Chapter 4

Perceptual Psychophysics

BibTex entry

This chapter appeared as:

```
@inbook{Parraga15sparse,
Author = {Parraga, C. Alejandro},
Title = {Perceptual Psychophysics},
Chapter = {4},
booktitle = {Biologically inspired computer vision},
Year = {2015}
Editor = {Crist{\'{o}}bal, Gabriel and Keil, Matthias S. and Perrinet, Laurent U.},
month = nov,
isbn = {9783527680863},
DOI = {10.1002/9783527680863.ch4},
url={http://onlinelibrary.wiley.com/doi/10.1002/9783527680863.ch4/summary},
publisher = {Wiley-VCH Verlag GmbH {\&} Co. KGaA},
}
```

4

Perceptual Psychophysics

1.1 Introduction

Since only a few decades ago, and in particular since the arrival of functional magnetic resonance techniques, there has been an explosion in our understanding of the functional brain structures and mechanisms that result in our conscious perception of the world. Although we are far from a complete understanding of cortical and extra-cortical processing, we know a fair amount of detail, perhaps enough to create computational models that replicate the workings of the best understood parts of the brain. We also know a fair amount about the environment and the evolutionary/neo-natal constraints that shaped the workings of the perceptual machinery so that, even when the information reaching us is incomplete and subject to noise, our perception of the world is mostly stable: our brain filters out variability and attempts to recreate the environmental information based on millions of years of evolution and its own life experience. All this is done within the neural signal processing framework that characterises our internal processes and drive our subjective responses. In this context, applying hard science methods to study how we humans reach subjective judgements has the appearance of an art or a craft, where each decision needs to be carefully weighted, experience is crucial and small mistakes might render experimental results impossible to interpret.

1.1.1 What is psychophysics and why do we need it?

Psychophysics is the branch of experimental psychology concerned with the relationship between a given stimulus and the sensation elicited by that stimulus. It engages problems by measuring the performance of observers in predetermined sensory (visual, auditory, olfactory, haptic, etc.) tasks. It has two remarkable roles, to describe and specify the sensory mechanisms underlying human performance and to test the hypotheses and models explaining these same mechanisms. Because of its nature, psychophysics inhabits two worlds; the world of physics where objects and interactions can be measured with ever increasing accuracy and the world of psychology,

where human performance is irregular and answers are partially determined by stochastic neural response processes. As a discipline, psychophysics was born from the need of psychologists to empirically study sensorial processes as a tool to understand more complex psychological processes. Its 19th century origins are intertwined with those of experimental psychology (it occupies a very central position in that discipline) and its theoretical foundation dates from the publication of *Elemente der Psychophysik* by German physicist Gustav Fechner (Fechner, 1860). In this book, Fechner developed the theoretical and methodological background required to measure sensation and gave experimental psychology the tools necessary to begin its study of the mind. Subsequent progress such as the *theory of signal detection* have broadened the scope of psychophysics, enabling its outstanding contributions to such different areas as visual perception, language processing, memory, learning, social behaviour, etc.

Although it is generally possible to accurately measure the physical properties of a given stimulus, task performance measures are much more difficult to interpret and are usually expressed in terms of probabilities. To simplify the problem, the first psychophysical studies consisted in modifying a stimulus' strength until the subjects perceived it and responded behaviourally. These measurements are called "threshold measurements", and reveal the strength of a signal necessary to cause a determined performance. Different set-ups are needed to quantity performances when threshold measurements cannot be obtained or when the physical characteristic of interest are not properly defined.

Psychophysical results can be directly related to the output of computational models, since models predict observers' performance given a stimulus signal. For example, if a visual difference model predicts that picture A and picture B are indistinguishable, we can design a psychophysical experiment where the same images are presented to human observers in a setup where they have to distinguish them. Moreover, visual psychophysicists are not content with just asking observers whether they see any differences between the pictures, since such answers would be tainted by expectations and subjectivity. They will design a task that can only be performed if observers do distinguish between the pictures. The ability to design, execute and interpret the results of such tests is the main skill of a psychophysicist.

1.2 Laboratory methods

Psychophysical experiments are usually conducted in a laboratory where all stimuli are controlled and their properties quantified. All laboratory settings (chairs, temperature, illumination, etc.) should be carefully considered to

make subjects comfortable while keeping them awake and fully attentive. Visual psychophysics setups are particularly complex since illumination should not interfere with the visual stimulus and in most cases; light reflected from other objects and the walls should be minimized. All relevant dimensions like stimulus luminance, its distance to the subject, its subtended visual angle, the observer's head position, etc. often need to be recorded and taken into account in the evaluation of the experiment.

1.2.1 Accuracy and precision

Experimenters dealing with physical magnitudes need to know the notions behind estimating measurement uncertainty (measurement errors) and error propagation. Every measurement has an associated set of confidence limits, denoting its proximity to the true physical value (also called its *accuracy*). *Precision* of a measurement of a physical variable refers to its reproducibility: I can use a calliper to measure very accurately the size of a grain of sand, but for bigger samples (e.g. a handful from my local beach) this measure will vary greatly from one grain to the next. In this case, the largest error is linked to the statistical probability of obtaining a certain size when randomly picking grains of sand from the sample. This uncertainty is called the *absolute error* of the measure and in our case is defined as the *standard deviation* of all the grains. Suppose that we want to have an estimation of the size of the grains: to be precise we have to measure all of them, calculate the mean $\bar{m}$ and the standard deviation s and express the size as $\bar{m} \pm s$. The second term of this expression s is the absolute error (also called Δm) and denotes the confidence interval of our measure: the size of a randomly picked grain has a 68% probability to lie within the region $[\bar{m} - s, \bar{m} + s]$.

Measuring length provides a good example of how to deal with these concepts: suppose we want to measure the length of a table with a ruler whose smaller unit is a centimetre. After we set one end of the table against the zero of the ruler, it is likely that the other end will show up between any two marks so we would "round" our results to the closest mark. In this case my confidence level is the size of the largest "rounding", which is half a cm. If the border is closest to the 91 cm mark I can express the size of the table as 91 ± 0.5 cm (the real length of the table is somewhere between 90.5 and 91.5 cm).

It is a common convention in science to use significant figures to express accuracy (the last significant place in our 91 cm measure is the 1 cm unit). When the error is not explicitly stated, it is understood to be half of the last significant value (in our case, half a cm). Using this convention, our measure could be simply written as 91 cm, and adding more significant

places, e.g. writing 91.00 cm is confusing, since it also states that our precision was less than a hundredth of a centimetre.

If I am given a better rule (let's say with millimetre marks) and asked to measure the width of the table, I can apply the same criterion and obtain a second measurement, e.g. 612 ± 1 mm. This time I will adopt 1 mm as my uncertainty (my eyesight does not allow me to see anything smaller than that). In summary I could safely state that my table is 91 ± 0.5 cm long and 61.2 ± 0.1 cm wide.

1.2.2 Error propagation

Now, suppose we want to express the perimeter of the table. It is definitely the sum of all four sides ($91 + 91 + 61.2 + 61.2 = 304.4$ cm), but what about its uncertainty? In our case it is easy; in the event of sum and subtraction of two independent measurements we can estimate uncertainty using eq. 1:

$$\begin{cases} L = a + b \\ L = a - b \end{cases} ; \qquad \Delta L = \Delta a + \Delta b. \tag{1}$$

where Δa and Δb represent the absolute error of each of the components of the sum. According to eq. 1, the combined uncertainty is $0.5 + 0.5 + 0.1 + 0.1 = 1.2$ cm. The perimeter may be expressed as 304.4 ± 1.2 cm. However, if we follow our convention, by writing the last decimal (number 4) as the last significant digit we imply that we can estimate millimetre units although our error is twelve millimetres! The correct approach is to round our last figure and express the perimeter as 304 ± 1.2 cm. Note that in this case the error associated to the length predominates over the error associated to the width, making the millimetre ruler rather useless.

Now, let's calculate and write the area of the table. A simple product of length times width gives an area of 5569.2 cm^2 however; to calculate its uncertainty we need to apply the following rule:

$$\begin{cases} A = a.b \\ A = \frac{a}{b} \end{cases} ; \qquad \frac{\Delta A}{A} = \frac{\Delta a}{a} + \frac{\Delta b}{b}. \tag{2}$$

where the terms $\Delta a/a$ and $\Delta b/b$ are called *relative uncertainties* of *a* and *b*. Eq. 2 applies to independent measures and describes a method for calculating uncertainty in the cases of multiplication and division by operating with relative uncertainties. In our example, the absolute uncertainty of the table surface *A* is calculated by replacing in eq. 2:

$$\Delta A = A \cdot \left(\frac{\Delta l}{l} + \frac{\Delta w}{w}\right) = \quad 5569.2 \cdot \left(\frac{0.5}{91} + \frac{0.1}{61.2}\right) = \quad 39.7\ cm^2. \quad (3)$$

therefore the error associated to A is approximately 40 cm^2, which makes the last three digits of 55<u>69.2</u> of little value. In view of this we can round them up and write $A = (55.7 \pm 0.4) \cdot 10^2 cm^2$, indicating our confidence limits.

