathbf{R}) \underset{H_0}{\overset{H_1}{\gtrless}} \eta. \tag{15}$$

Van Trees, H. L. (1968). *Detection, Estimation, and Modulation Theory.* New York: John Wiley & Sons, p. 26.

Appeals of SDT

- Allows separation of sensory ("hardware") and motivational ("psychological") components of a perceptual decision.

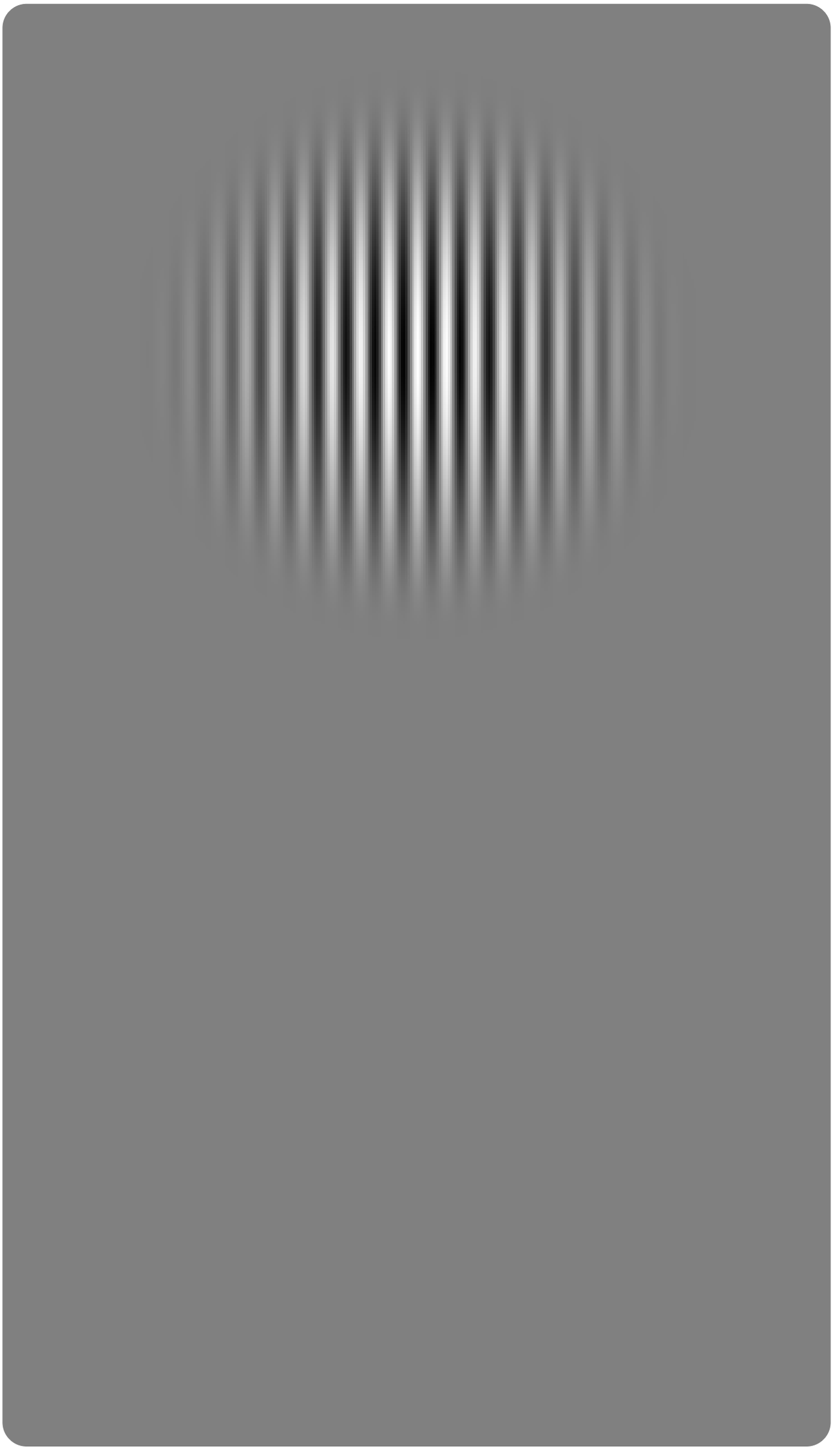
- Optimal placement of λ to maximize proportion correct can be easily calculated: If prior probabilities of signal and noise trials are equal and costs are assumed equal, too, then simply place λ where likelihood ratio is equal. (Independent of exact shape of the densities!)

