Received 13 February 1970 4.1, 4.5

Transformed Up-Down Methods in Psychoacoustics

H. Levitt

Doctoral Program in Speech, The City University of New York Graduate Center, 33 West 42 Street, New York, New York 10036

During the past decade a number of variations in the simple up–down procedure have been used in psychoacoustic testing. A broad class of these methods is described with due emphasis on the related problems of parameter estimation and the efficient placing of observations. The advantages of up–down methods are many, including simplicity, high efficiency, robustness, small-sample reliability, and relative freedom from restrictive assumptions. Several applications of these procedures in psychoacoustics are described, including examples where conventional techniques are inapplicable.

INTRODUCTION

A. Adaptive Procedures in Psychophysics

An adaptive procedure is one in which the stimulus level on any one trial is determined by the preceding stimuli and responses. This concept is not new. Many well-known psychophysical techniques are essentially adaptive procedures. The von Békésy technique and the method of limits are two obvious examples. Recent developments in automated testing, however, have brought about a reconsideration of these methods. In particular, the up–down procedure and variations of it have received much attention in recent years (see References). The purpose of this paper is to describe the principles underlying up–down testing and to review briefly the relative advantages and shortcomings of several well-known procedures. The conventional method of constants is also included by way of comparison.

Up–down methods of testing form a subset of a broader class of testing procedures generally known as sequential experiments. A sequential experiment may be defined as one in which the course of the experiment is dependent on the experimental data. Thus far, two separate classes of sequential experiments have received considerable attention: those in which the number of observations is determined by the data and those in which the choice of stimulus levels is determined by the data. Although the former class of methods, first described by Wald (1947), has found some application in psychoacoustics (Hawley, Wong, and Meeker, 1953), it is the latter class of methods, and the up–down method in particular, that is finding widespread application in psychoacoustics. The current trend towards computer-assisted testing will undoubtedly lead to wider application of these techniques.

B. The Psychometric Function

Figure 1 shows a typical psychometric function. The stimulus level is plotted on the abscissa; the ordinate shows the proportion of "positive" responses. The definition of *positive* response depends on the type of

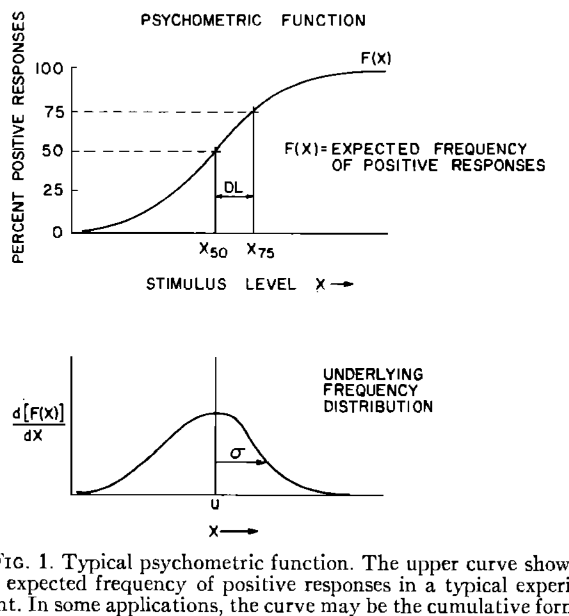

Fig. 1. Typical psychometric function. The upper curve shows the expected frequency of positive responses in a typical experiment. In some applications, the curve may be the cumulative form of the frequency distribution shown in the lower portion of the figure.

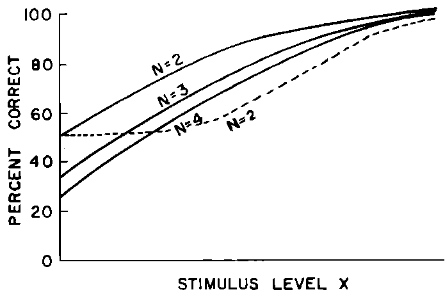

FIG. 2. The solid curves show the expected proportion of correct responses, assuming the stimulus level is directly proportional to d'. The parameter N is the number of alternatives. The dashed curve shows a psychometric function derived from the normal ogive in which an adjustment for chance responses has been made.

experiment. It may be, for example, a "signal present" response in a signal-detection experiment, or the correct identification of a word in an intelligibility test. For the case of forced-choiced experiments, a positive response may be defined as identifying the correct observation interval. Typical psychometric functions for this type of experiment are shown in Fig. 2. The solid curves show the expected proportion of correct responses, assuming the stimulus level is directly proportional to d'. The dashed curve shows a psychometric function derived from the normal ogive in which an adjustment for chance responses has been allowed. Note that the proportion of positive responses at zero stimulus level (i.e., no controlled difference between observation intervals) corresponds to random guessing and is determined by the number of possible alternatives.

For the curve shown in Fig. 1, two parameters are generally of interest: (1) the location of the curve, usually defined by the stimulus level corresponding to 50% positive responses, X_{50}, and (2) the spread of the curve, usually defined by the distance, DL, between two conveniently chosen points, such as $X_{75}-X_{50}$ or $X_{70}-X_{30}$. If the curve has a specific mathematical form, then interest may lie in determining the location and scale parameters defining the curve. For example, the psychometric function may be a cumulative normal, where the probability of a positive response at stimulus level X_1 is

$$F(X_1)=\int_{-\infty}^{X_1}\frac{1}{(2\pi)^{\frac{1}{2}}\sigma}e^{(x-\mu)^2/2\sigma^2}dX,$$

and the parameters of interest, μ and σ, are the mean and standard deviation, respectively, of the underlying normal distribution (see Fig. 1). The interpretation in psychophysical terms of μ and σ, or alternatively, X_{50} and DL, depends on the particular experiment. In an experimental comparison between a test and a reference stimulus, for example, X_{50} may be defined as *the point of subjective equality* and DL as the *difference limen* (Torgerson, 1958, p. 144). In other applications, X_{50} has sometimes been defined as the *threshold* and DL as the *differential threshold.*

For the forced-choice experiment (Fig. 2) it is frequently of interest to determine only one point on the psychometric function, such as the X_{75} level. This value has also been used to define the difference limen. Many experiments, of course, require that more than one point on the curve be determined.

