Vol. 18 The British Journal of Mathematical and Statistical Psychology May 1965
Part 1
1—10

SEQUENTIAL ESTIMATION OF POINTS ON A PSYCHOMETRIC FUNCTION

By G. B. WETHERILL and H. LEVITT[1]
Birkbeck College, London, and Bell Telephone Laboratories, New Jersey

A simple and efficient method of estimating points on the psychometric function, and thus of estimating absolute and difference limens, is described. An illustration of the method is given in which sensitivity to inter-aural time differences is measured.

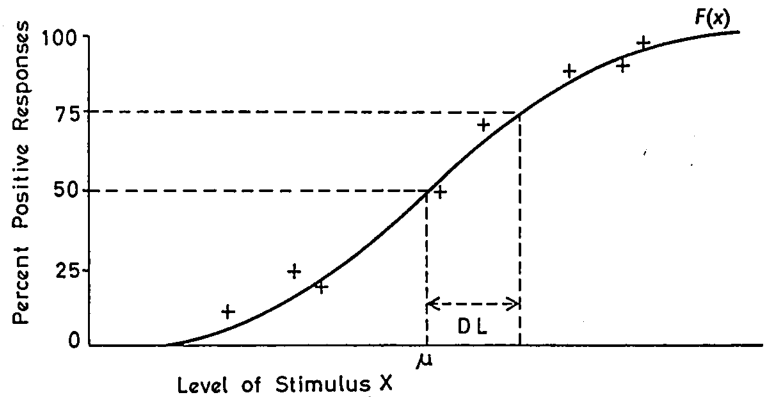
1. INTRODUCTION

Many psychophysical experiments are of the following kind: a stimulus is presented to the subject who is required to judge whether it is more or less intense than a given standard. In many experiments this standard stimulus is presented prior to, or simultaneously with, the test stimulus. There is usually a range of possible values (or levels) of the test stimulus. The subject is restricted to a simple binary decision; the resulting response (observation) is designated 'positive' or 'negative' according to some appropriate rule, and this will naturally be a function of the particular level of the stimulus which is used. These points are illustrated in an experiment carried out by Levitt (1964) in which it was desired to measure the sensitivity of a listener to inter-aural time differences (ITD) for binaurally presented signals. Subjects listened through a pair of headphones fitted to exclude extraneous noise, and signals were presented at each headphone with a range of levels for the ITD. Random white noise, subsequently band-limited to 4,560 c.p.s. at a signal level of 50 dB (*re* 0·002 dynes/cm^2) was used. The ITD was varied in steps of 5·8 μsec. At each presentation of the stimulus the subject was required to judge from which side the sound appeared to come. Answers took the form of "to the left" or "to the right", and no other answer was accepted. If, for example, we take "to the left" to be a positive response, and "to the right" to be a negative response, this experiment is of the form just described. No standard stimulus was used in this experiment, the subject being required to use his own estimate of the medial plane as a reference.

Suppose that observations of the kind just described are repeated a number of times at each of a number of levels of the stimulus (that is, at each of a number of values of ITD). Then the results can be plotted in the form of the familiar psychometric curve shown in Figure 1 (*a*). The ordinate represents the relative frequency of positive (or "to the left") responses, and the abscissa represents the level of the test stimulus (ITD). In most experiments of this type, it can

[1]The experiment described in Section 4 of this paper was part of a programme of research carried out by H. Levitt for a Ph.D. degree of London University, under the supervision of Professor E. C. Cherry.

be safely assumed that the responses tend to lie on a smooth curve which increases monotonically from zero to 100 per cent positive responses. That is, it may be assumed that the probability of a positive response at any level x of the stimulus is a function $F(x)$, called a psychometric function or a response curve, which is roughly of the shape shown.



+ = Observed frequency of positive responses as may be obtained in a typical experiment

$F(x)$ = Expected frequency of positive responses.

FIGURE 1 (a). The Psychometric Function.

