DOCUMENT RESUME

ED 135 837 TM 006 062

AUTHOR Livingston, Samuel A.
TITLE Choosing Minimum Passing Scores by Stochastic Approximation Techniques.
INSTITUTION Educational Testing Service, Princeton, N.J. Center for Occupational and Professional Assessment.
REPORT NO ETS-CCPA-76-02
PUB DATE Sep 76
NOTE 27p.

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS *Cutting Scores; *Performance Tests; *Statistical Analysis
IDENTIFIERS Robbings and Monro Method; *Stochastic Approximation; Up and Down Method

ABSTRACT

Often a written test is used as an inexpensive substitute for a performance measure. A specified minimum performance level or probability of successful performance can be translated into a minimum passing score for the written test most efficiently by measuring the performance of students whose written test scores are near the desired cutoff score. Stochastic approximation methods accomplish this purpose. The up-and-down method and the Robbins-Monro process are presented, discussed, and compared. (Author)

COPA-76-02

# Choosing Minimum Passing Scores by Stochastic Approximation Techniques

Samuel A. Livingston

Center for Occupational and Professional Assessment
Educational Testing Service

Princeton, New Jersey

September, 1976

This Bulletin is a draft for interoffice circulation. Corrections and suggestions for revision are solicited. The Bulletin should not be cited as a reference without the specific permission of the author. It is automatically superseded upon formal publication of the material.

Choosing Minimum Passing Scores by
Stochastic Approximation Techniques

## ABSTRACT

Often a written test is used as an inexpensive substitute for a performance measure. A specified minimum performance level or probability of successful performance can be translated into a minimum passing score for the written test most efficiently by measuring the performance of students whose written test scores are near the desired cutoff score. Stochastic approximation methods accomplish this purpose. The up-and-down method and the Robbins-Monro process are presented, discussed, and compared.

Acknowledgments

I wish to thank Frederic M. Lord for suggesting this application, and Henry Braun, Miles Davis, Vaclav Fabian, Roger H. Farrell, and Donald B. Rubin for their helpful advice and comments.

Written tests in education (and in other fields as well) are frequently used to make decisions that require the persons tested to be divided into two groups on the basis of their level of competence. In many cases the written test serves as an inexpensive substitute for an expensive individual assessment or performance test. For example, a school might want to determine which students need instruction in basic reading skills. The school cannot afford to have a group of experts assess the skills of each pupil individually, but the school can afford to have all the pupils take a written test. Those who score below a certain level on the written test will be given the basic instruction. But how should the school determine that score level?

A similar problem often arises in the case of professional certification and licensing examinations. Cost considerations rule out the possibility of having each applicant take a full-scale performance test covering an adequate sample of the tasks involved in the practice of the profession. Therefore written tests are commonly used. In this case, the setting of standards for acceptable performance would seem to be a simple exercise of professional judgment by the licensing agency. However, the written test is only an indirect measure of the skills to be tested. How can the agency's experts translate their judgment of a minimum acceptable level of <u>actual performance</u> into a minimum passing score for the written test?

These problems correspond closely to some common problems in biological and industrial testing, and the techniques that have been developed for those fields can be applied to education as well. For example, biologists frequently

want to know how large a dose of a drug is required to produce an observable effect on an animal. Individual animals vary in their response to the drug, and either the drug or the animals may be too expensive for large-sample tests. Engineers often need to know what level of an input variable in an industrial process (possibly an amount of an expensive chemical) will produce a finished product of a specified flexibility, impact resistance, etc. Samples of the product will vary even when the input is constant, and measurements of the finished product can be quite expensive.

In general, the problem is to determine what level of input (written test score) is necessary to produce a given response (performance), when measurements of the response are difficult or expensive. While the educator, unlike the biologist or the engineer, cannot control his input directly, he can control it indirectly by first administering the written test to a large and diverse sample of persons and then using these written test scores as a basis for choosing those few persons whose performance will be individually assessed.

The class of techniques used to solve problems of this type is called stochastic approximation, and the basic method, as applied to educational testing, is as follows.

1. Select any person. Record his written test score and measure his actual performance.

2. If the first person succeeds on the performance measure (if his performance is above the minimum acceptable), choose next a person with a somewhat lower written test score. If the first person fails on the performance measure, choose next a person with a somewhat higher written test score.