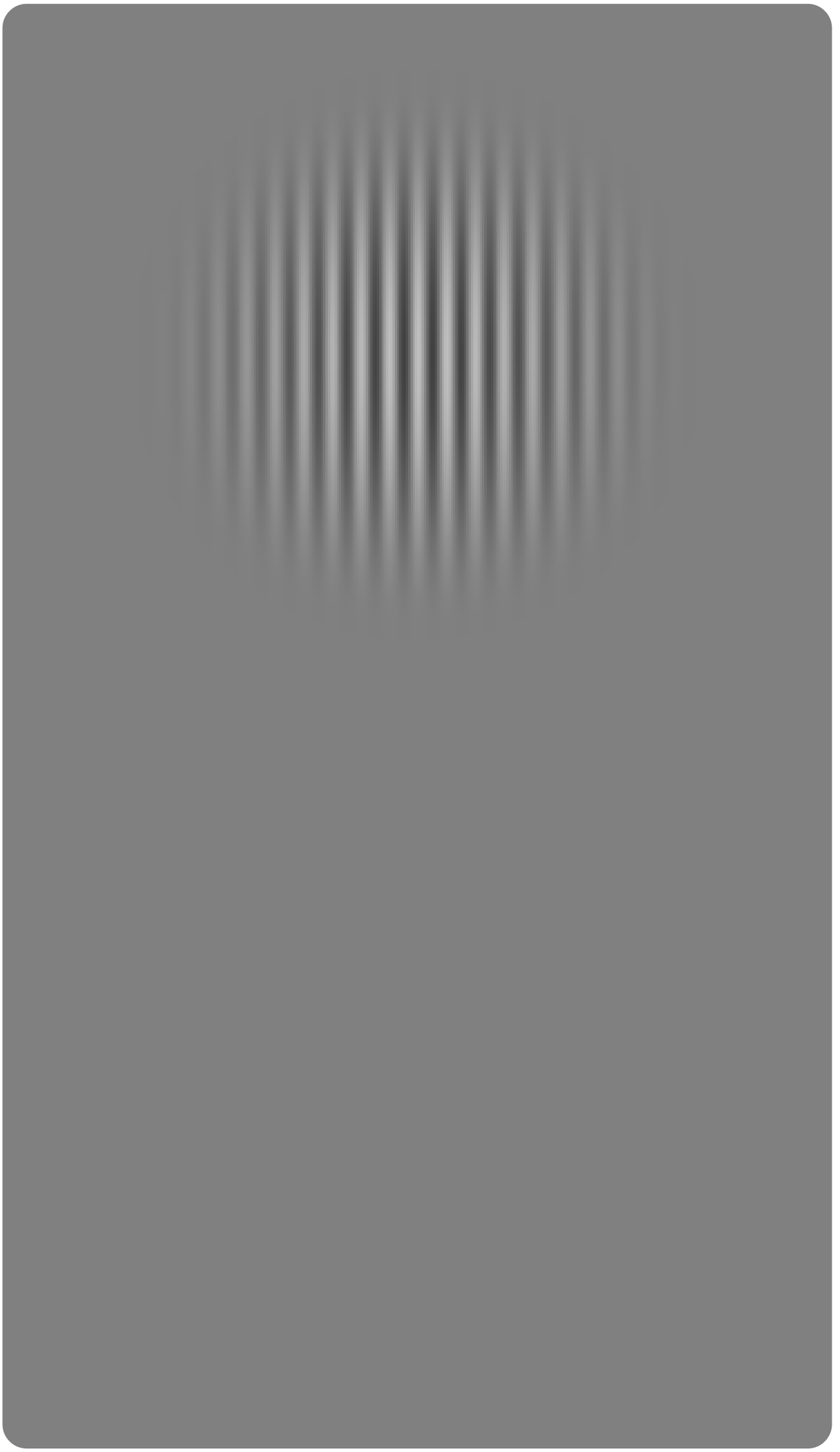
- Assuming signal and noise densities to be Gaussians with the same variance, the normalized distance between the means of the densities is called d' (d-prime). This is a criterion-free measure of task difficulty; λ informs about the motivational state. (Many nice experiments in the 50s and 60s changing costs of hits, false alarms etc.)

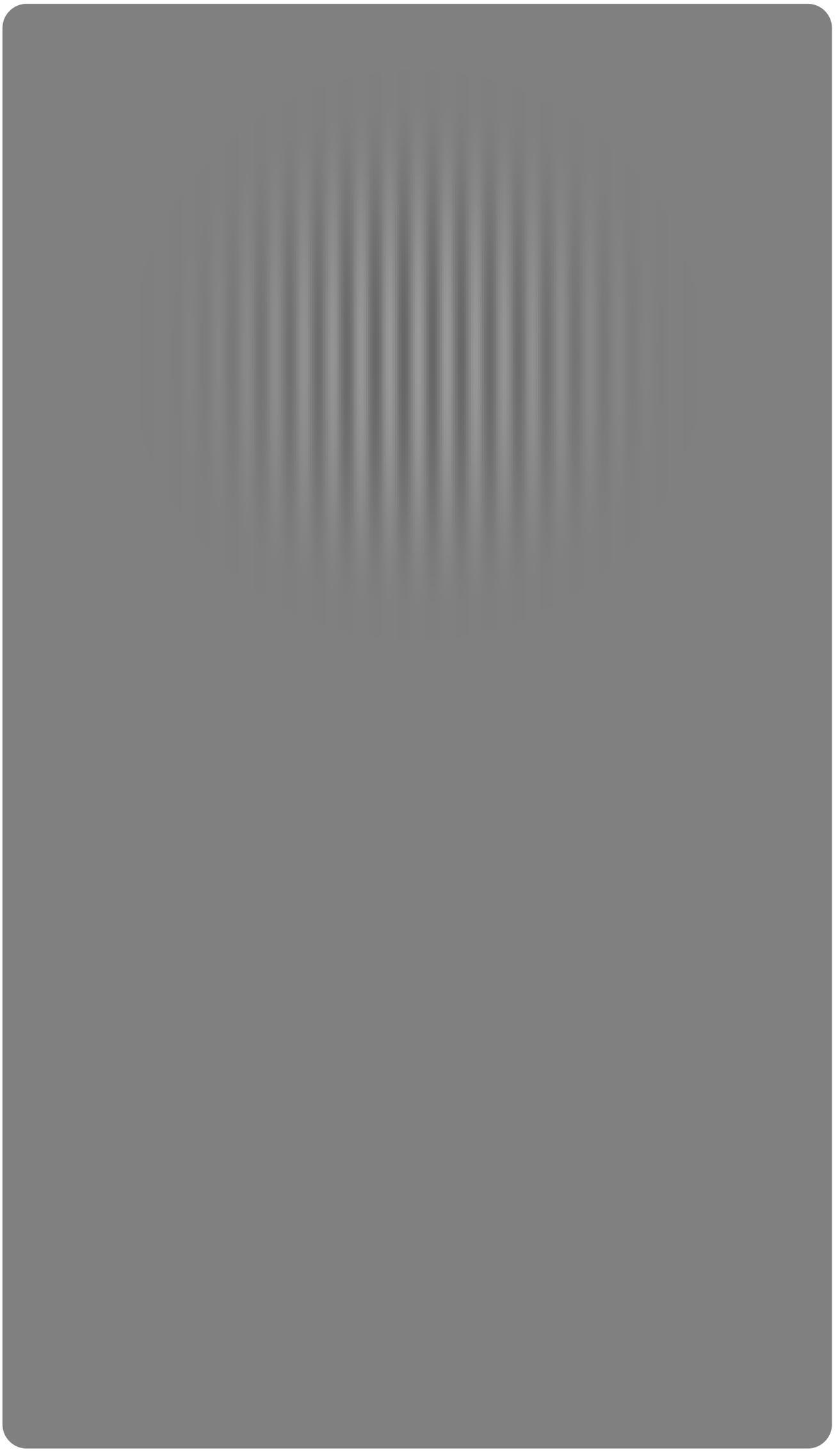
- Warning: Proportion correct is a function of the criterion λ for single-interval designs.

