CHAPTER 7

Methods in Psychophysics

FELIX A. WICHMANN AND FRANK JÄKEL

INTRODUCTION

In 1860 the experimental physicist Gustav Theodor Fechner published *Die Elemente der Psychophysik*, the work that made him famous (Fechner, 1860). The year 1860 and *Die Elemente* are now widely regarded not only as the birth of the scientific discipline of psychophysics, but as the beginning of the quantitative, scientific study of psychology. Quantitative science establishes precise regularities between measurable quantities in order to be able to make predictions for hitherto unseen or unobserved situations or stimuli: *Measure what is measurable, and make measurable what is not so*, the dictum of modern (natural) science, frequently—but very likely erroneously—attributed to Galileo Galilei (Kleinert, 2009).

Fechner wanted no less than to approach the mind using the rigorous measurement approach so successful in the natural sciences. However, Fechner's research program set out in *Die Elemente* was as much concerned with metaphysics, consciousness, and dreams as with what is now known as (sensory) psychophysics. His vision was clearly grander than to apply the measurement approach only to sensation (Hawkins, 2011; Heidelberger, 2004; Mausfeld, 2000):

> As a matter of course, we cannot in general deny that the mind is subject to quantitative principles. This is because apart from distinguishing stronger and weaker sensations, we can also distinguish stronger and weaker intensities of drive, higher and lower degrees of attention or vividness of recollections and fantasies, as well as different stages of consciousness and different intensities of individual thoughts.... Consequently, the higher mental processes can—in much the same way as sensory processes—be measured quantitatively, and the activity of the mind can be measured quantitatively in its entirety as well as in its components.[1]

We would like to thank Tom Wallis for helpful comments on the manuscript. This work was supported by the German Research Foundation (DFG; Sachbeihilfe Wi 2103/4-1 to FAW) and the German Federal Ministry of Education and Research (BMBF) through the Bernstein Computational Neuroscience Program Tübingen (FKZ: 01GQ1002).

[1] Translation by Karin Liebeskind from the original German text presented in a keynote by Rainer Mausfeld (2000). The original German text reads: Von vorn herein und im Allgemeinen kann nicht bestritten werden, dass das Geistige überhaupt quantitativen Verhältnissen unterliegt. Denn nicht nur lässt sich von einer grösseren und geringeren Stärke von Empfindungen sprechen, es giebt auch eine verschiedene Stärke von Trieben, es giebt grössere und geringere Grade der Aufmerksamkeit, der Lebhaftigkeit von Erinnerungs- und Phantasiebildern, der Helligkeit des Bewusstseins im Ganzen, wie der Intensität einzelner Gedanken.... Somit unterliegt das höhere Geistige nicht minder als das sinnliche, die Thätigkeit des Geistes im Ganzen nicht minder als im Einzelnen quantitativer Bestimmung (Fechner, 1860, p. 55).

Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience, Fourth Edition, edited by John T. Wixted.
Copyright © 2018 John Wiley & Sons, Inc.
DOI: 10.1002/9781119170174.epcn507

Whatever Fechner himself may have thought about psychophysics, he is nowadays often regarded as the founder of psychophysics in the sense of a narrow reduction program, attempting to explain sensation exclusively in terms of stimulus-response patterns, and relating (simple) detection or discrimination behavior on the ordinate to a (simple) physical variable on the abscissa (Irtel, 1996). But despite the current view that psychophysics is only concerned with quantifying the relation between simple external physical stimuli and the resulting internal sensations, *psychophysical methods* enjoy a widespread use in all areas of psychology and the neurosciences. The reason for their widespread appeal is aptly summarized by Baird and Noma:

> Psychophysics is commonly defined as the quantitative branch of the study of perception, examining the relations between observed stimuli and responses and the reasons for those relations. This is, however, a very narrow view of the influence it has had on much of psychology.... Because of its long history (over 100 years), its experimental methods, data analyses, and models of underlying perceptual and cognitive processes have reached a high level of refinement. For this reason, many techniques originally developed in psychophysics have been used to unravel problems in learning, memory, attitude measurement, and social psychology. In addition, scaling and measurement theory have adapted these methods and models to analyse decision making in contexts entirely divorced from perception. (Baird & Noma, 1978, p. 1)

Thus, and perhaps ironically, we now see the psychophysical methods inspired by Fechner being applied to quantify human and animal behavior, neural responses, as well as neural correlates of a vast range of mental phenomena—presumably more in line with what Fechner had hoped or envisaged back in the middle of the 19th century, instead of confining his influence to the narrow field now known as psychophysics.

Thanks to Fechner's legacy, in the 21st century it may appear entirely obvious to researchers and students in psychology and related fields that we should quantify behavior and use psychophysical (experimental) methods to make progress in understanding the mind. It should not be left unmentioned, however, that Fechner's experimental measurement approach was, at the time, not met with universal approval. William James, another heroic figure in the history of psychology, was at least ambivalent. The following passage from William James warrants at least two comments. First, we think that this may constitute the most beautifully written, sarcastic, and funny description of the experimental method. Second, William James displays his almost prophetic vision again, as he—grudgingly—foresees the success of the experimental method in the future. As witnessed by the enormous advances in so much of psychology and the neurosciences, we now know that he was correct.

> 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. (James, 1890, Chapter VII, The Methods and Snares of Psychology, pp. 192–193)

Scope

So what is the business of those researchers who cannot be bored? At its very core the quantitative analysis of behavior is based on single trials or questions in an experiment. These provide the classical triad of behavioral measurements:

1. The open behavioral response—be it a response and its associated "correctness" or "subjective equality," or a judgment of appearance or magnitude.
2. The time it took to make the open behavioral response—that is, the response or reaction time (RT).
3. The degree of belief in the accuracy of the response—that is, the meta-cognitive feeling of certainty in the accuracy or appropriateness of one's response.

In the following we provide a short overview of these three possible quantitative measurements, and we specifiy which measurements, or aspects thereof, we will cover in detail in this chapter.

The Open Behavioral Response

Within the measurement of open behavioral responses there exist two distinct traditions: First, a tradition dating back to Fechner (1860), concerned with *just noticeable differences* (JNDs)—that is, with measuring the minimal stimulus difference an observer is able to tell apart, often referred to as *threshold*.[2] Second, a tradition most often associated with Stevens (1957, 1960), concerned with the subjective magnitude of one's experience, with perceived loudness or perceived brightness, with clearly perceptible or *supra-threshold* stimuli. Some authors refer to the first tradition as a *sensory discrimination* tradition, to the second as one of *sensory judgment* (Laming, 2001).

With absolute threshold measurements researchers can determine, for example, how much sound pressure is needed before a sound can be heard, or how many light quanta are necessary before a light flash can be seen. With difference threshold methods one can also measure what the difference in sound pressure or luminance needs to be for stimuli to be reliably discriminated. With threshold

[2] The term *threshold* should be treated as a useful construct to summarize data, not an endorsement of (high-) threshold theory postulating a "real" threshold within the nervous system. Signal detection theory (SDT) argues against a threshold (e.g., Swets, 1961), and this position is widely, but not universally, accepted. A detailed and critical discussion is provided in Chapter 5 in this volume.

measurements one cannot, however, determine how loud a sound sounds or how bright a light appears. This is the domain of the second tradition in psychophysics, the domain of magnitude estimation and supra-threshold measurements, where the presented stimuli are far above the absolute threshold and stimuli are easily discriminated.

Physical measurement devices allow us to measure sound pressure (in *Pa*) and radiance (in $\mathrm{W/m^2/sr}$). But how do the physical measurements of sound pressure relate to perceived loudness? How do the physical measurements of radiance relate to perceived brightness? More fundamentally one might ask: How can subjective quantities, like perceived loudness or brightness, be measured at all? Are they measurable in the same way that physical quantities, like sound pressure and radiance, are measurable?

The original aim of supra-threshold methods, or *scaling methods* as they are also called, is to measure psychological quantities in analogy to measurement in physics. Scaling methods (e.g., magnitude estimation, magnitude production, ratio production) are described in a number of excellent books and book chapters; for example, the 2002 edition of the *Stevens' Handbook* has an entire chapter devoted to scaling by Marks & Gescheider (2002); mathematical treatments with links to measurement theory can be found in, for example, Coombs, Dawes, and Tversky (1970).

We will *not* discuss scaling methods in this chapter; instead we concentrate on the many different experimental designs and data acquisition methods used to measure thresholds or JNDs.[3] We attempt to present a useful taxonomy of the methods available, how they are related to each other, and which ones may be preferable. Furthermore, we describe rarely explicitly discussed best practices in psychophysics, as well as a number of problems, and experimental as well as statistical countermeasures: from interval bias and serial dependencies to temporal order effects and more.

Furthermore, we adopt a rather pragmatic approach: In this chapter we do not worry about the exact relation between external stimuli and (internal) sensation (Laming, 1997), whether or when distinctions between “sensation” and “perception” are meaningful, or what the putative internal scale may look like. Finally, we certainly make no attempt at forging the thus far elusive unified theory of psychophysics (Krueger, 1989; Ross, 1997)—that is, finding a mathematical relation between JND-style discriminability and subjective magnitude. Instead we attempt to stay as much as possible theoretically agnostic, and instead focus on which experimental designs lead to more reliable behavioral data, and how to best analyze such data statistically—for an overview of the most often applied general theoretical framework providing a statistical characterization of sensory decisions, SDT, see Chapter 5 in this volume or Green & Swets (1988) as the classic reference.

Response or Reaction Time

The second potential measurement—the response or reaction time—has been intensely studied at least since Donders (1969/1868). Empirical findings, methods, and models are covered in detail in Chapter 9 in this volume.

Confidence

The third potential measurement—the subjective confidence in one's response—has not seen as much attention over the past 150 years as the other two, despite early investigations by Peirce & Jastrow (1885) on the

[3] Thus in terms of the book by Gescheider (1997) we cover the methods described in the first four chapters but not the later chapters (9–14).

sense of pressure, measuring observer confidence judgments on a scale of four levels. One contributing factor might have been the dominance of behaviorism in much of scientific psychology from the 1920s into at least the 1960s. Behaviorism explicitly rejected non–directly observable mental events as being nonscientific, so the metacognitive confidence in one's own open behavioral response was not on its agenda. Another factor was perhaps the belief that confidence is merely a noisy correlate of threshold. In any case, the advent of SDT rekindled the interest in confidence, and recently there have emerged a number of promising avenues and developments into this meta-cognitive ability. For example, research is exploring in more detail whether confidence only reflects accumulated sensory evidence or is an independent internal variable, and how and where the confidence computations may be implemented in the human brain (Aitchison, Bang, Bahrami, & Latham, 2015; Barthelmé & Mamassian, 2010; Boldt & Yeung, 2015; Keane, Spence, Yarrow, & Arnold, 2015; Meyniel, Schlunegger, & Dehaene, 2015; Spence, Dux, & Arnold, 2015).[4]

In this chapter we do not focus on this metacognitive aspect, however, but instead focus entirely on the JND aspect of the first of the three measurements listed previously, the open behavioral responses of the threshold type.

Structure

Earlier we quoted William James's portrayal of the experimental method. In the *Principles of Psychology* James continues immediately following that passage:

> No general description of the methods of experimental psychology would be instructive to one unfamiliar with the instances of their application, so we will waste no words upon the attempt. (James, 1890, Chapter VII, The Methods and Snares of Psychology, p. 193)

We agree with James, and thus follow his recommendation and start our exposition with a number of concrete and representative examples ("instances of their application"). Following from the examples the remainder of the chapter is structured into two further major sections, on *data collection* and on *data analysis*. The section on data collection contains a description and critical assessment of common experimental designs to measure JNDs or thresholds, including their frequently neglected nonstatistical features such as their demands on memory, learning, or attention, as well as on experimental hardware ("know thy stimulus"). The second section on data analysis has a strong focus on how to perform Bayesian inference for the psychometric function. The psychometric function is typically often only one-dimensional; that is, only one experimental parameter is varied at a time. At the end of that section we briefly discuss generalizations to the multidimensional case. This will conclude our general description of the most important psychophysical methods—we hope without having wasted too many words in the attempt.

SOME EXAMPLES

Contrast Sensitivity Functions

Everyone knows the eye charts with letters of varying sizes from an eye doctor or a driving test. A more refined measurement of visual acuity is provided by the contrast sensitivity function (CSF; see, e.g., Hou, Lesmes, Bex, Dorr, & Lu, 2015; Pelli & Bex,

[4] In addition, the interested reader is referred to Vickers (1979), Chapter 6, "Confidence," as well as the introduction of Fetsch, Kiani, and Shadlen (2014) for a brief historical overview in the neuroscience context.

2013). The CSF provides a measurement of the visual sensitivity of an observer at varying spatial frequencies. In order to measure the sensitivity at one spatial frequency, the observer is presented with a grating of that frequency and asked whether she can see the stripes. The contrast of the grating is varied systematically in order to find the threshold at which the observer can only just detect the stripes. The lower this threshold, the higher the sensitivity for the respective spatial frequency. In order to determine the threshold, on each trial the observer is randomly presented with a contrast of zero (no stripes) or a nonzero contrast (stripes). The observer then has to answer: do you see stripes? As on each trial there is an equal probability of there being stripes, chance level is 50%. As each contrast is presented many times, the resulting data are the proportion of correct responses as a function of stimulus contrast. Figure 7.1A shows hypothetical data. By definition, the contrast level at which the observer has a performance of 75% correct is taken as the threshold. Observers with a higher sensitivity need a lower contrast to achieve this performance.

Visual-Haptic Integration

When you hold an object in your hand and look at it, you can judge its size visually and haptically. The information from both senses is integrated to form a modality-independent percept of size (Ernst & Banks, 2002). Visual-haptic integration can be studied by measuring how well subjects can discriminate the size of objects only visually and only haptically. For example, a bar of a certain size, say 3 cm, is taken as a standard stimulus. On each trial, this standard stimulus is shown together with a comparison stimulus that can be shorter or longer. The subject does not know which is which, and the task is to decide which of the two presented stimuli is longer.

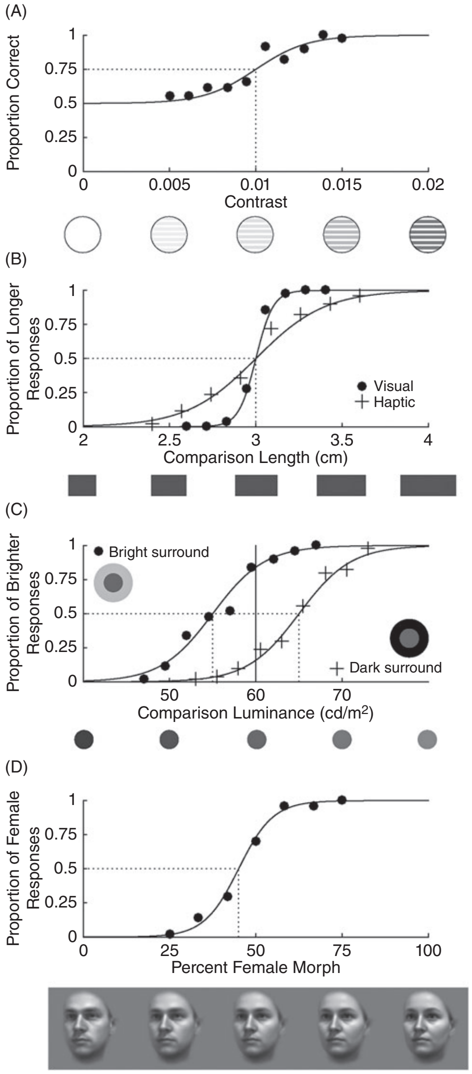

Figure 7.1 (A) Contrast detection. (B) Visuo-haptic comparison. (C) Context effects in brightness perception. (D) Gender classification of human faces.