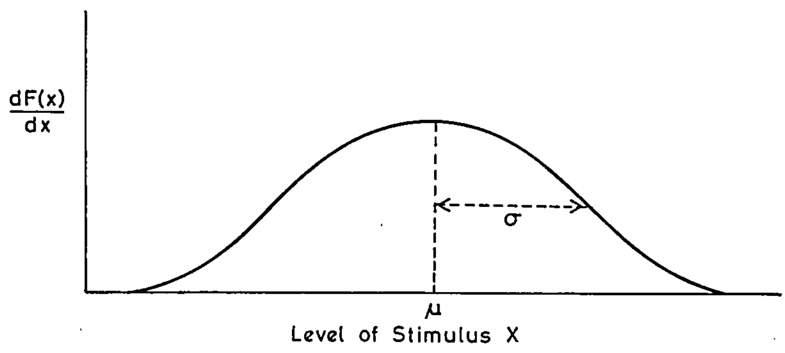


FIGURE 1 (b). The derived frequency distribution.

The psychometric function is often taken to be a cumulative normal curve

$$F(x)=\int_{-\infty}^{(x-\mu)/\sigma}\frac{1}{\sqrt{(2\pi)}}\,e^{-\frac{1}{2}t^2}\,dt,$$

where μ and σ are parameters which define, respectively, the location and slope of the psychometric function. A curve which is almost identical in shape, and which leads to much simpler mathematics in the estimation of parameters, is the logistic curve

$$F(x) = \{1 + e^{-(x-\mu')/\sigma'}\}^{-1}.$$

In fact, the methods proposed in this paper do not depend on a particular assumption about the psychometric function.

The parameters μ and σ (or μ', σ') can be interpreted in terms of a derived distribution; see Figure 1 (*b*). Figure 1 (*a*) gives the percentage area of a curve such as Figure 1 (*b*) up to points x, as a function of x. Thus, mathematically, a judgement as to whether a particular stimulus at a level x is positive or negative can be described as follows. Take a random variable from the distribution Figure 1 (*b*), and note a right or left response according to whether the observed random variable is to the right or left of the applied stimulus level x.

The interpretation of μ and σ in psychometric terms depends on the particular phenomenon being studied. In an experiment involving a comparison between a test stimulus and a standard stimulus, μ is usually referred to as the *point of subjective equality* and σ as the *difference limen*. There are alternative measures of dispersion which can be used instead of σ, such as the semi-interquartile range, which is the distance DL in Figure 1 (*a*). For Levitt's experiment the parameters μ and σ were interpreted as follows: μ was the ITD threshold—that is, that value of the inter-aural time delay which will place the sound image exactly in the medial plane; σ was the difference limen—that is, a measure of the sensitivity of the listener to changes in ITD about the threshold.

2. Estimation Problems

Let the stimulus level at which the probability of a positive response is p be denoted by L_p. Thus for the cumulative normal or logistic models μ is $L_{0\cdot 50}$: the stimulus level giving 50 per cent of positive responses in the long run. For experiments such as the one described above, an important practical problem is to estimate the location and slope of the psychometric function in some region. For example, we may require estimates of location and slope in the region of $L_{0\cdot 50}$. Either a parametric or a non-parametric formulation of the estimation problem may be adopted.

The parametric formulation is to assume that the psychometric function is of a particular form, say cumulative normal, and then estimates of μ and σ are obtained from the data by maximum likelihood methods (see Finney, 1952). This makes the estimates depend on the form of the assumed psychometric function, and further, a large number of observations are required to fit and test any given model over an extensive range of levels. In practice it is usually necessary to concentrate on a limited region, such as the immediate neighbourhood of $L_{0\cdot 50}$.

The non-parametric approach is to search for statistical methods which estimate, say $L_{0\cdot 25}$ and $L_{0\cdot 75}$. These estimates could be regarded as describing the position and slope near $L_{0\cdot 50}$, or alternatively, they could be easily converted into estimates of μ and σ when a parametric formulation is assumed. This paper discusses methods for estimating general percentage points L_p.

A classical design for the kind of experiment under discussion is for several observations to be made at each of five or six stimulus levels, the order of the observations being randomized. This approach is wasteful in that not all the stimulus levels will be well placed for the purpose of efficient estimation. In psychological experiments the number of observations per subject must usually be severely limited, and it is important to use these observations so that they shall be as informative as possible. For example, if it is desired to estimate $L_{0\cdot 50}$, then most observations should be placed as close as possible to the $L_{0\cdot 50}$ stimulus level. If it is desired to estimate σ, then it is most efficient to place observations in two groups which are towards the ends of the range of the psychometric function. (See Wetherill, 1963, for some discussion of this point.) Irrespective of whether or not a parametric approach is adopted, the use of techniques for simple and efficient estimation of selected percentage points L_p will provide a more efficient way of estimating μ and σ than the classical design referred to above, unless there is very complete prior knowledge of the psychometric function.