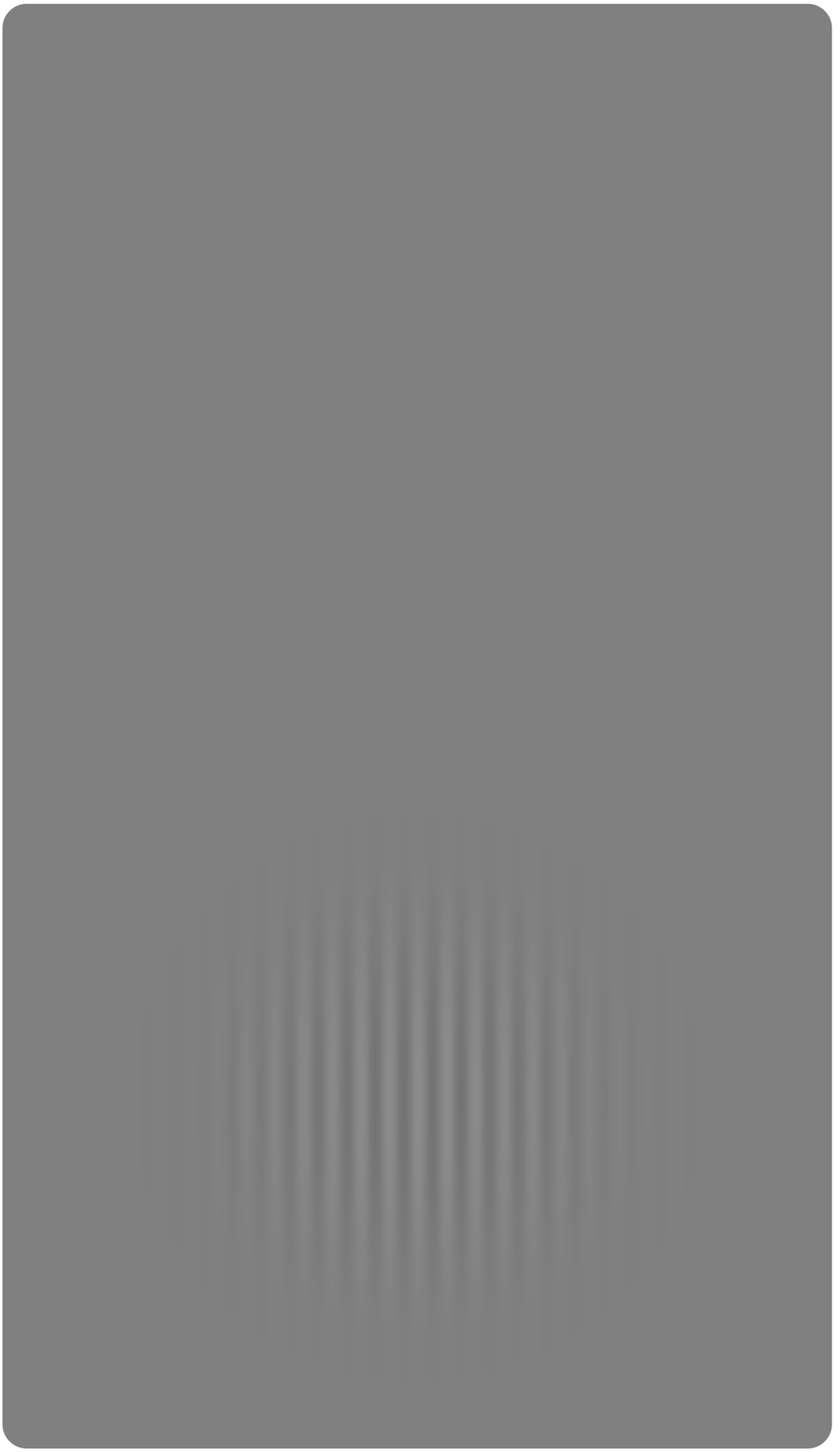
Forced-Choice Procedures

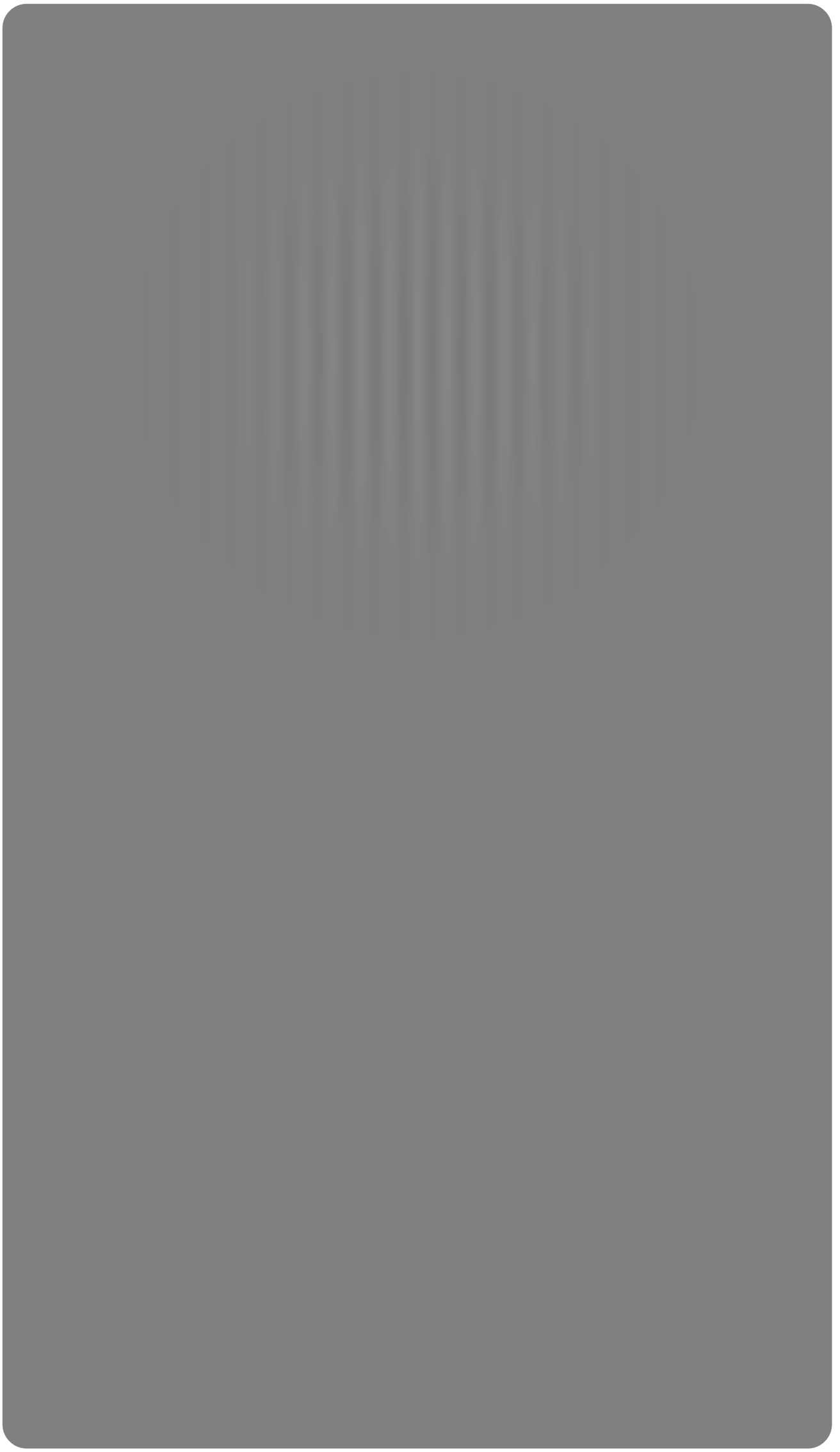
- 2-alternative forced choice (2-AFC) and 2-interval forced choice (2-IFC) procedures are theoretically best as they are criterion free (assuming observers adhere to a differencing rule/pick the interval with the larger activity on the decision-axis).