Hypothetical data for one participant in the visual-only and the haptic-only conditions are shown in Figure 7.1B. The *x*-axis shows the size of the comparison stimulus, and the *y*-axis the proportion of “longer” responses. If the comparison stimulus is as long as the standard stimulus (3 cm), the subject will not be able to decide which one is longer and accordingly the subject will choose either with equal probability. The steeper slope of the visual-only condition compared to the haptic-only condition indicates that the visual system is better at discriminating size than the haptic system: A smaller difference in size is required for the same discrimination performance. Different models for multimodal integration make different predictions about the performance that can be expected if both modalities are available simultaneously. In order to select between different models of multimodal integration, it is therefore important to measure precisely the discrimination ability of subjects in different conditions.

Context Effects in Brightness Perception

Many phenomena in perception show striking context effects. In vision, what is seen at a particular part of the visual field not only is a function of the light that hits the respective parts of the retina but may also depend on the rest of the visual scene. We owe this insight to Gestalt psychologists who thus overturned preceding structuralist conceptions of perception (Hochberg, 1964). A classic example is simultaneous lightness contrast: The perceived brightness of a circular gray patch (say with a luminance of 60 $\mathrm{cd/m^2}$) depends on the gray value of a surrounding ring (inset in Figure 7.1C). The size of this context effect can be measured by presenting a subject, on each trial, with this stimulus as a standard. In addition, the subject is shown patches of varying gray values without the outer ring and is asked whether the standard (with the outer ring as context) is darker or lighter than the comparison stimulus (without a ring). Different comparison stimuli are repeatedly compared to the standard stimulus, and the measurements consist of the proportion of "lighter" responses as a function of the gray value of the comparison patches. Of interest here is the *point of subjective equality* (PSE)—that is, the gray value of the comparison patch for which the subject is undecided about which one is darker or lighter and both responses are equally probable. The difference between the PSE and the actual gray value of the central patch of the standard stimulus can now be taken as a quantitative measure of the effect of the surrounding ring.

Gender Classification

Although many successful applications of psychophysical methods use simple stimuli (gratings, bars, or color patches), these methods are not confined to simple stimuli. However, the amount of effort required for experimental designs with many independent variables quickly becomes prohibitive, and complex stimuli usually have many dimensions that might potentially influence the dependent variable. Sometimes it is nevertheless possible to reduce the number of independent variables considerably. For example, human faces vary in many different dimensions, and human subjects can easily classify them into male or female. Which dimensions are the most relevant ones for gender classification? Systematically trying all dimensions and their combinations is prohibitive. However, by fitting different models of categorization to human classifications of a sample of faces, one can identify which dimension is the most relevant one for each model. For example, if the prototype theory of categorization was correct (for a suitable face space), then the axis between the mean of all male faces and the mean of all female faces should be the decisive dimension. By systematically varying images of faces along the crucial dimensions for each theory, one can test which model best predicts human behavior (Macke & Wichmann, 2010;

Wichmann, Graf, Simoncelli, Bülthoff, & Schölkopf, 2005). Thus, in the case of the prototype theory one can generate artificial, morphed faces that lie on the axis between the male and the female prototype and show these stimuli to subjects. Figure 7.1D shows the hypothetical proportion of "female" responses of one subject when seeing such morphs. Different models of categorization will make different predictions about the responses to these morphs and can thus be compared.

Psychometric Functions

Looking at the four preceding examples, it is clear that they address a wide range of different questions within vision research (we have concentrated on vision because this is the area we are most familiar with, but also because it is easier to show stimuli in print). Are there substantial methodological differences between these experiments? The first experiment uses a detection task to measure contrast sensitivity. On each trial the subject's task is to state whether a certain stimulus is there or not. In contrast, the visual-haptic integration experiment asks subjects to discriminate between two stimuli. In a discrimination task we have a *standard stimulus* (in some contexts called the pedestal) and a variable *comparison stimulus* (in some contexts called the increment or decrement of the signal). Notably, however, detection and discrimination tasks are not substantially different from each other—detection is simply discrimination from zero. Although the first experiment measures performance and the second experiment measures the probability of each possible response, these two measures can easily be translated into each other ("longer" responses are correct to the right of the standard stimulus and incorrect to the left of it). In contrast, measuring performance makes no sense in the brightness perception experiment or the gender classification experiment where we are interested in the participant's subjective impression and not the objective performance. Nevertheless, all these examples have in common that they vary the stimulus systematically along a single dimension and count a subject's open behavioral responses to form a proportion. The probability to see a specific response as a function of the stimulus level is called a *psychometric function*. As the concept of a psychometric function is so widely applicable, it is one of the central concepts for understanding psychophysical methods. Figure 7.1, panels A to D, has already shown the psychometric functions along with data, and the data analysis section will discuss how psychometric functions can be fitted to data (Figure 7.2 shows the psychometric function with its parameters). But first we will cover more systematically how the data are collected in the following section.

DATA COLLECTION

There are two aspects to psychophysical measurement (i.e., data collection): the psychological and the physical. Though we will be mainly focusing on the psychological aspects in the remainder of this chapter, it is equally important to state that psychophysicists also need to control the physical aspects of their stimuli. As both authors of this chapter mainly work in vision research, we restrict our discussion to physical and technical aspects in vision research. Of course, similar issues arise in all sensory modalities.

Setting Up the Hardware

Only 50 years ago, setting up a psychophysical experiment in vision research was no mean feat, as vividly described by Jan Koenderink (1999):

> I took up visual psychophysics in the mid sixties. Setting up a new experiment was considered to be a major undertaking. . . . Often an experimental setup would fill the best part of a laboratory room, typically dominated by a huge metal table with pneumatic dampers in the legs. . . . On this table, optical rails would be mounted on which the various parts moved on gliders. Assorted parts were such things as lenses, prisms, beamsplitters, mirrors, filters, shutters, mirror boxes, integrating spheres, and so forth. At the business end of the apparatus one mounted at least a head and/or chin rest, but even better a "bite board." I still shudder at the thought. The bite board might have been invented by the Inquisition. At the far end of the table one had one or more light sources. . . . It might easily take half a year to a year to really get things going and have it all properly lined up and calibrated. (p. 669)

Clearly, things have changed considerably in the meantime—almost all visual and auditory psychophysical experiments are now computer-controlled, and often researchers can make use of purpose-written psychophysical software packages like the truly excellent *Psychtoolbox-3* (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007; Pelli, 1997b).

However, the ease and speed with which even psychophysical novices are now able to program experiments and display stimuli on the liquid-crystal display (LCD) of their laptop—or present their auditory stimuli through headphones plugged into the headphone jack of it—may not be ideal, either: The programming and display ease belies the technical difficulty behind proper physical stimulus control. Every now and then we believe it to be helpful to remind ourselves of Wilson Geisler's "[f]irst commandment of psychophysics: 'Know thy stimulus'" (Geisler, 1987, p. 30), because *know thy stimulus* reminds us of just how important it is to know one's hardware and technical setup if one wants to perform a proper psychophysical experiment.

After the laborious setting up of optical benches, mirrors, and light sources and until the mid-2000s, cathode-ray tube (CRT) displays were the standard monitors or visual display units (VDUs) used in visual psychophysics. Some of the technical problems of CRTs, like geometric distortions, were so plainly obvious that experimentalists simply could not ignore them. Furthermore, most of the problems could be corrected or at least minimized through comparatively simple means, as CRTs were (largely) analogue devices. As a corollary, the issue of display measurement, calibration, and bit depth beyond the 8 bits provided by the digital-to-analogue converters (DACs) of the graphics cards was a much-discussed topic both at conferences and in journals (see, e.g., Bach, Meigen, & Strasburger, 1997; Brainard, Pelli, & Robson, 2002; Di Lollo, Seiffert, Burchett, Rabeeh, & Ruman, 1997; Golz & MacLeod, 2003; Naiman & Makous, 1992; Pelli, 1997a; Pelli & Zhang, 1991; Tyler, 1997, to name just a few of the relevant references)—it did not take half a year to a year to get a CRT experiment going, but researchers were aware of the technical issues and limitations of their equipment: The community (still) knew their stimuli.

The advent and subsequent widespread use of LCDs in laptops and consumer displays effectively ended the reign of the CRT in visual psychophysics,[5] and, unfortunately,

[5] Some vision laboratories bought sufficient quantities of CRTs in the late 2000s to allow them to continue using them in the near future.

the widespread knowledge of the limits and importance of VDUs. Standard consumer LCDs are optimized to display stimuli to look good—after all, this is why they get bought. LCDs do not suffer from obvious geometric distortions like CRTs; furthermore, the engineers certainly have done their homework, and they make images displayed on LCDs often really look good! However, "looking good" is not what visual psychophysicists should want—they should want stimulus control. Unfortunately, and despite their endearing image quality, consumer LCDs are beset with technical limitations that restrict their general usefulness as displays in vision experiments (e.g., Elze & Tanner, 2009, 2012).[6] Unlike in the case of CRTs, almost all the technical problems of LCDs cannot be corrected by the researcher, however, because they stem from often quite elaborate digital signal processing applied to the input signal before image display. The complexity and ubiquity of the signal processing applied inside LCDs prevents any attempt to counteract it through reverse engineering. The combination of being powerless to correct LCDs, and their beguiling image quality, has, we think, led to a loss of appreciation for the importance of *know thy stimulus*.

Fortunately, the future may look bright again for vision research: Organic light-emitting diode (OLED) displays are a promising technology believed to supersede LCDs even for the largish displays commonly used in vision research.[7] In principle, OLED displays should not suffer from the documented temporal artifacts of LCDs while retaining their advantages like zero geometric distortion. As of late 2016, large OLED displays are, alas, not yet free of serious technical problems hindering their usefulness for vision research (Elze, Taylor, & Bex, 2013; Hoffman, Johnson, Kim, Vargas, & Banks, 2015). Furthermore, large OLED displays are still prohibitively expensive. But we are hopeful that a psychophysical methods chapter in the next edition of the *Stevens' Handbook of Experimental Psychology* will be able to report that the difficult years after the demise of the CRT have finally ended, and vision researchers can use truly high-fidelity OLED displays to present their stimuli—and a lack of technical knowledge about the used OLED display by the individual researcher does not potentially undermine the soundness of the science as is the case with LCD displays.

However, we would like to draw attention to another important stimulus-fidelity issue in visual psychophysics, perhaps one that will be the dominant issue once high-fidelity OLED displays are common: the sharply growing interest in using realistic stimuli in vision research. Instead of just using photographs of natural images, the past two decades have seen a keen interest in using naturally appearing images under experimenter control as stimuli—that is, using advanced computer graphics to render photorealistic images in software. We strongly recommend anyone considering using rendered images in their research to first read the often-scathing critique entitled "Virtual Psychophysics" by Koenderink (1999). Luckily, the situation has changed considerably since 1999, and the exploration of the physical accuracy of rendering software (e.g., Heasly, Cottaris, Lichtman, Xiao, & Brainard, 2014) or the perceptual relevance of different physical light-scattering types (e.g., Zhang, de Ridder,

[6] Obviously, sometimes the lack of temporal and luminance precision may not matter, as it was shown by Kihara, Kawahara, and Takeda (2010) for metacontrast masking and the attentional blink *for the parameters explored by the authors*. But it is difficult to know a priori whether this will be the case for a new experiment. Dedicated LCDs developed for vision research are clearly better, but, unfortunately, clearly much more expensive, too (Ghodrati, Morris, & Price, 2015).

[7] Many small displays in smartwatches and smartphones are already OLED displays.

Fleming, & Pont, 2016), as well as the fidelity of (low-cost) virtual reality equipment (e.g., Scarfe & Glennerster, 2015) is clearly underway and appreciated in the vision community.

Still, we would like to emphasize that experimental reseachers should perhaps not be too reliant on, or perhaps not succumb too easily to, advanced automatic rendering or OpenGL tools when generating their stimuli unless they understand what the powerful tools they use do to their stimuli. Thus knowledge of the physical or technical aspect of data collection is and will always be crucial, even if using a putative "error-free" OLED display in the future. In the succinct words of Wilson Geisler: *Know thy stimulus*.

Experimental Tasks

After the hardware is set up and calibrated, there are still many decisions to be made about a psychophysical experiment. Most of those decisions are dictated by the scientific question under consideration and the nature of the stimuli. There are nevertheless many degrees of freedom in how to actually run the experiment. Luckily, as the aim of most psychophysical experiments is to measure a psychometric function, it is therefore possible to systematize different tasks and give some general recommendations on how to run such psychophysical studies. The nomenclature for these tasks is, unfortunately, not standardized in the literature and often differs significantly, especially between human and animal research. Stüttgen, Schwarz, and Jäkel (2011) give a useful overview of the most common tasks, which will be built on here. The tasks have in common that subjects are presented with a series of trials, and the comparison stimulus (the *x*-axis of the psychometric function) is varied systematically. Table 7.1 shows how each trial is structured in the different tasks in a discrimination experiment with standard and comparison stimuli.

Table 7.1 The Structure of a Trial in Different Tasks. Each line is a possible sequence of stimuli in a trial for the respective tasks. In a yes-no task the subject is asked whether the presented stimulus is different from the standard stimulus. In a yes-no with reminder task and the ABX task the subject is asked whether the stimulus is in the second or third interval, respectively. In the 2IFC and 3IFC tasks the subject is asked which of the intervals contains the stimulus that is different from the standard. In the same-different task, in contrast, the subject is asked whether the two intervals show the same or different stimuli. In the oddity task, the subject is asked to pick the interval that contains the stimulus that differs from the other stimuli in this trial.

Task	1st Interval	2nd Interval	3rd Interval
Yes-no (without catch)	Comparison		
Yes-no (with catch)	Standard		
	Comparison		
Yes-no with reference	Standard	Standard	
	Standard	Comparison	
Same-different	Standard	Standard	
	Standard	Comparison	
	Comparison	Comparison	
	Comparison	Standard	
ABX	Standard	Comparison	Standard
	Standard	Comparison	Comparison
2IFC	Standard	Comparison	
	Comparison	Standard	
3IFC	Comparison	Standard	Standard
	Standard	Comparison	Standard
	Standard	Standard	Comparison
Oddity	Comparison	Standard	Standard
	Standard	Comparison	Standard
	Standard	Standard	Comparison
	Standard	Comparison	Comparison
	Comparison	Standard	Comparison
	Comparison	Comparison	Standard

Yes-No

The simplest task, commonly used in detection experiments, is the *yes-no* task. In each trial there is a clearly marked presentation interval. During this interval the subject is presented with a stimulus, say a sound, a grating, or a flash of light. The subject is asked whether she perceives the stimulus. The strength of the stimulus is varied over trials in order to determine which stimulus-levels can or cannot be perceived. In this way, for example, an ear doctor can determine the sound pressure level that is required for a tone to be heard. The trouble with this task is that there is no objective measure as to whether the subject has really heard the sound. As the subject knows that there is a stimulus on each trial, she could easily cheat and always say yes.

Hence, in modern applications of the yes-no task, catch trials are almost always interspersed with the stimulus presentations (as for the example experiment shown in Figure 7.1A). On each trial participants are presented with a stimulus, or not. Whether a stimulus is shown or not depends on a coin flip. Usually the coin is fair and each possibility equally probable. The participants' task is to report whether they could detect a stimulus or not, answering yes or no by pressing one of two clearly labeled buttons. Even though the subject might be uncertain as to whether there was a stimulus, she still has to decide for one of the two options, guessing if necessary. Chance level for a correct response is thus at 50% for stimulus levels that cannot be perceived, and nonchance performance can be measured objectively.

Go/No-Go

In animal experiments the yes-no task is often replaced with a *go/no-go* task. Instead of choosing between two differential responses (yes and no) the task it to respond or not respond within a certain time frame. Training animals to follow this protocol can sometimes be easier than training them to respond differentially. However, unless there is more than enough time to respond, the trials where the subject did not respond are ambiguous. Therefore, the go/no-go task is rarely used in psychophysical studies with humans unless reaction time measurements are of primary interest.