In the general case, e.g. when our final measure x is a function of many dependent variables (a, b, c, etc. also measured with their respective Δa, Δb, Δc, etc) the absolute error Δx can be derived from the covariance matrix of the variables [1; 18]. In the case where fluctuations between a, b, c, etc are mostly independent and the number of observations is large, we can expect eq. 4 to be a reasonable approximation:

$$Given: \quad x = f(a, b, c, \ldots);$$

$$|\Delta x| = \left|\frac{\partial f}{\partial a}\right| \cdot |\Delta a| + \left|\frac{\partial f}{\partial b}\right| \cdot |\Delta b| + \left|\frac{\partial f}{\partial c}\right| \cdot |\Delta c| + \cdots; \qquad (4)$$

Although eq. 4, neglects the effects of correlations between the different variables, it is commonly used for calculating the effects of uncertainties on the final results.

1.3 Psychophysical threshold measurement

Once we know how to properly quantify all relevant physical magnitudes; we need to turn our attention to the problem of measuring sensation. The minimum amount of stimulus energy necessary to elicit a sensation is called *absolute threshold*. Given that our neural system is subject to noise, sensation measurements tend to fluctuate from a moment to the next and several of them are necessary to obtain an accurate estimation by averaging. A second important concept in psychophysics is the *difference threshold* or the amount of change in the stimulus necessary to elicit a *just noticeable* increment in the sensation (just noticeable difference or *jnd*). For example, if I close my eyes and hold a bunch of paperclips in my hand, how many paperclips can you add until I notice a change in weight? Let's say 2 paperclips: their combined weight is the difference threshold.

1.3.1 Weber's law

One of the first stimulus-sensation dependencies to be discovered was the relationship between difference threshold and the intensity level of a stimulus. Going back to the paperclips example, suppose that I hold 50 paperclips in my hand and we find out that the minimum number of

paperclips necessary for me to notice a change in weight is 2. Is it going to be the same if I hold 150 paperclips in my hand? And if I hold 300? The answer was found by the German physiologist E.H. Weber in 1834 who determined that the amount of difference threshold is a linear function of the initial stimulus intensity [23]. In summary, if I hold 150 paperclips in my hand I will need about 6 more to notice any change in weight. *Weber's law* can be expressed mathematically as:

$$\Delta\omega = c\omega; \qquad or \qquad \frac{\Delta\omega}{\omega} = c \ . \qquad (5)$$

where ω is the initial intensity of the stimulus, $\Delta\omega$ is the amount of change required to produce a just noticeable difference and c is a constant dependent of the sensory modality studied. The relationship in eq. 5 holds for a wide range of stimulus intensities (except for very weak stimuli) and appears to be true for many types of signals: visual, auditive, tactile, olfactory, etc. At very low levels of stimulation, the noise inherent to our neural system is of a magnitude similar to the stimulus necessary to produce a *jnd*, and a larger $\Delta\omega$ is needed for it to rise above the noise. Fechner [3] extended Weber's ideas by relating physical magnitudes of $\Delta\omega$ to their corresponding sensation *jnds*, assuming that all *jnds* where psychophysically equal. After a series of measurements, he came up with the following equation (known as *Fechner's law*):

$$S = k \cdot \log(\omega) \ ; \qquad (6)$$

where *S* is the magnitude of the sensation experienced; ω is stimulus intensity and *k* is a constant that also depends on sensory modality. Fechner's law can be obtained by writing Weber's law as a differential equation and integrating over $d\omega$ and both constants are be related by $c = -k \log(\omega_0)$, where ω_0 is the stimulus threshold below which there is no perceived sensation.

Later studies have challenged the validity of Fechner's assumption on the uniformity of *jnd* increments for any level of stimulus sensation, however, his equation continues to be one of the most significant attempts to measure sensation in the history of science.

1.3.2 Sensitivity functions

The equations above belong to the small group of laws discovered in perceptual psychology, and their historical relevance cannot be underestimated. However, it is perhaps the measurement of *absolute thresholds* which is responsible for the most significant advances in our

understanding of sensory mechanisms. Take the case of vision, where sensory thresholds are determined by properties of both, the optics of the eye and the neural machinery of the visual system. Visual sensation starts at the retina, where photoreceptors (i.e. cones and rods) are excited by photons, sending signals that in the case of rods, are spatially pooled together (*spatial summation*) and summed over time (*temporal summation*). In many aspects their performance seems to match the limits imposed by the nature of light. For example, Hecht et al [9] tested 7 observers for several months and concluded that about 56 to 148 photons are necessary to elicit a sensation 60% of the time in absolute darkness and optimal conditions. After considering the optics of the eye [10], Hecht et al estimated this to be equivalent to just 5 to 14 photons absorbed by a single rod photopigment. If a rod absorbed a smaller number of photons, the resulting signal could be confounded with signals due to spontaneous activity and therefore the output from many rods need to be pooled together. The interaction between signal and noise in the visual system is studied with the tools of signal detection theory and tasks such as visual discrimination and visual detection become inherently similar in the presence of noise: to detect a signal is the same as discriminating it from the noise.

1.4 Classic psychophysics: theory and methods

Psychophysics has the dual role of providing us with quantitative descriptions of the limits, features and capabilities of our senses and of offering ways of testing our hypothesis about the neural mechanisms behind these sensory capabilities. In the first role, it has contributed to the design of rooms and equipment optimised for human use such as auditoriums, concert halls, digitally compressed audio-visual imagery, ergonomically designed cabins and cockpits, LCD screens, etc. In the second role of testing hypothetical neural mechanisms and models, psychophysicists must assume a biunivocal correspondence between the neural activity generated by a stimulus and its perception. This principle (which has been termed "*Principle of Nomination*" [12]) is what allows us to link psychophysical to biological data and computer models and is perhaps the single most relevant fact relating psychophysics and computer science. Computer models often attempt to match the performance of human observers in predetermined tasks, either by recreating the workings of neural processing ("mechanistic models") or by other means ("functional models"). In mechanistic models, the need for psychophysics and signal detection theory is clear: they provide the tools to infer the properties of neural mechanisms from observer's data. Functional models often rely in pattern inference theory to provide statistical descriptions of the scene structure, but there is always the final

requirement of contrasting results against observer's performance. In the end, both approaches need to relate a measure of the observer's internal state (e.g. the perceived image as represented in the electronic machinery) to a physical quantity (e.g. the real image): this is what psychophysics does best though the Principle of Nomination.

1.4.1 Theory

To fully appreciate the power behind the Principle of Nomination we should consider two different stimuli: if they produce identical neural response, their perceptual effects should be the identical too. The reflexive form of the principle states that, if the sensory experience from two stimuli is identical, the neural responses should also be the same. This second form of the principle allows us to hypothesise and model the neural events that underlie our sensory experience by studying combinations of stimuli that elicit identical responses. It also allows us to combine findings from different disciplines such as neurophysiology, computational neuroscience and psychophysics by linking physics to neural response and sensation. Furthermore, it is possible to classify psychophysical experiments according to their compliance with the reflexive principle of nomination [2] (i.e. by whether their stimuli are aimed at producing identical responses in the observer or not). One example of full compliance is the case of matching the surfaces of real objects to those produced by a computer monitor. Colour monitors work by exploiting a property called "metamerism", i.e. colours with different spectral power distributions are perceptually similarity under certain conditions. Although the light produced by the three phosphors on the monitor screen has different properties from the light reflected by objects, monitors are routinely used to represent objects in psychophysical experiments and simulations. The "non-compliant" type of psychophysical experiments uses stimuli that are perceived as different except in a single perceptual dimension (e.g. chromaticity, brightness, speed, loudness, etc.) For example, experimenters may allow observers to manipulate the physical properties of one stimulus so that it matches the brightness of another, while they are still different in all other perceptual dimensions like chromaticity, texture, etc. These observations are less rigorous than "fully compliant" observations and their implications are limited (since there is no certainty that all neural responses are the same). However, they are still valid and provide the basis for many hypothesis-testing experiments.