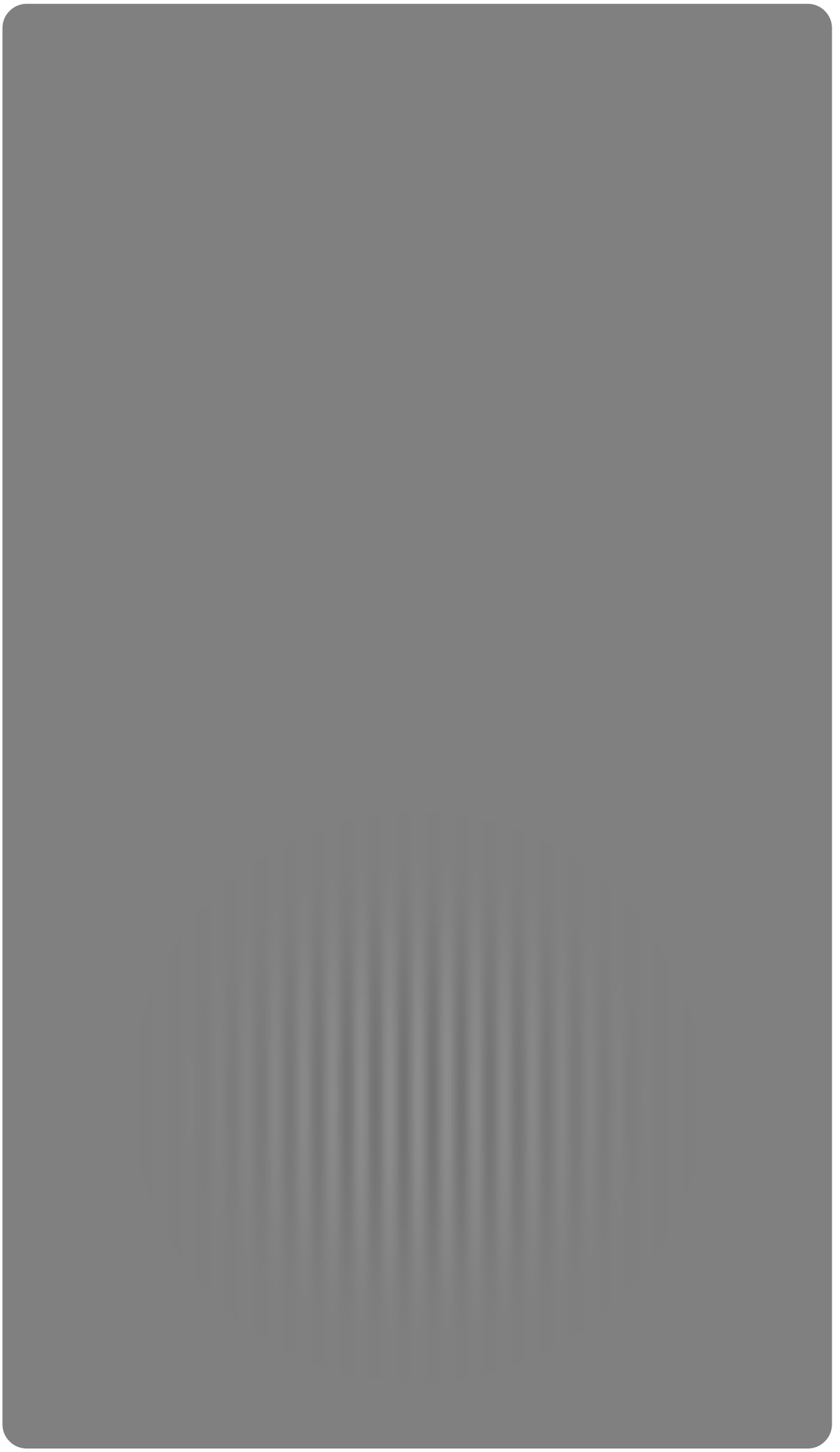












Forced-Choice Procedures

- 2-alternative forced choice (2-AFC) and 2-interval forced choice (2-IFC) procedures are theoretically best as they are criterion free (assuming observers adhere to a differencing rule/pick the interval with the larger activity on the decision-axis).

- If using a differencing rule, percent correct in 2-AFC/2-IFC corresponds to the area under the receiver-operating curve (AUC): ideal!

- Unfortunately (?) "takes perception out of perception research."

- Interval bias in forced-choice procedures?

Interval-Bias in 2AFC

	T_1	T_2	
R_1	$p(R_1\|T_1) \cdot p(T_1)$	$p(R_1\|T_2) \cdot p(T_2)$	
R_2	$(1 - p(R_1\|T_1)) \cdot p(T_1)$	$(1 - p(R_1\|T_2)) \cdot p(T_2)$	

TABLE 1. Parameterisation of the full probability table in a 2-AFC task.