Single-Interval Identification

Perhaps surprisingly, while the yes-no task with catch trials seems most natural in the context of detection experiments, it can also be used in studies that measure discrimination ability. In this case the stimulus in each trial can be either the standard or the comparison stimulus, and the subject responds "no" to the former and "yes" to the latter, effectively answering the question: Is the stimulus different from the standard stimulus? Alternatively, the standard stimulus is labeled "A" and the comparison stimulus "B" and the subject's task becomes to identify the stimulus that was shown. Although the change in the response labels does not really change the structure of the yes-no task, it is less confusing to refer to it as *single-interval identification* because the subject has to identify the stimulus in the single interval she is presented with. This nomenclature also obviously includes the generalization to more than two stimuli and more than two responses.

Yes-No With Reference

The yes-no task will often be used in discrimination tasks if the standard stimulus provides a natural reference—for example, when the standard is vertical motion and the task is to discriminate it from nonvertical motion. However, the standard stimulus is not always a natural reference point. A standard bar with a length of 3 cm in a length-discrimination task is a good example. On each trial the subject is shown a bar and has to decide whether it is different from the standard bar (that is not shown). In such single-interval discrimination tasks the subject has to store the standard bar in long-term memory in order to be able to discriminate the stimuli. That is, the task can be performed only if at least the standard stimulus is sufficiently familiar to the subject—for example, through extensive training. In order to alleviate this memory problem, an alternative to a yes-no task is to provide a reference stimulus on each trial. In such a *yes-no with reference* task, subjects are presented with two clearly marked presentation intervals in each trial. The first interval always contains the standard stimulus as a reference. The second interval is a standard yes-no interval and subjects then answer the question: Is the second stimulus different from the standard stimulus that was shown first as a reminder?

Same-Different and ABX

Essentially, the question in a yes-no with reference task is whether the two stimuli that are presented are the same or different. Hence, a yes-no with reference task might be called a *same-different* task. However, usually this term is reserved for tasks where one interval always contains the standard stimulus and the other interval contains the standard or the comparison stimulus, but it is chosen randomly which interval is which. Contrary to the yes-no with reminder task, in a same-different task the subject therefore does not know beforehand which of the two intervals contains the standard stimulus. Sometimes it is prudent to remind subjects on each trial not just about the standard stimulus but also about the comparison stimulus, as in the so-called *ABX* task. In each ABX trial subjects are presented with three stimulus intervals: A, B, and X. “A” is the standard stimulus, “B” the comparison stimulus, and “X” a normal yes-no interval. The order of A and B is often randomized, and the subject’s task is to report whether the third interval is the same as the first or the second interval.

2IFC and 2AFC

In the yes-no with reference task and the same-different task there are two presentation intervals in each trial. The same is true for two-interval forced-choice (2IFC). However, the subject in the two former tasks is effectively answering the question of whether the two stimuli are the same or are different. In contrast, in 2IFC the two intervals are always different: One contains the standard stimulus and one the comparison stimulus. Or in the case of detection, one contains a stimulus and one does not. It is decided randomly which one is shown first. Hence, as the subject knows this, the question that is being answered is which interval is which. This question assumes that the subject knows the stimuli and tries to identify them. However, the question can often equivalently also be phrased as which of the two intervals has the higher value on the dimension of interest (e.g., contrast, loudness, brightness, etc.), thus turning the task into a comparative judgment task (Thurstone, 1927).

Tasks that present subjects with several stimuli in one trial often do so in subsequent presentation intervals. This is usually the case in hearing research or haptics. In vision it is, however, also common to present the two stimuli side by side rather than one after the other. For another example, consider Fechner’s classic experiments on weight discrimination where the standard stimulus and the comparison stimulus may be held simultaneously in different hands. All tasks that are like a 2IFC task, but with the potential difference that the two stimuli in each trial are not presented in temporal intervals one after the other but in another way, are called two-alternative forced-choice (2AFC) tasks. Note that the “two alternatives” refer to two possible stimulus presentations for the comparison stimulus (e.g., first or second interval, bottom or top on the screen, left or right hand). This nomenclature has occasionally led to confusion in the literature, as the single-interval identification task that was just described has also wrongly been called 2AFC in the belief that the "two alternatives" refer to the two possible responses rather than the two presentation alternatives in each trial.

m*AFC and Oddity*

The 2AFC task can easily be generalized to more than two alternatives. For example, instead of showing only two gratings on a screen in each trial and asking which one has a higher (or lower) contrast, one can show more stimuli and ask which one has a higher (or lower) contrast than the others. Increasing the number of alternatives increases the efficiency of the method considerably (Hou et al., 2015; Jäkel & Wichmann, 2006). With more than two alternatives on each trial, the task can also be conceptualized as finding the one stimulus that is different from among the *m* alternatives. In an *m*AFC task the odd stimulus is always the comparison stimulus. In an oddity task the odd stimulus could also be the standard. The formulation as an oddity task has the advantage that it does not require an explicit ordering of stimuli along one dimension (such as contrast) and does also not assume that the subject already knows the stimuli.

Relating Different Tasks

Researchers in sensory psychophysics early on realized that different tasks and designs produce different results (Blackwell, 1952). One of the reasons for the popularity of signal detection theory (SDT) in psychophysics is that it relates a subject's performances in different tasks to each other (Green, 1960; Jang, Wixted, & Huber, 2009; Swets, 1959). Although, overall, SDT is extremely successful in doing so, unfortunately, even the supposedly simple tasks like yes-no and 2IFC sometimes do not relate to each other in the predicted way (Jäkel & Wichmann, 2006; Yeshurun, Carrasco, & Maloney, 2008). The ideal observer postulated by SDT for 2IFC compares the two stimuli on each trial by taking the difference on a putative internal evidence axis, but subjects might have a different, suboptimal strategy. For example, they might merely do two yes-no decisions, one for each interval, and answer randomly in cases of ambiguity. In addition, SDT does not model many of the extra-sensory factors that influence the results. For example, take a visual contrast discrimination task. Although the logical structure of 2IFC, where two stimuli are presented sequentially, and spatial 2AFC, where two stimuli are presented simultaneously side by side, is exactly the same, very different cognitive processes might be involved. To be able to compare the first to the second interval in 2IFC requires some kind of visual memory. In contrast, in spatial 2AFC the limiting factor beyond sensory discriminability might not be memory but spatial attention. Thus, not surprisingly, it has been reported that 2IFC and 2AFC can indeed lead to different results (Blackwell, 1952; Jäkel & Wichmann, 2006). Hence, even for simple psychophysical tasks there might be different mechanisms and strategies underlying the observed performance, and without understanding these mechanisms and strategies it may be difficult to compare results across tasks. This holds even more so for more complex tasks, like the method of adjustment that is traditionally covered in discussions of psychophysical methods (see, e.g., Laming, 2013; Wier, Jesteadt, & Green, 1976).

Experimental Design

Fixed Versus Adaptive Designs

Choosing a task is obviously not enough. One common problem when designing a psychophysical study is to choose the stimulus levels (e.g., the contrasts, lengths, luminances, or morphs in the previous examples) that will be shown to each subject. Unfortunately, there are often big interindividual differences, and different subjects will need to be presented with different stimulus levels in order to sample the psychometric function well over its whole range. This makes it hard to choose one experimental design (i.e. a set of stimuli and the order in which they are presented for each condition) for all subjects before the experiment. Instead, experimenters can do a quick pretest to get a rough idea where the psychometric function for a subject lies and then choose a *fixed* set of stimuli, usually 6 to 10 stimulus levels over the whole range of the psychometric function, appropriately for each subject separately.

Alternatively and frequently, experimenters choose the stimulus levels *adaptively* (for an oversiew, see, e.g., Treutwein, 1995). A classic method is *staircases*. For example, in the brightness perception experiment described earlier, the objective was to find the PSE using a 2AFC task (see the previous sections on brightness perception and on 2IFC and 2AFC). The experimenter has to choose comparison patches of different luminances but may not know a priori which luminance levels are appropriate. Using the staircase method, the PSE can be found quickly using the following automatic procedure: If the subject judges the comparison stimulus to be lighter than the standard stimulus, the luminance of the comparison stimulus will be decreased on the next trial. If it is judged darker, the luminance will be increased. As the PSE is the point where the subject is undecided about which stimulus is lighter, the luminance of the comparison stimulus will hover around this point where it is equally likely to be decreased or increased (cf. García-Pérez, 1998). Hence, staircases are a variant of the classic method of limits, but by interleaving several staircases it is possible to avoid, for example, the worrisome starting point bias (Cornsweet, 1962; Nachmias and Steinman, 1965).

Staircases are an example of a nonparametric adaptive method because they make no assumption about the underlying psychometric function. If one is willing to assume that the psychometric function comes from a parametric family, such as logistic or normal ogive, then the efficiency of adaptive methods can be improved further through the use of parametric adaptive methods (Gu, Myung, Pitt, & Lu, 2013; Kontsevich & Tyler, 1999; Shen & Richards, 2012; Watson & Pelli, 1983). After each trial a psychometric function is fit to the available data (see the data analysis section), and the stimulus for the next trial is determined such that it will be as informative as possible about the parameters of interest. Parametric adaptive methods vary in the assumptions they make about the psychometric function, the fitting and inference procedures, and the criteria for informativeness.

Blocked Versus Nonblocked Designs

Another aspect of the experimental design to consider is whether the stimuli are blocked. For example, if the experimenter decides to use a specific set of stimulus levels for the comparison stimulus in a 2AFC discrimination experiment, then there will be two obvious possibilities for the design. The standard fixed design in psychophysics is *blocked*: In a block of *N* trials that follow each other the subject sees only *one* comparison stimulus; that is, on each trial in this block the subject is presented with only the standard stimulus and this one comparison stimulus. Hence, after a few trials the subject knows precisely which two stimuli to expect in each 2AFC trial. In contrast, in a *nonblocked* randomized design with a fixed set of stimulus levels, the comparison stimulus is chosen randomly on each trial and the subject does not know which stimulus to expect. Also, staircases that potentially change the stimuli from trial to trial are therefore nonblocked designs. However, adaptive procedures can also be blocked; that is, it is possible to choose the next stimulus level after a block of trials rather than after every trial. In practice, many researchers do not choose the stimulus levels in a blocked design before the experiment but also do not use an automatic adaptive procedure. Instead they choose a new stimulus level by hand after every block until the whole psychometric function is measured in sufficient detail. Although this method is clearly not statistically optimal, it is very robust.

All tasks that we have presented can be used in a blocked or a nonblocked design. However, not all combinations are equally useful. Most obviously, the yes-no task without catch trials is downright silly if the stimulus is always the same in a block and the subject is aware of that. The yes-no task with catch trials, on the other hand, can be problematic in a nonblocked design. For example, for detection experiments standard SDT assumes that the subject knows which stimulus to expect on each trial, adjusting the criterion to optimize the proportion of correct responses. Violating this assumption in a nonblocked yes-no task makes it hard to analyze such data using SDT. The yes-no with reference task is designed to circumvent this problem by reminding the subject of the stimulus on each trial in a nonblocked design. It is therefore odd to use it in a blocked design where the optimal strategy from SDT (assuming the stimulus is known) would prescribe that the reference should be ignored (Stüttgen et al., 2011). In general, it makes a big difference for a signal detection analysis whether the stimuli can vary from trial to trial or it can safely be assumed that they are known—as in blocked designs. Macmillan and Creelman (1991) review the standard tools of SDT that are available for different tasks in blocked and nonblocked designs beyond the simple and well-understood yes-no and 2AFC tasks. Unfortunately, for many interesting and natural tasks, like *m*AFC or the oddity task, a signal detection analysis is complicated irrespective of whether the stimuli are blocked (Luce, 1963; Versfeld, Dai, & Green, 1996), leaving us with analyzing the psychometric functions that do not distinguish between blocked and nonblocked designs.

Best Practice

In order to get objective data, experimentalists often seem to treat human subjects in a psychophysical study like rats in a Skinner box. Although reading the literature may lend this impression, in our experience, treating human subjects like rats will lead to misunderstandings about the task, noisy and inconsistent data, and often no way to diagnose problems with the task or the design. There is an art to running psychophysical studies that is hard to automatize and is rarely ever discussed in textbooks (a noteworthy exception is Green & Swets, 1988, Appendix III). As William James has so aptly observed, psychophysical tasks are often extremely boring. Experimenters are dependent not only on subjects learning the often challenging tasks but also on them concentrating for hours. Hence, every effort has to be made to ensure that subjects are doing their best.[8] In the following, we describe what we do to this end in our labs.

[8] It is as well to note that the experimenter's attempt at getting the observers to "do their best" not only applies to the JND-style experimental tasks at the heart of this chapter, but equally applies to supra-threshold stimuli in sensory judgment settings: Koenderink, van Doorn, Albertazzi, and Wagemans (2015) report that some of their observers learned to "look mindfully" at images and make subjective judgments about pictorial relief. But some observers appear not to have looked mindfully, resulting in somewhat noisy and more difficult to interpret data.

The effort starts by extensively piloting the experiment on yourself and/or colleagues. You can be sure that if you, as a highly motivated subject who is really interested in obtaining these data, find it hard to do the task, a subject with less scientific zeal will quickly give up. This piloting involves getting the flow of the experiment right. Having to wait for the next trial to start when you are long ready or having to press unnecessary or hard-to-reach buttons quickly becomes taxing if you have to do thousands of trials. We often find that subjects like to get into a certain rhythm when doing repetitive trials. If the task is too fast-paced or too slow, this may be annoying for them. It thus often helps if the task is to some degree self-paced and the not self-paced presentation times (or waiting times for masks, etc.) are fixed and not variable (which may clash with demands for response time measurements). We usually take great care to tweak the timing for presentation intervals and interstimulus and intertrial intervals accordingly. On the one hand these times should be as short as possible so that subjects can do as many trials as possible, but on the other hand one often has to exclude effects of afterimages or leave enough time for multiple eye movements or long response latencies. We pay particular attention to possible effects of afterimages, unwanted memory or attention overload, or stimulus artifacts. Based on our own experiences with the task, we also consider potential strategies that subjects might be using. These considerations often lead to modifications of the task that prevent the execution of unwanted strategies. In our experience, carefully piloting the experiment on yourself is an important first step for obtaining clean data.

In addition, one should remember that subjects will need some time to learn the task. This involves understanding the task and practicing to routinely execute aspects of the task that are usually not of interest in psychophysical studies—for example, which buttons to press when. Here it certainly helps if the subject is a trained subject with prior experience in psychophysical studies. However, also the specifics that are of interest need to be learned. In particular, subjects need to learn which aspects of the stimulus they should pay attention to. In a contrast discrimination task we may tell the subject beforehand that in each block there will be two stimuli that differ only in contrast, but it is unlikely that the subject knows what this means until she has seen the difference. And even then there will be improvements due to perceptual learning. Since in most psychophysical studies we are interested in the limiting sensitivity (i.e., the highest performance that is achievable under the most favorable conditions), we usually give subjects ample practice before the actual experiment. Giving feedback is, of course, crucial if the best possible performance is to be achieved. A blocked design is therefore usually the right choice since it will help learning. In fact, if different stimulus levels are intermixed, the subject will have a harder time to learn the cues that are necessary for each stimulus to achieve a high performance (cf. Blackwell, 1952).

Whereas learning improves a subject's performance over time, fatigue usually decreases performance over time. The harder the task, the more worried one may get about the subject's vigilance and wonder about whether time of day, caffeine intake, or general motivation will have an effect. Our policy to deal with these issues is to create the best possible conditions for the subject to perform the task. We tell subjects that we know that the task is tiring and repetitive but that it is important for us to measure their best possible performance. We then instruct subjects to carefully monitor their vigilance, take breaks as needed, and provide coffee and sweets as required. Depending on the task, it may even be acceptable and beneficial to have subjects listen to their favourite music while doing the experiment (Blackwell, 1952). The end of every block, usually not lasting longer than 5 minutes, provides a possibility for a break, and we usually include a few warm-up trials at the beginning of each block. In a typical experiment we do blocks of 55 trials, including the five warm-up trials. In addition, some easy warm-up blocks at the beginning of each session are also advisable. For motivational reasons, we have found it important that the subjects are always aware of how many more blocks they still have to do and how well they are doing. An additional benefit of the experimenter choosing the stimulus level for the next block by hand, as is often done, is that as the experimenter is present throughout the experiment she signals to the subject that the data collection is important enough for her to stick around and that the work of the subject is valued. She also gets a direct impression of whether the subject was cooperative or dozed off during the task. Trained and experienced observers are, however, often trusted to collect the data without any supervision (Green & Swets, 1988, Appendix III).

