Pelli, D. G., & Farell, B. (2010). Psychophysical methods. In M. Bass, C. DeCusatis, J. Enoch, V. Lakshminarayanan, G. Li, C. MacDonald, V. Mahajan & E. V. Stryland (Eds.), *Handbook of Optics, Third Edition, Volume III: Vision and Vision Optics* (pp. 3.1-3.12). New York: McGraw-Hill.
http://psych.nyu.edu/pelli/pubs/pelli2010methods.pdf

3

PSYCHOPHYSICAL METHODS

Denis G. Pelli

Psychology Department and Center for Neural Science
New York University
New York

Bart Farell

Institute for Sensory Research
Syracuse University
Syracuse, New York

3.1 INTRODUCTION

Psychophysical methods are the tools for measuring perception and performance. These tools are used to reveal basic perceptual processes, to assess observer performance, and to specify the required characteristics of a display. We are going to ignore this field's long and interesting history,[1] and much theory as well.[2,3] Here we present a formal treatment, emphasizing the theoretical concepts of psychophysical measurement. For practical advice in setting up an experiment, please turn to our user's guide.[4] Use the supplied references for further reading.

Consider the psychophysical evaluation of the suitability of a visual display for a particular purpose. A home television to be used for entertainment is most reasonably assessed in a "beauty contest" of subjective preference,[5] whereas a medical imaging display must lead to accurate diagnoses[6,7] and military aerial reconnaissance must lead to accurate vehicle identifications.[8] In our experience, the first step toward defining a psychophysically answerable question is to formulate the problem as a task that the observer must perform. One can then assess the contribution of various display parameters toward that performance. Where precise parametric assessment is desired it is often useful to substitute a simple laboratory task for the complex real-life activity, provided one can either demonstrate, or at least reasonably argue, that the laboratory results are predictive.

Psychophysical measurement is usually understood to mean measurement of behavior to reveal internal processes. The experimenter is typically not interested in the behavior itself, such as pressing a button, which merely communicates a decision by the observer about the stimulus.* This chapter reviews the various decision tasks that may be used to measure perception and performance and evaluates their strengths and weaknesses. We begin with definitions and a brief review of visual stimuli. We then explain and evaluate the various psychophysical tasks, and end with some practical tips.

*Psychophysical measurement can also be understood to include noncommunicative physiological responses such as pupil size, eye position, electrical potentials measured on the scalp and face, and even BOLD fMRI responses in the brain, which might be called "unintended" responses. (These examples are merely suggestive, not definitive. Observers can decide to move their eyes and, with feedback, can learn to control many other physiological responses. Responses are "unintended" only when they are not used for overt communication by the observer.) Whether these unintended responses are called psychophysical or physiological is a matter of taste. In any case, decisions are usually easier to measure and interpret, but unintended responses may be preferred in certain cases, as when assessing noncommunicative infants and animals.

3.2 DEFINITIONS

At the highest level, an *experiment* answers a question about how certain "experimental conditions" affect observer performance. *Experimental conditions* include stimulus parameters, observer instruction, and anything else that may affect the observer's state. Experiments are usually made up of many individual measurements, called "trials," under each experimental condition. Each trial presents a stimulus and collects a response—a decision—from the observer.

There are two kinds of decision tasks: judgments and adjustments. It is useful to think of one as the inverse of the other. In one case the experimenter gives the observer a stimulus and asks for a classification of the stimulus or percept; in the other case the experimenter, in effect, gives the observer a classification and asks for an appropriate stimulus back. Either the experimenter controls the stimulus and the observer makes a *judgment* based on the resulting percept, or the observer *adjusts* the stimulus to satisfy a perceptual criterion specified by the experimenter (e.g., match a sample). Both techniques are powerful. Adjustments are intrinsically subjective (because they depend on the observers' understanding of the perceptual criterion), yet they can often provide good data quickly and are to be preferred when applicable. But not all questions can be formulated as adjustment tasks. Besides being more generally applicable, judgments are often easier to analyze, because the stimulus is under the experimenter's control and the task may be objectively defined. Observers typically like doing adjustments and find judgments tedious, partly because judgment experiments usually take much longer.

An obvious advantage of adjustment experiments is that they measure physical stimulus parameters, which may span an enormous dynamic range and typically have a straightforward physical interpretation. Judgment tasks measure human performance (e.g., frequency of seeing) as a function of experimental parameters (e.g., contrast). This is appropriate if the problem at hand concerns human performance per se. For other purposes, however, raw measures of judgment performance typically have a very limited useful range, and a scale that is hard to interpret. Having noted that adjustment and judgment tasks may be thought of as inverses of one another, we hasten to add that in practice they are often used in similar ways. Judgment experiments often vary a stimulus parameter on successive trials in order to find the value that yields a criterion judgment. These "sequential estimation methods" are discussed in Sec. 3.6. The functional inversion offered by sequential estimation allows judgment experiments to measure a physical parameter as a function of experimental condition, like adjustment tasks, while retaining the judgment task's more rigorous control and interpretation.