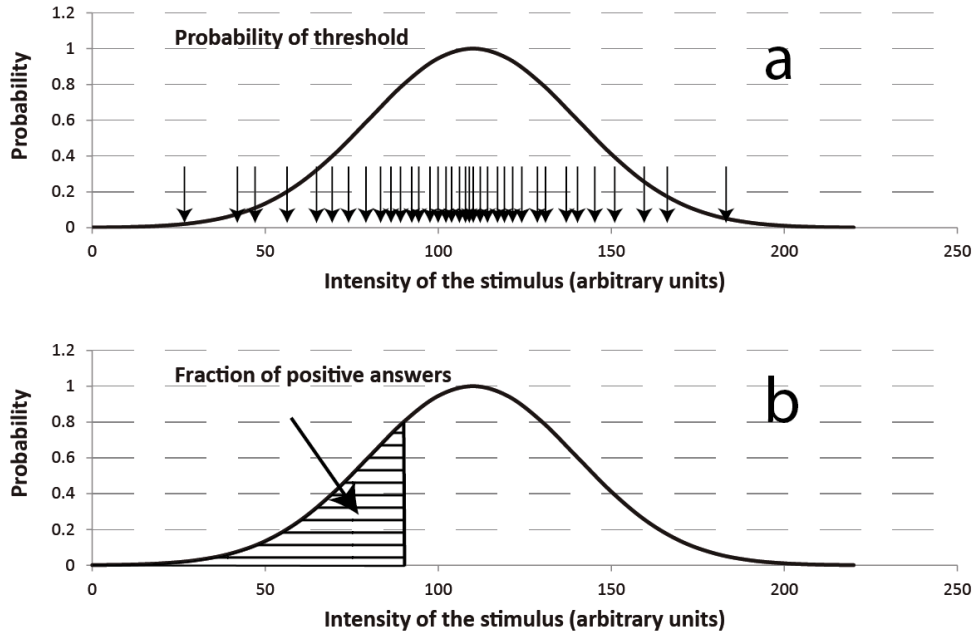

Figure 1: Gaussian distribution of perceived thresholds after a number of measures. This variability is the result of neural noise present in biological systems, experimental noise and in certain cases, quantal fluctuations of the stimulus. According to this, every time we look at the stimulus our "internal" threshold is different and we respond accordingly. These internal thresholds follow a normal distribution with a peak in the most probable value as shown in panel (a). For any given stimulus intensity, the ratio of positive to negative answers to the question "do you see the stimulus?" will be determined by these internal threshold fluctuations and corresponds to the area under the Gaussian shown in panel (b).

Asking observers direct questions on whether or not they perceive a given stimulus while varying its intensity is the most basic psychophysical method for measuring thresholds. In many cases, the answers are likely to contain all the uncertainties common to experiments involving biological systems: when presented several times with the same stimulus our "internal detector" will vary according to stochastic noise. For this reason, it is clear that thresholds need to be defined in terms of their statistical probability: in most cases a *threshold* is the stimulus level that is perceived in 50% of the trials. Figure 1 represents the results from a hypothetical threshold-measurement experiment where arrows in panel (a) denote the variability of "internal thresholds" in our neural system. The probability of the presence of these internal thresholds follows a normal (Gaussian) distribution and the area under the Gaussian in panel (b) represents the percentage of positive answers for any given stimulus. The peak of the Gaussian corresponds to the stimulus that elicits 50% of positive responses in our observers: half the time observers will say "no" because noise will shift the internal threshold above that value and half the time they will say "yes" because noise will

shift the internal threshold to the same value or below. The peak of the Gaussian meets our definition of threshold.

Fechner devised three different methods with their corresponding data analysis techniques for threshold measurement (Fechner, 1860). These methods, whose properties make them useful in different situations, are described below.

1.4.2 Method of constant stimuli

As its name indicates, the set of stimuli presented to the observer in the *method of constant stimuli* is always the same. The procedure starts by selecting stimuli, usually between 5 to 9 samples that range from the easily detectable to the almost impossible to detect. Their intensity (measured in physical units along some dimension) should be separated by equal steps. A simple experiment consists in presenting all the stimulus intensity steps to our observers in random order, repeating each presentation a number of times (e.g. 10 times) and noting the fraction of positive answers to the question "do you see the stimulus?" For large numbers of measurements, results tend to have a typical sigmoidal shape called *psychometric function* (see Figure 2) and are often best fit by a mathematical function called *ogive* which represents the area below a bell-shaped normal distribution curve.

Following our definition of threshold as the stimulus level that is detected in 50% of the trials, we just need to find the position of the curve where the ordinate y (fraction of positive answers) is equal to 0.5 and project to the abscissa x (stimulus intensity). Since the curve in Figure 2 is the integral of a normal distribution, the middle point of its "S-shape" (the threshold magnitude or where 50% of the answers are positive) corresponds to the peak of the Gaussian in Figure 1. Slimmer and taller bell-shaped Gaussians will result in steeper threshold transitions in Figure 2 and flattish Gaussians will result in smoother threshold transitions. The choice of unit measurement for the stimulus intensity is very important, since the abscissae should be measured in a linear scale. For example, monitors produce stimuli whose brightness is a non-linear function of the grey-level values stored in memory (the so called "Gamma function"). For this reason, specifying the stimulus in terms of grey-level values is unlikely to produce curves like those of Figure 1 and Figure 2.

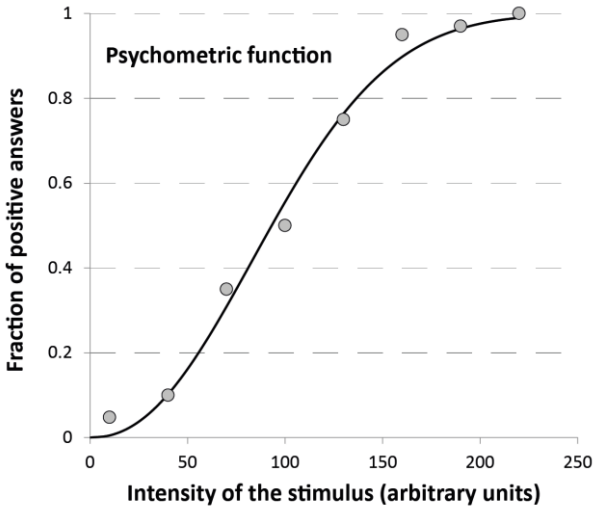

Figure 2: the psychometric function of a typical set of results using the method of constant stimuli. *Y*-axis values show the proportion of "yes" answers to the question "do you see the stimulus?" Each stimulus was presented an arbitrary number of times *n* (typically *n*=10) and each grey circle is the average result of these presentations.

Paradigms for measuring detection thresholds can be easily modified to measure *difference thresholds* between pairs of stimuli (i.e. observers are exposed to two stimuli and have to decide which produces the sensation of greatest magnitude). In this case, one of the stimuli is assigned the role of "reference" and the other the role of "test" stimulus. While the reference stimulus is kept constant, the test stimulus varies from trial to trial from values below those of the reference to values above in constant steps (usually 5 to 9). The test stimuli range need to be chosen so that the stimulus of lowest magnitude is always judged smaller than the reference, the stimulus of largest magnitude always judged larger, and the steps are equal. Reference and test stimuli are presented in pairs several times, with the test stimuli randomly chosen from the predetermined set and observers report which of the two produces the greatest sensation. Ideally, both stimuli should activate the same set of receptors and neural mechanism at the same time. However, in practice this is likely to be impossible and a compromise needs to be reached by either presenting stimuli at the same time but in different physical positions or in the same position but at different times.

1.4.3 Method of limits

This method is less precise but much faster to implement than other methods, making it perhaps the most popular technique for determining sensory thresholds. In the *method of limits*, it is generally indistinct whether the trials start from the top or the bottom of the stimulus staircase. A typical

experiment begins by presenting stimulus well below (or above) the observer's threshold, which is then sequentially increased (or decreased) in small amounts. The experiment ends when the observer reports the presence (or absence) of the expected sensation. The transition point is considered to be midway between the last two stimulus responses. The experiment is usually repeated using both ascending and descending stimuli series and a threshold is obtained by averaging the responses of these (see Figure 3).

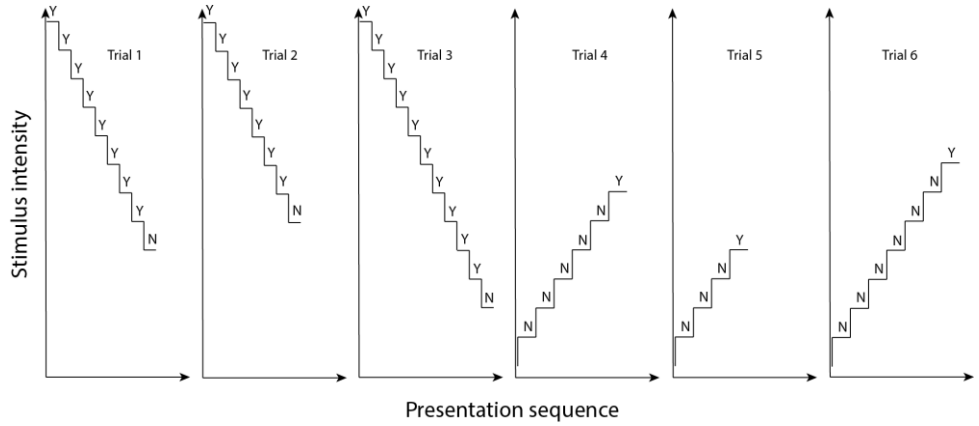

Figure 3: fictional set of results for a typical method of limits experiment. There are 6 trials consisting of staircases (3 ascending and three descending) of stimulus intensities. Observers answer "yes" (Y) or "no" (N) to indicate whether or not they have detected the stimulus. Individual trials start from one extreme (either from perfectly detectable or undetectable stimuli) and end as soon as observers alter their answers. In the example, no provision has been taken to counter habituation and expectation errors.