Independent of the experience of the observers, the importance of careful instructions and debriefing should not be underestimated. Inexperienced subjects are often discouraged by a performance that is considerably less than 90%. Hence, one should explain carefully to them that there are hard and easy blocks and that measuring their performance in the hard blocks is important for the experiment. In fact, the experiment is designed to be hard, because the hard trials are most informative, as we want to measure the subject's sensitivity. Similarly, inexperienced observers are often unsure how to react in forced-choice situations. Hence, they need to understand that not only is it okay to guess, but it is the purpose of the experiment. We usually state explicitly that they should not overthink their responses and that it is best if they follow their first inclination. Otherwise some subjects will resort to cognitive strategies and, e.g., not press one button unless they are absolutely sure of their response or alternate between response options. It is good to have standardized instructions, but the manifold ways of misunderstanding instructions lead us to believe that reacting flexibly to the subject's questions is more important than standardization. Interestingly, we have had the biggest misunderstandings with experienced observers because we thought that they already understood when they did not. After the experiment, we talk with each subject about the experiment; in this way we have sometimes discovered grave misunderstandings or unusual cognitive strategies—observers can be very inventive!—that otherwise would have just led to mysterious data.

Even if subjects do not use unusual cognitive strategies to solve a task, a common problem in psychophysical studies is that we do not know which cognitive processes are really involved in a task and whether there are unknown factors that contaminate the data. This is of particular concern when measuring a subject's sensitivity where the interest is usually in low-level sensory processes and not in memory, attention, or cognitive control. For this reason, we usually prefer to use simple and well-understood tasks, like yes-no or 2IFC, for which SDT is extremely well developed.

But SDT gives little guidance in dealing with task learning, fatigue, or other fluctuations in performance that are usually not of interest in a psychophysical study on perception. Therefore, we do not think a psychophysical experiment should be rushed. Subjects need enough time to learn the task and take breaks as needed. Unless one works in a clinical setting or with children, efficiency and time should be of less concern than obtaining accurate measurements of the whole psychometric function. Typically an experiment will thus take several sessions spread over several days (sometimes weeks or months). Despite the temptation to speed things up, in experiments where it is to be expected that learning or other performance fluctuations are involved, we do not recommend using the commonly used nonblocked adaptive procedures. The nonblocked design makes it harder for subjects to learn, and it also complicates the detection of the effects of learning or fatigue (Wichmann and Hill (2001a) describe standard tests for that purpose). In the tradition of those who value accuracy over time and therefore cannot be bored, Dallenbach (1966, p. 656) reminds users of adaptive methods that “[t]he search for short-hand methods in technology is laudable enough, but it is entirely out of place in science, where new trails are being blazed and attempts are made to reduce and to eliminate all of the errors of observation. Ease and convenience are poor experimental guides.” In our labs, we like to refer to this dictum as Dallenbach’s second commandment of psychophysics: “Thou shalt not be lazy (when gathering data).”

DATA ANALYSIS

Once the data are collected, they need to be analyzed. Luckily, the statistical analysis of the open behavorial response in all the tasks discussed previously is very similar and extremely well developed.[9] Consider a 2IFC or 2AFC JND-style task. If a stimulus condition is easy, the performance of an ideal subject will be close to perfect: In, say, 50 out of 50 trials the subject can identify the interval (2IFC) or position (2AFC) containing the stimulus correctly. If the task is impossibly difficult and the observer is just guessing between the two alternatives, the performance will (ideally) be binomially distributed with probability 0.5; that is, one expects the subject to get—on average—25 out of 50 correct. The observer’s sensitivity is measured by recording the performance for varying difficulty levels (i.e., several levels between chance and perfect performance). The function mapping some physical stimulus property to a response probability is called the *psychometric function*.

Empirically, the psychometric function is virtually never a step function with a hard threshold above which the subject can do the task and otherwise not. Almost always the change of the response probability as a function of stimulus intensity—or whatever stimulus property is changed—is well described by a smooth and monotonic function: a sigmoid function (“S”-shaped function).[10] In psychometric function estimation the standard choice for the sigmoid function is a cumulative distribution function (CDF) such as the cumulative Gaussian, the logistic, or the Weibull.

The most common summary statistic for the full psychometric function is termed the *threshold*; this is the stimulus level for which the response probability takes on a particular value. Common choices for the value are 0.75 for 2IFC and 2AFC, and 0.5 for yes-no

[9] Please note that this section is concerned only with the statistical analysis of behavioral data. Statistical analysis is most often a necessary first step in the analysis of behavioral data; however, it rarely if ever can serve as a substitute for a proper mechanistic model as the ultimate goal.

[10] In extremely rare cases no sigmoid may be a good model for the psychometric function, for example, for the unusually shaped psychometric functions in Henning, Millar, and Hill (2000). In such cases a nonparametric fit may be appropriate as proposed by Zychaluk & Foster (2009).

tasks *without catch trials*[11]: halfway between chance performance and perfect performance in the respective tasks. Calling the stimulus level for the (arbitrarily) chosen response probability the "threshold" is misleading, as there is no hard threshold. But the term *threshold* is universally used throughout experimental psychology and the behavioral neurosciences, and researchers thus have to continually remind themselves that it simply refers to a particular response probability. Finally, while it is convenient to summarize the psychometric function with a single number, it is important not to forget that the subject's behavior is fully described only by the entire psychometric function: Focusing on one threshold level only may lead to the neglect of other changes in the underlying psychometric function, most importantly its width: how quickly the response probability changes with the physical stimulus property. We will return to this issue in a later section.

Given the central importance of the psychometric function for the analysis of virtually all JND-style data (i.e., all examples and experimental designs covered in the preceding sections), it is not surprising that many researchers studied the estimation of the psychometric function—for example, Foster and Bischof (1987, 1991, 1997); Fründ, Haenel, and Wichmann (2011); Knoblauch and Maloney (2012); Kuss, Jäkel, and Wichmann (2005); Maloney (1990); O'Regan and Humbert (1989); Schütt, Harmeling, Macke, and Wichmann (2016); Treutwein and Strasburger (1999); Wichmann and Hill (2001a, 2001b). In the following sections we introduce the most common models for the psychometric function, and argue that in general Bayesian inference for the *beta-binomial mixture model* is the method of choice for analyzing data from JND-style experiments.

The Psychometric Function I: The Binomial Model

The standard assumption when fitting a psychometric function to data from a JND-style experiment is that at a given stimulus intensity—or difficulty level—the responses of an observer are *independent and identically distributed* (i.i.d.) Bernoulli trials (i.e., *independent* coin flips with a particular—*constant*—success probability p). As a consequence, the number of successes in a set of trials at a given stimulus level is binomially distributed.[12]

As mentioned before, the psychometric function ψ is usually modeled as a—scaled—sigmoid function S. S is a strictly monotonic function from the stimulus level x, $x \in \mathbb{R}$, to the unit interval $[0, 1]$. The shape of the sigmoid function is determined by the parameters m, the threshold, and w, the width: $S(x; m, w)$. The threshold m is typically taken to be the level at which the *unscaled* sigmoid function has value 0.5: $S(x = m, m, w) = 0.5$. The width w is the difference between the x-levels at which the function reaches 0.05 and 0.95; $x_1 = S^{-1}(0.05, m, w)$, $x_2 = S^{-1}(0.95, m, w)$, and $w = |x_2 - x_1|$.

Perhaps more commonly, the sigmoid S is parameterized by a shift parameter α

[11] Above we strongly argued against the classical version of yes-no without catch trials; in our yes-no CSF example above half the trials were catch trials, thus chance performance was 0.5, as for 2IFC and 2AFC, and consequently the threshold in this case would typically chosen to be 0.75—as for 2IFC and 2AFC. The thought experiment was still not a 2IFC or 2AFC experiment, however, as there was always only a single interval in the experiment.

[12] Statistically it does not matter whether the trials where blocked during a constant stimulus experiment, or randomly interleaved, or collected using an adaptive procedure. The critical assumption is that the data are a collection of i.i.d. Bernoulli trials.

and parameter β, related to the slope of S.[13] García-Pérez (1998) proposed the width parameterization, however, as it has several advantages: First, the threshold and the width have an easily accessible meaning measured in the same units—particularly the width is easier to interpret and understand than the slope parameter β being directly, or, depending on the particular sigmoid S chosen, even inversely, proportional to the slope of S. Second, m and w have the same meaning for all sigmoid families, fostering comparisons between data fitted with any sigmoid. If one uses Bayesian methods to estimate the psychometric function, as we advocate here, there are two additional advantages: Specifying an appropriate prior is much more difficult for the parameter β than for the width w (this point is clearly made by Kuss et al. (2005) on pp. 484–485 and shown in their Figure 3). Finally, because m and w have the same meaning independent of the choice of sigmoid S, the same priors can be specified and used whatever choice is made for the sigmoid. Thus we adopt the threshold m and width w parameterization throughout this chapter.

Depending on the experimental design, the psychometric function ψ needs to be scaled since performance in the absence of a signal is not necessarily zero. In forced-choice paradigms, for example, the lower bound of performance will be the reciprocal of the number of alternatives per trial. Thus with the chosen sigmoid family S, the psychometric function $\psi(x; m, w)$ is defined as:

$$\psi(x; m, w) = \gamma + (1 - \gamma)S(x; m, w) \qquad (1)$$

In this most basic form, γ denotes the probability of a correct answer in the absence of a stimulus, often termed the *guess rate*, and here it is not a parameter but a constant: It scales the psychometric function to be between γ and 1.0. Figure 7.2 illustrates the psychometric function and its parameters. Note that Figure 7.2 contains an additional parameter, λ, which will be introduced in the next section on the binomial mixture model.

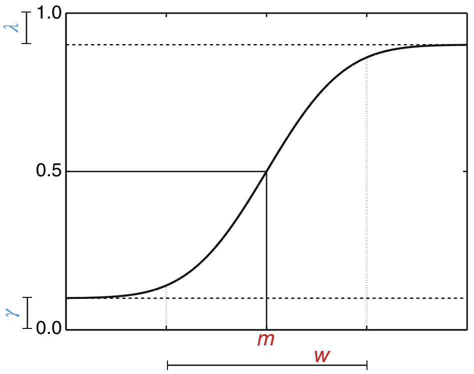

Figure 7.2 Example psychometric function with the definition of the parameters: threshold m, the stimulus level at which the unscaled psychometric function reaches 0.5; width w, the difference between the stimulus levels for which the unscaled function reaches 0.05 and 0.95, respectively; the guess rate γ, the difference between the lower asymptote and 0.0. In forced-choice paradigms, γ is a constant, not a parameter, and is fixed at $1/N$ if there are N stimulus alternatives to choose from; the lapse rate λ, the difference between the upper asymptote and 1.0. In the standard binomial model λ is zero and thus left out in Equation (1). λ is a parameter in the binomial mixture model and the beta-binomial mixture model (later section).

In this most simple formulation the psychometric function is an example of the thoroughly understood generalized linear model (GLM) (McCullagh & Nelder, 1989). Estimation of the parameters and the confidence intervals is straightforward, and software to perform this is available in many programming languages. Unfortunately, as we will see, it is inadequate to model data from psychophysical experiments properly.

[13] If the cumulative Gaussian is taken as the sigmoid, then μ and σ are commonly employed symbols for shift and slope, respectively.

The Psychometric Function II: The Binomial Mixture Model

The problem is that observers are rarely perfect binomial response machines: They sometimes exhibit lapses of attention, they may blink during a visual stimulus presentation, they forget which button they wanted to press, or they simply press the wrong button by accident. Thus they may not respond at 100% even for very large stimulus intensities. Wichmann and Hill (2001a) showed that maximum-likelihood estimates of threshold m—and in particular the width w—may be severely biased if one fits the standard binomial psychometric function to data contaminated by stimulus-independent lapses. Their solution was to introduce an additional parameter, λ, to allow the upper asymptote to vary. Then the psychometric function $\psi(x; m, w, \lambda, \gamma)$ is defined as the scaled sigmoid function:

$$\psi(x; m, w, \lambda, \gamma) = \gamma + (1 - \lambda - \gamma)S(x, m, w) \tag{2}$$

In this parameterization λ denotes the probability of an incorrect answer at infinitely high stimulus levels. For an ideal observer λ would always be zero, but in practice human (or animal) observers are rarely ideal.

The interpretation of a scaled sigmoid function following Wichmann and Hill (2001a) is mathematically equivalent to a *binomial mixture model* (Kuss et al., 2005). In the binomial mixture model the proportion of correct answers of the observer results from two independent Bernoulli processes: First, with a probability π_l the observer guesses independently of stimulus intensity, and has a probability π_c of guessing correctly. The probability π_c depends, as described earlier, on the experimental design, and is typically $\pi_c = 1/N$ in cases where the y-axis of the psychometric function is the probability for a correct response and N denotes the number of response alternatives.[14] Second, with a probability $1 - \pi_l$ observers attempt to solve the task the best they can; in this case the probability for a correct answer is $\psi(x, m, w, \pi_c) = \pi_c + (1 - \pi_c)S(x, m, w)$. Together this results is a Bernoulli variable again, with a probability of success as a function of the stimulus intensity x given by

$$\begin{aligned}
\psi(x; m, w, \pi_l, \pi_c)& \\
&= (1 - \pi_l)[\pi_c + (1 - \pi_c)S(x, m, w)] + \pi_l\pi_c \\
&= (1 - \pi_l - \pi_c + \pi_l\pi_c)S(x, m, w) \\
&\quad + (1 - \pi_l)\pi_c + \pi_l\pi_c \\
&= (1 - \pi_l - \pi_c + \pi_l\pi_c)S(x, m, w) + \pi_c \\
&= \pi_c + (1 - \pi_c - \pi_l(1 - \pi_c))S(x, m, w) \quad (3)
\end{aligned}$$

If, in the last line, we substitute $\gamma = \pi_c$ and $\lambda = \pi_l(1 - \pi_c)$, we obtain Equation (2), demonstrating the equivalence of the binomial mixture model to the scaled asymptote formulation.

We prefer to think about the psychometric function as a binomial mixture model, as this makes the underlying data generation process more explicit than to think of it only in terms of an upper asymptote less than one. Note that the lapsing proportion of the binomial mixture model is $\pi_l = \lambda/(1 - \pi_c)$. Thus if one estimates a psychometric function in 2AFC ($\pi_c = 0.5$) with a small lapsing rate, say $\lambda = 0.04$, this implies that the observer made stimulus independent responses on nearly 10% of all trials ($\pi_l = 2\lambda = 0.08$).

Unfortunately, standard GLM software implementations are typically not designed to deal with mixture models, and thus cannot be used to estimate the psychometric functions with varying asymptotes. In fact, finding the best parameter vector $\theta = \{m, w, \lambda, \gamma\}$ is a nonconvex optimization problem. Luck-

[14] In a yes-no task without catch trials, the y-axis is the proportion of "yes" responses, and for an unbiased subject we expect π_c to be near 0.

ily, special purpose software performing psychometric function estimation is freely available—for example, the Bayesian estimation package *psignifit 4* by Schütt et al. (2016).[15]

Violations of the i.i.d. Assumption

The binomial mixture model is a more realistic model of decision making in JND-style experiments than the binomial model. Frequently, however, the binomial mixture model itself is still too much of an idealization: The binomial mixture model, like the binomial model, assumes data to be *independent and identically distributed* (i.i.d.); that is, the model assumes the response of the observer to be solely determined by the external stimulus on the current trial, and not by the history (sequence) of previous stimuli, responses, or internal state fluctuations. Unfortunately, a large number of studies have demonstrated the varying degree to which data in psychophysical experiments exhibit such *serial dependencies* (i.e., violations of the i.i.d. assumption) (Green, 1964; Lages & Treisman, 1998; Maloney, Dal Martello, Sahm, & Spillman, 2005; Senders & Sowards, 1952; Treisman & Williams, 1984; Verplanck, Collier, & Cotton, 1952; Wagenmakers, Farrell, & Ratcliff, 2004).