Distinguishing between judgment and adjustment tasks emphasizes the kind of response that the observer makes. It is also possible to subdivide tasks in a way that emphasizes the stimuli and the question posed. In a *detection* task there may be any number of alternative stimuli, but one is a blank, and the observer is asked only to distinguish between the blank and the other stimuli. Slightly more general, a *discrimination* task may also have any number of alternative stimuli, but one of the stimuli, which need not be blank, is designated as the reference, and the observer is asked only to distinguish between the reference and other stimuli. A decision that distinguishes among more than two categories is usually called an *identification* or *classification*.[9] All decision tasks allow for alternative responses, but two alternatives is an important special case.[10]

As normally used, the choice of term, *detection* or *discrimination*, says more about the experimenter's way of thinking than it does about the actual task faced by the observer. This is because theoretical treatments of detection and discrimination usually allow for manipulation of the experimental condition by introduction of an extraneous element, often called a "mask" or "pedestal," that is added to every stimulus. Thus, one is always free to consider a discrimination task as detection in the presence of a mask. This shift in perspective can yield new insights (e.g., Refs. 11–15). Since there is no fundamental difference between detection and discrimination,[16] we have simplified the presentation below by letting detection stand in for both. The reader may freely substitute "reference" for "blank" (or suppose the presence of an extraneous mask) in order to consider the discrimination paradigm.

The idea of "threshold" plays a large role in psychophysics. Originally deterministic, *threshold* once referred to the stimulus intensity above which the stimulus was always distinguishable from blank, and below which it was indistinguishable from blank. In a discrimination task one might refer to a "discrimination threshold" or a "just-noticeable difference." Nowadays the idea is statistical; we know that the observer's probability of correct classification rises as a continuous function of stimulus

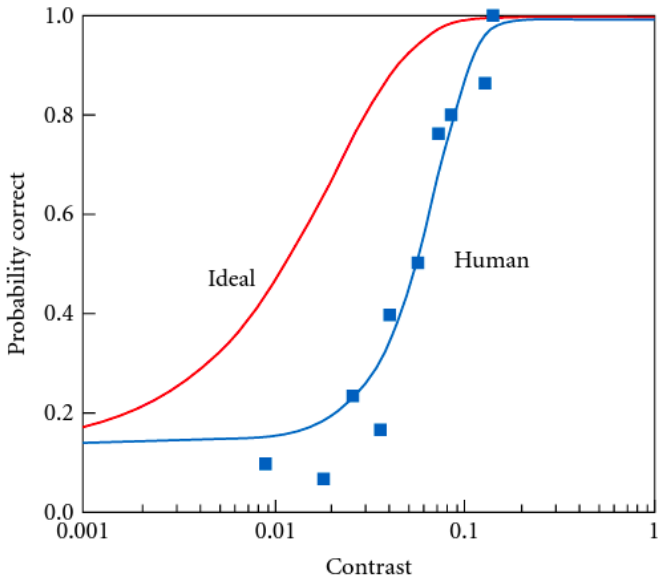

FIGURE 1 Probability of correctly identifying a letter in noise, as a function of letter contrast. The letters are bandpass filtered. Gaussian noise was added independently to each pixel. Each symbol represents the proportion correct in 30 trials. The solid curve through the points is a maximum likelihood fit of a Weibull function. The other curve represents a similar maximum likelihood fit to the performance of a computer program that implements the ideal letter classifier.[40] Efficiency, the squared ratio of threshold contrasts, is 9 percent. (*Courtesy of Joshua A. Solomon.*)

intensity (see Fig. 1). Threshold is defined as the stimulus intensity (e.g., contrast) corresponding to an arbitrary level of performance (e.g., 82 percent correct). However, the old intuition, now called a "high threshold," still retains a strong hold on everyone's thinking for the good reason that the transition from invisible to visible, though continuous, is quite abrupt, less than a factor of two in contrast.

Most psychophysical research has concentrated on measuring thresholds. This has been motivated by a desire to isolate low-level sensory mechanisms by using operationally defined tasks that are intended to minimize the roles of perception and cognition. This program is generally regarded as successful—visual detection is well understood (e.g., Ref. 17)—but leaves most of our visual experience and ability unexplained. This has stimulated a great deal of experimentation with suprathreshold stimuli and nondetection tasks in recent years.

3.3 VISUAL STIMULI

Before presenting the tasks, which are general to all sense modalities (not just vision), it may be helpful to briefly review the most commonly used visual stimuli. Until the 1960s most vision research used a spot as the visual stimulus (e.g., Ref. 18). Then cathode ray tube displays made it easy to generate more complex stimuli, especially sinusoidal gratings, which provided the first evidence for multiple "spatial frequency channels" in vision.[19] Sinusoidal grating patches have two virtues. A sinusoid at the display always produces a sinusoidal image on the retina.* And most visual mechanisms are selective in space and in spatial frequency, so it is useful to have a stimulus that is restricted in both domains.

*This is strictly true only within an isoplanatic patch, i.e., a retinal area over which the eye's optical point spread function is unchanged.