C. Basic Assumptions

Several basic assumptions are usually made when using up–down, or conventional testing procedures. These are

(1) The expected proportion of positive responses is a monotonic function of stimulus level (at least over the region in which observations are made).

(2) The psychometric function is stationary with time, i.e., there is no change in the shape or location of the function during the course of a test.

(3) The psychometric function has a specific parametric form, e.g., cumulative normal.

(4) Responses obtained from the observer are independent of each other and of the preceding stimuli.

Of the above four assumptions, (1) is the least restrictive and the only one that is essential in using the up–down procedure. Assumptions (2), (3), and (4), although not essential, often facilitate experimental design. Conventional procedures, such as the method of constants, usually require assumptions (2), (3), and (4).

D. Some General Principles

Two basic considerations govern the use of any experimental procedure: (1) the placing of observations and (2) estimation based on the resulting data. The most desirable situation is obviously good placing of observations followed by a good method of estimation. The least desirable situation is bad placing of observations followed by a bad estimation procedure. In many practical cases it is sufficient to have good placing of observations followed by a simple but adequate method of estimation. This situation is far better than poor placing of observations followed by a highly sophisticated, highly efficient estimation procedure. A good[1] procedure is defined here as one that is highly efficient, robust, and relatively free of bias.

Good placing of observations depends on the parameters to be estimated. If one is interested in estimating X_p, for example,[2] one should place observations as close to X_p as possible. If one is interested only in estimating σ, observations should be placed on either side of, but at some distance from, X_{50}. Figure 3 shows how the placing of observations affects the expected error variance of the maximum-likelihood estimates of μ and σ for a cumulative normal response curve. In this example, $\mu=X_{50}$; it is assumed that the observations are roughly symmetrically placed about the midpoint and that the total number of observations is not unduly small. The ordinate shows the increase in the reciprocal of the error variance with each new observation. The abscissa shows the location of that observation relative to the midpoint of the curve. If interest centers only on estimating the slope constant $1/\sigma$ with precision, assuming μ (i.e., X_{50}) is known, then observations should be placed at a distance of 1.57σ on either side of μ. If interest centers only on estimating μ with precision, then data should be placed at μ. A good compromise for estimating both μ and σ with relative precision is to place observations at a distance σ on either side of μ, i.e., at the $X_{15.9}$ and $X_{84.1}$ levels. In this case, the reciprocal of the error variance for both estimates is roughly 70% of its maximum value.

The value of an adaptive testing procedure should now be obvious. In order to get good placing of observations, it is necessary to have some knowledge of the quantities to be estimated. As we gather information during the course of an experiment, so we may use this information to improve our placing of observations on future trials.

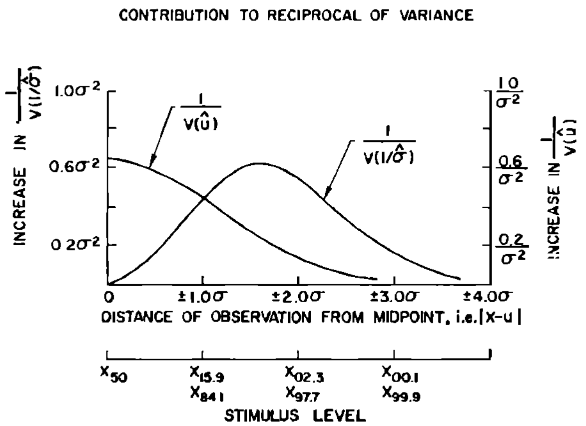

FIG. 3. Contribution to reciprocal of error variance. The two curves show the expected increase, with each new observation, in the reciprocal of the error variance for the estimates of $1/\sigma$ and μ, respectively. The predictions are for the large-sample maximum-likelihood estimates assuming a cumulative normal response curve. Ideally, the observations should be symmetrically placed about μ.

I. CONVENTIONAL PROCEDURES

A. The Method of Constants

In this experiment several stimulus levels are chosen beforehand by the experimenter, and groups of observations are placed at each of these stimulus levels. The order of the observations is randomized. A conventional method of estimation is used in fitting the psychometric function to the resulting data, e.g., probit analysis (Finney, 1952) for the cumulative normal, or Berkson's logit (1951) method for the logistic function. Simpler but less efficient methods such as the Spearman–Karber (Natrella, 1963) method may also be used.

The advantage of this procedure is that the data generally cover a wide range, and additional tests on the validity of the parametric assumptions can be included. A typical rule of thumb is for the experimenter to place the observations so as to roughly cover the range X_{10} to X_{90}. Frequently, however, a preliminary experiment is necessary in order to get some idea of the required range. The stimuli are presented in random order so that the subject cannot anticipate the stimulus sequence. The shortcomings of this technique are several, however. If one is interested in estimating only one point on the curve, then the method is inefficient in that a large proportion of the observations are placed at some distance from the region of interest. A second shortcoming is that the data for each test are pooled and a curve fitted to the pooled data. This procedure does not allow for the possibility of gradual changes in parameter values during the course of a test. Finally, difficulties arise with small samples. Slope estimates, in particular, are highly variable and subject to substantial biasing effects with small samples (Wetherill, 1963; Levitt, 1964).

B. Method of Limits

According to this method, a stimulus having a high probability of a positive response is presented to the subject. If a positive response is obtained, then the stimulus for the next trial is reduced in level. If again a positive response is obtained, the stimulus level is again reduced by the same amount (the step size). This procedure is continued until a negative response is obtained. The average value of the last two stimulus levels is used as an estimate of X_{50}. Additional measurements are sometimes obtained with the procedure inverted, i.e., starting with a stimulus level well below X_{50} and then increasing level in fixed steps until a positive response is obtained. The method of limits, or variations of it, are used in audiology and in other applications where a rapid estimate of the X_{50} level is required. It has several disadvantages. The observations are badly placed; i.e., although interest centers on X_{50}, most of the observations are placed at some distance from X_{50}. Secondly, the estimates may be substantially biased where the bias is dependent on both the step size and the choice of the initial stimulus level (Anderson, McCarthy, and Tukey, 1946; Brown and Cane, 1959).

II. UP–DOWN PROCEDURES

A. The Simple Up–Down or Staircase Method

A relatively efficient method of estimating the 50% level is the simple up–down or staircase method. It is similar to the method of limits in that the stimulus level is decreased after a positive response (or increased after a negative response), but unlike the method of limits the test is not terminated after the first reversal. A recommended procedure is to continue testing until at least six or eight reversals are obtained (Wetherill and Levitt, 1965). A typical data record is shown in Fig. 4.