3. Repeat step 2, choosing the third person on the basis of the second person's measured performance. Continue by choosing each person on the basis of the previous person's measured performance.

The advantage of this method of choosing persons for performance measurement is that it does not spread these expensive measurements over the full range of ability, but concentrates them in that portion of the range where they are needed to determine a cutoff score. Therefore stochastic approximation methods are not appropriate for determining the validity of the written test. Validation requires a sample that is representative of the population of interest, while the purpose of stochastic approximation is to produce a sample that is unrepresentative, in a way that is particularly useful for determining a cutoff score.

Stochastic approximation techniques can be classified into two types, according to the way in which the input is varied. In one type the input is varied by fixed steps. After each observation, we move up one step or down one step for our next observation. If the observation is a success (the person succeeds on the performance measure) we move down (we try a person with a written test score one step lower). If the observation is a failure, we move up. This technique is called the "up-and-down method" (Dixon and Mood, 1948). There are several variations of the up-and-down method which are intended to make it either more flexible or more efficient; some of these will be discussed later in this paper.

In the other type of stochastic approximation technique, the input is varied by an amount that depends on the difference between the observed performance and the minimum acceptable performance. For example, if the first person succeeds on the performance measure by a wide margin, we will

move down fairly far on the written test scale to choose the second person. But if the first person barely succeeds on the performance measure, we will choose for the second observation a person with a written test score only slightly lower than the first person's. The best known and most thoroughly investigated of these techniques is the Robbins-Monro process (Robbins and Monro, 1951). It would seem best suited to situations in which the written test has a large number of items, since it is based on the assumption that the input variable is continuous.

The test user who has decided to use a stochastic approximation technique for choosing a minimum passing score finds himself confronted with some specific problems and decisions:

1. Which stochastic approximation method should he use?

2. How large should the steps be?

3. How many persons should he select for the performance measure?

4. Given the data, how should he choose the minimum passing score?

5. What is the sampling variability of the minimum passing score chosen in this manner? How good is it as an estimate of the "true" minimum passing score -- the score he would choose if he could obtain written test and performance scores for all persons in the population?

These questions are all interrelated. They have been answered in several different ways and are still being investigated by mathematical statisticians. The remainder of this paper is an attempt to present some of the answers in a form that will be accessible and useful to educators with some knowledge

of basic statistical concepts. Derivations and proofs will be omitted; references will be provided for the reader who wishes to investigate the subject more deeply.[1]

Because stochastic approximation techniques were developed for situations other than educational and occupational testing, the more general terms "input variable" and "response variable" will sometimes be used in place of the terms "written test" and "performance measure", respectively. In addition, the term "response curve" will be used to refer to the function that gives the expected performance score for any given written test score.

## The up-and-down method

The up-and-down method was devised for use with a dichotomous response variable (performance measure). To use it with a continuous response variable we must impose an artificial dichotomy. To do so, we specify a particular performance level as the minimum acceptable. We then classify any performance at or above that level as a success and any performance below that level as a failure.

The up-and-down method also requires that the input variable (written test score) scale be divided into discrete levels, or "steps". The basic up-and-down rule directs us to move up one step on the input scale after a failure, and down one step after a success. This will cause the written test scores of the persons we select to center around the score that corresponds to a fifty per cent probability of success on the performance measure. (If we are interested in some other probability of success, we

[1] A good starting point for such an investigation is the excellent review by Scheber (1973).

must use a variation of the method described later in this paper.)

Table 1 presents the notation we will use in describing statistical procedures for the up-and-down method. Notice that if the performance measure is continuous, the decision-maker must specify both the minimum acceptable level of performance and the minimum acceptable probability of achieving this level. For example, he might want to estimate the written test score that corresponds to an eighty per cent probability of achieving a performance score of 125 or better. In the notation of Table 1, he would then specify $y_* = 125$ and $p = .80$. Also notice that when we specify a minimum acceptable probability of success, we are referring to the probability of success for the lowest-scoring person we will accept — one whose written test score is exactly equal to the minimum passing score.

Estimating the true minimum passing score

At least five distinct procedures have been recommended for estimating the true minimum passing score. The estimates they yield tend to be close to each other, as might be expected, but no two of the procedures yield the exactly same estimate in all cases. Four of these procedures will be presented here for the basic up-and-down method with $p = .50$; their adaptation to variations of the method with $p \neq .50$ will be discussed later, in connection with those variations.

