Farell, B. & Pelli, D. G. (1999) Psychophysical methods, or how to measure a threshold and why. In R. H. S. Carpenter & J. G. Robson (Eds.), *Vision Research: A Practical Guide to Laboratory Methods*, New York: Oxford University Press. http://psych.nyu.edu/pelli/

5

Psychophysical methods, or how to measure a threshold, and why

BART FARELL and DENIS G. PELLI

5.1 Introduction

This chapter explains how to measure visual effects. Psychophysical methods are usually described in a historical context, starting with Weber, Fechner, and Wundt in the 1800s and the development of the theoretical foundations; here we take a practical approach, focusing on what is most useful to know. Drawing conclusions about visual perception is difficult—not all questions are answerable. Psychophysics only considers questions that can be answered by measuring an observer's performance of a visual task. The art of psychophysical measurement is to channel one's curiosity into designing a question that retains the motivating interest and yet can be convincingly answered by measuring task performance. This chapter describes those tasks and measures that have proven to be most useful in vision research, and explains what kinds of question they answer.

Consider the complications in what might seem the simplest question, 'Do you see it?' One can simply present visual signals and put the question directly to the observer. But, on reflection, are we really interested in whether the observer says 'yes', or are we interested instead in whether the observer can prove that he or she has seen the signal, e.g. by correctly identifying it or locating it in time or space? In either case, when we collect the responses we find that the answer is probabilistic: in practice one measures the probability of each kind of allowed response to the signal. But then what does one do with these probabilities? Such complications need to be carefully considered on an experiment-by-experiment basis, but we will share with the reader the guidance offered by existing theory and practical experience about the most generally useful approaches to the most commonly encountered experimental problems.

5.2 Threshold

Probability measures of task performance, e.g. proportion correct, are usually much harder to interpret than the physical parameters of the stimulus. For example, theories of visual acuity directly relate known optical and anatomical properties of the eye—physical parameters—to the size of identifiable letters of an eye chart—another physical parameter—but would require speculative ancillary assumptions in order to predict the proportion correct. Consequently, the experimenter will almost always want to measure a 'threshold'. *Threshold* is the strength of the signal, as controlled by a particular stimulus dimension, that is required to attain a given level of task performance.

Fundamentally, there are two kinds of task that are used to obtain thresholds: adjustment and classification (Pelli and Farell, 1995). In *adjustment tasks* the observer is asked to adjust a knob controlling the stimulus to achieve some verbally described criterion, e.g. 'so you can just barely see it' or 'so it matches the standard stimulus'. Here, the observer directly sets the physical parameter of the stimulus. In *classification tasks* the observer is merely asked to identify the signal by placing it in one of a number of predetermined categories, e.g. 'is the screen displaying a pattern or a blank?' or 'is the test stimulus larger or smaller than the standard?' Repeated testing of the classification of a set of stimuli varying in signal strength measures the proportion of the identification responses to each stimulus that were correct, but it is usually most useful to find the threshold value of the stimulus parameter that would yield a certain proportion correct. So, in practice, both adjustment and classification tasks are used to estimate the threshold value of the signal parameter, i.e. the value that achieves a specified criterion, subjective in the case of adjustment, objective in the case of classification.

Thus, while one can imagine a wide variety of questions that might reasonably be asked to obtain a measure of psychophysical performance, the most useful methods that current vision science has to offer, and the most widely practised, are those that measure threshold.

5.3 Adjustment

In the days when vision research labs used analog function generators to synthesize their stimuli, it was very easy to continuously display a stimulus while adjusting it. Today, digital computers synthesize a vastly greater range of images to be used as stimuli, but, unfortunately, it takes some effort to get a computer dynamically to recompute the stimulus in response to the observer's adjustments. Nevertheless, computers are getting faster, and there are shortcuts to synthesizing certain stimuli. When feasible, adjustment tasks offer a quick direct measurement of a subjective match between the variable stimulus and a standard. Because of the subjective nature of the adjustment settings, this method is ideally suited for experimenters using themselves as observers, allowing them to quickly experience the full range of effects of a stimulus parameter on appearance.

When testing others, the instructions given to the observer are crucial. They are so crucial that if a published experiment relies on the method of adjustment, then the discussion section should convince the reader that the instructions used indeed bear on the aspect of perception that is the nominal topic of the paper. *Matching* instructions are particularly easy for observers to understand and are the most commonly used. In the matching paradigm, there are two objects, a standard and a test, and the observer is asked to adjust the test object to 'match' the standard. The criterion for matching is all important—what one asks the observer to do and what the observer actually does are not at all the same thing—and should not only be conceptually clear, but also, if at all possible, perceptually salient. For this reason, a particularly effective matching instruction is 'nulling'. This applies to cases where one presumes that the observer understands what the stimulus 'ought' to look like when it is 'undistorted' or 'neutral' and adjusts it to achieve that appearance. For example, adjusting a patch to eliminate any colour or motion or pattern; or adjusting a line to be straight.