One way to deal with serial dependencies is to explicitly include them as additional covariates in one's model (Schönfelder & Wichmann, 2013), or to find a statistical model to explicitly describe the structure of the serial dependencies (Fründ, Wichmann, & Macke, 2014; Gökaydin, Navarro, Ma-Wyatt, & Perfors, 2016). In addition, Chapter 2 in this volume describes how to use Bayesian statistics to derive the psychometric function for contaminated data using latent mixture models. With an explicit model one could, for example, attempt to correct for them.

However, in the context of psychometric function estimation, the exact statistical nature of the serial dependencies is frequently of little interest. Serial dependencies are a nuisance, a contamination of one's data, rather than a parameter of interest worth modeling explicitly. What one requires are accurate estimates of the parameters of the psychometric function, as well as confidence intervals with appropriate coverage, *even if the data exhibit serial dependencies*. Schütt et al. (2016) showed that the use of the beta-binomial model makes it possible to determine accurate credible intervals even in such cases.

The Psychometric Function III: The Beta-Binomial Mixture Model

The standard binomial as well as the binomial mixture models assume each trial to be a Bernoulli trial with a fixed success probability independent of all other trials. As discussed before it is well known that observers show fluctuations in performance due to fatigue or changes in their attentional state on longer timescales, and that trials in psychophysics are not independent of each other on short timescales, either. Typically, fluctuations in performance and serial dependencies result in data with variances larger than the variance of the binomial distribution: Data are *overdispersed*. Thus both the standard binomial model as well as the binomial mixture model provide only a *lower bound* for the variance of the actual variance inherent in behavioral data. Hence the confidence intervals—credible intervals in Bayesian analyses—derived from these

[15] The software can be obtained from https://github.com/wichmann-lab/psignifit

models might be too narrow, as shown by Fründ et al. (2011).

Schütt et al. (2016) developed methods for full Bayesian inference in a beta-binomial mixture model (Prentice, 1986), in which overdispersion is treated as an additional parameter. The beta-binomial model assumes that the success probability p per block at a constant intensity x is itself a beta-distributed random variable with mean $p = \psi(x)$. Thus the success probability is not fixed at $\psi(x)$ as in the standard binomial and binomial mixture models but is drawn randomly once for each block. The variance of the success probability is scaled by a new scale parameter η—ranging from 0 to 1—such that the variance of the success probability, $\mathrm{VAR}_p(x;\eta)$, equals

$$\mathrm{VAR}_p(x;\eta) = \eta^2\psi(x)(1-\psi(x)). \qquad (4)$$

For the beta-binomial model with scale parameter η the mean percent correct remains $p = \psi(x)$ as in the standard binomial and binomial mixture models, but the variance of percent correct, $\sigma^2_{\beta\text{-bin.}}(x;\eta)$, at a *fixed* stimulus level becomes

$$\sigma^2_{\beta\text{-bin.}}(x;\eta) = \left(\eta^2 + \frac{1-\eta^2}{N}\right)\psi(x)(1-\psi(x)) \qquad (5)$$

for a block of N trials.

For $\eta = 0$ the beta-binomial reduces to the standard binomial model: The variance of the success probability reduces to zero (Equation (4)), and the variance reaches its lower limit given by the variance of a binomial, $\frac{1}{N}\psi(x)(1-\psi(x))$ (Equation (5)).

For $\eta = 1$ both the variance of the success probability as well as the variance around percent correct reach their respective maxima: $\mathrm{VAR}_p(x) = \psi(x)(1-\psi(x))$ for the success probability (Equation (4)), and the variance of percent correct becomes $\sigma^2_{\beta\text{-bin.}}(x) = \psi(x)(1-\psi(x))$ (Equation (5)). Note that the variance of percent correct becomes independent of N. Thus in this most severe form of overdispersion, increasing the number of trials, N, does not lead to any reduction in the uncertainty about the location of the mean in the beta-binomial model. Expressed in a different way, the variance of percent correct according to the beta-binomial model is $1 + (N-1)\eta^2$ times the one of the binomial distribution: The beta-binomial distribution is overdispersed by this factor.

For $0 < \eta \leq 1$ the beta-binomial model becomes progressively more overdispersed. Note that the factor of overdispersion depends on the number of trials per block, as η scales the standard deviation of a distribution that is drawn from once per block. Thus any interpretation of η depends on the number of trials per block.[16] Please note that the likelihood for the beta-binomial model can still be calculated directly for each observation given the (now) five parameters: $\boldsymbol{\theta} = (m, w, \lambda, \gamma, \eta)$.

Schütt et al. (2016) used the beta-binomial model to provide essentially unbiased estimates of the parameters of the psychometric function, as well as credible intervals with reasonable coverage for binomially distributed as well as overdispersed data (employing Bayesian inference; see next section). Figure 7.3 shows some of the key results of their simulations for a moderately overdispersed observer (beta-binomial with $\eta = 0.2$). Threshold estimates are unbiased (panel A), and width estimates are unbiased, too (panel B). The coverages of the credible intervals for threshold and width are in the range of 92% to 97%; see panels C and D.

Clearly, the beta-binomial model implies a specific form of overdispersion, namely that the underlying performance level is constant within each block of data and changes randomly from block to block. For data obtained

[16] This is an important issue, discussed and explored at length by Schütt et al. (2016).

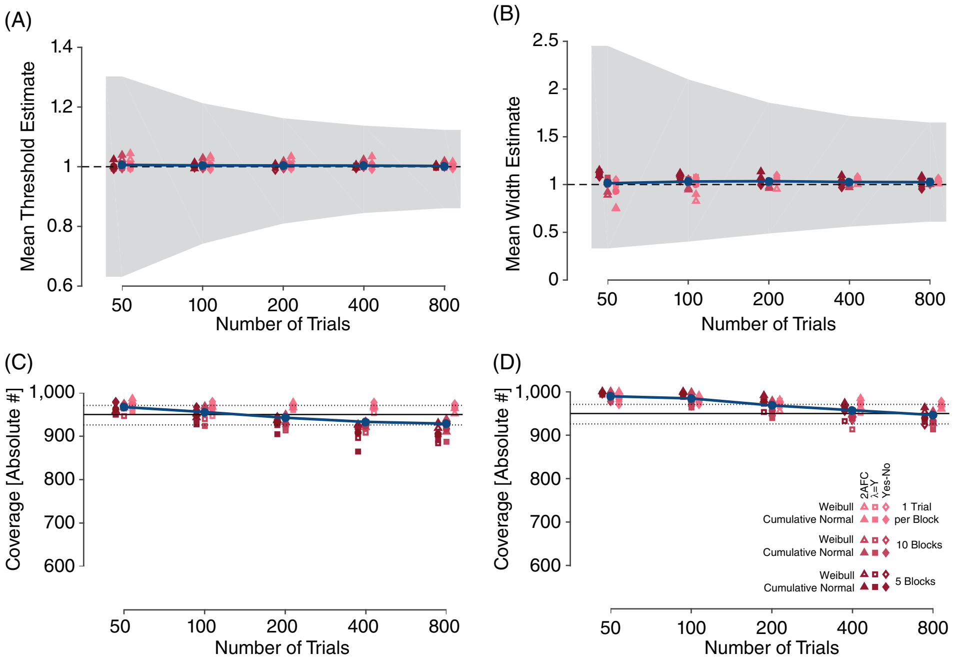

Figure 7.3 Simulation results for the moderately overdispersed observer with $\eta = 0.2$ for linearly spaced constant stimulus designs. Different symbols and colors correspond to different experimental designs, how many trials were simulated in a single block at each stimulus level, and which sigmoid was chosen. (A) Average MAP estimates against the number of trials. The thick line marks the grand average, the colored symbols individual conditions computed from 1,000 simulations each. The dashed line marks the true parameter value of 1.0 and the gray shade the average 95% credible interval over all conditions. (B) MAP estimates of the width, plotting conventions as in A. (C) Coverage of the 95% credible intervals for the threshold. The continuous black line marks the nominal value of 950 of the 1,000 simulations, the dashed lines mark the interval [926,971], which would contain the measured coverage in 99.9% of cases if the true one was exactly 95%. (D) Coverage of the credible intervals for the width. A perfectly unbiased estimate would lie at exactly 1 in A and B, and perfect credible interval size would produce a coverage of 950 in all conditions in C and D.

Source: From Schütt, Harmeling, Macke, and Wichmann (2016). Reproduced with permission of Elsevier.

in a traditional blocked design—that is, stimuli are grouped into blocks of the same magnitude rather than being randomized (see previous section on blocked versus nonblocked designs)—this is a reasonable assumption. If the data are not collected in blocks, however, the situation is not as straightforward, but it is not irredeemable: we refer the interested reader to sections 3.4 (p. 115) and 4.2 (pp. 117–118) in Schütt et al. (2016).

In addition, Schütt et al. (2016) showed that the beta-binomial model can still be used to model overdispersion in realistic real-world scenarios other than truly beta-binomial overdispersion: First, they modeled a *fluctuating* observer, whose threshold m varied over time according to an autoregression model resulting in slow fluctuations of the threshold. This could, for example, be caused by drifts in attention, concentration, or sleepiness. Second, they

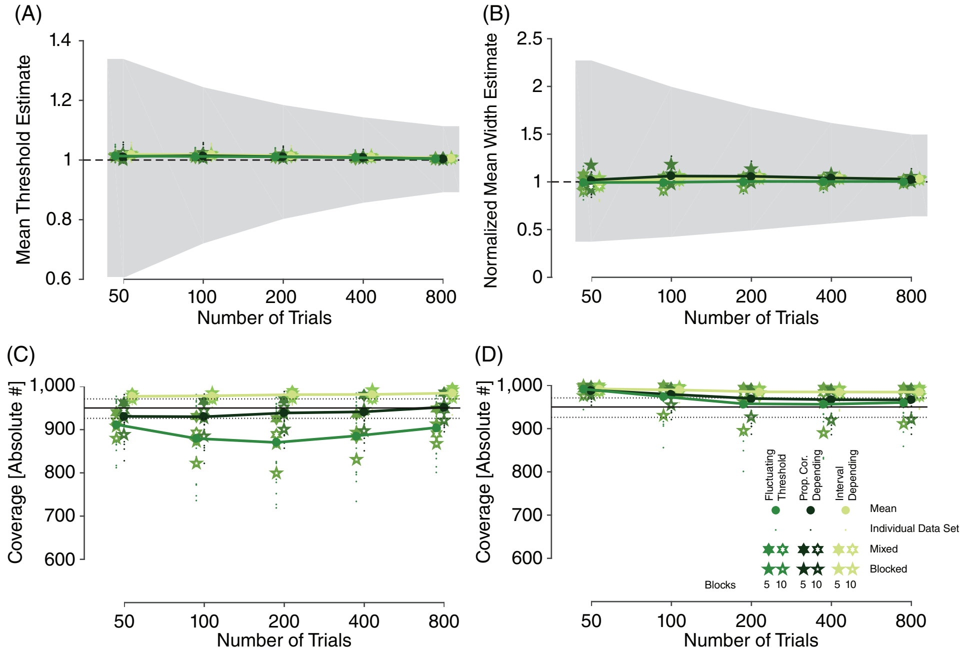

Figure 7.4 Results from three nonbinomial observers differentiated by shading: one observer whose threshold fluctuates over time, one whose probability correct depends on the outcome of the previous trial, and one whose bias depends on the previous trial. Plotting conventions as in Figure 7.3, but with different symbol shapes: pentagrams and hexagrams show whether stimuli were blocked or not ("Mixed"), respectively; small dots indicate worst case outliers; circles show the means across all conditions.

SOURCE: From Schütt, Harmeling, Macke, and Wichmann (2016). Reproduced with permission of Elsevier.

modeled a *sequential dependent* observer, whose threshold was lower after correct trials and higher after incorrect trials—that is, whose performance depended on the previous trial, introducing positive correlations for the correctness of trials and thus overdispersion (Prentice, 1986). Third, they modeled an *interval biased* observer, who preferred the interval that was the correct one in the previous trial.[17]

The key results from Schütt et al. (2016) for the three i.i.d.–violating observers are reproduced in Figure 7.4. As before, threshold estimates are unbiased (panel A), and width estimates are essentially unbiased (panel B). The coverages of the credible intervals for the threshold are in the range of 75% to 98% (panel C); however, the fluctuating observer is covered worse than the other observers. Note the interactions between the number of trials and blocks for the sequential dependent observers. The coverage for the width is reasonable for all observers (panel D).

Although clearly not perfect—the coverage is not always at 95% independent of the exact type of i.i.d. violation—the beta-binomial mixture model is always superior to the standard binomial model and the binomial mixture model (results not shown; see Schütt et al., 2016, for details).

[17] Note that only the first two i.i.d.-violating observers have a higher variance in their responses at a given *x*-level.

Earlier publications on psychometric functions emphasized the necessity to check the goodness of fit of the estimated psychometric function—that is, how well the model explains the data (Wichmann & Hill, 2001a). The aim of these checks is to detect when a model does not fit the data. In modeling of empirical data in general this is obviously sound advice: If a model does not fit the data, it is very often of little use. However, in the context of psychometric function fitting this recommendation may be too strict, or at least of little practical value, as often there is no alternative to fitting a psychometric function, and the goodness-of-fit rejection leaves researchers without a viable option to proceed. With the beta-binomial model, however, researchers can draw valid conclusions from overdispersed data, providing them with hitherto missing viable option. The overdispersion parameter of the beta-binomial model measures a very similar property to classical measures of goodness of fit, namely the additional variance around the function. But in the case of the beta-binomial model the overdispersion parameter can be used to increase the uncertainty until the data are consistent with the fitted model, instead of rejecting the binomial or binomial mixture model.

In summary, the beta-binomial mixture model works accurately and with reasonable credible interval coverage not only for the beta-binomial type of overdispersion (Figure 7.3), but also for at least three types of sequentially dependent observers (Figure 7.4).

The Psychometric Function IV: Bayesian Inference

Unless one has collected infinitely many trials per psychometric function, the parameters of the psychometric function are not fully constrained by the data and there remains uncertainty regarding the estimated parameters. To draw conclusions about thresholds and widths from different experimental conditions, it is therefore essential that the remaining uncertainty is quantified. Typically, the remaining uncertainty is expressed in the form of confidence intervals (frequentist statistics) or credible intervals (Bayesian statistics) around the point estimates.[18] Unfortunately, a reliable and accurate characterization of this uncertainty is harder to obtain than the estimates themselves.[19]

We advocate the use of Bayesian inference methods[20] to quantify the remaining uncertainty (i.e., for estimating the posterior distribution of the parameters of the psychometric function). Bayesian inference methods were previously shown to produce credible intervals with better coverage than those obtained from the bootstrap (Fründ et al., 2011; Kuss et al., 2005),[21] and Schütt et al. (2016) showed Bayesian point estimates and credible intervals to be accurate even for overdispersed data (see preceding section).

The estimation of psychometric functions particularly profits from Bayesian inference because it provides a principled way of integrating out the uncertainty over parameters that are not well constrained by the data. For example, data sets with few blocks hardly

[18] For a recent discussion on the superiority of credible intervals over confidence intervals, see Morey, Hoekstra, Rouder, Lee, and Wagenmakers (2015) and Morey, Hoekstra, Rouder, and Wagenmakers (2015).

[19] Much of conventional statistics relies on the asymptotic behavior of estimators and probability distributions; that is, it relies on ultimately infinitely large data sets. However, Wichmann and Hill (2001a, 2001b) showed that, for the typical size of psychophysical data sets, methods based on asymptotic theory are not always reliable. Thus it is not generally advisable to derive confidence intervals from standard, asymptotic methods.

[20] For an introduction to Bayesian statistics in psychology, see, e.g., Kruschke (2014).

[21] Note that this implies that for psychometric functions Bayesian statistics yields better intervals according to a *frequentist* evaluation criterion: coverage.

constrain η, and data sets with little data in the asymptotic range hardly constrain λ. Fitting a single value for these parameters—as in frequentist statistics—does not take the uncertainty about other parameters into account, yielding unstable estimates for *all* parameters. In contrast, Bayesian inference integrates the results for the different possible values for the unconstrained parameters, yielding sensible estimates and credible intervals for the other parameters.