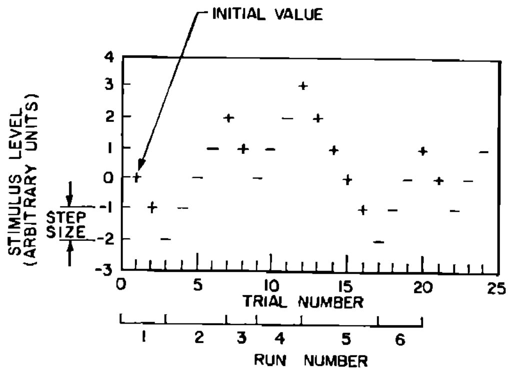

FIG. 4. Typical data for simple up–down procedure. A typical set of data using a fixed step size is shown. The initial value is usually the best *a priori* estimate of X_{50}. A run consists of a sequence of changes in stimulus level in one direction only. A highly efficient estimation procedure is to use the midpoint of every second run as an estimate of X_{50}. These estimates would be 0, 1.5, and −0.5 for runs 2, 4, and 6, respectively.

The increments by which the stimulus is either increased or decreased are referred to as *steps*. For the experiment shown in Fig. 4, a constant step size was used throughout. A series of steps in one direction only is defined as a *run*. Thus, in the example given, trials 1 to 3 constitute the first run, 3 to 7 the second run, 7 to 9 the third run, and so on. The level used on the very first trial is the *initial value*.

The simple up–down technique has the following advantages. Most of the observations are placed at or near X_{50}, which is good placing of observations for estimating the 50% level. Secondly, if there is a gradual drift during the test, the placing of observations will follow this drift. The technique, however, has the following disadvantages. The data are not well placed for estimating points other than X_{50}. Secondly, difficulties occur with very large or very small step sizes. With too large a step size, some of the data will be badly placed relative to X_{50}. If too small a step size is used, then many observations are wasted in converging on X_{50}. This may be a serious problem if a poor initial value is used. A third shortcoming is peculiar to psychophysical testing in that the subject, realizing that a sequential rule is being used, can anticipate the stimuli and adjust his responses accordingly.

Thus far, only the placing of observations has been considered. The analysis of the data is a separate problem. There are several methods of analyzing data obtained using an up–down testing procedure. One method is to pool all the data obtained in the experiment and fit the psychometric function using conventional techniques (e.g., probit analysis). This procedure is based on essentially the same assumptions as the method of constants. An alternative method of analysis developed specifically for up–down data and which is computationally simpler than probit analysis, although based on the same basic assumptions, is described by Dixon and Mood (1948).These methods, however, are for use with pooled data and, as such, require that there be no change in parameter values during an experiment.

An extremely simple method of estimation having distinct advantages is that developed by Wetherill (1963), in which the peaks and valleys of all the runs are averaged to provide an estimate of X_{50}. In order to reduce estimation bias, it is recommended that an even number of runs be used. This method of estimation is equivalent to taking the midpoint of every second run as an estimate of X_{50}. These are defined as mid-run estimates.[3] Apart from their simplicity, empirical sampling studies have shown that the mid-run estimates are robust, relatively efficient, and low in estimation bias provided the response curve is reasonably symmetric about the X_{50} level. The estimation bias with asymmetric response curves has not, as yet, been determined. It is apparent, however, that estimation bias will increase with increased asymmetry.

The precision of the mid-run estimates for the symmetric response curve was found to be excellent. For large experiments, the precision of the mid-run estimates was very close to the maximum-likelihood estimates,[4] and for small experiments (less than 30 trials) the technique was in fact more efficient than the maximum-likelihood method (Wetherill, Chen, and Vasudeva, 1966). The mid-run estimates have an additional advantage in that the sequence of contiguous estimates provides a direct indication of any significant drifts with time in the location of the response curve. The precision with which the mid-run estimates can track gradual drifts has not yet been determined. Preliminary studies indicate that it is a complicated function of the step size, the rate at which the point being estimated is changing in value, and the extent to which these changes are predictable.

A difficulty with *w* estimates is that the estimates obtained from adjacent runs are correlated. This makes the estimation of within-test variability a little more complicated than would be the case if the midpoints of successive runs were uncorrelated. Simple methods of estimating within-test variability are currently being evaluated.

Some of the difficulties encountered with the simple up–down procedure can be obviated by minor modifications of the technique. In the event that little is known about either the spread or location of the psychometric function, then it is recommended that at the start of an experiment a large step size be used which is gradually decreased during the course of the experiment. Robbins and Monroe (1951) suggested that the step size on trial n be equal to c/n where c is a constant. It has since been shown (Chung, 1954) that, under fairly general conditions, this method of reducing step size leads to a maximal, or near maximal, rate of convergence on the target stimulus level. The variance of the asymptotic distribution of stimulus values about the target value X_{50} is minimized if the constant c is equal to $0.5/b$, where b is the slope of the response curve in the region of X_{50}

TABLE I. Response groupings for transformed up–down strategies. Several simple response groupings are shown. Entry 1 corresponds to the simple up–down procedure. Entry 2 corresponds to the method used by Zwislocki *et al.* (1968) and Heinemann (1961). Entries 2 and 3, and 5 and 6, with random interleaving, were used by Levitt (1964). Entry 7 is typical of the BUDTIF procedure proposed by Campbell (1963). Entry 8 was used by Levitt and Rabiner (1967).

	Response sequences		Response groupings	
Entry	UP group increase level after:	DOWN group decrease level after:	Probability of a sequence from DOWN group = P[DOWN]	Probability of positive response at convergence
1	−	+	$P(X)$	$P(X)=0.5$
2	+ − or −	+ +	$[P(X)]^2$	$P(X)=0.707$
3	− −	− + or +	$[1-P(X)]P(X)+P(X)$	$P(X)=0.293$
4	+ + − or + − or −	+ + +	$[P(X)]^3$	$P(X)=0.794$
5	+ + + − or + + − or + − or −	+ + + +	$[P(X)]^4$	$P(X)=0.841$
6	− − − −	− − − + or − − + or − + or +	$1-[1-P(X)]^4$	$P(X)=0.159$
7	Any group of 4 responses with 1 or more negative responses	+ + + +	$[P(X)]^4$	$P(X)=0.841$
8	− − − + − + − −	+ + + − + − + +	$[P(X)]^2[3-2P(X)]$	$P(X)=0.5$

(Wetherill, 1963). This result, however, is based on the assumption that the response curve is approximately linear in the region of the target value. In practice, this is not usually a serious restriction since, with the exception of the first few trials, most observations are placed close to the target value with a step size that is small compared to the curvature of the response curve. Difficulties may occur if the response curve has a sharp discontinuity at or near the target value, but this is not a common situation in psychoacoustics.