Snellen,[20] in describing his classic eye chart, noted the virtue of letters as visual stimuli—they offer a large number of stimulus alternatives that are readily identifiable.[21,22] Other commonly used stimuli include annuli, lines, arrays of such elements, and actual photographs of faces, nature, and military vehicles. There are several useful texts on image quality, emphasizing signal-to-noise ratio.[23–26] Finally, there has been some psychophysical investigation of practical tasks such as reading,[27] flying an airplane,[28] or shopping in a supermarket.[29]

The stimulus alternatives used in vision experiments are usually parametric variations along a single dimension, most commonly contrast, but frequently size and position in the visual field. *Contrast* is a dimensionless ratio: the amplitude of the luminance variation within the stimulus, normalized by the background luminance. Michelson contrast (used for gratings) is the maximum minus the minimum luminance divided by the maximum plus the minimum. Weber contrast (used for spots and letters) is the maximum deviation from the uniform background divided by the background luminance. RMS contrast is the root-mean-square deviation of the stimulus luminance from the mean luminance, divided by the mean luminance.

3.4 ADJUSTMENTS

Adjustment tasks require that the experimenter specify a perceptual criterion to the observer, who adjusts the stimulus to satisfy the criterion. Doubts about the observer's interpretation of the criterion may confound interpretation of the results. The adjustment technique is only as useful as the criterion is clear.

Threshold

Figure 2 shows contrast sensitivity (the reciprocal of the threshold contrast) for a sinusoidal grating as a function of spatial and temporal frequency.[30] These thresholds were measured by what is probably the most common form of the adjustment task, which asks the observer to adjust the stimulus contrast up and down to the point where it is "just barely detectable." While some important studies have collected their data in this way, one should bear in mind that this is a vaguely specified criterion. What should the observer understand by "barely" detectable? Seen half the time? In order to adjust to threshold, the observer must form a subjective interpretation and apply it to the changing percept. It is well known that observers can be induced (e.g., by coaching) to raise or lower their criterion, and when comparing among different observers it is important to bear in mind that social and personality factors may lead to systematically different interpretations of the same vague instructions. Nevertheless, these subjective effects are relatively small (about a factor of two in contrast) and many questions can usefully be addressed, in at least a preliminary way, by quick method-of-adjustment threshold settings. Alternatively, one might ignore the mean of the settings and instead use the standard deviation to estimate the observer's discrimination threshold.[31]

Nulling

Of all the many kinds of adjustments, nulling is the most powerful. Typically, there is a simple basic stimulus that is distorted by some experimental manipulation, and the observer is given control over the stimulus and asked to adjust it so as to cancel the distortion (e.g., Ref. 32). The absence of a specific kind of distortion is usually unambiguous and easy for the observer to understand, and the observer's null setting is typically very reliable.

Matching

Two stimuli are presented, and the observer is asked to adjust one to match the other. Sometimes the experiment can be designed so that the observer can achieve a perfect match in which the stimuli are utterly indistinguishable, which Brindley[33] calls a "Class A" match. Usually, however, the stimuli

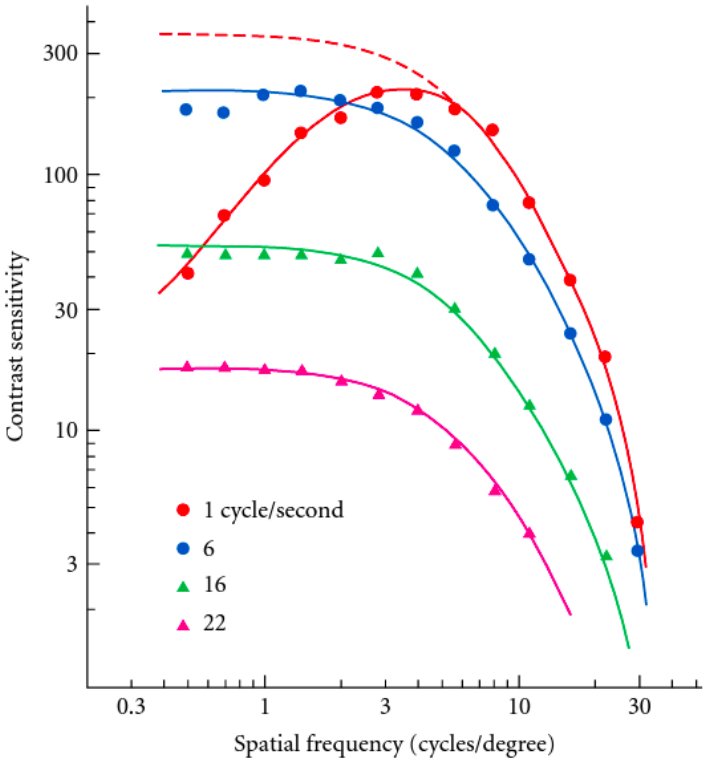

FIGURE 2 Spatial contrast sensitivity (reciprocal of threshold contrast) functions for sinusoidal gratings temporally modulated (flickered) at several temporal frequencies. The points are the means of four method-of-adjustment measurements and the curves (one with a dashed low-frequency section) differ only in their positions along the contrast-sensitivity scale. (*From Robson.*[30])

are obviously different and the observer is asked to match only a particular aspect of the stimuli, which is called a "Class B" match. For example, the observer might be shown two grating patches, one fine and one coarse, and asked to adjust the contrast of one to match the contrast of the other.[34] Or the observer might see two uniform patches of different colors and be asked to match their brightnesses.[35] Observers (and reviewers for publication) are usually comfortable with matching tasks, but, as Brindley points out, it is amazing that observers can seemingly abstract and compare a particular parameter of the multidimensional stimuli in order to make a Class B match. Matching tasks are extremely useful, but conclusions based on Class B matches may be less secure than those based on Class A matches because our understanding of how the observer does the task is less certain.