Given the beta-binomial mixture model described in the preceding section, we can calculate the likelihood $L(\boldsymbol{\theta} \mid \text{data}) = P(data \mid \boldsymbol{\theta})$. Together with the prior $P(\boldsymbol{\theta})$, the five-dimensional posterior $P(\boldsymbol{\theta}|\text{data})$ over all parameters $\boldsymbol{\theta} = (m, w, \lambda, \gamma, \eta)$ can be computed as:

$$P(\boldsymbol{\theta}|\text{data}) = \frac{P(\boldsymbol{\theta})\ L(\boldsymbol{\theta} \mid \text{data})}{\int_{\Omega} P(\boldsymbol{\theta})\ L(\boldsymbol{\theta} \mid \text{data})\ d\boldsymbol{\theta}}. \quad (6)$$

The specification of the prior $P(\boldsymbol{\theta})$ is often regarded as a thorny issue, particularly by frequentist statisticians (see, e.g., the discussion in Efron, 2013). Schütt et al. (2016) show how Bayesian inference can be accurate even according to the frequentist coverage criterion using carefully predetermined default priors extracted from the x-values of a data set to be analyzed. We will abstain from any further discussion on this debate here and instead refer the interested reader to the sections on the prior and its specification in Bayesian inference of the psychometric function provided by Kuss et al. (2005) and Schütt et al. (2016).

The right-hand side of Equation 6 cannot be evaluated analytically, and one instead has to rely on numerical methods to obtain the posterior distribution of the parameters given the data. Typically Markov chain Monte Carlo (MCMC) methods are used to generate samples from the posterior distribution over parameters (Andrieu, de Freitas, Doucet, & Jordan. 2003; Gilks, Richardson, & Spiegelhalter, 1996). MCMC is a standard method in Bayesian inference in general, and, in principle, it allows Bayesian inference to be performed on many statistical problems. Unfortunately MCMC requires considerable statistical expertise from the user to fine-tune the proposal distribution and the sampling step size, and especially to detect when the sampling fails. MCMC methods thus rarely work automatically with little or no user intervention the way analytical methods and the bootstrap do. However, Schütt et al. (2016) introduced a software implementation—*psignifit 4*—performing numerical integration of the posterior within automatically determined bounds, avoiding the use of MCMC methods and making their methods and software particularly user friendly.

Figure 7.5 shows an example of the result of a Bayesian inference to estimate the psychometric function, emphasizing that the result of the analysis is not only a single parameter value but a distribution over parameter values. For reasons of clarity, the figure restricts the presentation to the threshold and width parameters; the method of course returns a five-dimensional posterior, extending over the other three parameters not shown.

A final advantage of Bayesian inference we would like to draw attention to is that it allows the posterior distribution to be used in further statistical analyses. We are aware that such inference is often not performed routinely, and much statistical inference in psychology is still performed using hypothesis tests, although they are at best a weak solution (e.g., Nuzzo, 2014; Wagenmakers, 2007). The posterior distribution returned by Bayesian inference, on the other hand, could be used by the user, for example, for further Bayesian inference on a hierarchical model, predicting the parameters of the psychometric functions across conditions or observers.

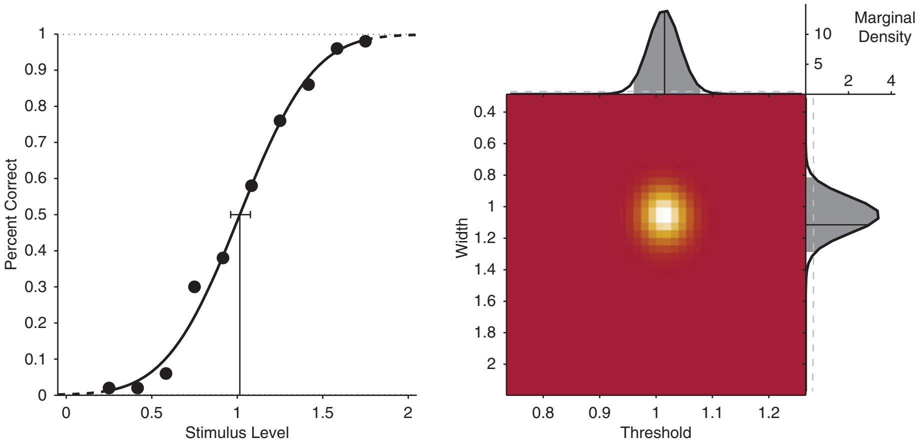

Figure 7.5 Bayesian analysis for a psychometric function based on 500 trials. The data with the maximum a posteriori (MAP) estimate of the function shape are plotted on the left. The marginal posterior for the threshold and the width is displayed in the right panel. Attached to it are marginal distributions for the single parameters. Priors for threshold and width are shown as dashed gray lines; the gray shading filling most of the marginal distributions corresponds to the extracted 95% credible intervals.

SOURCE: From Schütt, Harmeling, Macke, and Wichmann (2016). Reproduced with permission of Elsevier.

Importance of the Width of the Psychometric Function

In the section "Shape of the Psychophysical Function" Green (1960) stressed the importance of obtaining the entire psychometric function for theorizing and modeling in psychology:

> Obviously, it is extremely important for the model to specify the exact transformation of the physical stimulus which is used as the abscissa of the psychophysical function; without such specification the theory is incomplete.

Unfortunately, Green's recommendation has not been followed widely, as witnessed by the popularity of standard adaptive procedures attempting to estimate the threshold only (see section on experimental design). However, as Green said, neglecting the slope—or width, as we prefer—may be detrimental, and we illustrate the problem using sinusoidal contrast discrimination as a concrete example. Nachmias and Sansbury (1974) showed that sinusoidal contrast discrimination is sometimes easier than sinusoidal contrast detection, violating Weber's law.[22] Because of this initial decrease of the discrimination threshold Δx with increasing comparison (or pedestal) contrast x, the threshold-versus-contrast (TvC) function relating them is frequently termed the *dipper function*. Nachmias and Sansbury already noted that the slope of the psychometric function for detection was steeper than that for discrimination, particularly around the contrasts where discrimination was very sensitive. This pattern of slope change with increasing pedestal contrast was subsequently confirmed a number of times (Burton, 1981; Foley & Legge, 1981; Wichmann, 1999). Figure 7.6 shows a combined contour and density plot of an accurate descriptive model

[22] Detection is viewed as contrast discrimination with a comparison contrast of zero.

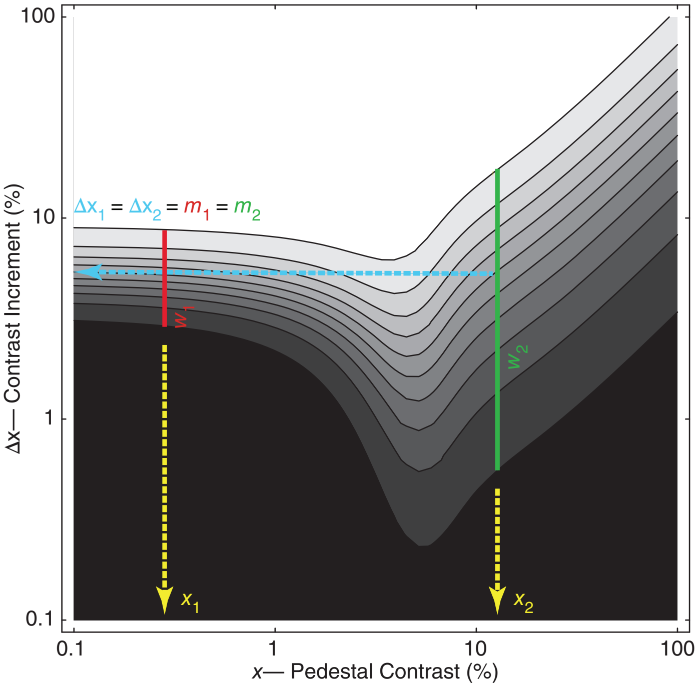

Figure 7.6 Combined contour and density plot of an accurate descriptive model of the dipper function (2IFC); the contrast increment Δx required for discrimination on the y-axis is plotted against the pedestal contrast x on the x-axis; white corresponds to better than 95% correct discrimination, black to below 55% correct discrimination, and intermediate performance is indicated by gray level. Two pedestal contrast x_1 and x_2 are indicated by yellow vertical arrows. The vertical lines w_1 and w_2 mark the very different widths of the psychometric functions at x_1 and x_2, respectively. The *threshold*, defined as 75% correct, however, is identical: $\Delta x_1 = \Delta x_2 = m_1 = m_2$ (horizontal arrow).

of the dipper function from Wichmann (1999). Plotted is the contrast increment Δx required for discrimination on the y-axis against the pedestal contrast x on the x-axis; white corresponds to better than 95% correct discriminations, black to discrimination performance below 55%; intermediate performance is indicated by gray level.[23] In the figure, pedestal contrast x_1 and x_2 are indicated by vertical arrows. The vertical lines w_1 and w_2 mark the widths of the psychometric functions one would obtain from a contrast discrimination experiment at x_1 and x_2. Note that the *threshold*, defined as 75% correct, is identical: $\Delta x_1 = \Delta x_2 = m_1 = m_2$ (horizontal arrow).

Clearly, discrimination performance is *not* the same at x_1 and x_2, despite the thresholds being the same. The width w_2 is more than three times that of w_1—compared to the lower pedestal contrast x_1, at the higher pedestal contrast x_2 observers require *more* contrast to be 90% correct, but much *less* to be 60% correct. A corollary of the slope change is that the shape of the TvC function is not invariant with the arbitrarily chosen threshold: At 90% correct there is only a modest dip visible, and maximal facilitation

[23] The model shown fits the data from a 2IFC experiment of one observer (GBH) at very short presentation times (t = 20 msec) at a spatial frequency of 8.37 cpd.

occurs for pedestals smaller than the detection threshold, $x^{(90)} < \Delta x^{(90)}$. For 60% correct, on the other hand, maximal facilitation (the dip) is very pronounced—up to a log-unit—and it occurs at pedestal contrasts much larger than the corresponding detection threshold $x^{(60)} > \Delta x^{(60)}$. Frequently made assertions that maximal facilitation occurs for pedestals equaling the detection threshold, and that the magnitude of the facilitation is around 2 to 4, are thus erroneous: The statement happens to be true for the TvC function corresponding to 75% correct, but it is certainly true for neither higher nor lower definitions of threshold—a clear example where reliance on single thresholds obtained from adaptive procedures has resulted in spurious "knowledge." Furthermore, Wichmann (1999) showed that competing models of contrast discrimination are almost always capable of fitting any single TvC shape, but that only a few models are able to predict the *change* in the shape of the TvC functions (i.e., the change in the width of the psychometric functions with pedestal contrast). A similar argument was recently made by Goris, Putzeys, Wagemans, and Wichmann (2013) in the context of deciding between competing models attempting to explain uncertainty effects in spatial frequency detection.

Bias and Sensitivity Differences in 2IFC and 2AFC

According to SDT, in 2IFC (2AFC) an ideal observer takes the difference of the sensory evidence—or the internal representations—of the presented stimuli, and chooses the interval (location) with more evidence. It is precisely the differencing rule that makes 2IFC (2AFC) so desirable from an SDT point of view: The decision is criterion free, and proportion correct exactly corresponds to the area under the ROC curve if one were to perform a yes-no experiment with the same stimuli. Implicit in the ideal SDT conception of decision making in 2IFC and 2AFC is that temporal order (2IFC) or spatial location (2AFC) is irrelevant: Observers perform equally in the temporal intervals or spatial locations. As a result, the data are almost always aggregated across intervals for analysis, and very rarely analyzed separately. However, this idealization of the decision process in SDT neglects potential attentional and memory influences, the role of adaptation, and so on; we describe the many psychological factors that may influence animate decision makers in the best practice section. It is not inconceivable that constraints in the temporal deployment of attention, for example, could cause sensitivity differences in the two intervals; the same is true for adaptation or gain-control and, in fact, for most of the factors we describe. As a result, it may be that for some tasks, or some observers, the two intervals in 2IFC or the two locations in 2AFC may not be equivalent.

Figure 7.7 columns A to C show unpublished data for three observers from a 2IFC experiment measuring the so-called uncertainty effect in sinusoidal contrast detection (Schütt & Wichmann, 2014).[24] The top row across columns shows estimated psychometric functions for the entire, aggregated data set (black), or separate for the first or second interval. Rows 2 to 5 show the marginal densities of four estimated parameters, threshold m, width w, lapse rate λ, and guessing rate γ, respectively. Column A shows data from a highly experienced psychophysical observer. The data are ideal in the sense that they are not afflicted by any significant temporal order effects in any of the estimated parameters. Column B shows data from a

[24] The *psignifit 4* toolbox by Schütt et al. (2016) contains a function to assess temporal (or spatial) order effects in 2IFC or 2AFC data as shown in Figure 7.7.

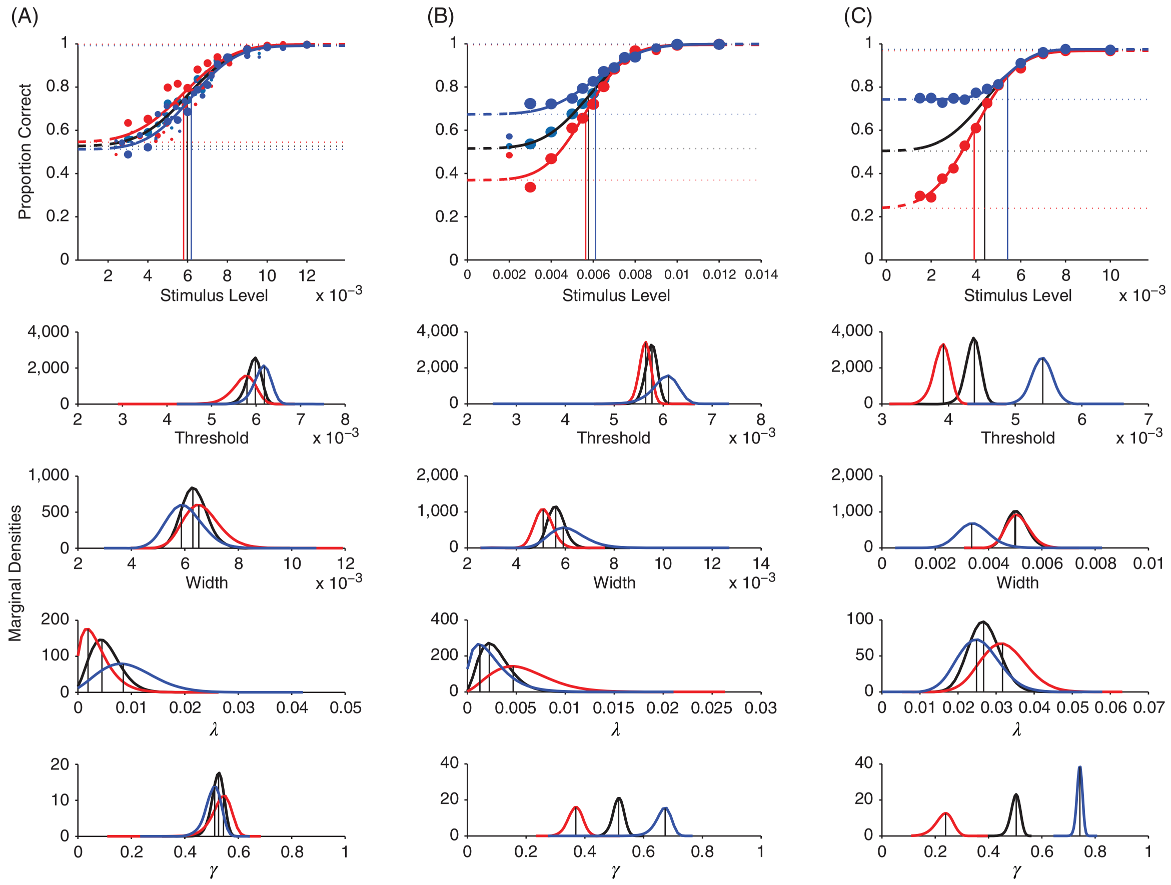

Figure 7.7 Psychometric functions estimated for the entire, aggregated data set (black), or separate for the first or second interval. Across columns A to C the top row shows the estimated psychometric functions; rows 2 to 5 show the marginal densities of four estimated parameters, threshold m, width w, lapse rate λ, and guessing rate γ, respectively. (A) Ideal data without any significant temporal order effects in any of the estimated parameters. (B) Comparatively innocuous or benign interval or finger bias; that is, the observer is biased toward the second interval, as shown by the difference in γ. However, neither threshold nor width differs between the intervals. (C) Example of potentially problematic data set for which not only γ but also the threshold is significantly different between the first and second intervals. (Note that in this data set the observer is actually more sensitive to signals in the first, *nonpreferred*, interval.)

less experienced observer with significant "interval" or "finger" bias (cf. an alternative analysis of interval bias in Jäkel & Wichmann, 2006). As can be seen from rows 2 to 5 for column B, neither threshold nor width differs between the intervals. The observer is biased toward the second interval, as shown by the difference in γ; this is not an uncommon finding for right-handed observers, pressing the button indicating the first interval using their left hand, and the second interval using their right hand. We deem this to be a form of—comparatively—innocuous or benign bias because, in balanced designs such as 2IFC or 2AFC, it effectively cancels around the commonly used threshold halfway up the psychometric function.[25] If required, it can be modeled simply by an additive bias term π_b—with $0 \leq \pi_b \leq \frac{1}{2} - \lambda$—to Equation 2, positively added to the preferred interval (location), and subtracted from the nonpreferred interval (location):

$$\psi(x; m, w, \lambda, \pi_b) = \begin{cases} \frac{1}{2} + \pi_b + \left(\frac{1}{2} - \lambda - \pi_b\right) S(x; m, w) \\ \quad \text{preferred interval} \\ \frac{1}{2} - \pi_b + \left(\frac{1}{2} - \lambda + \pi_b\right) S(x; m, w) \\ \quad \text{non-preferred interval} \end{cases} \tag{7}$$

In our experience, feedback to observers about their interval or location bias often helps to reduce it.