An important practical problem is deciding on the size of the first step. If many repeated measurements are being made, then an estimate of slope as obtained from previous experiments may be used to determine the initial step size. If this is not known, it is necessary to guess at the initial step size. The reduction in efficiency is not very great if too large an initial step size is used, but it can be disastrous if the initial step size is too small. For example, if the initial step size is twice the optimum value, the efficiency is reduced by 25%. If the initial step size is half the optimum value, the reduction in efficiency is close to 100%. Hence, when in doubt, aim at a larger initial step size.

Although the method of reducing step size after every trial has distinct theoretical advantages, its practical implementation may require fairly complicated equipment. A simple approximation is to reduce the step size by a proportionate amount after one or more complete runs. A convenient practical procedure is to halve the step size after the first, third, seventh, fifteenth, etc., runs. After each halving of the step size, the next run is started at the average of the peaks and valleys for the preceding step size. Empirical sampling trials (Wetherill, 1963) indicate that, if the spread and location of the response curve are known within reasonable limits (e.g., for a symmetric ogive, the standard deviation σ is known within 0.5–2.0 of its true value and X_{50} is known within $\pm 2\sigma$), then a very good approximation to the Robbins and Monro efficiency is obtained by halving the step size only once, preferably after the third run.

The recommendation that the step size be systematically decreased during a test is based on the assumption that the response curve is fixed throughout the experiment. If this assumption is not valid (e.g., there may be a gradual drift in either the shape or location of the curve during a test), then a more sophisticated rule allowing for a possible increase as well as a decrease in step size may be necessary. One set of rules incorporating this possibility has been proposed by Taylor and Creelman (1967). The optimum strategy for increasing or decreasing step size depends on the type and the extent of the changes that are likely to occur during a test, and this is not always known.

In order to prevent the subject from anticipating the stimuli used on each trial, the presentations for two or more strategies can be interleaved at random

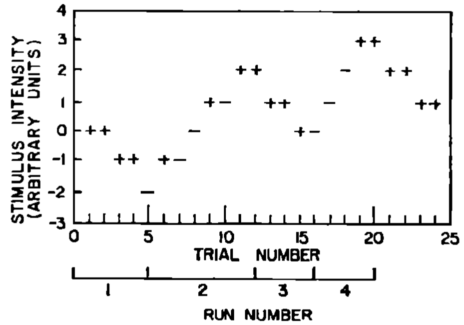

FIG. 5. Typical data for transformed up–down procedure. The data are typical of the strategy corresponding to Entry 2 of Table I, which converges on the $X_{70.7}$ level. The rule for changing stimulus level is analogous to that for the simple up–down procedure, except that one of a set of response sequences is required before a change in stimulus level. As with the simple up–down procedure, a run is a sequence of changes in stimulus level in one direction only.

(Cornsweet, 1962; Smith, 1961); that is, two or more tests are performed simultaneously where observations for each of the tests are assigned at random.

B. The Transformed Up–Down Procedure

The simple up–down procedure is designed primarily to place observations in the region of X_{50} and hence is not well suited for estimating points other than X_{50}. A general method for estimating points on a psychometric function is the transformed up–down procedure. Observations, or sequences of observations, are categorized into two mutually exclusive groups. These are termed the UP group and the DOWN group, respectively. The method of grouping the observations depends on the point to be estimated. Table I shows some typical groupings including several proposed in other contexts (Zwislocki *et al.*, 1958; Heinemann, 1961; Campbell, 1963). The rule for controlling the stimulus level is analogous to the simple up–down rule, except that the stimulus level is changed only after a sequence of observations belonging to either the UP or the DOWN group is obtained. The stimulus level is not changed until such a sequence is obtained. For example, according to Entry 2 in Table I, the stimulus level would be increased after either a negative response or a positive response followed by a negative response on the next trial. The stimulus level is decreased after two consecutive trials yielding positive responses. Note that, as the test progresses, one or other of these sequences must be obtained.

Some typical data for this strategy are shown in Fig. 5. In this illustrative example, the changes in stimulus level resemble those for the simple up–down strategy shown in Fig. 4. That is, if in Fig. 5 the $-$ and $+ -$ response sequences belonging to the UP group are replaced by a single $-$ response and the $+ +$ sequence belonging to the DOWN group is replaced by a single $+$ response, then the resulting set of *transformed responses* is identical to that of the simple up–down strategy shown in Fig. 4. In the case of the transformed up–down strategy, however, the average number of trials per run is increased. Runs 2, 3, and 4, for example, consist of trials 5–12, 11–16, and 15–20, respectively. The midpoints of these runs are 0, 1, and 1.5 units, respectively.

The transformed up–down strategy tends to converge on that stimulus level at which the probability of a DOWN response sequence equals the probability of an UP responses equence (i.e., the probability of either equals 0.5). It is a relatively simple matter to compute the probability of obtaining either an UP or a DOWN sequence. For example, referring again to Entry 2 of Table I, the probability of obtaining an UP sequence (i.e., either $+-$ or $-$) is $P(X)[1-P(X)]+[1-P(X)]$, where $P(X)$ is the probability of a positive response at stimulus level X. The probability of getting a DOWN response sequence (i.e., $++$) is simply $[P(X)]^2$. The strategy therefore converges on that value of X at which $[P(X)]^2=0.5$, i.e., $P(X)=0.707$.