The Up-and-Down (UD) rule Dixon and Mood (1948) suggested a sequential technique for estimating $L_{0\cdot 50}$. The level of the test stimulus is varied in steps of a constant size. The first observation is made at the best guess of $L_{0\cdot 50}$ available, and trials are made sequentially. When a positive response is obtained the following observation is taken at the next lower level, and when a negative response is obtained the following observation is taken at the next higher level. If we denote positive and negative results by X and O respectively, this technique follows the form illustrated in Figure 2. For

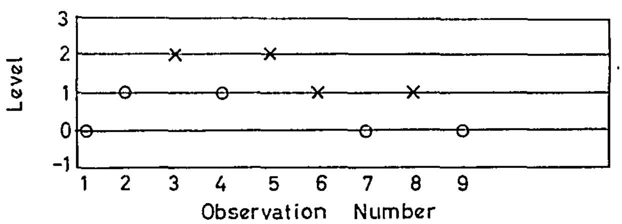


FIGURE 2. Typical pattern of results from the UD rule.

example, in the experiment of Levitt the ITD levels were varied in steps of 5·8 μsec. In Figure 2, observations 1, 2 and 3, ending in a change of response type are called a run. Thus in Figure 2 there are six runs altogether.

If the UD rule is terminated at a set number n of observations, and if n is small, then sometimes only one or two runs are involved and very wild estimates of μ result. A safer procedure in small samples is to stop after a given number of runs have been completed, and this leads to more observations being taken if the initial guess was well away from μ.

To estimate μ from a set of results, Brownlee, Hodges and Rosenblatt (1953) showed that the average of all levels used from observation 2 to the level at which trial $(n+1)$ would have been made, can be used. Alternatively, they suggested that the first run be excluded, except for the final observation. Wetherill and Chen (1964) showed that an estimator with slightly better properties than Brownlee's method is simply to average the peaks and valleys for the sequence. Thus for Figure 2 there are three peaks, at observations 3, 5 and 8, and three valleys at observations 4, 7 and 9, and the average of the six peaks and valleys is

$$\hat{\mu}=\frac{1}{6}\{2+1+2+0+1+0\}=1.$$

A summary of the properties of the UD rule is given in Wetherill (1963). In very large sample sizes, small step sizes are best and give a very efficient placing of observations for estimating μ. In small sample sizes any error in the initial guess of μ gives rise to bias in the estimator, and the bias is much worse for small step sizes. After exploring a number of possibilities Wetherill (1963) suggested the following plan. Operate the UD rule for six or eight runs, then make an estimate of $L_{0\cdot 50}$ from the data, and restart the sequence at or near this estimate, with a step size half that used previously. A second division of the step size could be used if desired, but it is doubtful if it would often be worth while. This plan combines extreme simplicity with very high efficiency. To estimate $L_{0\cdot 50}$ from the results, simply average all peaks and valleys. The best choice of the initial step size is between $\frac{1}{2}\sigma$ and σ.

If desired, maximum likelihood estimation can be carried out on all the results, but the resulting estimate of $L_{0\cdot 50}$ is no more accurate than the average of peaks and valleys estimate. However, by using maximum likelihood methods, an estimate of σ can be obtained, but it is subject to large and unknown biases and is not reliable (see Wetherill, 1963).

3. The Up-and-Down Transformed Response (UDTR) Rule

Wetherill (1963), and Wetherill and Chen (1964) have proposed the Up-and-Down transformed response rule. This is a simple generalization of the UD rule, to estimate points other than $L_{0\cdot 50}$. It is best explained by a simple example.