Column C of Figure 7.7, finally, shows an example of a potentially problematic temporal order effect. Here the estimated psychometric functions for the first and second intervals differ not only in γ—that is, show an innocuous finger bias—but also the threshold is significantly different. Note that in this data set the observer is actually more sensitive to signals in the first, *nonpreferred*, interval. The data are from a right-handed psychophysical novice, and one may want to speculate that they are the combination of the observer mainly attending to the first interval together with the typical finger bias for the second interval exhibited by right-handers.

Multidimensional Psychometric Functions

We have emphasized in the preceding example section that psychometric functions are a fundamental concept for understanding psychophysical methods. The data obtained in detection, discrimination, paired comparison, categorization, and many other tasks can be modeled by a psychometric function. For all tasks where experimenters manipulate a single continuous independent variable and collect relative proportions of behavioral observations as a dependent variable, the statistical tools described in the previous sections are applicable. Though many experiments in psychophysics fall in this category, an obvious question is: What if it is not clear whether there exists only a single independent variable, and, even if does exist, which is it? What if the stimuli are multidimensional? For example, in the previous gender categorization example, the stimuli are images of faces—very high-dimensional stimuli indeed. How do the different dimensions along which faces can vary influence the probability for a male or a female response? There is not just one independent variable that systematically varies; there are different dimensions that may or may not influence the open behavioral response to varying degrees. Can the machinery of psychometric functions

[25] *Effectively* here means in terms of percent correct and in purely practical terms; that is, the residual error from assuming cancellation is typically much smaller than the credible interval around threshold. Note that theoretically the cancellation would be exactly correct only if the psychometric function were linear. Thus it is only approximately correct for the almost linear parts of the typical sigmoidal psychometric functions around their point of inflexion.

also be used under these circumstances (Macke & Wichmann, 2010)?

The only aspect of the psychometric function ψ that depends on the stimulus x is $S(x, m, w)$ (Equation (1) or Equation (2)). Let us assume we knew which features of a face were relevant for gender categorizations and we also knew how these features are combined by an observer to form a percept of "femaleness"; then we could just take the multidimensional stimulus x and compute the femaleness f of this image and input this value into $S(f(x), m, w)$, hence reducing the high-dimensional problem to a one-dimensional one. The trouble is, of course, that in general we will not know what this function f looks like. We might, however, have k candidate features $\phi_1, ..., \phi_k$ that we want to consider. If we assume that these candidate features are combined linearly, then we can express $f(x)$ as

$$f(x) = \sum_{i=1}^{k} \alpha_i \phi_i(x).$$

We have mentioned before that the standard binomial model is a generalized linear model, and with this formulation the α_i can easily be estimated using standard GLM packages. The α_i might then be interpreted as the relative importance of the different features (Mineault, Barthelmé, & Pack, 2009).

There are, however, several complications. First of all, if the number of features is large, the number of free parameters is large, too, and the amount of data will not be enough to constrain these parameters. Second, the features are usually correlated with each other, making it hard to interpret the parameters. Luckily, modern machine learning and statistics have developed tools to deal with these problems. So-called regularization techniques allow a stable estimation of high-dimensional parameters by trading off the size of the α_i with the model fit (Macke & Wichmann, 2010). In fact, sparse regularization techniques that try to keep as many of the α_i at zero as possible also improve the interpretability of the results tremendously (Mineault et al., 2009; Schönfelder & Wichmann, 2012). An application of this approach to signal-in-noise detection in audition could settle a long-standing debate in the literature about which features are actually used by human listeners (Schönfelder & Wichmann, 2013).[26]

It might seem that even this scenario where there are candidate features that might influence the open behavioral response is very limiting, simply because one might not have a nearly complete list of candidate features. In such a case one wants to discover or identify the features used by human observers in an exploratory way. However, using a big enough set of standard features it is in fact possible to approximate any nonlinear function f. Kernel methods in machine learning, nonparametric statistics, as well as convolutional deep neural networks (CDNNs) (Kriegeskorte, 2015; Krizhevsky, Sutskever, & Hinton, 2012) make use of this idea in cases where the relevant features are unknown—and the same can be done here. In fact, it is sometimes even possible to recover the features without making strong assumptions about them (Jäkel, Schölkopf, & Wichmann, 2009; Kienzle, Franz, Schölkopf, & Wichmann, 2009). Thus, in recent years, machine learning techniques have extended tremendously the applicability of the simple psychometric function to multimensional problems. However, at the moment these extensions

[26] We motivate the multidimensional psychometric function via GLMs and techniques from machine learning to make the estimation feasible. In vision science a related approach is termed classification images, and they too are used for feature identification (Abbey & Eckstein, 2002; Ahumada, 2002; Knoblauch & Maloney, 2008; Murray, 2012, 2016).

are almost exclusively limited to the standard binomial model, and the binomial mixture model or the beta-binomial mixture model have hardly been considered in the multidimensional case.

CONCLUSION

In this chapter on methods in psychophysics we stress the importance of a time-proven, careful, and somewhat conservative—William James would have called it boring—approach to the acquistion of JND-style behavioral data. Subsequent to the acquistion, we suggest to use Bayesian inference to estimate the psychometric function of a beta-binomial model as the default data analysis if there is a single independent variable. For multidimensional data and exploratory data analyses, we introduced multidimensional psychometric function estimation techniques making use of regularization techniques from machine learning. However, in the latter case the methods are, to our knowledge, still limited to the standard binomial model; that is, they do not consider—and thus cannot adequately deal with—either signal-independent errors (lapses) or overdispersion.

Our approach may be succinctly summarized in the following two commandments of psychophysics:

1. *Know thy stimulus.*
   Wilson Geisler; see section on setting up the hardware.
2. *Thou shalt not be lazy (when gathering data).*
   Karl Dallenbach; see section on best practice.

We explicitly stress the importance of time-proven and time-consuming experimental methods because there is a—we believe: worrying—trend in the behavioral sciences to substitute the careful and slow psychophysical methods we describe in this chapter by outsourcing one's experiments to the Internet (most often Amazon's Mechanical Turk). However, outsorcing one's experiments clearly and undeniably violates the first and second commandments of psychophysics: There is little to no control over the quality of the hardware or software used to present the stimuli, the timing accuracy of the interplay between hardware and software, the attentional state of the participants, their concentration, the general illumination (or sound isolation) of the room, and the participants' age, health, and whatnot.[27]

Psychology is currently experiencing a crisis of confidence and a disconcerting lack of replicability (e.g., Earp & Trafimow, 2015; Open Science Collaboration, 2015; Pashler & Harris, 2012). There are multiple reasons for the crises, from plain statistical issues (e.g., Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2015; Rouder, Morey, Verhagen, Province, & Wagenmakers, 2016), to issues stemming from the experimenter's flexiblity in experimental design and post hoc data analysis, often subsumed under the heading of *questionable research practices* (e.g., Gelman & Loken, 2014; Ioannidis,

[27] We are not denying that the Mechanical Turk may be permissible for experiments that only require participants to respond to stimuli where the stimulus fidelity or presentation is irrelevant—for example, very high-level cognitive experiments using a multiple choice questionnaire (cf. the more positive discussion of using the Mechanical Turk in experiments by, e.g., Chandler, Mueller, & Paolacci, 2014; J. Marder & Fritz, 2015). For some such experiments there may perhaps even be an advantage in terms of broader demographics as compared to the typical university student sample (cf. Henrich, Heine, & Norenzayan, 2010). But even in such cases there remains a worry about concentration, attention, or even language competence in the case of questionnaires. Perhaps experiments using the Mechanical Turk should always be replicated on a small sample of participants in a proper psychophysical lab setting before publication.

Munafò, Fusar-Poli, Nosek, & David, 2014; John, Loewenstein, & Prelec, 2012; Simmons, Nelson, & Simonsohn, 2011), to skewed incentive structures and hiring policies in science (e.g., Higginson & Munafò, 2016; Himmelstein, Ariely, & Woolhandler, 2014; E. Marder, Kettenmann, & Grillner, 2010; Nosek, Spies, & Motyl, 2012). Furthermore, in areas of psychology making use of highly complex hardware and software such as functional magnetic resonance imaging (fMRI), the frequently ill-understood software tools may exacerbate the problem (Eklund, Nichols, & Knutsson, 2016; Kriegeskorte, Simmons, & Bellgowan, 2009).

We think that a lack of knowledge about psychophysical methods—as distilled into the two commandments of psychophysics—may also contribute to the problem of replicability in some areas of the behavioral sciences. We should resist the urge to design and run our experiments as quickly as possible; we should ensure that students in experimental laboratories acquire the necessary technical and hardware skills as well as the necessary statistical training to do proper science. Above all, we should be wary not to degrade behavioral experiments to the cranking of a handle, a conveyor belt production line without human inspection of all aspects of an experiment, from technical issues like the experimental hardware to the raw data and the statistical analysis.

REFERENCES

Abbey, C. K., & Eckstein, M. P. (2002). Classification image analysis: Estimation and statistical inference for two-alternative forced-choice experiments. *Journal of Vision*, *2*(1), 66–78.

Ahumada, A. J. (2002). Classification image weights and internal noise level estimation. *Journal of Vision*, *2*(1), 121–131.

Aitchison, L., Bang, D., Bahrami, B., & Latham, P. E. (2015). Doubly Bayesian analysis of confidence in perceptual decision-making. *PLoS Computational Biology*, *11*(10), e1004519. doi:10.1371/journal.pcbi.1004519.

Andrieu, C., de Freitas, N., Doucet, A., & Jordan, M. I. (2003). An introduction to MCMC for machine learning. *Machine Learning*, *50*(1–2), 5–43.

Bach, M., Meigen, T., & Strasburger, H. (1997). Raster-scan cathode-ray tubes for vision research limits of resolution in space, time and intensity, and some solutions. *Spatial Vision*, *10*(4), 403–414.

Baird, J. C. & Noma, E. (1978). *Fundamentals of scaling and psychophysics*. New York, NY: Wiley.

Barthelmé, S. & Mamassian, P. (2010). Flexible mechanisms underlie the evaluation of visual confidence. *Proceedings of the National Academy of Sciences, USA*, *107*(48), 20834–20839.

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

Boldt, A. & Yeung, N. (2015). Shared neural markers of decision confidence and error detection. *Journal of Neuroscience*, *35*(8), 3478–3484.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*(4), 433–436.

Brainard, D. H., Pelli, D. G., & Robson, T. (2002). Display characterization. In J. Hornak (Ed.), *Encyclopedia of Imaging Science and Technology* (pp. 172–188). New York, NY: Wiley.

Burton, G. J. (1981). Contrast discrimination by the human visual system. *Biological Cybernetics*, *40*(1), 27–38.

Chandler, J., Mueller, P. A., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*, *46*(1), 112–130.

Coombs, C. H., Dawes, R. M., & Tversky, A. (1970). *Mathematical psychology*. Englewood Cliffs, NJ: Prentice-Hall.

Cornsweet, T. N. (1962). The staircase-method in psychophysics. *American Journal of Psychology*, *75*(3), 485–491.

Dallenbach, K. M. (1966). The staircase-method critically examined. *American Journal of Psychology*, *79*(4), 654–656.

Di Lollo, V., Seiffert, A. E., Burchett, G., Rabeeh, R., & Ruman, T. A. (1997). Phosphor persistence of oscilloscopic displays: A comparison of four phosphors. *Spatial Vision*, *10*(4), 353–360.

Donders, F. C. (1969/1868). On the speed of mental processes. *Acta Psychologica*, *30*:412–431 [Translation of the original article "Over de snelheid van psychische processen" published in Dutch in 1868].

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, *6*, 621.

Efron, B. (2013). A 250-year argument: Belief, behavior, and the bootstrap. *Bulletin of the American Mathematical Society*, *50*, 129–146.

Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences, USA*, *113*(28), 7900–7905.

Elze, T., & Tanner, T. G. (2009). Liquid crystal display response time estimation for medical applications. *Medical Physics*, *36*(11), 4984–4990.

Elze, T., & Tanner, T. G. (2012). Temporal properties of liquid crystal displays: Implications for vision science experiments. *PLoS ONE*, *7*(9), e44048/1–20.

Elze, T., Taylor, C., & Bex, P. J. (2013). An evaluation of organic light emitting diode monitors for medical applications: Great timing, but luminance artifacts. *Medical Physics*, *40*(9), 092701, 1–6.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*, 429–433.

Fechner, G. T. (1860). *Elemente der Psychophysik*. Leipzig, Germany: Breitkopf und Härtel.

Fetsch, C. R., Kiani, R., & Shadlen, M. N. (2014). Predicting the accuracy of a decision: A neural mechanism of confidence. *Cold Spring Harbor Symposia on Quantitative Biology* (pp. 185–197). doi:10.1101/sqb.2014.79.024893

Foley, J. M., & Legge, G. E. (1981). Contrast detection and near-threshold discrimination in human vision. *Vision Research*, *21*, 1041–1053.

Foster, D. H., & Bischof, W. F. (1987). Bootstrap variance estimators for the parameters of small-sample sensory-performance functions. *Biological Cybernetics*, *57*, 341–347.

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

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

Fründ, I., Haenel, N. V., & Wichmann, F. A. (2011). Inference for psychometric functions in the presence of nonstationary behavior. *Journal of Vision*, *11*(6:16), 1–19.

Fründ, I., Wichmann, F. A., & Macke, J. H. (2014). Quantifiying the effect of intertrial dependence on perceptual decisions. *Journal of Vision*, *14*(7), 9, 1–16.

García-Pérez, M. A. (1998). Forced-choice staircases with fixed step sizes: Asymptotic and small-sample properties. *Vision Research*, *38*(12), 1861–1881.

Geisler, W. S. (1987). Ideal observer analysis of visual discrimination. In *Frontiers of Visual Science: Proceedings of the 1985 Symposium* (pp. 17–31). Washington, DC: National Academies Press.