The curve relating the probability of a DOWN response sequence to stimulus level is the *transformed response curve.* Figure 6 shows the transformed response curve for Entry 2 of Table I. Note that, for the simple up–down procedure, the transformed response curve is numerically identical to the subject's psychometric function. The performance of a transformed up–down (or UDTR)[5] procedure in terms of the transformed response curve is analogous to that of the simple up–down procedure in terms of the subject's psychometric function. For example, the transformed up–down strategy converges on the 50% point of the transformed response curve, which for the example shown in Fig. 6, corresponds to the 70.7% point of the subject's psychometric function. Essentially the same methods of estimation may be used with the transformed up down procedure as with the simple up–down procedure. In particular, the simple yet relatively efficient mid-run estimates may be used. In this case, the midpoint of a run is an estimate of the 50% point of the transformed response curve. As with the simple up–down procedure, efficiency of estimation may be increased by systematically decreasing step size during the test. The analogy between the simple and transformed up–down procedures is particularly valuable in that transformed up–down procedures may be evaluated in terms of the extensive results, including empirical sampling trials, already obtained for the simple up–down procedure.

A useful feature of the transformations listed in Table I is that for many typical psychometric functions the transformed response curves are approximately normal in the region of the 50% point. This approximation was made use of by Heinemann (1961), who applied the simple computational procedure of Dixon and Mood (1948) based on the normal distribution to estimate the 50% point of the transformed response curve.

Entries 2–6 in Table I are based on the method of inverse binomial sampling (Haldane, 1945) and are relatively efficient and easy to instrument (Levitt and Bock, 1967). Manual control of the stimulus level is also possible, and special control charts have been developed to facilitate the procedure (Levitt and Treisman, 1969). Entry 7 of the table is typical of the BUDTIF testing procedure proposed by Campbell (1963). It is also easy to instrument, but converges on the target value at a slightly less rapid rate, since more trials are required on the average before changing level. More complicated strategies than those shown in Table I may of course be used, including a wide range of possible transformations based on Wald's (1947) probability-ratio rule.

Wald's method of sequential testing is designed specifically for minimizing the expected number of trials required for determining, with specified error probabilities, whether or not a given hypothesis is true. For the case of binary data, the technique can be used to determine if the probability of a positive response lies within prescribed bounds.[6] The probability-ratio rule may be applied to up-down testing (Taylor and Creelman, 1967) by increasing the stimulus level whenever the rule indicates that the probability of a positive response lies below the lower bound, and decreasing stimulus level whenever the rule indicates that the probability of a positive response lies above the upper bound. Given an intelligent choice of the α and β error probabilities and of the prescribed bounds within which the measured probability should lie, an up-down procedure of this type will converge rapidly on the target value and concentrate observations within the prescribed region. Simplifying assumptions, such as setting $\alpha=\beta$, $p-p_1=p_2-p$, may be used, leading to simple practical procedures for deriving the UP or DOWN response sequences. The special advantages of this procedure depend on the intended application and the choice made for α, β, p_1, and p_2. In general, the technique provides the experimenter with greater power and flexibility in controlling the placing of observations. The relevant response sequences, however, may be more complicated than those shown in Table I.

The transformed response curve may be computed for the probability-ratio rule as for any other grouping of response sequences. The simple mid-run estimates are also applicable to this technique, provided the step size is fixed over one or more runs. The PEST procedure as specified by Taylor and Creelman (1967) is an amalgam of the probability-ratio rule and some special rules for changing step size, including the possibility of an increase in step size. As such, it is difficult to evaluate in terms of known results for the simple up-down procedure.

A technique which has important implications for up-down testing is that of interleaving trials from two or more sequential strategies being run concurrently. One important practical application is that when two points symmetrically placed about X_{50} are to be estimated. Trials for the two strategies may be interleaved at random. The particular strategies to be used

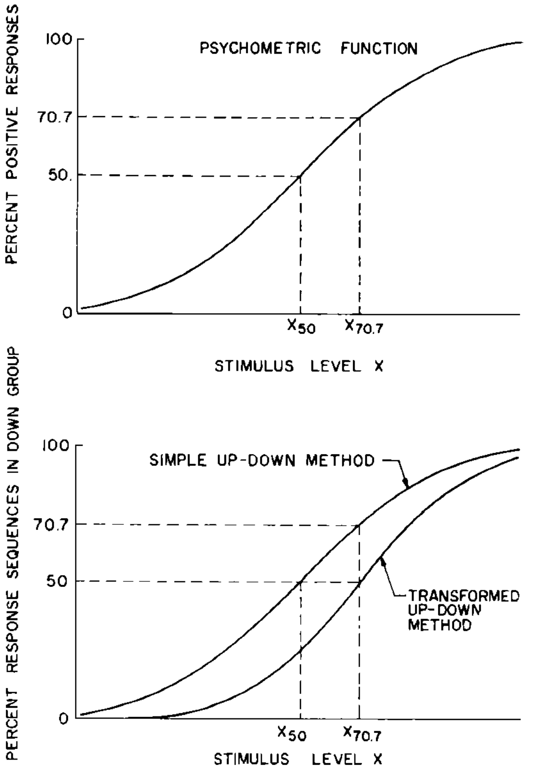

FIG. 6. Transformed response curves. The upper half of the figure shows the psychometric function. The lower half of the figure shows the transformed response curves for Entries 1 and 2 of Table I. The curve for the simple up-down method is not transformed in any way and is numerically identical to the psychometric function.

depend, of course, on the requirements of the experiment. If the pyschometric function is approximately normal and both the location and scale parameters (μ and σ) are to be estimated with reasonable precision, then strategies for estimating $X_{15.9}$ and $X_{84.1}$ (Entries 5 and 6 of Table I) may be used. These two strategies concentrate observations at two symmetrical points at a distance of σ on either side of μ. According to Fig. 3, observations placed in this region provide a practical compromise for estimating both μ and σ with relatively high precision. Although the curves of Fig. 3 are for large sample maximum-likelihood estimates, these predictions are not very different for other estimation methods of comparable precision. If a slightly less precise estimate of σ can be tolerated, then a very practical combination of strategies would be Entries 2 and 3 of Table I for estimating the $X_{70.7}$ and $X_{29.3}$ stimulus levels, respectively. In this case, data would be concentrated at roughly 0.54σ on either side of μ. Compared to other transformed up-down methods, the strategies for estimating $X_{29.3}$ and $X_{70.7}$ require relatively few trials per run and are more sensitive to tracking gradual drifts in parameter values during a test.