The procedure for estimating $x_*$ originally suggested by Dixon and Mood (1948) can be expressed as follows. If there have been more successes than failures, take the mean written test score for all persons who failed the performance measure and subtract half the step size. If there have been

Table 1. Notation

$x_i$ written test score of the i th person selected for performance measurement.

$y_i$ observed performance of the i th person.

$Y_i$ random variable resulting from variation in performance between persons with written test score $x_i$, from instability of performance, and from unreliability of performance measurement.

$X_i$ random variable resulting from the fact that the selection of person i depends on the observed performance of person i - 1.

$y_*$ minimum acceptable performance level (performance level required for success).

p minimum acceptable probability of success.

$x_*$ true minimum passing score : the written test score such that, in the entire population of interest, Prob $(Y_i \geq y_* \mid X = x_*) = p$ .

$\hat{x}_*$ minimum passing score estimated from observed sample data.

$\hat{X}_*$ random variable resulting from variability in the data used to estimate $x_*$ .

more failures than successes, take the mean written test score for those persons who succeeded on the performance measure and add half the step size.

A second estimation procedure was suggested by Brownlee, Hodges, and Rosenblatt (1953). The procedure they recommended is to disregard the first run of successes or failures, except for the last observation in that run, and average the written test scores of all the rest of the persons selected (including that of the person whose performance would be measured next if the procedure were continued). If the first k persons all succeed (or all fail) and an additional n persons are selected, then the estimate of $x_*$ is

$$\hat{x}_* = \frac{1}{n+1} \sum_{i=k}^{k+n} x_i$$

Notice that only k + n - 1 persons will actually have had their performance measured. However, the (k + n)th person is considered to have been selected, because his written test score will have been determined by the (k + n - 1)st person's performance.

A third estimation procedure is Wetherill's "peaks-and-valleys" method, suggested by Wetherill and Levitt (1965) and Wetherill (1975). A "peak" is any failure preceded by a success; a "valley" is any success preceded by a failure. The descriptive terms derive from the fact that a "peak" represents a person with a written test score higher than those of the persons selected just before and just after him; a "valley" is exactly the opposite. The estimate of $x_*$ is simply the mean written test score for all the "peaks" and "valleys".

A fourth estimation procedure is the use of the "Spearman-Karber estimate".

This procedure was originally devised before the introduction of stochastic approximation techniques; its use in connection with the up-and-down method was investigated by Tsutakawa (1967). The estimate is

$$\hat{x}_* = x_{min} - \frac{1}{2} d + d \sum (1 - \hat{p}_j)$$

where $x_{min}$ is the lowest written test score among all persons actually measured with the performance measure, d is the step size, $\hat{p}_j$ is the proportion of success at the jth written test score level, and the sum is over all the different written test score levels at which persons were selected and measured for actual performance. For example, if the persons whose performance was measured all had written test scores of 70, 80, 90, or 100, then $x_{min}$ would be 70 and d would be 10. To find the $\hat{p}_j$ we would compute the proportion of successes at each of the four levels. An equivalent expression for this estimate, which may sometimes be more convenient, is

$$\hat{x}_* = x_{max} + \frac{1}{2} d - d \sum \hat{p}_j$$

Table 2 presents a set of hypothetical data illustrating the estimation of $x_*$ by each of the four procedures. For this particular set of data, the Dixon-Mood estimate and the Spearman-Karber estimate yield the same result. However, if the ninth person took the performance measure and succeeded, the Dixon-Mood estimate would remain unchanged, while the Spearman-Karber estimate would decrease from 51.67 to 50. (Brownlee's estimate would decrease from 51.43 to 50, while Wetherill's would remain unchanged at 52.5.)

A fifth estimation procedure suggested by Dixon (1965) requires the use of tables contained in his article and is not presented here.