Magnitude Production

The observer is asked to adjust a stimulus to match a numerically specified perceptual criterion, e.g., "as bright as a 60-watt light bulb." The number may have a scale (watts in this case) or be a pure number.[2] The use of pure numbers, without any scale, to specify a perceptual criterion is obviously formally ambiguous, but in practice many experimenters report that observers seem comfortable with such instructions and produce stable results that are even reasonably consistent among different observers. Magnitude production, however, is rarely used in visual psychophysics research.

3.5 JUDGMENTS

Judgment tasks ask the observer to classify the stimulus or percept. They differ primarily in the number of alternative stimuli that may be presented on a given trial and the number of alternative responses that the observer is allowed.

The Ideal Observer

When the observer is asked to classify the stimulus (not the percept) it may be useful to consider the mathematically defined ideal classifier that would yield the most accurate performance using only the information (the stimuli and their probabilities) available to the observer.[36–42] Obviously this would be an empty exercise unless there is some known factor that makes the stimuli hard to distinguish. Usually this will be visual noise: random variations in the stimulus, random statistics of photon absorptions in the observer's eyes, or random variations in neural processes in the observer's visual system. If the stimuli plus noise can be defined statistically at some site—at the display, as an image at the observer's retinae, as a pattern of photon absorptions, or as a spatiotemporal pattern of neural activity—then one can solve the problem mathematically and compute the highest attainable level of performance. This ideal often provides a useful point of comparison in thinking about the actual human observer's results. A popular way of expressing such a comparison is to compute the human observer's efficiency, which will be a number between 0 and 1. For example, in Fig. 1 at threshold the observer's efficiency for letter identification is 9 percent. As a general rule, the exercise of working out the ideal and computing the human observer's efficiency is usually instructive but, obviously, low human efficiencies should be interpreted as a negative result, suggesting that the ideal is not particularly relevant to understanding how the human observer does the task.

Yes-No

The best-known judgment task is yes-no. It is usually used for detection, although it is occasionally used for discrimination. The observer is either asked to classify the stimulus, "Was a nonblank stimulus present?" or classify the percept, "Did you see it?" The observer is allowed only two response alternatives: yes or no. There may be any number of alternative stimuli. If the results are to be compared with those of an ideal observer, then the kind of stimulus, blank or nonblank, must be unpredictable.

As with the method-of-adjustment thresholds discussed above, the question posed in a yes-no experiment is fundamentally ambiguous. Where is the dividing line between yes and no on the continuum of internal states between the typical percepts generated by the blank and nonblank stimuli? Theoretical considerations and available evidence suggest that observers act as if they reduced the percept to a "decision variable," a pure magnitude—a number if you like—and compared that magnitude with an internal criterion that is under their conscious control.[40,43] Normally we are not interested in the criterion, yet it is troublesome to remove its influence on the results, especially since the criterion may vary between experimental conditions and observers. For this reason, most investigators no longer use yes-no tasks.

As discussed next, this pesky problem of the observer's subjective criterion can be dealt with explicitly, by using "rating scale" tasks, or banished, by using unbiased "two-alternative forced choice" (2afc) tasks. Rating scale is much more work, and unless the ratings themselves are of interest, the end result of using either rating scale or 2afc is essentially the same.

Rating Scale

In a rating scale task the observer is asked to rate the likelihood that a nonblank stimulus was presented. There must be blank and nonblank stimulus alternatives, and there may be any number of alternative ratings—five is popular—but even a continuous scale may be allowed.[44] The endpoints of

the rating scale are "The stimulus was definitely blank" and "The stimulus was definitely nonblank," with intermediate degrees of confidence in between. The results are graphed as a receiver operating characteristic, or ROC, that plots one conditional probability against another. The observer's ratings are transformed into yes-no judgments by comparing them with an external criterion. Ratings above the criterion become "yes" and those below the criterion become "no." This transformation is repeated for all possible values of the external criterion. Finally, the experimenter plots—for each value of the criterion—the probability of a yes when a nonblank stimulus was present (a "hit") against the probability of a yes when a blank stimulus was present (a "false alarm"). In medical contexts the hit rate is called "sensitivity" and one minus the false alarm rate is called "specificity." Figure 3 shows an ROC curve for a medical diagnosis;[45] radiologists examined mammograms and rated the likelihood that a lesion was benign or malignant.