The process of interleaving at random ensures that the subject cannot anticipate the rule used by the experi-

TABLE II. Estimates of σ obtained from pairs of transformed up–down strategies. The data are estimates of the spread (σ) of a cumulative normal lateralization curve (Levitt, 1964). Left–right judgments were obtained for a binaural sound image, where the controlled variable was interaural time delay varied in steps of 5.8 μsec. The replication error (i.e., the within-cell error variance) is shown at the bottom of the table. The relative precision of the two estimates is comparable to that predicted by Fig. 3 for data placed near $\mu \pm 0.54\sigma$ and $\mu \pm \sigma$, respectively.

Subject	Estimates of σ (μsec) $(X_{70.7}-X_{29.3})/1.09$	$(X_{84.1}-X_{15.9})/2$
1	12.2	13.7
2	20.0	15.0
3	21.7	17.1
4	8.2	14.2
5	7.3	7.4
Mean	13.9	13.5
Within-cell error variance	39.5	22.0
Average No. of trials	60	76
Error variance/100 trials	23.7	16.7

menter. By choosing two points symmetrically placed about X_{50}, the average number of positive responses tends to equal the average number of negative responses. If there is reason to believe that the psychometric function is not symmetric about the X_{50} point, a third strategy for estimating X_{50} may be run concurrently to check for symmetry.

III. APPLICATIONS

One of the most useful applications of up–down procedures is the tracking of gradual drifts in parameter values. Zwislocki *et al.* (1958) used the *forced-choice tracking method* (equivalent to Entry 2 of Table I) to track variations in auditory detection. In another application (Levitt, 1964), precision of lateralization was measured despite gradual changes in the subject's reference plane. In this study the parameter of interest, the spread of the lateralization curve, remained relatively constant, but the effect of gradual changes in the location of the lateralization curve had to be compensated for.

During the lateralization study, a comparison was made between measurements of spread (i.e., σ) obtained by estimating $(X_{84.1}-X_{15.9})/2$ and by estimating $(X_{70.7}-X_{29.3})/1.09$. Both of these estimates have an expected value of σ for the cumulative normal response curve. The results for five subjects are shown in Table II. The stimulus consisted of a continuous recording of male speech presented binaurally via headphones. The apparent lateral position of the sound image was controlled by an interaural delay which could be varied in steps of 5.8 μsec. The listener was restricted to a simple binary response: whether the binaural sound image appeared to come from the left-hand or right-hand side of the head. Each cell in Table II is the average of six tests. Fewer runs per test were obtained for the $X_{84.1}$ and $X_{15.9}$ estimates so as to compensate in part for the greater number of trials per run required with these estimates. On average, 60 trials per test were obtained for the $(X_{70.7}-X_{29.3})/1.09$ estimates and 76 trials per test for the $(X_{84.1}-X_{15.9})/2$ estimates. No significant differences were observed between the estimates of σ obtained from the two techniques. As predicted by the curves of Fig. 3, the precision of the σ estimates was greater when placing observations in the region of $X_{84.1}$ and $X_{15.9}$ than when placing observations in the region of $X_{70.7}$ and $X_{29.3}$. The error variances were compared on the basis of the expected variance for 100 trials. This was done so as to account for the differences in average number of trials per test

In the lateralization study there was every reason to expect the psychometric function to be symmetric about the midpoint; i.e., a delay to the left ear has a similar but opposite effect as a delay to the right ear. Not all pyschometric functions, however, exhibit the same degree of symmetry. In a recent study on the detection of a 250-Hz tone in noise using a YES–NO technique (Levitt and Bock, 1967), the slope of the response curve in the region of 29.3% detection was almost one-third less than that in the region of 70.7% detection. A concomitant result was that the precision of estimation for the $X_{70.7}$ point was correspondingly greater than that for the $X_{29.3}$ point. An analogous effect occurs with the Békésy audiometer in that the variability of valleys (minima in stimulus level) is greater than the variability of peaks (maxima in stimulus level).

A less obvious application of transformed up–down procedures is to extend the range over which the transformed response curve is approximately symmetric. In an experiment on intelligibility testing (Levitt and Rabiner, 1967), it was found that the intelligibility function flattened off sharply at 80% intelligibility. A simple transformation was used (Entry 8, Table I) which raised the flattening effect on the transformed-response curve to above the 90% level. As a result, a substantially greater portion of the data fell within the symmetric region of the intelligibility function, leading to fewer wasted observations and greater efficiency in estimating X_{50}. Note that, in this case, unlike other transformed up–down procedures, the strategy converged on the 50% level.

A key assumption in a large number of testing procedures is that each response is independent of preceding stimuli and responses. This is not always a safe assumption, and it would be convenient to test this assumption during the course of an experiment. A convenient method of doing this is to run two or more identical up–down strategies concurrently and to interleave the trials for each strategy according to a rule such that, should a sequential dependency exist, there would be a significant difference between the estimates obtained from each strategy (Levitt, 1968). For example, a check on first-order response dependencies is obtained if two strategies are interleaved such that the trials for one strategy always follow a positive response, and trials for the second strategy always follow a negative response. If there is no response dependency, then the two strategies should yield estimates that do not differ by an amount significantly greater than the combined sampling error of the two estimates. Note that the rule for interleaving strategies is independent of the rules used for changing stimulus level within each strategy. The above rule for interleaving was designed to detect first-order response dependencies. Similar rules may be developed for detecting higher-order dependencies, including dependencies that are a function of the preceding stimulus levels as well as preceding responses.

An important advantage of the method of interleaving by rule is that the effect of a response dependency is obtained directly in terms of the quantity being measured. Usually, data on sequential dependencies are specified in terms of transitional probabilities, and it is not always easy to gauge the possible effect on the measurements of wrongly assuming independent responses. In many psychoacoustic investigations, the assumption of independent responses is no more than a simplifying approximation, and it is important to know if the accuracy of the measurements is affected more than negligibly by a breakdown in this assumption.

A useful conceptual aid for evaluating the performance of up–down strategies when sequential dependencies are known to exist is the *transitional response curve*. Figure 7 shows the transitional response curves for a first-order response dependency. The upper curve shows probability of a positive response as a function of stimulus level given that the previous response was positive. The lower curve shows the probability of a positive response, given that the previous response was negative. The two curves of Fig. 7 are based on data obtained for one subject in an experiment on sequential response dependencies in the YES–NO detection of a 250-Hz tone (Levitt, 1968). Note that the two curves diverge at low detection levels. The difference between the two curves at the 50% level is roughly 1 dB. This is the expected difference between the estimates for two simple up–down strategies interleaved according to the rule described earlier. The transformed response curves for data having a known response dependency may be derived from the transitional response curves in a manner analogous to that shown in Fig. 6. A separate transformed response curve would be derived for each transitional response curve.