As in the UD rule the level of the test stimulus is varied only in steps of a given size, but no change in level is made until one of the following patterns of results is obtained at any level:

Move down after		Move up after	
X X X	Type D response	O	Type U response
		or X O	
		or X X O	

This generates a sequence of the kind shown in Figure 3. The sequence is terminated when a specified number of runs have been completed and an even number is preferred. Figure 3 shows a sequence of six runs. If the responses are replaced simply by D or U where appropriate, Figure 3 and Figure 2 look very similar (and ' peaks ' and ' valleys ' can be defined in an obvious way for Figure 3 as for Figure 2). That is, the UDTR is simply a UD rule on a trans-

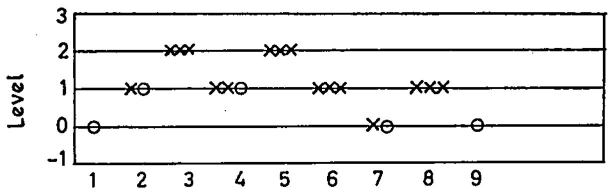


FIGURE 3. Typical pattern of results from the UDTR rule.

formed psychometric function. If the probability of a positive response at any level x is $F(x)$, then the probability of a type D response is $F^3(x)$. Thus the UDTR will estimate the level of x at which

$$F^3(x) = 0{\cdot}50$$

or $L_{0{\cdot}794}$, the level at which the per cent of positive responses is 79·4, since $\sqrt[3]{0{\cdot}50} = 0{\cdot}794$.

To estimate $L_{0{\cdot}794}$ from a set of data, the average of the peaks and valleys can be used and, up to samples of size 100 or more, this is better on a mean squared error basis than maximum likelihood. The estimate obtained by averaging the peaks and valleys is called a ' Wetherill estimate '.

Some care is needed in starting the scheme. If the starting level happened to be chosen well above the true level being estimated, then many observations would be wasted with a string of type D responses. It is better to proceed as follows: " If the first response is negative, start immediately with the UDTR. If the first response is positive, move down one level and try again, etc., until the first negative response is obtained. Treat the first negative response as the first observation of the UDTR." This rule avoids a waste of observations at the beginning. The level at which the first negative response is obtained is not included in the estimate.

The UDTR rule is extremely simple to operate. First draw some ' music paper ', in which lines correspond to levels of the test stimulus, and then plot each point with an X or O as is appropriate. Carry out an even number of runs, and then average the peaks and valleys to obtain an estimate.

Other patterns Table 1 lists some of the possible patterns for use with the UDTR rule, and the example given above is number 2 in this table. The reader can readily generate further entries for himself. There is a fair range of possible percentage points which can be estimated, but the extreme percentage points will require rather a large number of observations. If it is necessary to estimate a particular percentage point for which no strategy is given, then the nearest one can be used, but maximum likelihood estimation would have to be carried out on the results.

TABLE 1. SOME POSSIBLE UDTR RULE PATTERNS

Entry No.	Response Type *D*	*U*	*Transformation*	*Percentage point estimated*
1	XX	O,XO	$P=F^2$	0·7071
2	XXX	O,XO,XXO	$P=F^3$	0·7940
3	XXX,XXOX	O,XO,XXOO	$P=F^3(2-P)$	0·7336
4	XXXX	O,XO, etc.	$P=F^4$	0·8409
5	XXXXX	O,XO, etc.	$P=F^5$	0·8705
6	XXXXXX	O,XO, etc.	$P=F^6$	0·8908
7	X,OX	OO	$P=1-(1-F)^2$	0·2929
8	X,OX,OOX	OOO	$P=1-(1-F)^3$	0·2060

etc. by symmetry from entries 1 to 6.

X = positive response. O = negative response.

The number of observations required per run is on average slightly less than

$$1/(1-p) \quad \text{if } p>0{\cdot}50$$
$$\text{or } 1/p \quad \text{if } p<0{\cdot}50,$$

where p is the percentage point being estimated. Thus entry number 6 estimates the 89 per cent point and requires about 10 observations per run.

Approximate standard error The determination of the standard error of the percentage points is made difficult because successive peaks and valleys are very statistically dependent. However, the correlation between the averages of pairs of runs is much smaller, and an approximate standard error can be obtained as follows.

Suppose there are n peaks and n valleys in a sequence. If the sequence is broken up into pairs of runs, we can calculate the n averages of the peak and valley in each pair. The sample variance of these n averages, divided by n, is an obvious estimate of the variance of the estimated percentage point and its square root is, of course, the standard error.

Other methods of obtaining a standard error are being examined currently. For a more accurate standard error, maximum likelihood estimation should be used.