Consider the influence of form on brightness and how it might be quantified by brightness matching. Benary (1924) and Adelson (1993) showed that the brightness of a surface depends on the perceived object structure. They presented two fairly similar images made up of contiguous uniform patches arranged to produce different three-dimensional interpretations: one patch in one image was adjustable, and the observer was asked to adjust its luminance to match the brightness of a particular patch in the other image. To explain brightness, as opposed to lightness, Adelson asked his observers to 'judge the shade of ink on the page' rather than make any inference about the surfaces of the objects portrayed.

In an effective use of the nulling technique, Cavanagh and Anstis (1991) employed motion nulling to measure the contribution of colour to the perception of motion. They showed observers a rightward-moving luminance grating superimposed on a leftward-moving chromatic grating. The observers adjusted the contrast of a second luminance grating, moving leftward in phase with the chromatic grating, to null the motion of the entire pattern. The difference between the contrasts of the leftward and rightward luminance gratings is then a measure of the contribution of colour to perceived motion.

5.4 Classification

At present, three kinds of classification are widely used: yes/no, two-alternative forced choice (2afc), and identification. Each task asks the observer to reply to a query: 'Did you see it?' (yes/no); 'Was the signal in the first or in the second interval?' (2afc); or 'Which signal was it?' (identification). All three call for the observer to classify stimuli (or their subjective responses). Those 2afc tasks that present a signal and a blank on each trial are said to be 'detection' tasks. In a 'discrimination' task, the signal is added to a constant background stimulus that appears in both intervals. Yes/no, 2afc, and identification all have their special niches, and all three tasks have been used to measure thresholds and convincingly establish important scientific conclusions. All other things being equal, however, readers who value their time will use identification if possible, otherwise 2afc, and yes/no only as a last resort.

5.4.1 Yes/no

If one must, then with some effort a *frequency of seeing curve* can be measured, which plots the probability of saying 'yes' to the question—'Did you see it?'—as a function of a stimulus parameter (e.g. contrast). Unfortunately, the observer in a yes/no experiment can't avoid introducing an internal subjective criterion in deciding whether each faint ambiguous percept deserves a 'yes' or a 'no'. The observer's personality, the instructions, and other experimental details may all affect the internal criterion, and thereby threshold. This thorny problem has been thoroughly analysed and various remedies have been devised, of which we endorse only one, though it is by far the most time-consuming. Theory of signal detectability shows that the frequency of seeing or 'hit rate' is uninformative unless one also measures the *false-alarm rate*, i.e. the probability of saying 'yes' when a blank is presented. By systematically changing the instructions one can push the observer's criterion up or down, and measure both the hit and false-alarm rates at each criterion level. This yields an ROC (receiver-operating characteristic) graph of hit vs. false-alarm rate, parameterized by the (unknown) internal criterion. The area under the ROC curve is an excellent measure of the visibility of the signal (Swets and Pickett, 1982). The ROC curve can be obtained more quickly by asking the observer to give a 1-to-5 confidence rating instead of merely saying yes or no, but this still entails a substantial effort on the part of the experimenter to collect and analyse the results to obtain the ROC area.

5.4.2 Two-alternative forced choice

The 2afc task gives the observer one of two stimulus arrangements and asks the observer to identify which it is. The advantage of the 2afc task is that it can be designed to avoid criterion effects by presenting a symmetric unbiased choice. Typically, this is achieved by having two stimulus arrangements both containing, say, a signal and a blank, which differ solely by the interchange of signal and blank. One might present the signal in a first interval and the blank in a second interval, or vice versa (randomly), and ask the observer in which interval the signal was presented ('two-interval forced choice'). Or one might present them simultaneously, side by side, and ask the observer on which side the signal is. Under mild theoretical assumptions, the measured proportion correct will equal the ROC area described in the previous section (Green and Swets, 1974; Nachmias, 1981), thus obtaining a similar result with much less effort.

This is an excellent technique. Its only drawback is that, because the observer has only two alternatives and thus will be right half the time even if the signal is invisible, a relatively large number of trials (about 60) is required to obtain a good threshold estimate. For greatest efficiency in 2afc tasks one should use a sequential estimation procedure (described below) to adjust the signal strength systematically and estimate threshold directly. Or, with somewhat more effort, one can measure the proportion correct as a function of signal strength (the 'psychometric' function). In each trial, one strength value is presented from a fairly small number (usually 5 or 7) that span the performance range (50–100%). Then the threshold for any level of performance can be read off the psychometric function. The shape of this function, like the shape of the ROC curve, can also be used to infer the distributions of internal stimulus representations on which decision processes operate (see Graham, 1989). However, usually you will just want to know threshold.