Data from Naive Observers

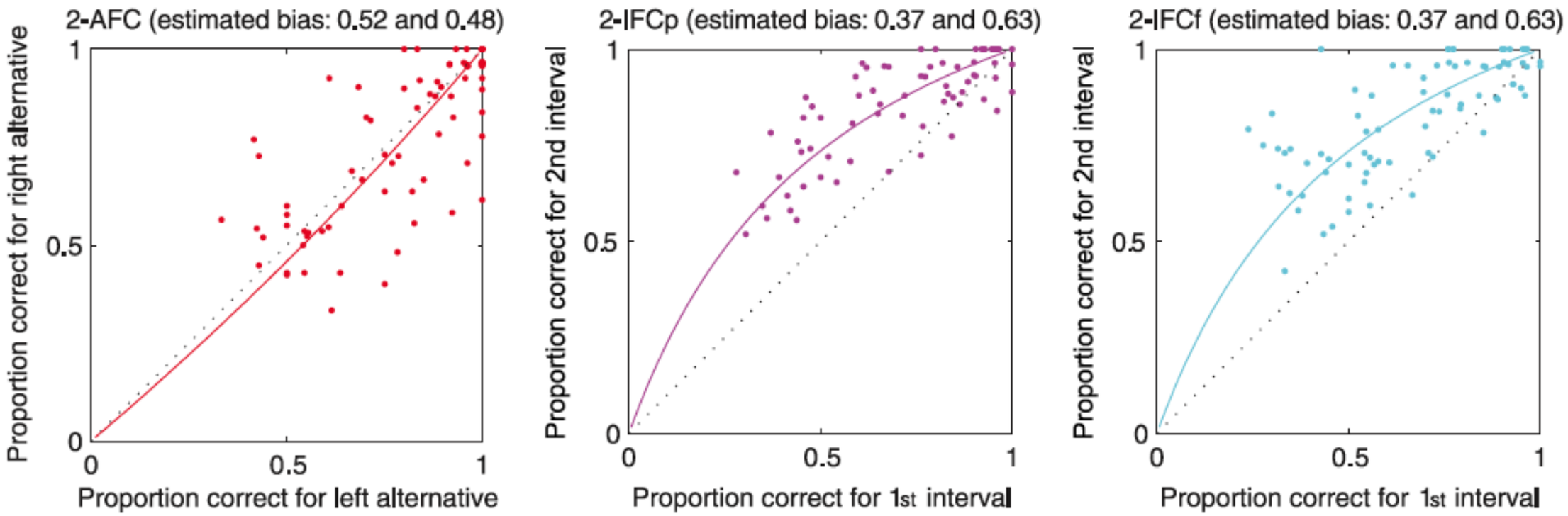


Figure 7. Response biases and isobias curves for participant D.C. for the three methods that involved only two possible responses. Each data point is a block of trials. The axes are the proportion of correct answers for a stimulus that is shown in the first or second interval or on the left or right 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.

from: Jäkel, F and Wichmann, FA. Spatial four-alternative forced-choice method is the preferred psychophysical method for naïve observers. *Journal of Vision* (2006) vol. 6 (11) pp. 1307-22.

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.

from p. 1317:

Jäkel, F and Wichmann, FA. Spatial four-alternative forced-choice method is the preferred psychophysical method for naïve observers. *Journal of Vision* (2006) vol. 6 (11) pp. 1307-22.

Literature on Signal Detection Theory

Tanner, W. P., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review*, 61, 401-409.

Swets, J. A. (1961). Is there a sensory threshold? *Science*, 134(3473), 168-177.

Nachmias, J. (1972). Signal detection theory and its application to problems in vision. In: *Handbook of Sensory Physiology* (Vol. VII/4, pp. 56-77). Berlin: Springer Verlag.

Wickens, T. D. (2002). *Elementary Signal Detection Theory*. Oxford: Oxford University Press.

Van Trees, H. L. (1968). *Detection, Estimation, and Modulation Theory.* New York: John Wiley & Sons.

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

Azzopardi, P. & Cowey, A. (1997). Is blindsight like normal, near-threshold vision? *Proceedings of the National Academy of Sciences*, 94, 14190-14194.

Jäkel, F., & Wichmann, F. A. (2006). Spatial four-alternative forced-choice method is the preferred psychophysical method for naive observers. *Journal of Vision*, 6, 1307-1322.

Essential Practical Implications of SDT

- Decision-axis is likely an abstraction; certainly not necessarily a simple and direct "sensory-signal-strength-axis"

- Stay in the binary world unless you really have to leave it!

- Never use single-interval designs (YesNo and pseudo-2AFC) with adaptive procedures.

- Best to use (proper!) forced-choice procedures unless the subjective aspect of the stimulus appearance is crucial.

- If you use single-interval designs measure AUC (ideal but very time consuming) or calculate d'—percent correct (bad) or hit rates (catastrophe) are insufficient as they depend on the criterion of the observer!

- If you use d' not only plot the psychometric function (converting d' into percent correct assuming an unbiased observer) but show each block of trials in ROC space to spot criterion shifts.

Response Times

History, Terminology & Major Findings

- F.C. Donders (1865) applying electric shocks to the L and R feet of his subjects.

- Technical difficulty: how to measure events with millisecond precision?

- Three components
    1. encoding (sensory): signals have to get to the brain
    2. response time proper
    3. execution (motor): any overt reaction has to involve muscles

- Different types:
    1. Simple RTs—there is only one stimulus and one response
    2. Recognition RTs—respond to one set of stimuli, not to another ("Go-NoGo")
    3. Choice RTs—distinct responses for each possible class of stimuli

- Considerable variability; minimum simple RTs for auditory response 160 msec, visual 190 msec

Influences on RTs

- Stimulus intensity (early 1900's)
- Arousal
- Age; get faster until late 20s, very slight decline to the 60s, then faster deterioration thereafter
- Practice, fatigue, distraction, punishment, drugs ... all in line with common sense
- Amazing (crazy?) amount of studies on whatever variable you want:
    i. no influence of fasting for up to three days on RTs
    ii. adult females have *same* RTs after six (!) cans of beer (but worse accuracy ...)
    iii. breathing cycle (shorter when exhaling)
    iv. Personality: neurotic College students have more RT variability than peers
    v. 22 (!) weeks of intense water-exercise did not improve RTs in elderly people
    vi. Brain injury and illness can slow RTs (heading a ball in soccer has no effect)
- Intelligence