Although it is easy to implement, the results obtained by this method are influenced by two systematic errors, called *error of habituation* and *error of expectation*. The first affects experiments using descending stimuli and derives from the fact that observers may become used to give the same response for a period of time, and continue to do so well after the threshold point has been reached. Conversely, in the case of ascending stimuli, observers may anticipate the arrival of the threshold by reporting its presence before they actually sense it. As a result, thresholds for ascending (or descending) stimuli will be misleadingly lower (or higher) than they actually are. Moreover, the magnitudes of these complementary errors are not necessary the same and it is not possible to average them out by repeating the experiment with alternating ascending (or descending) staircases. One possible improvement comes from varying the starting points of the staircases in each trial to stop observers "counting" steps or predicting the threshold from previous trials. Another improvement consists on alternating up and down staircases, switching direction as soon as the observer reports a threshold, and recording the reversing points. After a

number of staircase reversals, the final threshold is obtained by averaging these reversing points. The precision of this method largely depends on the size of the steps considered.

Forced-choice methods

Although habituation and expectation biases in the method of limits can be reduced by training, it is impossible to be certain that observers are always honest and extraordinary results are always a consequence of extraordinary low/high thresholds and not of these errors. Furthermore, results are strongly influenced by observers' expectations and the probability of the test stimulus appearing: observers tend to learn whether the test stimulus is presented often and try to predict its appearance. In response to these criticisms, a series of methodological variations were developed were observers do not report directly whether they sense the stimulus but have to pick an item from several possible instead. For example, in a typical trial the observer chooses an item from a set of stimuli, only one of which is different from the rest in some sensory dimension. In each trial, these items can be presented sequentially (temporal forced-choice) or simultaneously (spatial forced-choice), and the difference between the odd (test) stimulus and the others (references) is varied from one trial to the next following a staircase, making the task more difficult or easier. This method and its variations are termed "forced-choice" because the observer always has to pick an option among others. When asked to select the odd item from a set of stimulus that are nearly the same, observers are unlikely to answer correctly most of the time if they cannot sense the test stimulus' difference. If this difference if below the perceptual threshold, the answer will be determined by chance and if it is above, the answer will be above chance. Since this method relies on statistics, each step in the staircase procedure (each trial) needs to be repeated many times to rule out "lucky" answers. For example, in visual psychophysics it is common to present observers with a choice of several closely-looking images from which they have to pick one according to some instructions (the one where some signal is present, the odd one out, etc.). If the observer is presented with two choices whose difference is below threshold, about half the answers will be correct; however if the difference is clearly visible the test stimulus will be selected in all trials. At threshold, the test stimulus will be selected midway between 50% and 100% of the time, i.e 75%. In the case when observers are presented with two choices, the method is called 2AFC or 2-alternative forced choice.

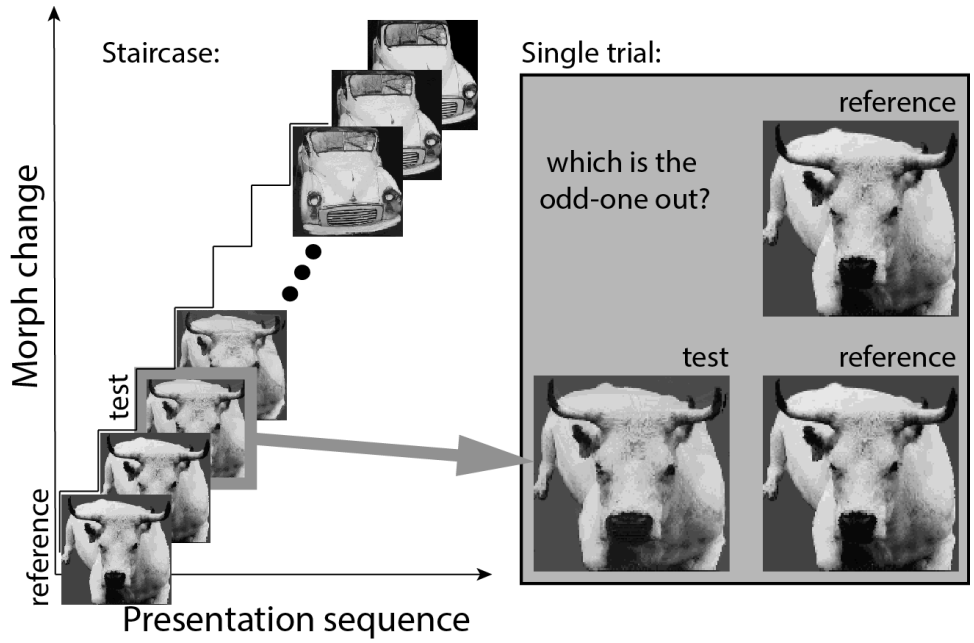

Figure 4: example of a forced-choice version of the method of limits. The staircase consisted of a sequence of morphed pictures (a bull that morphs into a car in small steps). The experiment looked for the point in the sequence where observers can just tell apart a modified picture from the original picture of the bull (discrimination threshold). To decide this, each step in the staircase originated a series of trials where observers had to select the corresponding test image (a morphed bull) from two references (the normal bull) presented in randomised order. By repeating each step a number of times, experimenters measured the morph change necessary for observers to tell the difference, without the inconvenience of habituation and expectation biases [15].

Figure 4 shows an exemplary forced-choice version of the method of limits used to measure the threshold for discriminating between the photograph of an object and a slightly different (morphed) version of the same photograph [15]. The number of choices offered to the observer determines many of the characteristics of the method, such as the proportion of right answers obtained by chance and the threshold position, since results are always fitted by an ogive (see Figure 2). In all cases, the ogive's lowest *y*-value will correspond to the fraction of right answers obtained by randomly selecting stimuli and its highest *y*-value to the fraction of right answers obtained by clearly distinguishing the test from the reference stimulus. For example, in a 2AFC the ogive's lowest value is always 0.5 and in theory, its highest value should be 1 but in practice experimenters allow for mistakes made by observers accidentally making wrong choices. These mistakes (also called "finger errors") lower the chance of getting 100% correct responses even when the test stimulus is clearly sensed. To account for mistakes the ogive's upper asymptote is usually set at 0.98 and the threshold to 0.74 (98% and 74% of correct responses respectively).

Despite involving longer experimental hours and tiredness, forced-choice procedures are perhaps the best techniques to obtain perceptual thresholds free from response and observer expectation biases.

1.4.4 Method of adjustments

The *method of adjustments* has much lower accuracy than the methods described above and tends to be selected in circumstances where there is an advantage in giving observers control over changes in the stimulus, making experiments more participative and arguably less tedious. In a single trial, the stimulus intensity is set far below (or above) the expected threshold and the observers increase (or decrease) the intensity until the sensation is just perceptible (or just disappears). This method has the same shortcomings as the method of limits in terms of response and expectations biases and therefore requires randomised starting points and a large number of trials to reach acceptable accuracy levels, always combining ascending and descending runs and averaging the results to obtain the threshold. There are also issues regarding the use of discrete instead of continuously variable stimuli which makes more difficult to calculate the method's error. In summary, the method of adjustments is seldom used in rigorous laboratory contexts. It might have advantages when a "quick and dirty" measure of threshold is required or when the task is too long and tedious for observers to keep a reasonably high performance throughout the session.

1.5.5 Estimating psychometric function parameters

Psychometric functions constructed from multiple-choice experiments differ from those constructed from yes-no experiments (see Figure 2) in that the position of the ogive is determined by the probability of a "chance" selection and "detection". For example, if the experiment contained 4 intervals and the observer cannot detect which interval contains the signal, results are likely to be correct in 25% of the trials, and this will determine the lower asymptote of the ogive.

The measurement of thresholds using multiple-choice paradigms with psychometric functions built from ascending and descending staircases could be long and tedious for observers. This is particularly true if the initial parameters of the procedure (step size, starting level, etc.) are not known in advance. Several algorithms have been proposed to optimize the collection of results, which reduces the number of experimental trials by aiming at the threshold magnitude. Two of the most popular are PEST [19] and QUEST [22], which consist of staircases that reverse their directions once the observer obtains results that are consistently different from the previous ones. They also start with a larger step size which is reduced as the

experiment reaches threshold, effectively “homing” on the most likely value and reducing the number of trials. A comprehensive analysis of popular staircase algorithms alongside with some practical recommendations can be found in the work of Garcıa-Perez [5].

1.5 Signal detection theory

Signal detection theory was developed to understand and model the behaviour of observers in detection tasks. It examines their actions by dissociating decision factors from sensory ones, incorporating ideas from the field of signal recovery in electronics which studies the separation of signal patterns from electronic noise [7; 11; 17]. In psychophysics, the signal is our stimulus and the noise refers to random patterns that may originate within both the stimulus itself and the neural activity in the observer’s nervous system.

