*Perception & Psychophysics*
*2001, 63 (8), 1377-1388*

# Adaptive threshold estimation with unforced-choice tasks

CHRISTIAN KAERNBACH
*Universität Leipzig, Leipzig, Germany*

This paper evaluates an adaptive staircase procedure for threshold estimation that is suitable for unforced-choice tasks—ones with the additional response alternative *don't know*. Within the framework of a theory of indecision, evidence is developed that fluctuations of the response criterion are much less detrimental to unforced-choice tasks than to yes/no tasks. An adaptive staircase procedure for unforced-choice tasks is presented. Computer simulations show a slight gain in efficiency if *don't know* responses are allowed, even if response criteria vary. A behavioral comparison with forced-choice and yes/no procedures shows that the new procedure outdoes the other two with respect to reliability. This is especially true for naive participants. For well-trained participants it is also slightly more efficient than the forced-choice procedure, and it produces a smaller systematic error than the yes/no procedure. Moreover, informal observations suggest that participants are more comfortable with unforced tasks than with forced ones.

Adaptive procedures for threshold estimation (Levitt, 1971; Treutwein, 1995) are used in the attempt to find the signal level corresponding to a prescribed response probability. With the use of these procedures, the signal level is decreased after a correct or *yes* response and is increased after an incorrect or *no* response. With these procedures, single-stimulus yes/no or *N*-alternative forced-choice (NAFC) tasks are usually employed.

For yes/no tasks, the threshold is often defined as the signal level for which the probability of *yes* responses is 50%. The simplest way to arrive at this point is the *simple up–down* rule introduced by Georg von Békésy: There the signal level is decreased one step after each *yes* response and increased one step after each *no* response. The well-known drawback of this procedure is its strong dependence on the participant's maintaining a stable response criterion. Fluctuations of the response criterion will lead to large fluctuations in the threshold estimates.

To circumvent this problem, one can employ the NAFC task. The threshold is then often defined as the signal level at which the probability for correct responses is halfway between perfect performance and chance performance (i.e., 75% for two-alternative forced-choice [2AFC] tasks). Levitt (1971) summarized several rules that converge to response probabilities close to 75%. Another effective way to find this signal level is the *weighted up–down* procedure (WUD; Kaernbach, 1991), in which the simple up–down rule is modified. Here the signal level is decreased one step after each correct response and increased three steps after each incorrect one. The theoretical convergence level of this procedure is 75%. Other ratios between upward and downward step size lead to other asymptotic response probabilities. It should be noted that the claimed theoretical convergence points of staircase procedures may not be attained under all circumstances (see, e.g., García-Pérez, 1998).

Such adaptive procedures are designed to concentrate data sampling around some interesting region of the psychometric function and to avoid data sampling at too high or too low signal levels. Consider a situation where the signal intensity is far from that region of interest. For high signal intensities, where the response probability is close to one, both yes/no and forced-choice procedures lead to a quasi-deterministic movement of the adaptive run toward the region of interest of the psychometric function. Thus, given the size, of the separation (between the momentary position and the region of interest) and the step size, one can predict the number of trials necessary before the region of interest is reached. For low signal intensities, only yes/no procedures show such a deterministic drift, whereas forced-choice procedures lead to random walk behavior with an average movement toward the region of interest. This stochastic behavior of parts of the adaptive run leads to stochastic variations of the threshold estimate. This may explain why, for some participants, yes/no procedures are sometimes found to be superior to forced-choice ones.

Another possible difficulty with forced-choice tasks is that at low signal intensities the participant is often truly uncertain about the correct answer. The fact of being forced to make a choice even when one does not have the slightest idea which response is correct may introduce an uncomfortable aspect to this task. This may be of special

I thank Søren Buus, Douglas Creelman, Stan Klein, Birger Kollmeier, and Bernhard Treutwein for helpful comments and discussions. Particular thanks are due R. Duncan Luce for a thorough review of the manuscript. Correspondence should be addressed to C. Kaernbach, Institut für Allgemeine Psychologie, Universität Leipzig, Seeburgstraße 14-20, 04 103 Leipzig, Germany (e-mail: christian@kaernbach.de).

 

importance in situations in which the participant is confronted for the first time with such a method (e.g., as a patient in a clinical study), or where it is infeasible to dispel the reluctance of the participants to act "randomly" because it would require too much explanation and might even reduce the confidence of the participants in the seriousness of the investigation.

An unforced-choice task with the additional response alternative *don't know* appears to bypass both problems. The participant is not forced to give an answer when he or she simply does not know one. The probabilistic behavior of adaptive runs (for a possible rule, see below) at low signal intensities is also reduced insofar as the participant makes use of the *don't know* response at these intensities. But there is a price. Unforced-choice tasks reintroduce a criterion that not only can differ from participant to participant, but might vary among, or even within, runs. Given that a major reason for using forced-choice tasks was its not requiring a criterion, does it make sense to reintroduce one by giving the participant the additional choice alternative *don't know*?

The present study demonstrates that the response criterion involved in unforced-choice tasks has quite different effects from that involved in yes/no tasks: It is less detrimental in the sense that its variations do not induce large variations of the threshold estimate, and it does not in fact reduce the efficiency of the adaptive procedure. Moreover, the tendency to have more deterministic behavior for low signal intensities increases the efficiency of the procedure, and the gain in comfort makes it an ideal procedure for clinical settings.

In the following sections, four things are done: (1) a signal detection theory model of unforced-choice tasks is presented, (2) an adaptive method, *unforced weighted up–down* (UWUD), that specifies level adaptations for each of the three response types of unforced-choice tasks is introduced, (3) Monte Carlo simulations of adaptive runs using forced- and unforced-choice tasks are presented, and (4) behavioral data of 6 human participants for yes/no, forced-choice, and unforced-choice tasks are reported.