5.4.3 Identification

An even more efficient method is to present one of many signals and ask the observer to identify it. How many? Simulations show that four (or more) alternatives suffice to achieve a high efficiency accruing from minimizing the chance of blind guessing (Pelli *et al.*, 1988). Theoretical consideration of the ideal observer suggests that, in order to obtain a steep psychometric function (to estimate threshold quickly), the signals should all have approximately equal contrast energy and similar, pairwise cross-correlation (van Trees, 1968). Of course, observers must learn to identify the signals. Experiments involving the identification of foreign and novel alphabets show that observers learn to identify new symbols quickly, requiring only 2000 trials to attain the same threshold for letter identification as fluent readers of the alphabet (Pelli *et al.*, 1998). The observer's responses implicitly divide the high-dimensional stimulus space into many regions (one per kind of response) separated by category boundaries. In principle, these category boundaries are subjective and movable, like the observer's internal criterion in the yes/no task, but, on the one hand, theoretical understanding of the high-dimensional case is still wanting (Ashby, 1992), so there is nothing one can do about it, while, on the other hand, the problem is less worrisome, because high probabilities of correct identification are not attainable by blind guessing.

5.4.4 Sequential estimation: QUEST

Given that the experimenter is willing to run a reasonable number of trials (e.g. 40), and has some prior knowledge of the psychometric function and its parameters, one would like an efficient procedure for threshold estimation—a procedure for running each trial at whatever signal strength would contribute most to minimizing the variance of the final threshold estimate. Such a procedure combines the experimenter's prior knowledge and the observer's responses on past trials in choosing the signal strength for the next trial, and, at the end, estimating threshold. The best current procedure is called ZEST (King-Smith *et al.*, 1994), but QUEST (Watson and Pelli, 1983), which is nearly as efficient, can be implemented by a tiny C program, which we present below.

The only unknown is threshold, which is treated as a random variable, X, to be estimated. The experimenter supplies an initial guess, by specifying the mean and SD of a Gaussian probability density function. For the reader's convenience, we supply a one-line simulation of an observer with threshold *tActual*, so the program can be run on its own. To run a real experiment, that line must be replaced by code that presents a stimulus (at intensity x) and collects the observer's response (1 if right, 0 if wrong). After each response, the probability density function, q, is updated by Bayes's rule. Each trial is placed at x, the current maximum-probability estimate of threshold, i.e. the mode. The final threshold estimate is also the mode.

```
#include <math.h>
#include <stdio.h>
#include <stdlib.h>
#define DIM 400
#define DIM2 (2*DIM)
#define GRAIN 0.01
#define xi(i) (((i)-DIM/2)*GRAIN)
#define xii(ii) (((ii)-DIM2/2)*GRAIN)
#define iix(x) (int) (0.5+DIM2/2+(x)/GRAIN)
void main(void)
```

```
{
  float p[DIM2+1],s[2][DIM2+1],q[DIM+1];
  int i,ii,trialsDesired=40,k,imode,right;
  double beta=3.5,delta=0.01,gamma=0.5; /* parameters of psychometric
   function */
  double x,tGuess=-2.0,tGuessSD=4.0,tActual;
  char wrongRight[2][]={"wrong","right"},string[64];

  for(ii=0;ii<=DIM2;ii++){
    p[ii]=delta*gamma+(1-delta)* (1-(1-gamma)*exp(-pow(10,beta*xii(ii))));
    s[0][DIM2-ii]=log(1-p[ii]);
    s[1][DIM2-ii]=log(p[ii]);
  }
  for(i=0;i<=DIM;i++){
    x=xi(i)/tGuessSD;
    q[i]=-0.5*x*x;
  }
  printf("Estimate threshold:");
  gets(string);
  sscanf(string,"%lf",&tGuess);
  printf("Specify true threshold of simulated observer:");
  gets(string);
  sscanf(string,"%lf",&tActual);
  for(k=1;k<=trialsDesired;k++){
    for(imode=0,i=0;i<=DIM;i++)if(q[i]>q[imode])imode=i;
    x=xi(imode)+tGuess;
    /* to test a real observer, */
    /* replace the next line with your experimental task */
    right=p[iix(x-tActual)] > rand()/(RAND_MAX+1.0);
    printf("Trial %3d at %4.1f is %s\n",k,x,wrongRight[right]);
    for(i=0;i<=DIM;i++)q[i]+=s[right][i-imode+DIM2/2];
  }
  for(imode=0,i=0;i<=DIM;i++)if(q[i]>q[imode])imode=i;
  x=xi(imode)+tGuess;
  printf("Final (mode) threshold estimate is %4.1f\n",x);
}
```