1.5.1 Signal and noise

Much of the early work on signal detection theory studied the detection of targets on radar screens by military radar operators where important decisions had to be taken quickly and in the presence of uncertainty. For example, imagine a rather critical situation where a radar controller is examining a scan display, looking for evidence of enemy missiles. The controller’s response is determined by intrinsic neural noise regardless whether there is an incoming missile or not. In this scenario, either there is an enemy missile (signal present) or not (signal absent) and the controller either sees it and activates countermeasures (fires the alarm) or not (does nothing). The possible outcomes are four: enemy missile detected and destroyed (hit), enemy missile incoming undetected (miss), no missile and no alarm (correct rejection) and no missile and alarm activation (false alarm). The first and third outcomes (hits and correct rejections) are desirable while the others (misses and false alarms) are not. There is also a ranking of bad outcomes where a false alarm is much more preferable than a miss.

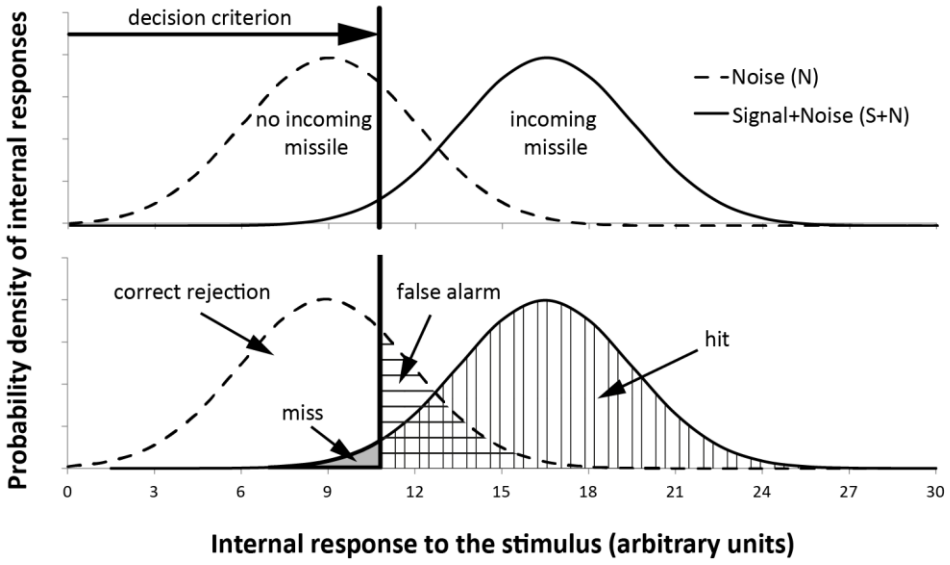

Figure 5: the hypothetical probability distribution of noise (N, in broken lines) and signal+noise (S+N, in solid lines) as a function of the observer's neural response. As expected, the S+N curve generates more neural activity (e.g. spikes per second) in average than the N curve; however there are sections where both curves overlap in x. The vertical line represents the observer's decision criterion, which could be anywhere along the *x*-axis. Depending on its position, four distinctive regions are determined: hit, miss, correct rejection and false alarm.

Figure 5 shows the probability distributions of the noise (broken lines) as a function of the observer's internal response and the same probability distribution when a signal is added (solid lines). The second is bigger in average, so the observer's task is to separate the stimulus containing signal from the pure noise by setting a criterion threshold: when his/her visual neural activity is larger than the criterion threshold the "signal" is acknowledged and vice-versa. This clearly defines the four sections that correspond to the possible outcomes (hits, misses, false alarms, and correct rejections) in Figure 5. On both hits and false alarms, our radar controller activates the countermeasures and correspondingly, misses occur when there is an incoming missile but the neural activity it generates is insufficient to reach the decision criterion, as indicated in the figure.

In our example, the radar operator's job is to adjust the decision criterion according to the importance of each one of the possible outcomes. A low decision criterion means that the alarm will be set to almost everything, be it internal noise or not. The operator will never miss an incoming missile but similarly will have a high rate of "false alarms" (potentially leading to a waste of resources). On the contrary, setting a high criterion will lead to very few false alarms with the occasional undetected incoming missile.

1.5.2 The receiver operating characteristic

In controlled lab conditions, our radar operator's task (judging the presence or not of a signal) would be treated as a yes-no procedure. In these procedures observers are told in advance the different probabilities and costs associated to the different experiment outcomes (in the same way a radar operator knows the conditions and consequences of his actions). Such experiments typically consist of a large number of trials where observers must judge whether the signal was present or not.

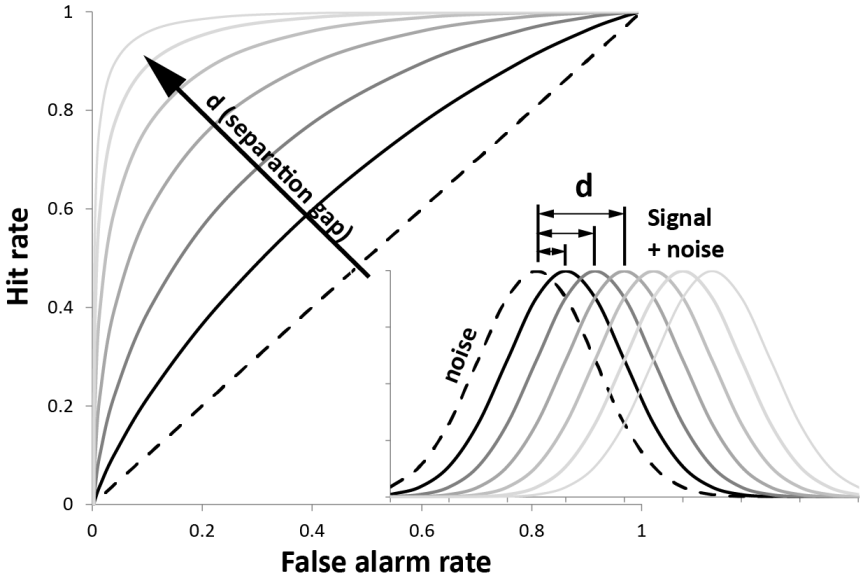

Figure 6: Receiver operating characteristic (ROC) curves describing the relationship between hits and false alarms for several decision criteria and signal strengths. As the observer lowers its decision criteria, both areas labelled "false alarm" and "hit" in Figure 5 increase in different proportions, depending of the proximity between the curves. The ROC family of curves describes all the options for our hypothetical radar controller regarding the separation between S and S+N curves. When both curves are clearly separated, it is possible to obtain a nearly perfect hit rate without false alarms.

Figure 6 shows how the results from a hypothetical yes-no experiment would look like if we consider the areas under the curves labelled "hit" and "false alarm" in Figure 5. If there is considerable overlap and the observer sets a low decision criterion, it is likely that both hits and false alarms will be high. As the decision criterion increases, both areas will decrease in a ratio that depends of their separation *d* and the overlap region (see insert in Figure 6). The curves in Figure 6 describing the possible relationships between hits and false alarms rates are called *receiver operating characteristic* (ROC) curves. They were first developed in Britain during World War II to facilitate the differentiation of enemy aircraft from noise (e.g. flocks of birds) by radar operators.

From the ROC curves that best fit experimental hit/false-alarm data, one can obtain a measure of the strength of the signal that is independent of the observer's decision criterion:

$$d' = \frac{(\overline{SN} - \overline{N})}{\sigma_N} \tag{7}$$

where d' is called the observer's discriminability index, $\overline{SN}$ and $\overline{N}$ are the mean values of the S+N and N distributions respectively and σ_N is the standard deviation of the N distribution.

The optimum decision criterion β_{opt} (i.e. the criterion that produces the largest winnings in the long run) can also be estimated from the probabilities of N and S+N and the costs of the various decision outcomes using:

$$\beta_{opt} = \left[\frac{P(N)}{P(SN)}\right] \cdot \left[\frac{value\ of\ correct\ rejection - cost\ of\ false\ alarm}{value\ of\ hit - cost\ of\ miss}\right] \tag{8}$$

where costs are entered as negative numbers. Eq. 8 has been found to reasonably represent the judgements made by actual observers [6].

ROC curves can be thought of as a plot of the fraction of true positives (also known in statistics as *sensitivity*) in the y-axis versus the fraction of false positives (also known as *specificity*) in the x-axis. Both measure the statistical performance of a binary classifier. While sensitivity measures the proportion of actual positives that are correctly identified (e.g. cancer patients correctly identified with the condition), specificity indicates the proportion of negatives correctly identified (e.g. healthy patients identified as not having the condition). Additionally, the area under the ROC curve is related to the accuracy of the method (how good is the classifier when discriminating between the two possible conditions). A perfect classifier has an area equal to one, while a classifier no better than chance has an area of 0.5. Because of its graphical nature, ROC analyses are excellent tools for showing the interrelations between statistical results and many algorithms/toolboxes have been developed [16].