In real-life applications the main value of ROC curves is that they can be used to optimize yes-no decisions based on ratings, e.g., whether to refer a patient for further diagnosis or treatment. However, this requires knowledge of the prior stimulus probabilities (e.g., in Fig. 3, the incidence of disease in the patient population), the benefit of a hit, and the cost of a false alarm.[6,7] These conditions are rarely met. One usually can estimate prior probability and assess the cost of the wasted effort caused by the false alarms, but it is hard to assign a commensurate value to the hits, which may save lives through timely treatment.

The shape of the ROC curve has received a great deal of attention in the theoretical detection literature, and there are various mathematical models of the observer's detection process that can account for the shape.[46–48] However, unless the actual situation demands rating-based decisions, the ROC shape has little or no practical significance, and the general practice is to summarize the ROC curve by the area under the curve. The area is 0.5 when the observers' ratings are independent of the stimuli (i.e., useless guessing). The area can be at most 1—when the observer makes no mistakes. The

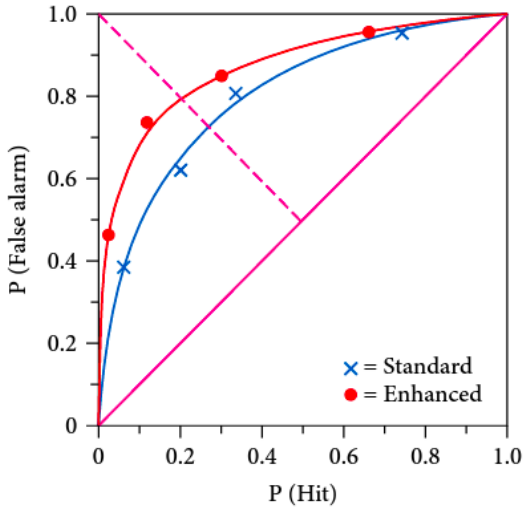

FIGURE 3 Example of an empirical ROC. Six radiologists attempted to distinguish between malignant and benign lesions in a set of 118 mammograms, 58 malignant and 60 benign, first when the mammograms were viewed in the usual manner ("standard"), and then—"enhanced"—when they were viewed with two aids, including a checklist of diagnostic features. The ratings were "very likely malignant," "probably malignant," "possibly malignant," "probably benign," and "very likely benign." The areas under the curves are 0.81 and 0.87. (*From Swets.*[45])

area can descend below 0.5 when the observer reverses the categories of blank and nonblank. We'll see in a moment that a result equivalent to ROC area can usually be obtained with much less effort by doing a two-alternative forced choice experiment instead.

Two-Alternative Forced Choice

This task is traditionally characterized by two separate stimulus presentations, one blank and one nonblank, in random order. The two stimuli may be presented successively or side by side. The observer is asked whether the nonblank stimulus was first or second (or on the left or right). We noted above that in yes-no tasks observers seem to reduce the stimulus to a decision variable, the magnitude upon which they base their decisions. The 2afc task is said to be "unbiased" because the observer presumably chooses the presentation that generated the higher magnitude, without referring to any subjective internal criterion. At the beginning of this section we said that all judgment tasks consist of the presentation of a stimulus followed by a judgment. In this view, we might consider the two presentations in the 2afc task to be a single stimulus. The two possible composite stimuli to be discriminated are reflections of one another, either in space or time. The symmetry of the two alternatives suggests that the observer's choice between them may be unbiased.

Other related tasks are often called "two-alternative forced choice" and are similarly claimed to be unbiased. There is some confusion in the literature over which tasks should be called "2afc." In our view, the "2afc" label is of little consequence. What matters is whether the task is unbiased, i.e., are the alternative stimuli symmetric for the observer? Thus a yes-no discrimination of blank and nonblank stimuli may be biased even though there are two response alternatives and the choice is forced, whereas it may be reasonable to say that the judgment of the orientation of a grating that is either horizontal or vertical is unbiased even though there is only a single presentation. We suggest that authors wishing to claim that their task is unbiased say so explicitly and state why. This claim might be based on a priori considerations of the symmetry between the stimuli to be discriminated, or on a post hoc analysis of relative frequencies of the observer's responses.

In theory, if we accept the assumptions that each stimulus presentation produces in the observer a unidimensional magnitude (one number, the decision variable), that the observer's ratings and 2afc decisions are based, in the proper way, on this magnitude, and that these magnitudes are stochastically independent between presentations, then the probability of a correct response on a 2afc trial must equal the area under the ROC curve.[47] Nachmias[43] compared 2afc proportion correct and ROC area empirically, finding that ROC area is slightly smaller, which might be explained by stimulus-induced variations in the observer's rating criteria.

MAGNITUDE ESTIMATION

In the inverse of magnitude production, a stimulus is presented and the observer is asked to rate it numerically.[2] Some practitioners provide a reference (e.g., a stimulus that rates 100), and some don't, allowing observers to use their own scale. Magnitude estimation and rating scale are fundamentally the same. Magnitude estimation experiments typically test many different stimulus intensities a few times to plot mean magnitude versus intensity, and rating-scale experiments typically test few intensities many times to plot an ROC curve at each intensity.