Diffusion Model—Basics

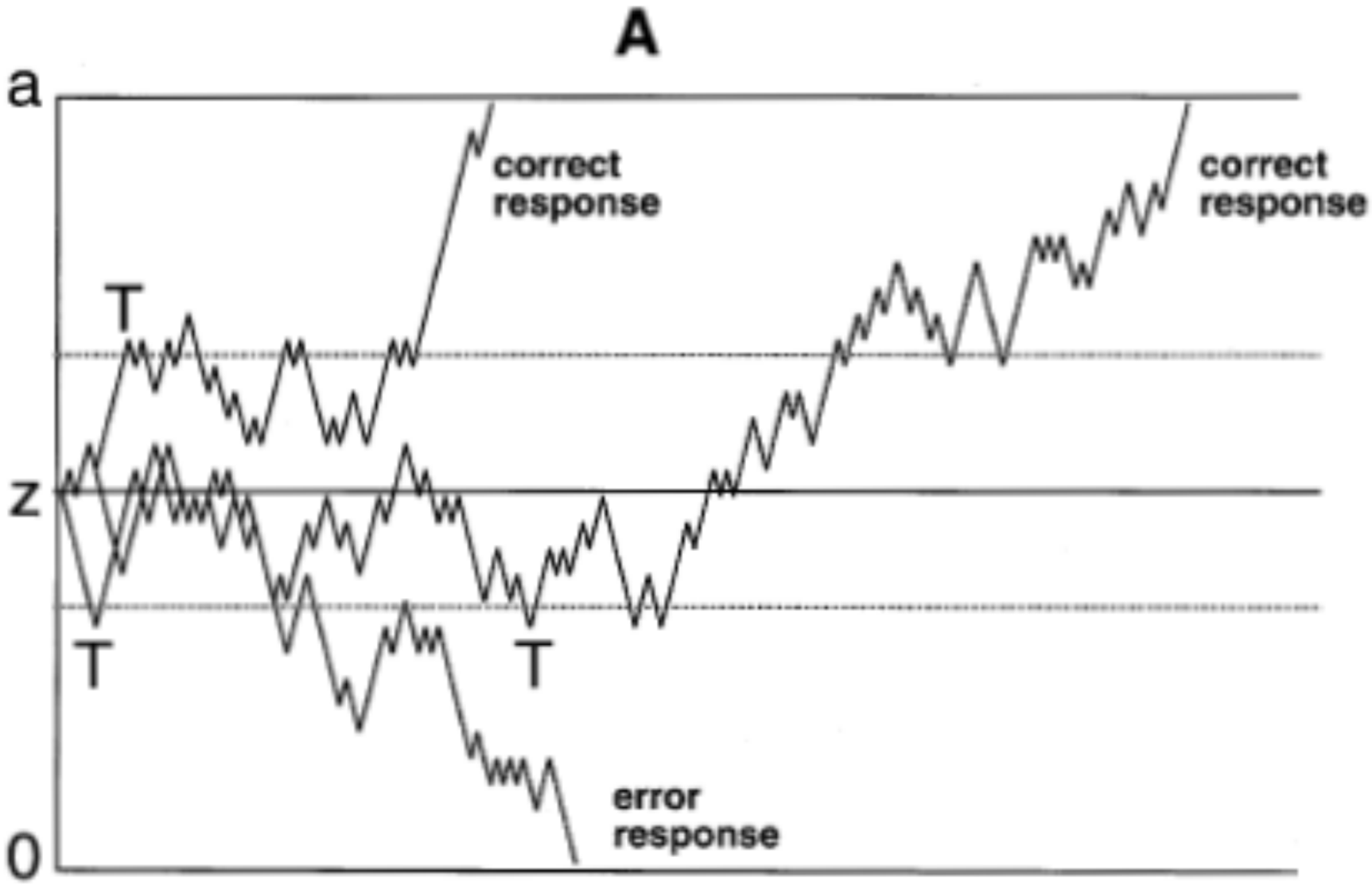
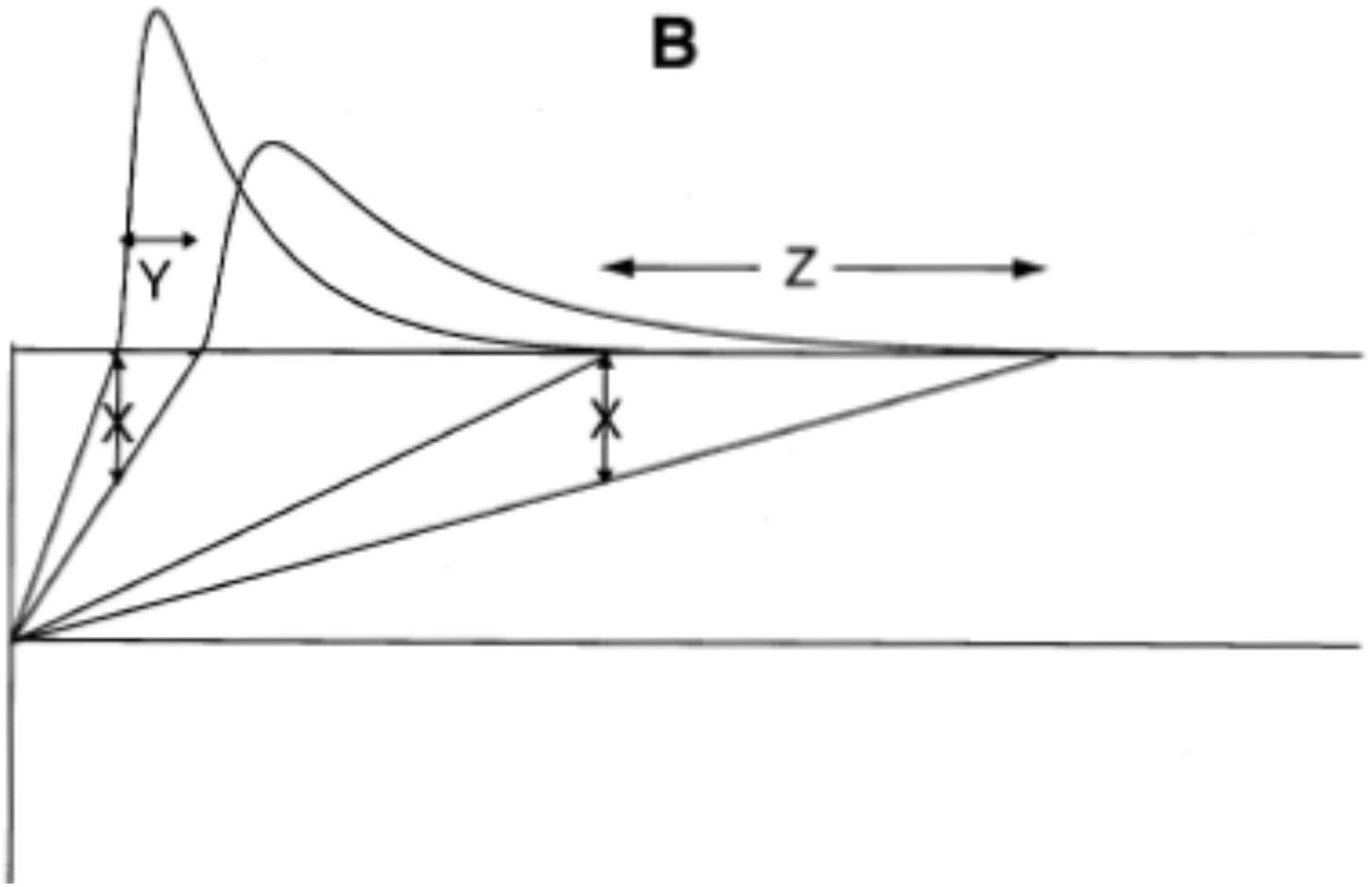