Traditional psychophysical research methods have been criticized because of their inability to separate subjects' sensitivity from other response biases like the various costs of subjects' decision outcomes [7]. In this view, threshold measurements may be contaminated by arbitrary changes in the observer's decision criterion which may lead to faulty results. This explains for example, the dependence of the method of constant stimuli's results with the fraction of trials that contain the signal: when the probability of a signal increases, observers lower their decision criterion and the fraction of "yes" answers increases. As we have seen, signal detection

theory provides a series of methods for computing d' and β which allows studying the effects of particular variables on each of them separately. In practice it is possible to remove fluctuations in the decision criterion by using forced-choice experiments where many stimuli intervals are presented and the observer has to choose the one containing the signal. As seen before, forced-choice paradigms are not influenced by response biases and d' measures can be obtained directly by measuring the proportion of correct responses [8]. The advantages of several-alternatives procedures over the yes-no procedure reside in the fact that in every trial observers are presented with more information (several intervals) whether in the yes-no procedure a single trial may contain the signal or not.

1.6 Psychophysical scaling methods

Early psychophysicists agreed that a complete mapping between physical stimulus and perceptual response could only be produced indirectly by measuring absolute and relative thresholds, with the *jnd* as the sensory magnitude unit. Their approach relied on two main assumptions: (1) that all *jnds* (for a given sensory task) are perceptually equal and (2) it is always possible to control and measure the physical properties of the stimulus. Given these assumptions, it is possible to characterize the relationship between stimulus and response by measuring differential thresholds to ever increasing stimulus and obtain the magnitude of a perceptual response simply by counting the number of *jnds* a stimulus is above absolute threshold. However, researchers working in the field of experimental aesthetics soon encountered situations that challenged these assumptions [4]. Indeed attributes such as “pleasantness”, “beauty”, “naturalness”, “repugnance”, etc. are the product of higher, cognitive cortical levels where processes like attention, memory, emotions, creativity and language play a much more important role that is absent at perceptual, pre-conscious levels. To obtain a stimulus-response relationship in such complex cases, a different set of techniques were developed which either treat observers’ judgements directly as sensory measurements or deal indirectly with stimuli whose relevant physical properties cannot be specified. This mapping of physical stimulus to perceptual (or most likely psychological) response resulted in the creation of many *psychophysical scales*: ranked collections of measurements of response to some property of the physical world.

1.6.1 Discrimination scales

Although the methods described in previous sections are designed to measure discrimination responses to small stimulus changes, it is possible to

extend them to construct psychophysical scales relating large changes in the physical stimulus to their corresponding psychological responses. This is a consequence of Weber's law (eq. 5), which states that Δω (in physical units) is proportional to the magnitude ω, and Fechner's law (eq. 6) which assumes equal increments of the corresponding sensation *jnds* regardless of stimulus intensity. For example, if one obtains the signal change Δω required to produce one *jnd* of sensation at many levels of signal intensity, it is possible construct a function over the whole range of the stimulus signal, relating the stimulus' physical units to psychological sensation. Panel (a) in Figure 7 shows an imaginary example of such construction. The *y*-axis represents the sensation (in number of *jnds*) elicited by the stimulus intensity (*x*-axis). As the stimulus intensity grows, larger increments ($\Delta\omega_2 > \Delta\omega_1$) are necessary to elicit the same sensation (1 *jnd*) until a saturation level is reached. Panel (b) shows a psychophysical scale obtained from adding the *jnds* in panel (a).

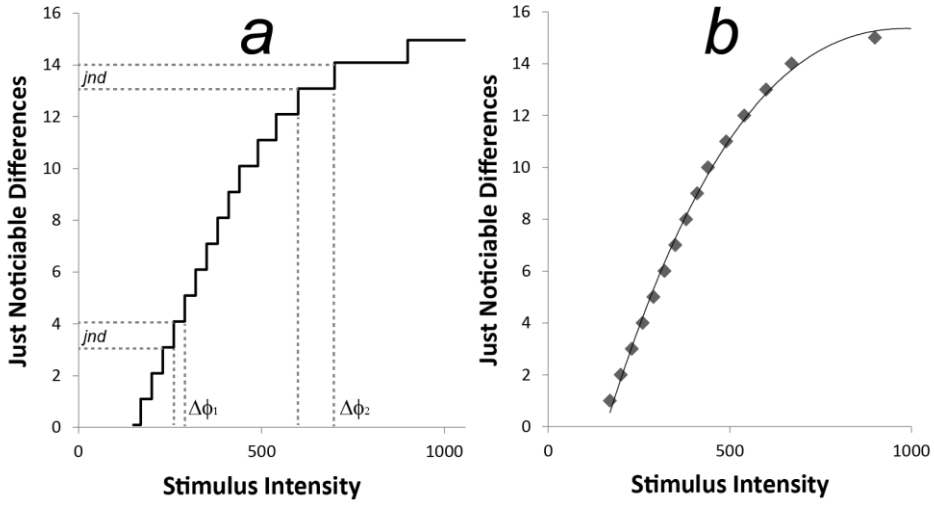

Figure 7: Hypothetical example of how to obtain a psychophysical sensation scale over a large stimulus range from measures of sensation discrimination at threshold (*jnd*). Panel (a) shows that to keep producing the same increment in sensation (equal steps in the *y*-axis), the stimulus intensity needs to be increased (unequal steps in the *x*-axis). Panel (b) represents the psychophysical scale obtained from the results in panel (a).

As we have seen before, Weber and Fechner's laws are related through their respective constants, meaning that the second is only valid if the first is correct. It has been shown [14] that the Weber fraction in eq. 5 is not strictly constant over the whole range of stimulus intensities, which means that the data-points of Figure 7 cannot be derived accurately from equations 5 and 6 and have to be measured. After a hundred years of being accepted by the psychology and engineering community, today Fechner's law is not

considered to be an accurate description of the relationship between stimulus and sensation; however his achievement still represents a good approximation and should not be underestimated.

1.6.2 Rating scales

In some situations, it is not possible to produce psychophysical scales from indirect judgments such as detection or discrimination responses and the only option is to use direct observer judgements to assign numerical values to stimulus attributes. For example, it might be necessary to psychophysically quantify the distortion produced by different compression levels in images or to arrange colours and textures in different categories (such as "red" "blue" or "coarse" or "smooth"), translating physical attributes like hue into perceptual categories such as "magenta". In other cases there is no quantifiable physical attribute to relate to, as in the task of ranking a series of objects according to a subjective judgement (e.g. beauty) and arranging their position in a numerical scale. In all these circumstances, it is convenient to produce a "*mapping*" that links physical to psychological attributes with steps matching the internal relationships of the psychological magnitudes considered.

Equipartition scales

A common task in psychophysical research is to partition a psychological continuum into interval scales of equal *distances*. For example, two grey patches might be provided, a bright patch "A" and a dark patch "B" and observers may have to evaluate whether the lightness distance between pair A-B is equal, more, or less than the distance between another pair of grey patches C and D. To systematize this task one might want to relate the continuum of physical luminances to the perceived sensation distances between A-B and C-D by dividing this continuum in perceptually equal steps [13].

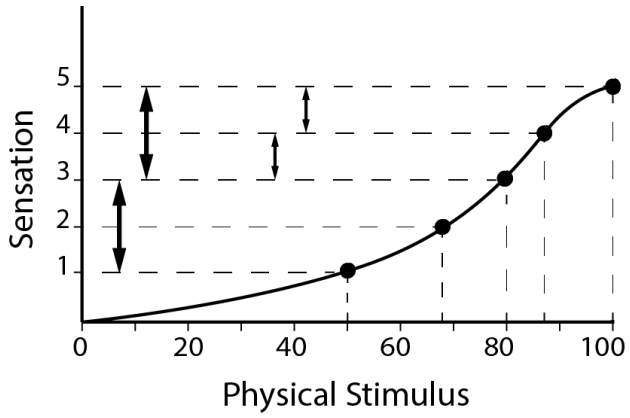

Figure 8: Hypothetical non-linear relationship between a physical stimulus and its corresponding psychological sensation. The arrows show how a perceptual scale with uniform measurement units can be produced by partitioning the scale into smaller intervals.

Figure 8 shows a hypothetical psychological sensation that varies as a non-linear function of the physical stimulus in the range [0, 100]. In one partitioning method, observers are first exposed to fixed stimulus values (20, 40, 60, etc.) and then asked to manipulate them until all scale steps are perceptually equal. A second method starts by exposing observers to the full range of physical stimulus and asking them to find the perceptual middle-point. The double arrows in Figure 8 show the equal intervals in which a given observer splits the sensation elicited by the physical signal in the interval [50, 100]. In our example, this perceptual middle-point corresponds to physical stimulus of strength 80. In a second iteration of the experiment, the observer splits the range of sensations elicited by the physical stimulus of strength 80 and one extreme (i.e. strength= 100), subdividing the scale in smaller evenly-spaced intervals. After a few iterations, the psychological continuum is subdivided in equally-spaced steps.