Table 2. Estimates of $x_*$ with the up-and-down method, for p = .50 (hypothetical data).

| Person | Written test score | Performance |
|---|---|---|
| 1 | 70 | S (success) |
| 2 | 60 | S |
| 3 | 50 | S |
| 4 | 40 | F (failure) |
| 5 | 50 | F |
| 6 | 60 | S |
| 7 | 50 | F |
| 8 | 60 | S |
| 9 | 50 | not measured |

Dixon-Mood: $\frac{1}{3}(40 + 50 + 50) + \frac{1}{2}(10) = 51.67$

Brownlee: $\frac{1}{7}(50 + 40 + 50 + 60 + 50 + 60 + 50) = 51.43$

Wetherill: $\frac{1}{4}(40 + 60 + 50 + 60) = 52.5$

Spearman-Karber: $40 - 5 + 10\,(1 + \frac{2}{3} + 0 + 0) = 51.67$

Variance of the up-and-down estimate

Procedures have been suggested for estimating the variance of $\hat{X}_*$ based on each of the four procedures discussed in the previous section. The technique suggested by Dixon and Mood (1948) for computing the variance of their estimate requires some strong assumptions not likely to be satisfied in practical applications to educational testing: the response curve is assumed to be a normal cumulative distribution function with known standard deviation. (Brownlee, et al, 1953, pointed out that estimation of this standard deviation from observed data would require very large samples for reasonable precision.)

A procedure for estimating the variance of Brownlee's sample-average estimate of $x_*$ was devised by Tsutakawa (1967). This procedure requires us to identify the most frequently occurring written test score level. We then divide the whole sequence of observations into subsequences, ending each subsequence as soon as this most frequent level is reached and beginning the next subsequence with the next person. Let $t_m$ be the number of persons in the mth subsequence, and let $U_m$ be the sum of their written test scores. Let s be the number of subsequences. Then we disregard the first subsequence and estimate the variance of $\hat{X}_*$ by

$$\frac{\sum_{m=2}^{s} (U_m - t_m \hat{x}_*)^2}{\left[\sum_{m=2}^{s} t_m\right]^2}$$

If there is more than one most frequent level (i.e., a tie), we estimate the variance of $\hat{X}_*$ separately for each of the most frequently occurring levels, and average these estimates (Tsutakawa, personal communication, 1975).

Wetherill and Levitt (1965) suggest a procedure for estimating the variance of Wetherill's peaks-and-valleys estimate which may be useful if the sample size is not too small. They suggest averaging the peaks and valleys in pairs, letting the first estimate of $x_*$ be the average of the first peak and the first valley, the second estimate be the average of the second peak and the second valley, and so on. The sample variance of these individual estimates of $x_*$ divided by the number of individual estimates, is an estimate of the variance of $\hat{X}_*$. If we let $P_k$ and $V_k$ represent the kth peak and valley, the formula for the estimated variance of $\hat{X}_*$ is

$$\frac{\sum_{k=1}^{n} [\ \frac{1}{2}\ (P_k - V_k) - \hat{x}_*\ ]^2}{n\ (n - 1)}\ .$$

The variance of the Spearman-Karber estimate was derived by Cornfield and Mantel (1950, p. 208). The procedure they suggested for estimating it can be described as follows. Let $p_j$ represent the proportion of successes at the jth written test score level, and let $n_j$ represent the number of persons observed at that level. Let d represent the step size. Then the variance of $\hat{X}_*$ is estimated by

$$d^2 \sum_j \frac{\hat{p}_j\ (1 - \hat{p}_j)}{n_j - 1}$$

where the sum is over all written test score levels from which persons were actually measured for performance.

Choosing the step size

The choice of step size in the up-and-down method represents a trade-off between speed and precision. Larger step sizes lead more quickly to the portion

of the written test score range containing $x_*$; smaller step sizes permit more precise estimation of $x_*$. The weaker the relationship between the written test and the performance measure, the larger the step size needed, and the less precise will be the resulting estimate. (Dixon and Mood, 1948; Wetherill, 1963; Dixon, 1965; Davis, 1971). Brownlee, et al (1953) suggested using large steps as long as only successes or only failures are observed, then switching to small steps with the first change of performance. Wetherill (1963, 1975) suggested a more general version of this method: use large steps until some specified number of changes of performance (runs of successes or failures) have been observed; then compute $X_*$ and begin again at this input (written test score) level, using smaller steps to produce a more precise estimate.

Stopping rules for the up-and-down method

The choice of a stopping rule will often be dictated by economic, rather than statistical considerations. The test user may have to specify his sample size before beginning to collect performance data. However, in many cases it may be possible to let the number of observations depend on the data, at least within limits. Brownlee, et al (1953) recommend taking a specified number of observations beyond the initial run of successes or failures. Wetherill and Levitt (1965) recommend stopping after a specified number of runs of successes or failures (i.e., a specified number of "peaks" and "valleys"). Another possibility is to compute the estimated variance of the estimate after each observation (or after each run of successes or failures). When this variance becomes less than a specified size, stop taking observations. The ideal method for choosing sample size would be

an application of decision theory, taking into account (at any stage of the procedure) the costs of additional performance measurement and the benefits of increased precision. However, the resulting computations might be cumbersome.