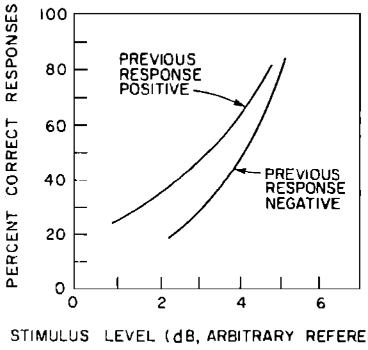

FIG. 7. Transitional response curves. The upper curve shows the expected frequency of positive responses, given that the preceding response was positive. The lower curve shows the expected frequency of positive responses, given that the preceding response was negative. The curves are based on data obtained in an experiment on sequential response dependencies (Levitt, 1968).

IV. DISCUSSION

Adaptive testing procedures offer many advantages over conventional procedures, including higher efficiency, greater flexibility, and less reliance on restrictive assumptions. Although higher efficiency (and hence greater precision for a fixed number of observations) is often thought of as the major advantage of adaptive procedures, the latter advantages may well be of greater practical importance. In many cases, increased efficiency represents an improvement in degree, whereas freedom from restrictive assumptions represents an improvement in kind. Thus, for example, it is possible to use the up–down procedure to track gradual drifts in parameter values or to compensate for an unpredictable trend in one parameter while estimating a second parameter. Both of these problems are beyond the scope of the conventional method of constants, which requires the assumption of a fixed response curve. In addition, up–down methods do not require any parametric assumptions regarding the form of the response curve. The only major restriction is that the probability of a positive response increase monotonically with stimulus level. A very large number of experiments in psychoacoustics satisfy this requirement. Also, the restriction of monoticity need only hold over the range in which data are to be placed.

Although up–down procedures are relatively free of restrictive assumptions, it is nevertheless to the experimenter's advantage to make use of additional assumptions if they are known to be reliable. The choice of initial-value step-size rules for controlling step size and methods of interleaving depend on the experimenter's prior knowledge, however rough, of the quantities to be measured. If reasonable assumptions can be made regarding the form of the response curve, the extent of possible drifts in parameter values and the existence (or absence) of any significant response dependencies, then an extremely efficient up–down strategy can be designed. If one or more of these assumptions turns out to be false, however, then the penalty is a loss in efficiency rather than an invalid result.

It is important to recognize the distinction between the problem of placing observations and the subsequent estimation problem. Although the same statistical methods can be used for both purposes (e.g., in estimating X_p, each observation could be placed at the best current estimate of X_p), there is a substantial difference in emphasis between the two operations. Whereas some latitude may be allowed in precision of placing, as shown by the fairly flat curves of Fig. 3, whatever mistakes are made during this operation are irreparable. No amount of statistical manipulation can overcome the effects of bad placing of observations. It is also possible that the final estimates may be obtained by using additional information which was not available for the placing of observations. Similarly, ancillary information may be used to assist in the placing of observations, but which is not used in the subsequent data analysis. A practical example of the latter situation is as follows. On each trial, the subject is required to give both a binary judgment and a confidence rating. The binary judgments are used to decide on the direction of change in the stimulus level, as in any standard up–down procedure, and the confidence ratings are used to decide on the step size; e.g., a large step size is used after a high confidence rating. The psychometric function, however, is fitted to the binary data only by means of conventional techniques (e.g., maximum likelihood). It is not permissible to use the simple mid-run estimates in this context, since steps of variable size are used. The role of the confidence ratings is to improve the placing of observations. The information obtained from the ratings appears potentially most useful during the early part of an experiment, where little is known *a priori* about the location or scale of the psychometric function. The experimenter can, at any stage, revert to a standard up–down procedure using binary data only.

Throughout this paper, the emphasis has been on simple, practical procedures that do not require complicated equipment. If, however, sophisticated instrumentation is available, then more complex adaptive strategies can be used to advantage. Hall (1968), for example, has developed an on-line procedure using a digital computer such that each observation is placed at the maximum-likelihood estimate of the target value derived from the data already at hand. In another application, Smith (1966) has derived a strategy in which the information gained on each trial is maximized. Predictions of precision based on the curves of Fig. 3 indicate that a well-designed up–down procedure (such as a good approximation to the Robbins–Monro procedure) will place observations sufficiently close to the target value so as to obtain a precision of estimation within about 30% of that obtainable, if all data were to be placed at the target value. Since no procedure can consistently place observations exactly at the target value, the potential gain in efficiency in going from a well-designed up–down procedure to the more sophisticated procedures is not very great. However, for more complicated experiments in which several variables are under the control of the experimenter, the cumulative gain in efficiency may be quite large. The major advantages of computer-assisted testing would seem to lie not so much in improving efficiency in fairly standardized experiments, but rather in developing new methods of experimentation.

In conclusion, it should be noted that there is no generally optimum testing procedure. Each technique has its own merits and shortcomings. Techniques that are theoretically very highly efficient are usually more complex and tend to place greater reliance on the underlying assumptions. Special problems also occur with small samples. Many of the theorems showing maximum efficiency or maximum rates of convergence are only asymptotically true, and testing procedures based on these results may be inferior in experiments of limited size. In psychoacoustics in particular, where experiments are of limited duration and the reliability of underlying assumptions are often suspect, it is valuable to have a flexible set of testing procedures that are not heavily dependent on underlying assumptions and are readily adapted to match the requirements of a given experiment. Transformed up–down methods provide an extensive family of such procedures. Furthermore, these techniques are simple, robust, and highly efficient.

[1] It is, of course, possible to speak of optimum procedures where some desirable property of the estimate (e.g , efficiency) is maximized. However, since over-all quality involves several different properties, it is more realistic to speak of "good" rather than "optimum" procedures.

[2] Xp is the stimulus level at which p% positive responses are obtained.

[3] Wetherill (1966, p. 171) defines the w estimate as the mid-value of the last step used in a run. The mid-run estimate, as defined here, is the average of two consecutive w estimates.