Another popular method of production of scales consists in adjusting stimuli so that their perceived intensity is a multiple of some original magnitude, extrapolating sensorial ratios rather than dividing a given interval. Since this task is also critically dependent on the observer's ability and training, results may vary and there are some consistency tests available to evaluate their performance in such tasks [6].

Paired comparison scales

Paired comparison scales are preferred when the psychological sensation to quantify does not have any evident correspondence with a measurable physical attribute. As an example, consider again the experiment described in Figure 4 but now, suppose we are not interested in discriminations made between pictures *at threshold* but want to know the perceptual difference between pictures in the *suprathreshold* region (for instance, between the two extremes of the morphed sequence: the car and the bull). When image differences are very large (e.g. they contain different objects or are subject to extreme colour, texture or digital effects distortions) a problem arises since there is no obvious "morph metric" or any convenient physical measure to link to our perceptions [21]. A viable solution was proposed by Gustave Fechner, who was the first to systematically study this problem (Fechner, 1876) whose theoretical foundations were established half a century later by Louis L. Thurstone as a *law of comparative judgment* [20]. Fechner's original aim was to quantify aesthetic preferences for objects, and to do that, he put forward the notion that the distance between two objects in some subjective pleasantness continuum was linked to the *proportion* of times an object was judged more pleasant than the other. By having subjects

compare pairs of objects many times he noticed that, if object A is judged more pleasant than object B half the time, then both objects are equally pleasant. Correspondingly, if object C is consistently judged more pleasant than object A, then by the same measure, C is likely to be the most pleasant object. Thurstone generalised this concept into a measurement model which uses simple comparisons between pairs of stimuli (pairwise) to create a perceptual measurement scale. To do that, he linked *p*, the proportion of times a stimulus was judged greater than another, to the number of psychological scale units separating the two sensations. Similarly to the case of threshold detection (see Figure 1), the comparative-judgement data dispersion in the psychological continuum was considered to be a product of the fluctuations of the neural system, and the psychological scale value was assumed to be the mean of these fluctuations (i.e. the mean of the frequency distribution). Since this frequency distribution is neural in origin and basically unknown, Thurstone devised an indirect method for calculating the psychological scale value from the proportion of positive comparison results *p*, following a series of careful considerations. In terms of signal detection theory, the psychological scale value that we want to obtain here is equivalent to the discriminability index (*d'*), but calculated as the distance between two S+N distributions instead of the classical S and S+N (see Figure 6 and eq. 7).

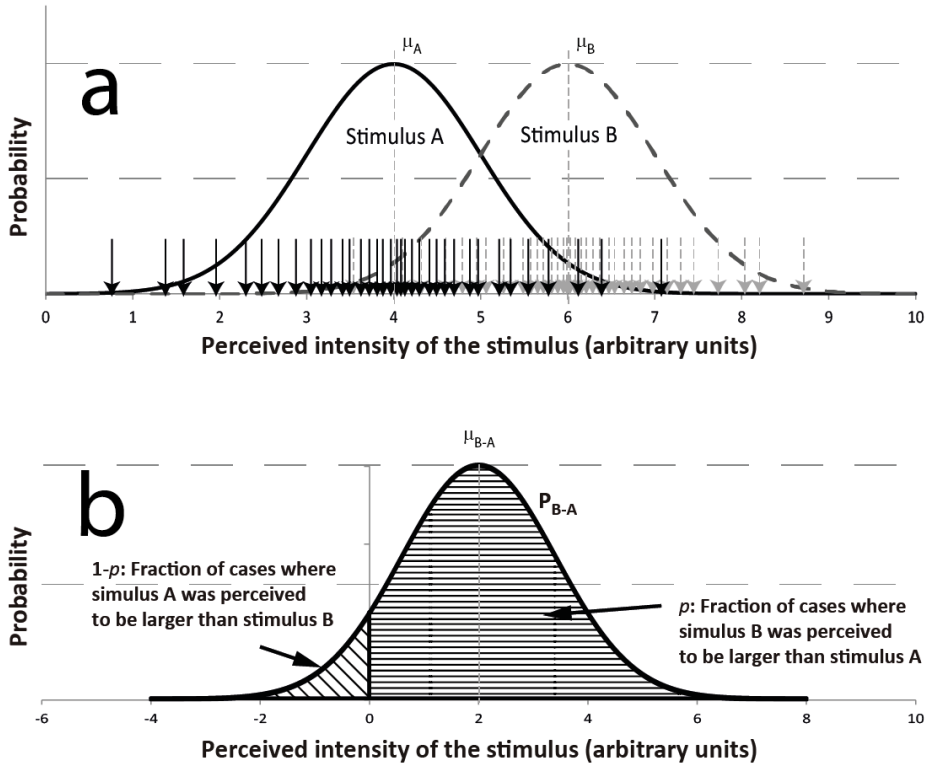

Figure 9: Panel (a)- Probability distribution of judgements for two different stimuli A and B in an arbitrary psychological continuum. Their average perceptual distance in determined by $\mu_B - \mu_A$, the distance between the two means. Panel (b)- Probability distribution of the differences in perceptual judgements for two different stimuli, derived from panel (a). The mean μ_{B-A} of this new Gaussian is equal to the average perceptual distance between A and B and its variance is also related to the variances of the distributions in panel (a).

Figure 9 shows the schematics of the statistical processes involved in a comparison between two different stimuli along an arbitrary psychological continuum. Panel (a) represents the distribution of a large number of judgements of stimulus A (solid curve) and stimulus B (broken lines curve). The downward-pointing arrows show the internal neural processes that determine judgement magnitudes corresponding to each stimulus (black arrows for stimulus A and grey broken arrows for stimulus B) in the psychological continuum. In a single pairwise comparison, a perceived value S_j for stimulus A is compared to the perceived value S_i for stimulus B and distance S_i-S_j is obtained. Because of the stochastical nature of neural processes, a tendency emerges after a number of comparisons with most judgements pairs concentrated around their mean values following the two Gaussian distributions of panel (a). In consequence, the average perceived distance between stimulus A and B is given by the difference between the means of the Gaussians ($\mu_B - \mu_A$).

The arrows in panel (a) show that in most cases, the value assigned to stimulus A will be larger than that of stimulus B, but in some cases the opposite will occur (some broken grey arrows are located to the left of some continuous black arrows). The proportion of times where S_i is larger than S_j is determined by the proximity between the two Gaussians and their corresponding variances, and can be calculated by integrating

$$P_{B-A}(u) = \iint_{-\infty}^{\infty} \frac{e^{-\frac{x^2}{2\sigma_A^2}}}{\sigma_A\sqrt{2\pi}} \frac{e^{-\frac{y^2}{2\sigma_B^2}}}{\sigma_B\sqrt{2\pi}} \delta((x-y)-u) \;\; dx \;\; dy \tag{9}$$

where P_{A-B} is the normal difference distribution, σ_A^2 and σ_B^2 are the variances of the Gaussians corresponding to stimulus A and B, and δ is a delta function [24]. The solution to eq. 9 can be conveniently written as:

$$P_{B-A}(u) = \frac{e^{\frac{-[u-(\mu_B-\mu_A)]^2}{2(\sigma_A^2+\sigma_B^2)}}}{\sqrt{2\pi(\sigma_A^2+\sigma_B^2)}} \tag{10}$$

which describes another normal distribution with mean value $\mu_{B-A} = \mu_B - \mu_A$ and variance $\sigma_{B-A}^2 = \sigma_A^2 + \sigma_B^2$. The relationship between this new Gaussian and the original pair is represented in Figure 9, where the difference between the means of the curves in panel (a) corresponds to the mean of the curve in panel (b) and the zero of the abscissa in panel (b) splits the shaded area under the curve according to the proportion of positive comparison results *p*. The case where the variables *x* and *y* in eq. 9 are jointly normally distributed random variables, is slightly more complex, however the results of the integral are still normally distributed with a mean equal to $\mu_B - \mu_A$. The main change is that variances σ_A^2 and σ_B^2 are now correlated (and therefore not additive) which results in a slightly more complex value for the standard deviation of the normal difference distribution:

$$\sigma_{B-A} = \sqrt{\sigma_A^2 + \sigma_B^2 + 2\rho_{AB}\sigma_A\sigma_B} \tag{11}$$