Response Time

In practical situations the time taken by the observer to produce a judgment usually matters, and it will be worthwhile recording it during the course of the experiment. Some psychophysical research has emphasized response time as a primary measure of performance in an effort to reveal mental processes.[49]

3.6 STIMULUS SEQUENCING

So far we have discussed a single trial yielding a single response from the observer. Most judgments are stochastic, so judgment experiments usually require many trials. An uninterrupted sequence of trials is called a *run* (or a *block*). There are two useful methods of sequencing trials within a run.

Method of Constant Stimuli

Experimenters have to worry about small, hard-to-measure variations in the observer's sensitivity that might contaminate comparisons of data collected at different times. It is therefore desirable to run the trials for the various conditions as nearly simultaneously as possible. One technique is to interleave trials for the various conditions. This is the classic "method of constant stimuli." Unpredictability of the experimental condition and equal numbers of trials for each condition are typically both desirable. These are achieved by using a randomly shuffled list of all desired trials to determine the sequence.

Sequential Estimation Methods

One can use the method of constant stimuli to measure performance as a function of a signal parameter—let us arbitrarily call it intensity—and determine, by interpolation, the threshold intensity that corresponds to a criterion level of performance.* This approach requires hundreds of trials to produce a precise threshold estimate. Various methods have been devised that obtain precise threshold estimates in fewer trials, by using the observer's previous responses to choose the stimulus intensity for the current trial. The first methods were simple enough for the experimenter to implement manually, but as computers appeared and then became faster, the algorithms have become more and more sophisticated. Even so, the requisite computer programs are very short.

In general, there are three stages to threshold estimation. First, all methods, implicitly or explicitly, require that the experimenter provide a confidence interval around a guess as to where threshold may lie. (This bounds the search. Lacking prior knowledge, we would have an infinite range of possible intensities. Without a guess, where would we place the first trial? Without a confidence interval, where would we place the second trial?) Second, one must select a test intensity for each trial based on the experimenter's guess and the responses to previous trials. Third, one must use the collected responses to estimate threshold. At the moment, the best algorithm is called ZEST,[50] which is an improvement over the popular QUEST.[51] The principal virtues of QUEST are that it formalizes the three distinct stages, and implements the first two stages efficiently. The principal improvement in ZEST is an optimally efficient third stage.

3.7 CONCLUSION

This chapter has reviewed the practical considerations that should guide the choice of psychophysical methods to quickly and definitely answer practical questions related to perception and performance. Theoretical issues, such as the nature of the observer's internal decision process, have been de-emphasized. The question of how well we see is answerable only after we reduce the question to measurable performance of a specific task. The task will be either an adjustment—for a quick answer when the perceptual criterion is unambiguous—or a judgment—typically to find threshold by sequential estimation.

*The best way to interpolate frequency-of-seeing data is to make a maximum likelihood fit by an S-shaped function.[52] Almost any S-shaped function will do, provided it has adjustable position and slope.[53]

The success of psychophysical measurements often depends on subtle details: the seemingly incidental properties of the visual display, whether the observers receive feedback about their responses, and the range of stimulus values encountered during a run. Decisions about these matters have to be taken on a case-by-case basis.

3.8 TIPS FROM THE PROS

We asked a number of colleagues for their favorite tips.

- Experiments often measure something quite different from what the experimenter intended. Talk to the observers. Be an observer yourself.
- Viewing distance is an often-neglected but powerful parameter, trivially easy to manipulate over a 100:1 range. Don't be limited by the length of your keyboard cable.
- Printed vision charts are readily available, offering objective measurement of visiblity, e.g., to characterize the performance of a night-vision system.[20,21,54,55]
- When generating images on a cathode ray tube, avoid generating very high video frequencies (e.g., alternating black and white pixels along a horizontal raster line) and very low video frequencies (hundreds of raster lines per cycle) since they are typically at the edges of the video amplifier's passband.[56]
- Liquid crystal displays (LCD) have largely replaced cathode ray tube (CRT) displays in the market place. LCDs are fine for static images, but have complicated temporal properties that are hard to characterize. Thus, CRTs are still preferable for presentation of dynamic images, as they allow you to know exactly what you are getting.[57]
- Consider the possibility of aftereffects, whereby past stimuli (e.g., at high contrast or different luminance) might affect the visibility of the current stimulus.[58–61]
- Drift of sensitivity typically is greatest at the beginning of a run. Do a few warm-up trials at the beginning of each run. Give the observer a break between runs.
- Allow the observer to see the stimulus once in a while. Sequential estimation methods tend to make all trials just barely detectable, and the observer may forget what to look for. Consider throwing in a few high-contrast trials, or defining threshold at a high level of performance.
- Calibrate your display before doing the experiment, rather than afterward when it may be too late.

3.9 ACKNOWLEDGMENTS

Josh Solomon provided Fig. 1. Tips were contributed by Al Ahumada, Mary Hayhoe, Mary Kaiser, Gordon Legge, Walt Makous, Suzanne McKee, Eugenio Martinez-Uriega, Beau Watson, David Williams, and Hugh Wilson. Al Ahumada, Katey Burns, and Manoj Raghavan provided helpful comments on the manuscript. Supported by National Eye Institute grants EY04432 and EY06270.