Fig. 1. Illustration of the diffusion model. The sample paths in (a) are derived from a random walk designed to mimic the diffusion process (the continuous version of the random walk). The bottom boundary is set to zero, the starting point of the walk to z, and the upper boundary to a. If the boundaries were moved in to the dotted lines, the processes would terminate at the points T. The straight diagonal lines in

Diffusion Model—Behaviour



(b) represent average paths for two conditions in which the fastest responses differ in mean drift by *X*, and the slowest responses differ in mean drift by *X*. The two curves at the upper decision boundary show illustrative distributions of reaction times for these two conditions. The distributions show that the same difference in mean drift leads to smaller differences between the shortest response times (*Y*) than between the longest response times (*Z*), illustrating the skewing of the response time distribution that is usually obtained empirically when conditions vary in difficulty.

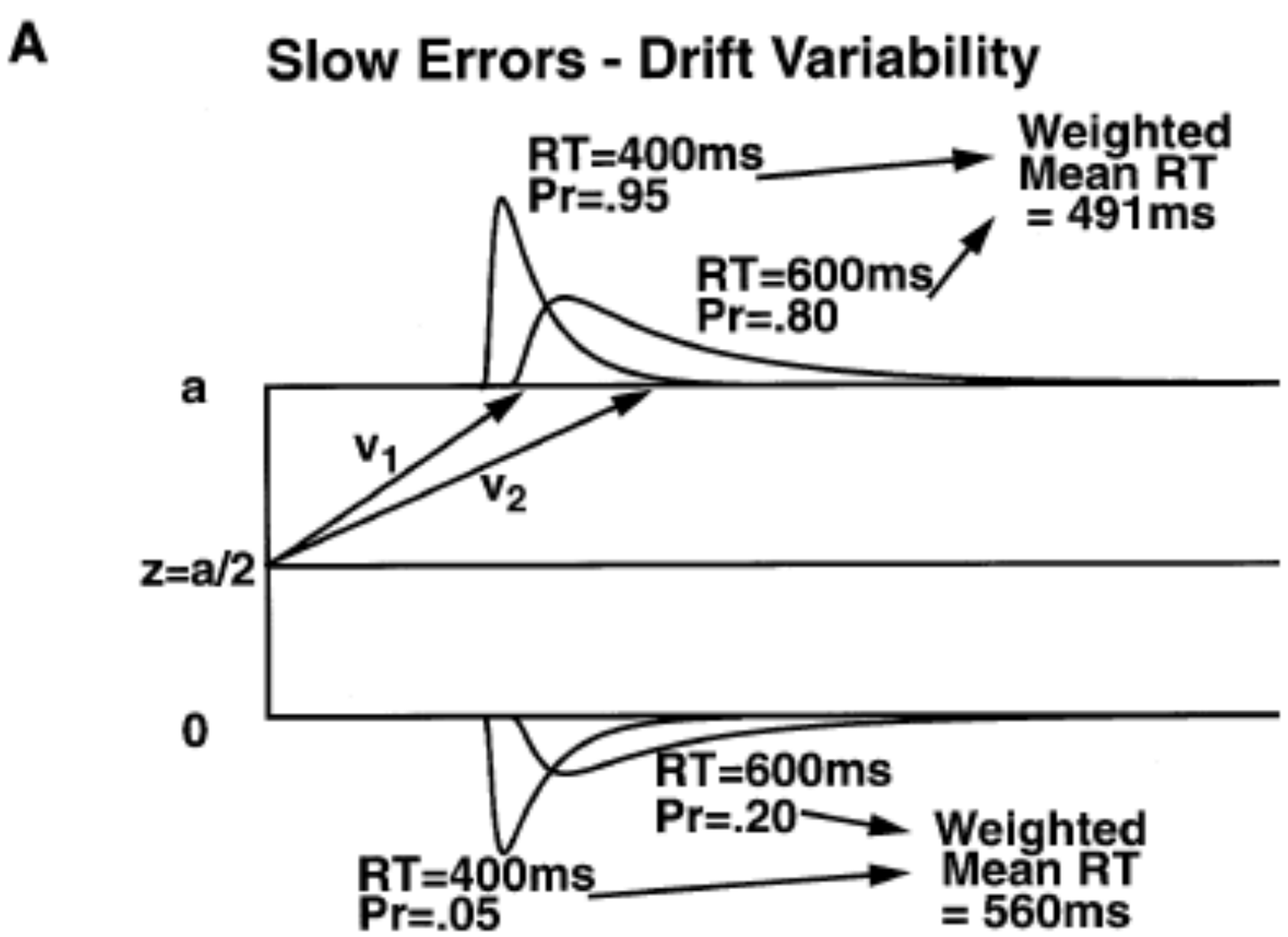
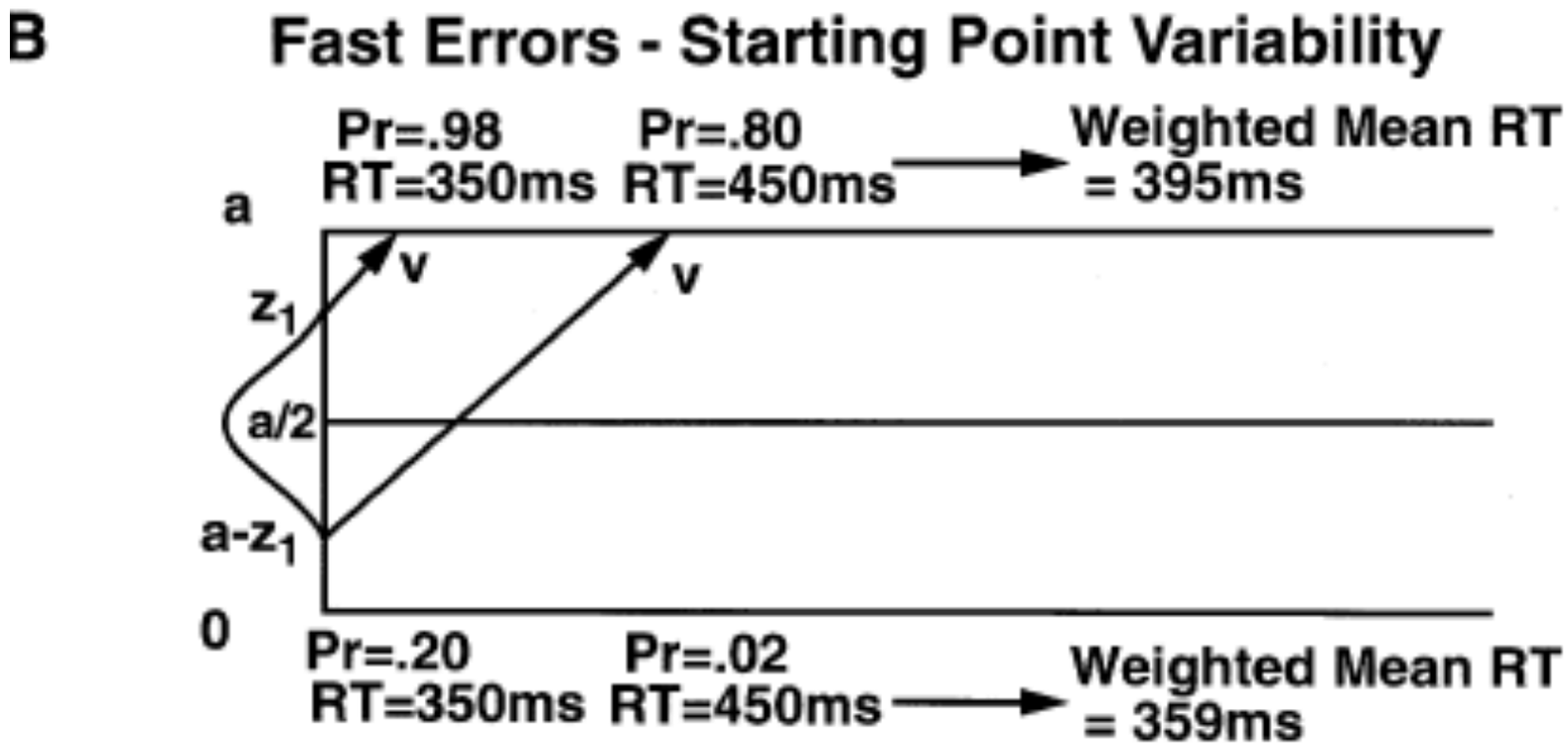
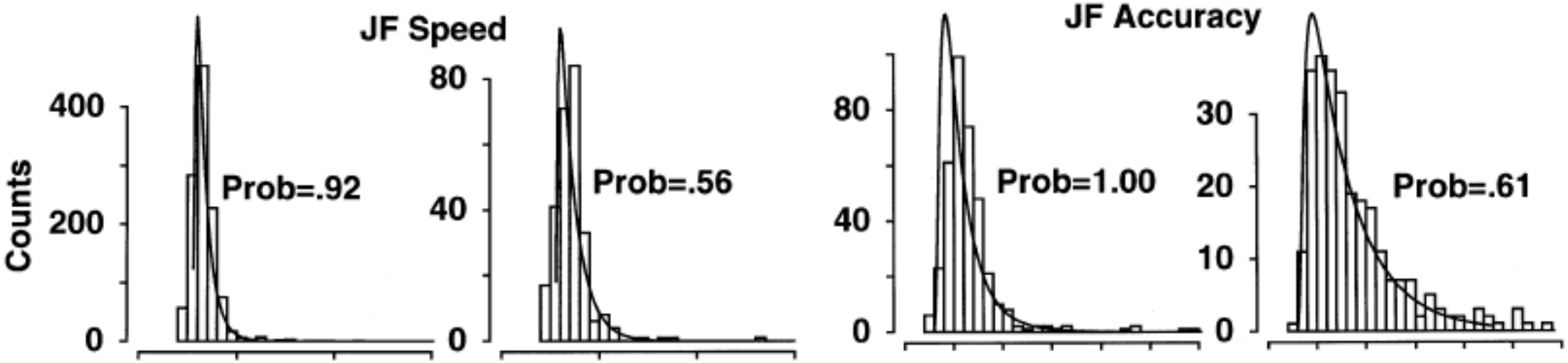