[4] It is a fairly common practice to use the maximum-likelihood procedure as a standard for the comparison of different methods of estimation. Although it can be shown that, under fairly general conditions, the maximum-likelihood estimates are asymptotically efficient (Kendall and Stuart, 1967), this result is not necessarily true for small samples. For the particular case of fitting binary response curves with relatively small samples, the minimum chi-squared estimates may be preferable to maximum likelihood in terms of efficiency and estimation bias (Berkson, 1955).

[5] UDTR stands for up–down transformed response (Wetherill and Levitt, 1965).

[6] For example, if p_1 and p_2 are the upper and lower bounds, respectively, and p is the expected proportion of positive responses, then the probability-ratio rule will minimize the number of trials required on average, to determine if $p>p_1$ or $p<p_2$ with error probability α of wrongly decreasing level when $p<p_2$, and error probability β of wrongly increasing level when $p>p_1$.

REFERENCES

Anderson, T. W., McCarthy, P. I., and Tukey, I. W. (1946). "Staircase Methods of Sensitivity Testing," NAVORD Rep. 65-46 (21 March).

Berkson, J. (1951). "Why I Prefer Logits to Probits," Biometrics *7*, 327–339.

Berkson, J. (1955). "Maximum-Likelihood and Minimum Chi-Square Estimates of the Logistic Function," J. Amer. Statist. Ass. *50*, 130–162.

Brown, A. E. (1965) "Measurement of Auditory Thresholds," J. Acoust. Soc. Amer. *38*, 86–92.

BROWN, J., and CANE, V. R. (1959). "An Analysis of the Limiting Method," Brit. J. Statist. Psychol. *12*, 119–126.

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.

CAMPBELL, R. A. (1963). "Detection of a Noise Signal of Varying Duration," J. Acoust. Soc. Amer. *35*, 1732–1737.

CARDOZO, B. L. (1966). "A Sequential Up-and-Down Method," Inst. voor Perceptie Onderzock, Ann. Prog. Rep. *1*, 110–114.

CHUNG, K. L. (1954). "On a Stochastic Approximation Method," Ann. Math. Statist. *25*, 463–483.

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* (Cambridge U. P., Cambridge, England), 2nd ed.

HALDANE, J. B. S. (1945). "A Labour-Saving Method of Sampling," Nature (London) *155*, 49–50.

HALL, J. L. (1968). "Maximum-Likelihood Sequential Procedure for Estimation of Psychometric Functions," J. Acoust. Soc. Amer. *44*, 370(A).

HAWLEY, M. E., WONG, W., and MEEKER, W. F. (1953). "Sequential Statistical Analysis for Word Articulation Tests," J. Acoust. Soc. Amer. *25*, 192(A).

HEINEMANN, E. G. (1961). "The Relation of Apparent Brightness to the Threshold for Differences in Luminance," J. Exp. Psychol. *61*, 389–399.

KAPPAUF, W. E. (1967). "Use of an On-Line Computer for Psychophysical Testing with the Up-and-Down Method," Amer. Psychol. *32*, 207–211.

KENDALL, M. G., and STUART, A. (1967). *The Advanced Theory of Statistics* (Hafner, New York), 2nd ed., Vol. 2.

LEVITT, H. (1964). "Discrimination of Sounds in Hearing," PhD thesis, University of London.

LEVITT, H. (1968). "Testing for Sequential Dependencies," J. Acoust. Soc. Amer. *43*, 65–69.

LEVITT, H., and BOCK, D. (1967). "A Sequential Programmer for Psychophysical Testing," J. Acoust. Soc. Amer. *42*, 911–913.

LEVITT, H., and RABINER, L. R. (1967). "Use of a Sequential Strategy in Intelligibility Testing," J. Acoust. Soc. Amer. *42*, 609–612.

LEVITT, H., and TREISMAN, M. (1969). "Control Charts for Sequential Testing," Psychometrika *34*, 509–518.

MCCARTHY, P. J. (1969). "A Class of Methods for Estimating Reaction to Stimuli of Varying Severity," J. Educ. Psychol. *40*, 143–156.

MILLER, G. A., and GARNER, W. R. (1944). "Effect of Random Presentation on the Psychometric Function: Implications for a Quantal Theory of Discrimination," Amer. J. Psychol. *57*, 451–467.

NATRELLA, M. G. (1963). *Experimental Statistics* (Nat. Bur. Stand. Handbook 91, Superintendent of Documents, U. S. Govt. Printing Office, Washington, D. C.), Sec. 10.2.

RAO, C. R. (1952). *Advanced Statistical Methods in Biometric Research* (Wiley, New York).

ROBBINS, H., and MONRO, S. (1951). "A Stochastic Approximation Method," Ann. Math. Statist. *22*, 400–407.

SMITH, J. E. K. (1961). "Stimulus Programming in Psychophysics," Psychometrika *26*, 27–31.

SMITH, J. E. K. (1966). "Discussion on Psychophysical Testing," Conf. Computer Assisted Testing, Educ. Testing Service, Princeton, N. J.

TAYLOR, M. M., and CREELMAN, C. D. (1967). "PEST: Efficient Estimates on Probability Function," J. Acoust. Soc. Amer. *41*, 782–787.

TORGERSON, W. S. (1958). *Theory and Methods of Scaling* (Wiley, New York).

WALD, A. (1947). *Sequential Anaylsis* (Wiley, New York).

WETHERILL, G. B. (1963). "Sequential Estimation of Quantal Response Curves," J. Roy. Statist. Soc. *B25*, 1–48.

WETHERILL, G. B. (1966). *Sequential Methods in Statistic* (Methuen, London).

WETHERILL, G. B., CHEN, H., and VASUDEVA, R. B. (1966). "Sequential Estimation of Quantal Response Curves: A New Method of Estimation," Biometrika *53*, 439–454.

WETHERILL, G. B., and LEVITT, H. (1965). "Sequential Estimation of Points on a Psychometric Function," Brit. J. Math. Statist. Psychol. *18*, 1–10.

ZWISLOCKI, J., MAIRE, F., FELDMAN, A. S., and RUBIN, H (1958). "On the Effect of Practice and Motivation on the Threshold of Audibility," J. Acoust. Soc. Amer. *30*, 254–262 (see Expt. V).