3.10 REFERENCES

1. E. G. Boring, *Sensation and Perception in the History of Experimental Psychology*, Irvington Publishers, New York, 1942.
2. G. A. Gescheider, *Psychophysics: Methods, Theory, and Application*, 2d ed., Lawrence Erlbaum and Associates, Hillsdale, N.J., 1985, pp. 174–191.
3. N. A. Macmillan and C. D. Creelman, *New Developments in Detection Theory*, Cambridge University Press, Cambridge, U.K., 1991.

4. B. Farell, and D. G. Pelli, "Psychophysical Methods, or How to Measure a Threshold and Why," R. H. S. Carpenter and J. G. Robson (eds.), *Vision Research: A Practical Guide to Laboratory Methods*, Oxford University Press, New York, 1999.
5. P. Mertz, A. D. Fowler, and H. N. Christopher, "Quality Rating of Television Images," *Proc. IRE* **38:**1269–1283 (1950).
6. J. A. Swets, and R. M. Pickett, *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*, Academic Press, New York, 1982.
7. C. E. Metz, "ROC Methodology in Radiologic Imaging," *Invest. Radiol.* **21:**720–733 (1986).
8. F. Scott, "The Search for a Summary Measure of Image Quality—A Progress Report," *Photographic Sci. and Eng.* **12:**154–164 (1968).
9. F. G. Ashby, "Multidimensional Models of Categorization," *Multidimensional Models of Perception and Cognition*, F. G. Ashby (ed.), Lawrence Erlbaum Associates, Hillsdale, N.J., 1992.
10. D. G. Pelli, N. J. Majaj, N. Raizman, C. J. Christian, E. Kim, and M. C. Palomares, "Grouping in Object Recognition: The Role of a Gestalt Law in Letter Identification," *Cognitive Neuropsychology* (2009) In press.
11. F. W. Campbell, E. R. Howell, and J. G. Robson, "The Appearance of Gratings with and without the Fundamental Fourier Component," *J. Physiol.* **217:**17–18 (1971).
12. B. A. Wandell, "Color Measurement and Discrimination," *J. Opt. Soc. Am. A* **2:**62–71 (1985).
13. A. B. Watson, A. Ahumada, Jr., and J. E. Farrell, "The Window of Visibility: A Psychophysical Theory of Fidelity in Time-Sampled Visual Motion Displays," *NASA Technical Paper, 2211*, National Technical Information Service, Springfield, Va. 1983.
14. E. H. Adelson, and J. R. Bergen, "Spatiotemporal Energy Models for the Perception of Motion," *J. Opt. Soc. Am. A* **2:**284–299 (1985).
15. S. A. Klein, E. Casson, and T. Carney, "Vernier Acuity as Line and Dipole Detection," *Vision Res.* **30:**1703–1719 (1990).
16. B. Farell, and D. G. Pelli, "Psychophysical Methods," *A Practical Guide to Vision Research*, J. G. Robson and R. H. S. Carpenter (eds.), Oxford University Press, New York, 1999.
17. N. V. S. Graham, *Visual Pattern Analyzers*, Oxford University Press, Oxford, 1989.
18. H. B. Barlow, "Temporal and Spatial Summation in Human Vision at Different Background Intensities," *J. Physiol.* **141:**337–350 (1958).
19. F. W. Campbell, and J. G. Robson, "Application of Fourier Analysis to the Visibility of Gratings," *J. Physiol.* **197:**551–566 (1968).
20. H. Snellen, *Test-Types for the Determination of the Acuteness of Vision*, London: Norgate and Williams, 1866.
21. D. G. Pelli, J. G. Robson, and A. J. Wilkins, "The Design of a New Letter Chart for Measuring Contrast Sensitivity," *Clin. Vis. Sci.* **2:**187–199 (1988).
22. D. G. Pelli, and J. G. Robson, "Are Letters Better than Gratings?," *Clin. Vis. Sci.* **6:**409–411 (1991).
23. J. C. Dainty, and R. Shaw, *Image Science*, Academic Press, New York, 1974.
24. E. H. Linfoot, *Fourier Methods in Optical Image Evaluation*, Focal Press, New York, 1964.
25. D. E. Pearson, *Transmission and Display of Pictorial Information*, John Wiley & Sons, New York, 1975.
26. O. H. Schade, Sr., *Image Quality: A Comparison of Photographic and Television Systems*, RCA Laboratories, Princeton, N.J., 1975.
27. G. E. Legge, D. G. Pelli, G. S. Rubin, and M. M. Schleske, "Psychophysics of Reading—I. Normal Vision," *Vision Res.* **25:**239–252 (1985).
28. J. M. Rolf and K. J. Staples, *Flight Simulation*, Cambridge University Press, Cambridge, U.K., 1986.
29. D. G. Pelli, "The Visual Requirements of Mobility," *Low Vision: Principles and Application*, G. C. Woo (ed.), Springer-Verlag, New York, 1987, pp. 134–146.
30. J. G. Robson, "Spatial and Temporal Contrast-Sensitivity Functions of the Visual System," *J. Acoust. Soc. Am.* **56:**1141–1142 (1966).
31. R. S. Woodworth and H. Schlosberg, *Experimental Psychology*, Holt, Rinehart, and Winston, New York, 1963, pp. 199–200.
32. P. Cavanagh and S. Anstis, "The Contribution of Color to Motion in Normal and Color-Deficient Observers," *Vision Res.* **31:**2109–2148 (1991).