4. ILLUSTRATION OF THE METHOD

In the experiment of Levitt quoted earlier, σ was of greater interest than μ and so the following plan was adopted. UDTR plans 1 and 7 were run concurrently, but independently, the presentations for each plan being interleaved at random. An initial estimate of μ was obtained from four runs on the UD rule, and then the UDTR rules were used without the starting rule. Suppose each sequence is split into pairs of runs, each consisting of a peak and valley, and the averages of the pairs calculated. This provides a sequence of estimates of the 70·7 per cent and 29·3 per cent points. If we assume a normal psychometric curve, the difference between these percentage points is 1·09 σ, and a sequence of estimates of the percentage points and σ can be calculated as the test progresses. A record of the data is shown in Figure 4 and the associated estimates in Table 2.

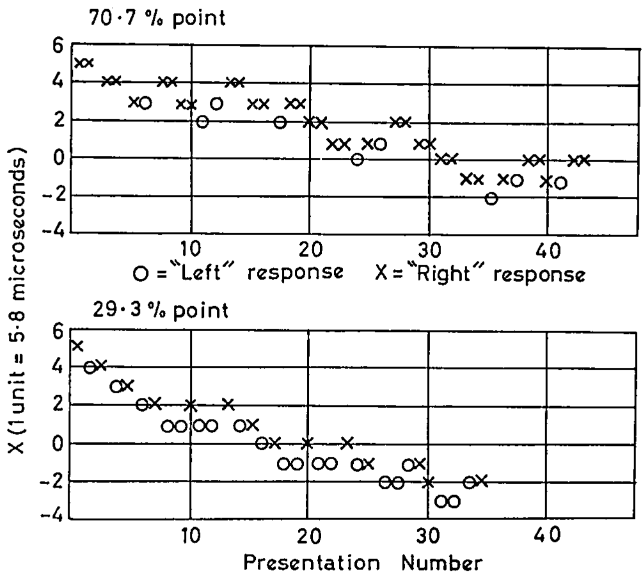


FIGURE 4. Use of UDTR rule to estimate 70·7 per cent and 29·3 per cent points.

TABLE 2. AVERAGES OF PAIRS OF PEAKS AND VALLEYS FOR THE DATA SHOWN IN FIGURE 4. UNITS μSECS

	Estimate number					
Estimates	1	2	3	4	5	6
$\hat{L}_{70\cdot7}$	20·3	17·4	14·5	5·8	−5·8	−2·9
$\hat{L}_{29\cdot3}$	8·7	8·7	−2·9	−2·9	−8·7	−14·5
$\hat{\sigma}=[\hat{L}_{70\cdot7}-\hat{L}_{29\cdot3}]/1\cdot09$	10·6	8·0	16·0	8·0	2·7	10·6
$\hat{\mu}=\frac{1}{2}[\hat{L}_{70\cdot7}+\hat{L}_{23\cdot9}]$	14·5	13·1	5·8	1·5	−7·3	−8·7

These results illustrate an interesting phenomenon which appeared several times in an experiment which involved a large amount of testing with different subjects. There appeared to be a gradual drift in the position of μ during the course of the experiment, although the difference limen σ remained reasonably constant. Since the peaks and valleys in each sequence are statistically dependent it is not easy to obtain a precise test for drift. However, it seems very unlikely that the drifts observed were the result of sampling fluctuations. The drifts were all fairly regular, and σ appeared to remain relatively constant. For some subjects, therefore, there is evidence of a gradual fluctuation in the value of ITD corresponding to the position of the listener's medial plane, although the differential sensitivity about this threshold appears to be constant.

5. Discussion

The advantages and disadvantages of the UDTR rule are now itemized.

Advantages

(i) The rule is the best currently available strategy for estimating general percentage points. (For other methods see Wetherill, 1963.) If a division of the step size is used after six runs it is very highly efficient, provided the initial step size is not too small, say not less than about $\sigma/3$.

(ii) It is very simple to use. Estimates are quickly and easily obtained and can be calculated during the course of a test. Hence the experimenter can follow the performance of a subject as the test progresses and a faulty or meaningless experiment can be detected at an early stage.

(iii) The technique is not critically dependent on the mathematical form of the psychometric function.