5.5 Reaction time

Threshold is the stimulus strength required for a specified probability of correct decision. We typically assume that the observer is making a simple decision about a single elementary stimulus. In many practical situations, however, people do not respond on the basis of a single elementary decision, but only after making multiple decisions about the many stimuli present in a complex display. Searching for a particular face in a crowd is a familiar example. Thresholds could be found for tasks like this, but often the interest is in different types of questions than threshold measures answer. A researcher might be interested in measuring how long it takes to perform a task, or in analysing the theoretically more challenging question of how the component decisions leading up to a task response are distributed in time (Sternberg, 1969). Response times can be measured in any task, but one must not forget to measure response accuracy at the same time, because of trade-offs between the two (e.g. McElree and Dosher, 1989).

5.6 Devilish details

Having decided on a task, implementing it can bring a flood of new questions. Some are easy, e.g.: 'How many threshold estimates?' Enough to make the standard error small. Many other questions are harder, with answers that depend on details of your theory and experiment. The issues include practice, who triggers the trial (experimenter or observer), cueing (to warn of impending stimulus), manner of response (button, speech), allowing the observer to not respond ('I blinked and missed it'), feedback ('right'), and frequency of rest breaks. In general, you should look for the easiest way to obtain a convincing answer to your experimental question. You'll want to be very sure that the task is obvious to the observers. Counting on the observer's intelligence to figure out what the task really is invites huge individual differences in the results that are probably unrelated to the perceptual questions you are really interested in.

References

Adelson, E. H. (1993). Perceptual organization and the judgment of brightness. *Science*, **262**, 2042–4.

Ashby, F. G. (1992). *Multidimensional models of perception and cognition*. Lawrence Erlbaum, Hillsdale, NJ.

Benary, W. (1924). [The influence of form on brightness contrast, translated in Ellis, 1938] Beobachtungen zu einem Experiment über Helligkeitskontrast. *Psychologische Forschung*, **5**, 131–42.

Cavanagh, P. and Anstis, S. (1991). The contribution of color to motion in normal and color-deficient observers. *Vision Research*, **31**, 2109–48.

Ellis, W. D. (1938). *A source book of gestalt psychology*. Harcourt, Brace, and Co., New York.

Graham, N. V. S. (1989). *Visual pattern analysers*. Oxford: Oxford University Press.

Green, D. M. and Swets, J. A. (1974). *Signal detection theory and psychophysics*. Krieger, Huntington, NY.

King-Smith, P. E., Grigsby, S. S., Vingrys, A. J., Benes, S. C., and Supowit, A. (1994). Efficient and unbiased modifications of the QUEST threshold method: theory, simulations, experimental evaluation and practical implementation. *Vision Research*, **34**, 885–912.

McElree, B. and Dosher, B. A. (1989). Serial position and set size in short-term memory: Time course of recognition. *Journal of Experimental Psychology: General*, **118**, 346–73.

Nachmias, J. (1981). On the psychometric function for contrast detection. *Vision Research*, **21**, 215–23.

Pelli, D. G. and Farell, B. (1995). Psychophysical methods. In *Handbook of optics* (2nd edn), Vol. I (ed. M. Bass, E. W. Van Stryland, D. R. Williams, and W. L. Wolfe), pp. 29.1–29.13. McGraw-Hill, New York.

Pelli, D. G., Robson, J. G., and Wilkins, A. J. (1988). The design of a new letter chart for measuring contrast sensitivity. *Clinical Vision Sciences*, **2**, 187–99.

Pelli, D. G., Burns, C. W., Farell, B., and Moore, D. C. (1998). Identifying letters. *Vision Research* (In press.)

Sternberg, S. (1969). The discovery of processing stages: extensions of Donder's method. In *Attention and performance* II (ed. W. G. Koster), pp. 276–315. North-Holland, Amsterdam.

Swets, J. A. and Pickett, R. M. (1982). *Evaluation of diagnostic systems: methods from signal detection theory*. Academic Press, New York.

Van Trees, H. L. (1968). *Detection, estimation, and modulation theory*. Wiley, New York.

Watson, A. B. and Pelli, D. G. (1983). QUEST: a Bayesian adaptive psychometric method. Percept *Psychophys*, **33**, 113–20.