33. G. A. Brindley, *Physiology of the Retina and the Visual Pathways*, Edward Arnold Ltd., London, 1960.
34. M. A. Georgeson and G. D. Sullivan, "Contrast Constancy: Deblurring in Human Vision by Spatial Frequency Channels," *J. Physiol.* **252:**627–656 (1975).
35. R. M. Boynton, *Human Color Vision*, Holt Rinehart and Winston, New York, 1979, pp. 299–301.
36. W. W. Peterson, T. G. Birdsall, and W. C Fox, "Theory of Signal Detectability," *Trans. IRE PGIT* **4:**171–212 (1954).
37. W. P. Tanner, Jr. and T. G. Birdsall, "Definitions of d' and η as Psychophysical Measures," *J. Acoust. Soc. Am.* **30:**922–928 (1958).
38. H. L. Van Trees, *Detection, Estimation, and Modulation Theory*, Wiley, New York, 1968.
39. W. S. Geisler, "Sequential Ideal-Observer Analysis of Visual Discriminations," *Psychol. Rev.* **96:**267–314 (1989).
40. D. G. Pelli, "Uncertainty Explains Many Aspects of Visual Contrast Detection and Discrimination," *J. Opt. Soc. Am. A* **2:**1508–1532 (1985).
41. D. G. Pelli, "The Quantum Efficiency of Vision," *Vision: Coding and Efficiency*, C. Blakemore (ed.), Cambridge University Press, Cambridge, U.K., 1990, pp. 3–24.
42. D. G. Pelli, C. W. Burns, B. Farell, and D. C. Moore-Page, "Feature Detection and Letter Identification," *Vision Res.* **46**(28)**:**4646–4674 (2006). See Appendix A.
43. J. Nachmias, "On the Psychometric Function for Contrast Detection," *Vision Res.* **21:**215–223 (1981).
44. H. E. Rockette, D. Gur and C. E. Metz, "The Use of Continuous and Discrete Confidence Judgments in Receiver Operating Characteristic Studies of Diagnostic-Imaging Techniques," *Invest. Radiol.* **27:**169–172 (1992).
45. J. A. Swets, "Measuring the Accuracy of Diagnostic Systems," *Science* **240:**1285–1293 (1988).
46. J. Nachmias and R. M. Steinman, "Brightness and Discriminability of Light Flashes," *Vision Res.* **5:**545–557 (1965).
47. D. M. Green and J. A. Swets, *Signal Detection Theory and Psychophysics*, Krieger Press, Huntington, N.Y., 1974.
48. L. W. Nolte and D. Jaarsma, "More on the Detection of One of M Orthogonal Signals," *J. Acoust. Soc. Am.* **41:**497–505 (1967).
49. R. D. Luce, *Response Times: Their Role in Inferring Elementary Mental Organization*, Oxford University Press, New York, 1986.
50. P. E. King-Smith, S. S. Grigsby, A. J. Vingrys, S. C Benes and A. Supowit, "Efficient and Unbiased Modifications of the QUEST Threshold Method: Theory, Simulations, Experimental Evaluation and Practical Implementation," *Vision Res.* **34:**885–912 (1994).
51. A. B. Watson and D. G. Pelli, "QUEST: A Bayesian Adaptive Psychometric Method," *Percept Psychophys.* **33:**113–120 (1983).
52. A. B. Watson, "Probability Summation Over Time," *Vision Res.* **19:**515–522 (1979).
53. D. G. Pelli, "On the Relation Between Summation and Facilitation," *Vision Res.* **27:**119–123 (1987).
54. S. Ishihara, *Tests for Color Blindness*, 11th ed., Kanehara Shuppan, Tokyo, 1954.
55. D. Regan and D. Neima, "Low-Contrast Letter Charts as a Test of Visual Function," *Ophthalmology* **90:**1192–1200 (1983).
56. D. G. Pelli and L. Zhang, "Accurate Control of Contrast on Microcomputer Displays," *Vision Res.* **31:**1337–1350 (1991).
57. D. H. Brainard, D. G. Pelli and T. Robson, "Display Characterization," In J. Hornak (ed.), *Encyclopedia of Imaging Science and Technology*, Wiley, 2002, pp. 172–188.
58. C. Blakemore and F. W. Campbell, "Adaptation to Spatial Stimuli," *J. Physiol.* **200**(1)**:**11–13 (1969).
59. C. Blakemore and F. W. Campbell, "On the Existence of Neurones in the Human Visual System Selectively Sensitive to the Orientation and Size of Retinal Images," *J. Physiol.* **203:**237–260 (1969).
60. T. N. Cornsweet, *Visual Perception*, Academic Press, New York, 1970.
61. F. S. Frome, D. I. A. MacLeod, S. L. Buck and D. R. Williams, "Large Loss of Visual Sensitivity to Flashed Peripheral Targets," *Vision Res.* **21:**1323–1328 (1981).