Variations of the up-and-down method for $p \neq .50$.

Since the basic up-and-down method leads to the selection of persons with written test scores corresponding to a 50 per cent probability of success on the performance measure, it is not well suited to estimating the written test score corresponding to a probability of success other than .50. However, there are a number of variations of the method which make it suitable for this more general situation. Derman (1957) suggested a probabilistic method that can be described as follows. If $p > 1/2$, move up after any failure, but after a success, move down with probability $1/(2p)$ and up with probability $(2p - 1)/(2p)$. Thus, the higher the value of p, the less the probability of moving down after a success. That is, the higher probability of success we require, the more we will concentrate on persons with high written test scores. Conversely, if $p < 1/2$, move down after any success, but after a failure, move up with probability $1/(2 - 2p)$ and down with probability $(1 - 2p)/(2 - 2p)$. The estimate of $x_*$ for Derman's procedure is simply the written test score that occurs most frequently (or, if there are two or more such scores, their average).

Wetherill (1963, p. 35) suggested that Derman's probabilistic technique would be "likely to produce some inefficiency in small samples," and discussed some alternative variations of the up-and-down method. One variation which he did not recommend was the obvious device of moving up more than one step

after a failure but down only one step after a success (for $p > 1/2$; vice versa for $p < 1/2$). His objections to this method were that it would lead to substantially biased estimates of $x_*$ and that the written test scores of persons selected would not be closely grouped around the true population value of $x_*$.

Wetherill (1963) did suggest two other variations of the up-and-down method which he considered preferable to either of the two variations described above. The first of these is as follows: After each observation on the performance measure compare the proportion of successes at that level (call it $p_j$) with p, the required probability of success. If $p_j > p$, move down; if $p_j < p$, move up; if $p_j = p$, remain at that level.

Wetherill's second suggested variation is one which he calls the "up-and-down transformed response rule" (Wetherill and Levitt, 1965; Wetherill, 1975). This variation requires the experimenter to choose a rule such that when the probability of success at a given level equals the desired probability (not necessarily .50), the probability of moving up is exactly equal to the probability of moving down. The rule is started anew after each change of levels. For example, consider the rule: "Move up after any failure; move down after two successes." This rule allows only three possible sequences before changing levels. If $p_j$ is the true probability of success at level j, the possible sequences, with their associated probabilities and results, are the following:

| Sequence | Probability | Result |
|---|---|---|
| SS | $p_j^2$ | move down |
| SF | $p_j(1 - p_j)$ | move up |
| F | $1 - p_j$ | move up |

For this rule, the probability of moving up equals the probability of moving down when $p_j^2 = .50$; that is, when $p_j = .71$. Therefore this rule would be appropriate for estimating $x_*$ when $p = .71$. One obvious limitation of this variation is that it offers the decision-maker a limited number of different choices of p for which the rule is reasonably simple. However, this limitation does not seem too severe in fields such as education, where measurement is not extremely precise. Table 3 lists up-and-down transformed response rules corresponding to several different values of p.

For estimating $x_*$ by means of the up-and-down transformed response rule, any of the four estimation procedures discussed previously would seem to apply, with the following revisions: Instead of counting individual responses, count sequences of responses at the same level. For "failure", substitute "sequence leading to a move up"; for "success", substitute "sequence leading to a move down". For example, a "peak" in Wetherill's peaks-and-valleys procedure would be redefined as any sequence leading to a move down which was preceded by a sequence leading to a move up. In the Spearman-Karber estimate, $p_j$ would be the proportion of sequences at level j which led to a move down, and so on.