(iv) The average difficulty of a test can be matched to the ability of the subject. For example, in certain cases a subject may perform more reliably if the decisions required of him are roughly of equal difficulty throughout the test. The UDTR rule provides a rapid and efficient method whereby the average difficulty of the decision task (as measured by the proportion of 'correct' responses at a given stimulus level) can be controlled according to the requirements of the experiment.

(v) The artifice of running two or more strategies concurrently allows the experimenter considerable latitude in the choice of a presentation sequence, without any significant sacrifice in statistical efficiency. Since in psychophysics there is no standardized 'best' procedure, it is of great value to have a technique which is easily modified to suit a given experimental stituation. The method of interleaving two strategies at random is particularly relevant to the measurement of difference limens, where at least two separate points on the psychometric function need to be estimated. This approach also has the advantage that, by selecting two points symmetrically placed about the mid-point, the average frequency of positive responses in a test approaches 50 per cent. Blackwell (1952) has pointed out that if relatively few positive responses occur in a test then the subject may, towards the end of the test, exhibit a bias favouring a negative response. The need for interleaving two or more strategies has been discussed by Cornsweet (1962) and Smith (1961).

Disadvantages

(i) The Wetherill estimates are slightly biased from two separate causes. Firstly, some bias arises from any error in the initial guess at the level being estimated. This bias decreases with the number of runs made, and is very small for eight or more runs, when step sizes of $\sigma/2$ or more are used. Secondly, some bias arises from the shape of the transformed psychometric function. If this has the same symmetry as the cumulative normal or logistic, this source of bias is zero. More usually, the original psychometric function will be approximately cumulative normal or logistic, and the transformed curve has a shape which gives rise to a small positive bias in the estimates (negative if $p < 0{\cdot}50$). The asymptotic bias for the rules in Table 1 is of the order of $1/20 \times$ step size.

(ii) In using the UDTR rule it is assumed that successive responses are independent. This is not usually a safe assumption in psychophysics, and this aspect of the technique is at present under investigation. It should be noted that this problem is not peculiar to the UDTR rule or to sequential strategies in general. For example, the use of a randomized presentation sequence (as in the Constant Method) does not eliminate the sequential dependencies in a subject's responses, but rather confounds these effects with a host of other experimental errors. In this respect a properly executed sequential strategy is of value in that it may be possible to study some of the effects of sequential dependencies.

(iii) Apart from serial effects, of which the subject may be totally unaware, there is also the danger that the subject may recognize a pattern in the sequence of presentations and hence adjust his responses accordingly. Interleaving two or more strategies at random provides a useful protection against this danger (see Cornsweet, 1962).

(iv) The response sequences required for estimating extreme points on the psychometric function ($p > 0{\cdot}90$, $p < 0{\cdot}10$) are lengthy and may be impractical to employ.

All of these points reflect to some extent difficulties inherent in the underlying problem, rather than disadvantages peculiar to the UDTR. This is especially true of the last point, (iv).

References

Blackwell, H. R. (1952). The influence of data collection procedures upon psychophysical measurement of two sensory functions. *J. exp. Psychol.* **44**, 306–315.

Brownlee, K. A., Hodges, J. L., and Rosenblatt, M. (1953). The Up-and-Down method with small samples. *J. Amer. statist. Ass.* **48**, 262–277.

Cornsweet, T. N. (1962). The Staircase-Method in psychophysics. *Amer. J. Psychol.* **75**, 485–491.

Dixon, W. J., and Mood, A. M. (1948). A method for obtaining and analyzing sensitivity data. *J. Amer. statist. Ass.* **43**, 109–126.

Finney, D. J. (1952). *Probit analysis.* 2nd ed. Cambridge: The University Press.

Levitt, H. (1964). Discrimination of sounds in hearing. Unpublished Ph.D. thesis, University of London.

Smith, J. E. K. (1961). Stimulus programming in Psychophysics. *Psychometrika*, **26**, 27–31.

Wetherill, G. B. (1963). Sequential estimation of quantal response curves. *J. Roy. statist. Soc.*, Series B, **25**, 1–48.

Wetherill, G. B., and Chen, H. (1964). Sequential estimation of quantal response curves. II. A new method of estimation. *Bell Telephone Laboratories Technical Memorandum.*