Fig. 2. Illustration of how parameter variability in the diffusion model leads to fast and slow error responses. In (a), two processes have drift rates v_1 and v_2, and the starting point, z, is halfway between the two boundaries. The diagonal lines ending in arrows represent average paths, and the curves at the decision boundaries show distributions of response times (RTs) for the two processes. Correct and error responses have equal RTs (400 ms and 600 ms, respectively). The average of these RTs (exemplifying variability in drift across trials) weighted by probability of response (Pr) leads to slow error responses relative to correct responses. In (b), the effect of variability in starting point is



correct responses. In (b), the effect of variability in starting point is illustrated. Each of the two average paths begins from an extreme of the distribution of starting points centered at $z/2$. Processes starting at z_1 hit the correct boundary with high accuracy and short RT, and errors are slow. Processes starting at $a - z_1$ hit the correct boundary with lower accuracy and longer RT, and errors are fast. The weighted average gives fast errors.

Diffusion Models—Predictions



each tick mark corresponds to 500msec

Literature on Response Times

Luce, D.R. (1986). *Response Times*. Oxford: Oxford University Press.

Laming, D.R.J. (1968). *Information Theory of Choice Reaction Times*. New York: Wiley.

Sternberg, S. (1969). Memory scanning: Mental processes revealed by reaction time experiments. *American Scientist*, 57, 421-457.

Ashby, F.G. & Maddox, W.T. (1994). A response time theory of separability and integrality in speeded classification. *Journal of Mathematical Psychology*, 38, 423-466.

Ratcliff, R. & Rouder, J.N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, 9, 347-356.

Ratcliff, R., Van Zandt, T. & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review*, 106, 261-300.

Epilog

A mathematical theory never guarantees any property of an empirical process; what it says is that if certain assumptions are true, then certain results will follow.

Green, D.M. & Swets, J.A. (1988). *Signal Detection Theory and Psychophysics*. New York: Wiley, p. 11.

Bernstein Center for
Computational Neuroscience
Berlin

Thank you very much!

Felix A. Wichmann

Modelling of Cognitive Processes Group
Bernstein Center for Computational Neuroscience
and
Technische Universität Berlin

felix.wichmann@tu-berlin.de

Bundesministerium
für Bildung
und Forschung