The "multiple-sample up-and-down method" (Hsi, 1969) is a generalized form of the up-and-down method. The rule can be stated as follows: At each input level, take response measures on k persons. If s or fewer succeed, move up. If r or more succeed, move down. Otherwise remain at the same input level. Of course, r must be greater than s. The basic up-and-down method can be described in this form by the values $k = 1$; $s = 0$; $r = 1$. When the desired success probability is .50, the three values will be

Table 3. Up-and-down transformed response rules for estimating written test scores corresponding to selected probabilities of success.

| P | Move up after | Move down after |
|---|---|---|
| .50 | F | S |
| .60 | F or SFF | SS or SFS |
| .71 | any F | SS |
| .79 | any F | SSS |
| .84 | any F | SSSS |
| .87 | any F | SSSSS |
| .89 | any F | SSSSSS |
| | | |
| .40 | FF or SFS | S or FSS |
| .29 | FF | any S |
| .21 | FFF | any S |
| .16 | FFFF | any S |
| .13 | FFFFF | any S |
| .11 | FFFFFF | any S |

related by the expression $r + s = k$. For success probabilities greater than .50, $r + s > k$; for success probabilities less than .50, $r + s < k$. The estimate of $x_*$ is Brownlee's sample-average estimate.

## The Robbins-Monro Process

The Robbins-Monro process was devised for use with a continuous response variable (performance measure) and a continuous input variable (written test score). It does not require the test user to dichotomize the response variable (the performance measure). For the continuous-response case, the test user specifies the minimum acceptable performance in terms of an expected score on the performance measure. Let $y_*$ represent this expected performance score. The minimum passing written test score $x_*$ is then defined by the expression

$$\mathcal{E}(Y_i \mid X_i = x_*) = y_*$$

where $y_*$ is specified by the test user and the symbol $\mathcal{E}$ indicates the expected value.

Notice that it is possible to use the Robbins-Monro process with a dichotomous response variable; in this case Y would be either 1 (for a success) or 0 (for a failure) and $y_*$ would be a specified probability of success. However, in this case one of the special advantages of the process is lost: the dependence of the step size on the size of the difference $(Y - y_*)$. Empirical results with artificial data indicate that the Robbins-Monro process works well with a dichotomous response variable <u>only</u> when the desired success probability is close to .50 (Wetherill, 1963, pp. 9-18).

The Robbins-Monro process is defined by the following rule for changing the input:

$$X_{i+1} = X_i - d_i (Y_i - y_*)$$

where the $d_i$ are a decreasing sequence of constants such that

$$\sum^{\infty} d_i = \infty \qquad \text{and} \qquad \sum^{\infty} d_i^2 < \infty \ .$$

These decreasing step coefficients cause the values of $X_i$ to converge to the true value of $x_*$ instead of bouncing back and forth around it as in the up-and-down method. Therefore the estimate of $x_*$ after n observations is simply $X_{n+1}$, the written test score of the student who would be selected next for performance measurement.

Robbins and Monro (1951) recommended choosing step coefficients according to the sequence

$$d_1 = C; \quad d_2 = \frac{C}{2}; \quad d_3 = \frac{C}{3}; \ \ldots; \quad d_n = \frac{C}{n} \ .$$

This choice of coefficients can be justified intuitively as follows: at any stage of the process we have a prior estimate, based on all the previous observations, which we will revise on the basis of one additional observation. If this additional observation is the nth observation, it contains 1/n of the information we have obtained. The rest of the information is contained in the prior estimate. Therefore we will weight the nth observation only 1/n as heavily as we would if it were our only piece of information.

There remains the problem of choosing a value for C, the initial step coefficient. The optimal choice of C depends on the slope of the response curve; ideally, C should be the inverse of the slope at the point $x_*$ (Venter, 1967). However, since $x_*$ is unknown, this result is useful only in placing

a lower limit on C, and then only if some prior information about the response curve is available. The results of a simulation study by avis (1971) suggest that it is better to have a value of C that is too large than one that is too small. If the response curve is a normal cumulative distribution function, a good value for C would be from two to four times its standard deviation. If the shape of the response curve is completely unknown, we have little guidance in choosing C.

One way to guard against the choice of too small a value for C is to use the "delayed Robbins-Monro process" (Davis, 1971), in which the step coefficients do not begin to decrease until there is a change of direction. From then on, the process continues as an ordinary Robbins-Monro process. For example, if the first three persons all have performance scores above $y_*$ and the fourth scores below $y_*$ the step coefficients would be C, C, C, C/2, C/3, ....

Since the Robbins-Monro process is an iterative process that converges, the test user may want to choose a stopping rule based on this convergence property. For example, he may want to stop performance testing when a specified number of estimates of $x_*$ all lie within a specified distance of each other.