Gelman, A. & Loken, E. (2014). The statistical crisis in science. *American Scientist*, *102*(6), 460–464.

Gescheider, G. A. (1997). *Psychophysics: The fundamentals*. Mahwah, NJ: Erlbaum.

Ghodrati, M., Morris, A. P., & Price, N. S. C. (2015). The (un)suitability of modern liquid crystal displays (LCDs) for vision research. *Frontiers in Psychology*, *6*(303), 1–11.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov chain Monte Carlo in practice*. Boca Raton, FL: CRC Press.

Gökaydin, D., Navarro, D. J., Ma-Wyatt, A., & Perfors, A. (2016). The structure of sequential effects. *Journal of Experimental Psychology: General*, *145*(1), 110–123.

Golz, J., & MacLeod, D. I. A. (2003). Colorimetry for CRT displays. *Journal of the Optical Society of America A*, *20*(5), 769–781.

Goris, R. L. T., Putzeys, T., Wagemans, J., & Wichmann, F. A. (2013). A neural population model for visual pattern detection. *Psychological Review*, *120*(3), 472–496.

Green, D. M. (1960). Psychoacoustics and detection theory. *Journal of the Acoustical Society of America*, *32*(10), 1189–1203.

Green, D. M. (1964). Consistency of auditory detection judgments. *Psychological Review*, *71*(5), 392–407.

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

Gu, H., Myung, I. J., Pitt, M. A., & Lu, Z.-L. (2013). Bayesian adaptive estimation of psychometric slope and threshold with differential evolution. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Meeting of the Cognitive Science Society* (pp. 2452–2457). Austin, TX: Cognitive Science Society.

Hawkins, S. L. (2011). William James, Gustav Fechner, and early psychophysics. *Frontiers in Physiology*, *2*(68), 1–10.

Heasly, B. S., Cottaris, N. P., Lichtman, D. P., Xiao, B., & Brainard, D. H. (2014). Rendertoolbox3: Matlab tools that facilitate physically based stimulus rendering for vision research. *Journal of Vision*, *14*(2), 6, 1–22.

Heidelberger, M. (2004). *Nature from within: Gustav Theodor Fechner and his psychophysical worldview*. Pittsburgh, PA: University of Pittsburgh Press.

Henning, G. B., Millar, R. W., & Hill, N. J. (2000). Detection of incremental and decremental bars at different locations across Mach bands and related stimuli. *Journal of the Optical Society of America A*, *17*(7), 1147–1159.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*, 61–135.

Higginson, A. D., & Munafò, M. R. (2016). Current incentives for scientists lead to underpowered studies with erroneous conclusions. *PLoS Biology*, *14*(11), e2000995.

Himmelstein, D. U., Ariely, D., & Woolhandler, S. (2014). Pay-for-performance: toxic to quality? Insights from behavioral economics. *International Journal of Health Services*, *44*(2), 203–214.

Hochberg, J. (1964). *Perception*. Foundations of Modern Psychology. Englewood Cliffs, NJ: Prentice-Hall.

Hoffman, D. M., Johnson, P. V., Kim, J. S., Vargas, A. D., & Banks, M. S. (2015). 240 Hz OLED technology properties that can enable improved image quality. *Journal of the Society for Information Display*, *22*(7), 346–356.

Hou, F., Lesmes, L., Bex, P., Dorr, M., & Lu, Z.-L. (2015). Using 10AFC to further improve the efficiency of the quick CSF method. *Journal of Vision*, *15*, 1–18.

Ioannidis, J. P. A., Munafò, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences: Detection, prevalence, and prevention. *Trends in Cognitive Sciences*, *18*(5), 235–241.

Irtel, H. (1996). Methoden der Psychophysik. In E. Erdfelder, R. Mausfeld, T. Meiser, & G. Rudinger (Eds.), *Handbuch quantitative Methoden* (pp. 479–489). Weinheim, Germany: Psychologie Verlags Union.

Jäkel, F., Schölkopf, B., & Wichmann, F. A. (2009). Does cognitive science need kernels? *Trends in Cognitive Sciences*, *13*(9), 381–388.

Jäkel, F., & Wichmann, F. A. (2006). Spatial four-alternative forced-choice method is the preferred psychophysical method for naïve observers. *Journal of Vision*, *6*(11), 1307–1322.

James, W. (1890). *The principles of psychology: Vol. 1*. New York, NY: Henry Holt.

Jang, Y., Wixted, J. T., & Huber, D. (2009). Testing signal-detection models of yes/no and two-alternative forced-choice recognition memory. *Journal of Experimental Psychology: General*, *138*, 291–306.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532.

Keane, B., Spence, M. L., Yarrow, K., & Arnold, D. H. (2015). Perceptual confidence demonstrates trial-by-trial insight into the precision of audio-visual timing encoding. *Consciousness and Cognition*, *38*, 107–117.

Kienzle, W., Franz, M. O., Schölkopf, B., & Wichmann, F. A. (2009). Center-surround patterns emerge as optimal predictors for human saccade targets. *Journal of Vision*, *9*(5:7), 1–15.

Kihara, K., Kawahara, J.-I., & Takeda, Y. (2010). Usability of liquid crystal displays for research in the temporal characteristics of perception and attention. *Behavior Research Methods*, *42*(4), 1105–1113.

Kleiner, M., Brainard, D. H., & Pelli, D. G. (2007). "What's new in Psychtoolbox-3?" *Perception*, *36* (ECVP Abstract Supplement).

Kleinert, A. (2009). Der messende Luchs. *NTM Zeitschrift für Geschichte der Wissenschaften, Technik und Medizin*, *17*(2), 199–206.

Knoblauch, K., & Maloney, L. T. (2008). Estimating classification images with generalized linear and additive models. *Journal of Vision*, *8*(16:10), 1–19.

Knoblauch, K., & Maloney, L. T. (2012). *Modeling psychophysical data in R*. New York, NY: Springer.

Koenderink, J. J. (1999). Virtual psychophysics. *Perception*, *28*(6), 669–674.

Koenderink, J. J., van Doorn, A. J., Albertazzi, L., & Wagemans, J. (2015). Relief articulation techniques. *Art and Perception*, *3*, 151–171.

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

Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, *1*, 417–446.

Kriegeskorte, N., Simmons, W. K., & Bellgowan, P. S. F. (2009). Circular analysis in systems neuroscience: The dangers of double dipping. *Nature Neuroscience*, *12*(5), 535–540.

Krizhevsky, A., Sutskever, I. I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Neural Information Processing Systems*, *25*, 1097–1105.

Krueger, L. E. (1989). Reconciling Fechner and Stevens: Toward a unified psychophysical law. *Behavioral and Brain Sciences*, *12*(6), 251–267.

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). Waltham, MA: Academic Press.

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

Lages, M., & Treisman, M. (1998). Spatial frequency discrimination: Visual long-term memory or criterion setting? *Vision Research*, *38*(4), 557–572.

Laming, D. (1997). *The measurement of sensation*. Oxford Psychology Series. Oxford, England: Oxford University Press.

Laming, D. (2001). Psychophysics. In N. J. Smelser & P. B. Baltes (Eds.), *International encyclopedia of the social and behavioral sciences* (Vol. 18, pp. 12444–12448). Killington, Oxford, England: Elsevier.

Laming, D. (2013). Contrast discrimination by the methods of adjustment and two-alternative forced choice. *Attention, Perception and Psychophysics*, *75*, 1774–1782.

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

Macke, J. H., & Wichmann, F. A. (2010). Estimating predictive stimulus features from psychophysical data: The decision image technique applied to human faces. *Journal of Vision*, *10*(5), 1–24.

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide*. Cambridge, United Kingdom: Cambridge University Press.

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

Maloney, L. T., Dal Martello, M. F., Sahm, C., & Spillmann, L. (2005). Past trials influence perception of ambiguous motion quartets through pattern completion. *Proceedings of the National Academy of Sciences USA*, *102*(8), 3164–3169.

Marder, E., Kettenmann, H., & Grillner, S. (2010). Impacting our young. *Proceedings of the National Academy of Sciences USA*, *107*(50), 21233.

Marder, J., & Fritz, M. (2015). The Internet's hidden science factory. *PBS Newshour*.

Marks, L. E., & Gescheider, G. A. (2002). Psychophysical scaling. In H. Pashler & J. Wixted (Eds.), *Stevens' handbook of experimental psychology: Vol. IV. Methodology in experimental psychology* (pp. 91–138). New York, NY: Wiley.

Mausfeld, R. (2000). Von der Zahlenmetapher zur Maßformel: Fechners Psychophysik in der Tradition der Mathesis universalis. In *Invited Keynote Address, International Symposium in Honour to G. Th. Fechner, Universität Leipzig*. International Society for Psychophysics, October 19–23, 2000.

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

Meyniel, F., Schlunegger, D., & Dehaene, S. (2015). The sense of confidence during probabilisitic learning: A normative account. *PLoS Computational Biology*, *11*(6), e1004305, doi:10.1371/journal.pcbi.1004305

Mineault, P., Barthelmé, S., & Pack, C. (2009). Improved classification images with sparse priors in a smooth basis. *Journal of Vision*, *9*(10), 17, 1–24.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2015). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin and Review*, *23*, 103–123.

Morey, R. D., Hoekstra, R., Rouder, J. N., & Wagenmakers, E.-J. (2015) Continued misinterpretation of confidence intervals: Response to Miller and Ulrich. *Psychological Bulletin and Review*, *23*, 131–140.

Murray, R. F. (2012). Classification images and bubbles images in the generalized linear model. *Journal of Vision*, *12*(7), 2, 1–8.

Murray, R. F. (2016). Classification images in a very general decision model. *Vision Research*, *123*, 26–32.

Nachmias, J., & Sansbury, R. V. (1974). Grating contrast: Discrimination may be better than detection. *Vision Research*, *14*(10), 1039–1042.

Nachmias, J., & Steinman, R. M. (1965). An experimental comparison of the method of limits and the double staircase-method. *American Journal of Psychology*, *78*, 112–115.

Naiman, A., & Makous, W. L. (1992). Spatial nonlinearities of grayscale CRT pixels. *Proceedings of the SPIE* (Vol. 1666, pp. 41–56).

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*(6), 615–631.

Nuzzo, R. (2014). Statistical errors. *Nature*, *506*(7487), 150–152.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), 943; aac4716–1–8.

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

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, *7*(6), 531–536.

Peirce, C. S., & Jastrow, J. (1885). On small differences in sensation. *Memoirs of the National Academy of Sciences*, *3*, 73–83.

Pelli, D. G. (1997a). Pixel independence: Measuring spatial interactions on a CRT display. *Spatial Vision*, *10*(4), 443–446.

Pelli, D. G. (1997b). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442.

Pelli, D. G., & Bex, P. J. (2013). Measuring contrast sensitivity. *Vision Research*, *90*, 10–14.

Pelli, D. G., & Zhang, L. (1991). Accurate control of contrast on microcomputer displays. *Vision Research*, *31*(7/8), 1337–1350.

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*(349), 321–327.

Ross, H. E. (1997). On the possible relations between discriminability and apparent magnitude. *British Journal of Mathematical and Statistical Psychology*, *50*, 187–203.

Rouder, J. N., Morey, R. D., Verhagen, J., Province, J. M., & Wagenmakers, E.-J. (2016). Is there a free lunch in inference? *Topics in Cognitive Science*, *8*(3), 520–547.

Scarfe, P., & Glennerster, A. (2015). Using high-fidelity virtual reality to study perception in freely moving observers. *Journal of Vision*, *15*(9), 3, 1–11.

Schönfelder, V. H., & Wichmann, F. A. (2012). Sparse regularized regression identifies behaviorally-relevant stimulus features from psychophysical data. *Journal of the Acoustical Society of America*, *131*(5), 3953–3969.

Schönfelder, V. H., & Wichmann, F. A. (2013). Identification of stimulus cues in narrow-band tone-in-noise detection using sparse observer models. *Journal of the Acoustical Society of America*, *134*(1), 447–463.

Schütt, H. H., Harmeling, S., Macke, J. H., & Wichmann, F. A. (2016). Painfree and accurate Bayesian estimation of psychometric functions for (potentially) overdispersed data. *Vision Research*, *122*, 105–123.

Schütt, H. H., & Wichmann, F. A. (2014). Uncertainty effects in visual psychophysics. In *Tagung experimentell arbeitender Psychologen (TeaP)*. Giessen, Germany: Universität Gießen.

Senders, V., & Sowards, A. (1952). Analysis of response sequences in the setting of a psychophysical experiment. *American Journal of Psychology*, *65*(3), 358–374.

Shen, Y., & Richards, V. M. (2012). A maximum-likelihood procedure for estimating psychometric functions: Thresholds, slopes, and lapses of attention. *Journal of the Acoustical Society of America*, *132*(2), 957–967.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366.

Spence, M. L., Dux, P. E., & Arnold, D. H. (2015). Computations underlying confidence in visual perception. *Journal of Experimental Psychology: Human Perception and Performance*, 1–12. doi:10.1037/xhp0000179

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review*, *64*, 153–181.

Stevens, S. S. (1960). The psychophysics of sensory function. *American Scientist*, *48*, 226–253.

Stüttgen, M. C., Schwarz, C., & Jäkel, F. (2011). Mapping spikes to sensations. *Frontiers in Neuroscience*, *5*(125), 1–17.

Swets, J. A. (1959). Indices of signal detectability obtained with various psychophysical procedures. *Journal of the Acoustical Society of America*, *31*(4), 511–513.

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

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*, 273–286.

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*(17), 2503–2522.

Treutwein, B., & Strasburger, H. (1999). Fitting the psychometric function. *Perception and Psychophysics*, *61*(1), 87–106.

Tyler, C. W. (1997). Colour bit-stealing to enhance the luminance resolution of digital displays on a single pixel basis. *Spatial Vision*, *10*(4), 369–377.

Verplanck, W. S., Collier, G. H., & Cotton, J. W. (1952). Non-independence of successive responses in measurements of the visual threshold. *Journal of Experimental Psychology*, *44*(4), 273–282.

Versfeld, N. J., Dai, H., & Green, D. M. (1996). Optimum decision rules for the oddity task. *Perception & Psychophysics*, *58*(1), 10–21.

Vickers, D. (1979). *Decision processes in visual perception*. New York, NY: Academic Press.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychological Bulletin and Review*, *14*(5), 779–804.

Wagenmakers, E.-J., Farrell, S., & Ratcliff, R. (2004). Estimation and interpretation of 1/f alpha noise in human cognition. *Psychological Bulletin and Review*, *11*(4), 579–615.

Watson, A. B. & Pelli, D. G. (1983). Quest: A Bayesian adaptive psychometric method. *Perception and Psychophysics*, *33*(2), 113–120.

Wichmann, F. A. (1999). *Some aspects of modelling human spatial vision: Contrast discrimination*. (PhD thesis). University of Oxford, Oxford, United Kingdom.

Wichmann, F. A., Graf, A. B. A., Simoncelli, E. P., Bülthoff, H. H., & Schölkopf, B. (2005). Machine learning applied to perception: Decision-images for gender classification. In L. K. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in neural information processing systems* (Vol. 17, pp. 1489–1496). Cambridge, MA: MIT Press.

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

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

Wier, C. C., Jesteadt, W., & Green, D. M. (1976). A comparison of method-of-adjustment and forced-choice procedures in frequency discrimination. *Perception & Psychophysics*, *19*, 75–79.

Yeshurun, Y., Carrasco, M., & Maloney, L. T. (2008). Bias and sensitivity in two-interval forced choice procedures: Tests of the difference model. *Vision Research*, *48*, 1837–1851.

Zhang, F., de Ridder, H., Fleming, R. W., & Pont, S. C. (2016). Matmix 1.0: Using optical mixing to probe visual material perception. *Journal of Vision*, *16*(6), 11, 1–18.

Zychaluk, K., & Foster, D. H. (2009). Model-free estimation of the psychometric function. *Attention, Perception and Psychophysics*, *71*(6), 1414–1425.