In summary, the probability distribution of differences plotted in panel (b) provides the link between the psychological distance $\mu_B - \mu_A$ and the quantity *p*, which is easily computed by noting the proportion of times subjects report that the sensation produced by stimulus A is larger than that

of stimulus B. To operate we just need the following property of normal Gaussian distributions applied to $P_{B-A}(u)$:

$$z = \frac{\mu_{B-A}}{\sigma_{B-A}} \tag{12}$$

which defines the standard score *z* (also known as the z-score) for $P_{B-A}(u)$. *z* quantifies the number of standard deviations that occur between our zero and the mean value μ_{B-A} in panel (b). Traditionally, the calculation of *z* involved large look-up tables but today it can be obtained numerically, by inverting the cumulative distribution function, which describes the probability that a random variable will be found to have a value less than or equal to *z*. Once *z* is obtained from *p*, there are some remaining considerations about eq. 11 before solving it for μ_{B-A} that were organized by Thurstone in five different cases [20]. Cases I and II consider the complete form of eq. 11; Cases II and IV consider $\rho_{AB} = 0$ (no correlation between variances σ_A^2 and σ_B^2); and Case V considers the simplest solution, where there is no correlation and variances σ_A^2 and σ_B^2 are arbitrarily given a value of one, effectively choosing unity as *de facto* unit of measurement. Eq. 13 shows the simplest and most common expression for Thurstone law of comparative judgement Case V:

$$\mu_B - \mu_A = z\sqrt{2} \tag{13}$$

Table 1 to Table **3** illustrate how to produce a comparison scale from the statistical results of an imagined paired comparison experiment consisting of 5 stimuli labelled A, B, C, D and E. These were compared against each other obtaining the results shown in Table 1, with rows showing the fraction of times they were perceived to be larger than columns. For example the value [row A, col B] shows that A>B in 48% of the trials. Table 2 shows the standard scores derived for the same values and Table 3 shows the results of eq. 13 applied to Table 2. Column 7 of Table 3 is just the average for each stimulus, and column 8 shows the definitive scale, where zero has been set to the lowest value.

Table 1: Hypothetical results of a pairwise comparison experiment.

	A	B	C	D	E
A		0.48	0.62	0.9	0.95
B	0.52		0.54	0.74	0.9
C	0.38	0.46		0.7	0.85
D	0.1	0.26	0.3		0.69
E	0.05	0.1	0.15	0.31	

Table 2: Standard scores derived from the results in Table 1

	A	B	C	D	E
A		-0.05	0.31	1.28	1.64
B	0.05		0.10	0.64	1.28
C	-0.31	-0.10		0.52	1.04
D	-1.28	-0.64	-0.52		0.50
E	-1.64	-1.28	-1.04	-0.50	

Table 3: Perceptual distances calculated for each comparison and their averages

	A	B	C	D	E	Average	Average-min
A		-0.07	0.43	1.81	2.33	1.12	2.70
B	0.1		0.14	0.91	1.81	0.73	2.31
C	-0.4	-0.14		0.74	1.47	0.41	1.98
D	-1.8	-0.91	-0.74		0.70	-0.69	0.89
E	-2.3	-1.81	-1.47	-0.70		-1.58	0.00

1.7 Conclusions

Although great progress has been made in the century and a half since the birth of the discipline, many challenges remain. For example, there is still no set of fundamental laws or concepts that explain all psychophysical results and to our days, different scaling methods may yield different results, with no certainty that these methods are not measuring different aspects of perceptual/cognitive processes. It might be the case that a single solution to all these problems might be developed by some kind of fundamental laws of psychophysics, but so far there is no agreement on what these laws should be based on.

In this chapter we have attempted to provide an overview of the most basic concepts necessary to both conduct and interpret psychophysical laboratory experiments. On a practical level, psychophysics experimentation

is not to be taken lightly. Designing and implementing a task could be extremely uncertain, and many questions arise during the experimental design: how many trials? How long should the experiment take? What should I say to the observers? How should I reward my observers? All these questions can be answered from your theory and if you consider that (1) you should try to make the measurement error small, (2) you do not want your observers to get tired or to lose interest, (3) you want to take the greatest advantage of your experimental hardware, (4) you do want your observers to understand exactly what they have to do: run mock training experiments until they feel confident with the task. Besides, you do not want to waste time (and presumably money) in an experiment that is likely to yield noise as a result. To avoid mistakes, it is often recommended to run a pilot to see whether all these conditions are fulfilled and the final results are likely to answer your questions. It is also recommended that you remove yourself as much as possible from the final data collection, to avoid introducing your own expectations and biases. As a general rule, assume that observers will always try to second-guess your intentions: only talk to your observers about the experiment after the experiment has finished to avoid giving them wrong cues and expectations.

As science progresses and communication among researchers from different disciplines increases exponentially, the boundaries between computer vision, computational neuroscience and experimental psychology are blurring. Today it is common to see computer vision laboratories that routinely collaborate with psychology departments establishing larger "Vision Research Centres" or even resort to equip their own psychophysics laboratories. This partnership has not always been frictionless, especially because of the lack of a common language and background; however the potential rewards are enormous. Today, psychophysics is behind much of the progress (through the use of ROC curves) in medical decision analysis, text retrieval and processing, the experimental underpinning of computational colour vision, primary visual cortex modelling (both functional and developmental), computational models of visual attention, segmentation and categorization, image retrieval and the list grows. Even larger, higher-level problems such as visual aesthetics are routinely explored applying statistical methods to "ground truth databases" where the performance of human observers has been recorded or/and analysed. Indeed, ground truth databases are perhaps the single most conspicuous manifestation of the influence of psychophysical approaches and methods in computer science. It is in this role of users of psychophysical data that computer scientists most need to understand all measurement limitations inherent to each experiment to properly interpret how they may influence their own results.

1.8 References

[1] Bevington, P. R., and Robinson, D. K. (2003). Data reduction and error analysis for the physical sciences, 3rd ed. edn (Boston, Mass. ; London: McGraw-Hill).
[2] Brindley, G. S. (1970). Physiology of the retina and visual pathway, 2nd edn (London,: Edward Arnold).
[3] Fechner, G. T. (1860). Elements of Psychophysics, 1966 edn (New York: Holt, Rinehart and Winston).
[4] Fechner, G. T. (1876). Vorschule der Aesthetik, 2. Aufl. edn (Leipzig,: Breitkopf & Härtel).
[5] Garcıa-Perez, M. A. (1998). Forced-choice staircases with fixed step sizes: asymptotic and small-sample properties. Vision Research *38*, 1861–1881.
[6] Gescheider, G. A. (1997). Psychophysics : the fundamentals, 3rd ed. edn (Mahwah, N.J. ; London: Lawrence Erlbaum Associates).
[7] Green, D. M., and Swets, J. A. (1966). Signal detection theory and psychophysics (New York,: Wiley).
[8] Hacker, M. J., and Ratcliff, R. (1979). Revised Table of D' for M-Alternative Forced Choice. Perception & Psychophysics *26*, 168-170.
[9] Hecht, S., Shlaer, S., and Pirenne, M. H. (1941). Energy at the threshold of vision. Science *93*, 585-587.
[10] Ludvigh, E., and McCarthy, E. F. (1938). Absorption of visible light by the refractive media of the human eye. Archives of Ophthalmology *20*, 37-51.
[11] Macmillan, N. A., and Creelman, C. D. (2005). Detection theory : a user's guide, 2nd edn (Mahwah, N.J.: Lawrence Erlbaum Associates).
[12] Marks, L. E. (1978). The unity of the senses : interrelations among the modalities (New York: Academic Press).
[13] Munsell, A. E. O., Sloan, L. L., and Godlove, I. H. (1933). Neutral value scales. I. Munsell neutral value scale. J Opt Soc Am *23*, 394-411.
[14] Murray, D. J. (1993). A Perspective for Viewing the History of Psychophysics. Behavioral and Brain Sciences *16*, 115-137.
[15] Parraga, C. A., Troscianko, T., and Tolhurst, D. J. (2000). The human visual system is optimised for processing the spatial information in natural visual images. Current Biology *10*, 35-38.
[16] Stephan, C., Wesseling, S., Schink, T., and Jung, K. (2003). Comparison of eight computer programs for receiver-operating characteristic analysis. Clinical Chemistry *49*, 433-439.
[17] Tanner, W. P., and Swets, J. A. (1954). A Decision-Making Theory of Visual Detection. Psychological Review *61*, 401-409.
[18] Taylor, J. R. (1997). An introduction to error analysis : the study of uncertainties in physical measurements, 2nd edn (Sausalito, Calif.: University Science Books).
[19] Taylor, M. M., and Creelman, C. D. (1967). PEST - Efficient Estimates on Probability Functions. Journal of the Acoustical Society of America *41*, 782-&.
[20] Thurstone, L. (1927). A Law of Comparative Judgment. Psychological Review *34*, 273-286.
[21] To, M. P., Lovell, P. G., Troscianko, T., and Tolhurst, D. J. (2010). Perception of suprathreshold naturalistic changes in colored natural images. J Vis *10*, 12 11-22.
[22] Watson, A. B., and Pelli, D. G. (1983). QUEST: A Bayesian adaptive psychometric method. Perception and Psychophysics *33*, 113-120.
[23] Weber, E. H. (1834). De Pulsu, resorptione, auditu et tactu: Annotationes anatomicae et physiologicae: C.F. Koehler).

[24] Weisstein, E. W. (2014). Normal Difference Distribution, In MathWorld - A Wolfram Web Resource (Wolfram Research, Inc.).