Variance of the Robbins-Monro estimate

Estimating the variance of $X_n$ in the Robbins-Monro process is a complex problem. There is an assymptotic result (Sacks, 1958) which states that as the number of observations increases, the distribution of the random variable $\sqrt{n}\ (X_{n+1} - x_*)$ converges to a normal distribution with mean zero and variance

$$\frac{\sigma^2}{a\ (2\alpha - a)}$$

where $\sigma^2$ is the conditional variance of Y, given $x = x_*$ ; $\alpha$ is the slope of of the regression of Y on X at the point $X = x_*$ ; and a is the inverse of the

first step coefficient. Venter (1967) has suggested techniques for estimating $\sigma^2$ and $\alpha$ by using a variation of the basic Robbins-Monro process. His suggested method involves taking two observations at each step, with input values above and below the most recent estimate of $x_*$. The distance between the two input values decreases at each step, but at a slower rate than the decrease in the step coefficients. This process has the additional advantage of converging faster than the basic Robbins-Monro process, but it is somewhat more complicated to administer.

Farrell (1962) devised nonparametric confidence interval procedures for both the Robbins-Monro process and the up-and-down method. However, these procedures are mathematically complex and (like other nonparametric confidence interval procedures) tend to produce very wide intervals (Fabian, personal communication, 1975).

## Which method to use?

Most test users will probably find the up-and-down method (or a variation of it) more practical than the Robbins-Monro process, for the reasons given by Wetherill (1975):

> Two difficulties arise in attempting to apply the Robbins-Monro procedure to a practical problem. Firstly, observations must be taken serially and a calculation performed in between each one, which is not always convenient. Secondly, it is nearly always impracticable to stick to step sizes of C/N.

One limitation of the up-and-down method is its lack of flexibility in estimating probabilities of success other than .50. The "up-and-down transformed response" rule helps to impart some of the needed flexibility,

but the researcher still must choose from a fairly small selection of success probabilities. However, the choice of values given in Table 3 should be sufficient for most applications in educational and occupational testing.

## REFERENCES


Brownlee, K. A., Hodges, J. L., and Rosenblatt, M. The up-and-down method with small samples. Journal of the American Statistical Association, 1953, 48, 202-277.

Cornfield, J., and Mantel, N. Some new aspects of the application of maximum likelihood to the calculation of the dosage response curve. American Statistical Association Journal, 1950, 45, 181-210.

Davis, M. Comparison of sequential bioassays in small samples. Journal of the Royal Statistical Society, Series B, 1971, 33, 78-87.

Derman, C. Nonparametric up-and-down experimentation. Annals of Mathematical Statistics, 1957, 28, 795-798.

Dixon, W. J. The up-and-down method for small samples. Journal of the American Statistical Association, 1965, 60, 967-978.

Dixon, W. J., and Mood, A. M. A method for obtaining and analyzing sensitivity data. Journal of the American Statistical Association, 1948, 43, 109-126.

Farrell, R. H. Bounded length confidence intervals for the zero of a regression function. Annals of Mathematical Statistics, 1962, 33, 237-247.

Hsi, B. P. The multiple-sample up-and-down method in bioassay. Journal of the American Statistical Association, 1969, 64, 147-162.

Robbins, H., and Monro, S. A stochastic approximation method. Annals of Mathematical Statistics, 1951, 22, 400-407.

Sacks, J. Asymptotic distribution of stochastic approximation procedures. Annals of Mathematical Statistics, 1958, 29, 373-405.

Scheber, T. K. Stochastic Approximation: A Survey. Document AD 761 766. Springfield, Virginia: National Technical Information Service, 1973.

Tsutakawa, R. K. Random walk design in bioassay. Journal of the American Statistical Association, 1967, 62, 842-856.

Venter, J. H. An extension of the Robbins-Monro procedure. Annals of Mathematical Statistics, 1967, 38, 181-190.

Wetherill, G. B. Sequential estimation of quantal response curves. Journal of the Royal Statistical Society, Series B, 1963, 25, 1-48.

Wetherill, G. B. Sequential Methods in Statistics. London: Chapman & Hall (New York: Halsted), 1975.

Wetherill, G. B., and Levitt, H. Sequential estimation of points on a psychometric function. British Journal of Mathematical and Statistical Psychology, 1965, 18, 1-10.