## Signal Detection Theory of Unforced-Choice Tasks: A Theory of Indecision

Signal detection theory models are used to explain response probabilities in both yes/no and forced-choice tasks (for a review, see Macmillan & Creelman, 1991). The Gaussian model of signal detection with equal variances (Green & Swets, 1974) is considered by many to be a good first-order approximation, and it is in wide use. According to this model, the stimuli elicit internal states on a one-dimensional decision axis, distributed following Gaussian normal distributions. The participant in a yes/no task has to fix a criterion on this decision axis and reply *yes* for all events greater than this criterion (producing hits, but also false alarms); whereas, in a forced-choice task, the participant need only report the interval or spatial region that elicited the greatest event.

In an unforced-choice task, the participant has to decide whether to decide at all. Just as with optimal decision strategies, an *optimal indecision strategy* should be based on Bayesian logic. Each trial of an *N*-alternative task corresponds to a set $\vec{e}$ of internal states $e_i$, $i = 1 \ldots N$, Gaussian distributed $\varphi_{\mu,1}$ with variance of one. The distribution is centered around mean value $\mu = 0$ for noise stimuli and around $\mu = d'$ for signal stimuli. Let $k$ be the pointer to the greatest internal representation. If answering at all, the participant will answer correctly if and only if the internal state $e_s$ corresponding to the signal stimulus is the greatest of all $e_i$ (i.e., if $k$ is a good clue because $e_k = e_s$). The optimal indecision strategy would be to maximize the probability $p(D)$ to decide in this case [i.e., maximize $p(D \mid e_k = e_s)$] and to minimize the probability $p(D \mid e_k \neq e_s)$ to decide in those cases where $k$ is an invalid clue. These two probabilities correspond to the hit rate and false alarm rate of decision theory. Please note that, in contrast to decision theory, these two probabilities cannot be assessed directly because the internal states $\vec{e}$ are not accessible.

Just as in decision theory, optimal indecision strategy is based on the calculation of the probabilities for the conditions of the conditional response probabilities $p(D \mid e_k = e_s)$ and $p(D \mid e_k \neq e_s)$: The participant wants to know the probability that the clue is a good one [i.e. $p(e_k = e_s)$]. The actual set $\vec{e}$ of all $N$ internal states is helpful: Certain combinations of internal states make it more plausible that $e_k = e_s$, and others make it less plausible. The task is therefore to calculate $p(e_k = e_s \mid \vec{e})$.Following a Bayesian approach, this can be done by calculating the probability densities $p(\vec{e} \mid e_j = e_s)$ that a specific set $\vec{e}$ of internal states results from a trial where a certain stimulus $j$ was the signal stimulus:

$$p(\vec{e} \mid e_j = e_s) = \varphi_{d',1}(e_j) \cdot \prod_{i=1\ldots N, i \neq j} \varphi_{0,1}(e_i) . \quad (1)$$

Here $\varphi(x)$ denotes the density of the normal distribution around internal state $x$. Please note that knowledge of $d'$ is required in order to evaluate $p_{\text{cor}}$. If the participant does not know the true value of $d'$, he/she has to operate with an estimate $d'^*$ instead. Given equal a priori probabilities $p(e_i = e_s) = 1/N$, the a posteriori probability to answer correctly if answering $k$ is then

$$p_{\text{cor}} = p(e_k = e_s \mid \vec{e}) = \frac{p(\vec{e} \mid e_k = e_s)}{\sum_{i=1\ldots N} p(\vec{e} \mid e_i = e_s)} . \quad (2)$$

This probability is greater than or equal to $1/N$. The optimal indecision strategy is to base the decision of whether to decide at all on this probability, or on the likelihood ratio $p_{\text{cor}}/(1-p_{\text{cor}})$, which is monotonically related with $p_{\text{cor}}$. The participant specifies a "safety margin" $\delta$ and refuses to designate the stimulus with the highest internal representation if $p_{\text{cor}}$ is not larger than $1/N + \delta$. The optimal value of $\delta$ depends on the costs for the possible response types (i.e., correct, incorrect, and *don't know* responses).

Let us consider the special case $N = 2$. Suppose the participant perceives event $e_1$ from the first stimulus and event $e_2$ from the second stimulus. The a priori probabil-

ity $p(e_i = e_s)$ is .5. In a forced-choice task, the participant responds "first stimulus" if $e_1 - e_2 > 0$ and "second stimulus" if $e_1 - e_2 < 0$. This strategy leads to the correct answer whenever the difference $e_s - e_n$ is positive (i.e., whenever the event elicited by the signal is greater than the event elicited by noise). Given the above distributions of $e_n$ and $e_s$, the distribution of $e_s - e_n$ is Gaussian $\varphi_{d',2}$, which is centered around $d'$ and with variance two (i.e., the standard deviation $\sigma = \sqrt{2}$, , see Figure 1A). The probability of a correct response given only order information is equal to the probability that $e_s - e_n > 0$ (Green & Swets, 1974):

$$p(e_k = e_s \mid e_k > e_l) = \int_0^\infty \varphi_{d',2}\,,$$

where $k$ denotes the pointer to the greater internal state, and $l$ the pointer to the smaller one. Given the exact values of $e_1$ and $e_2$ and full knowledge of $d'$, the participant can calculate the a posteriori probability for a correct response much more precisely:

$$\begin{aligned} p_{\text{cor}} &= p(e_k = e_s \mid \vec{e}) \\ &= \frac{\varphi_{d',1}(e_k)\varphi_{0,1}(e_l)}{\varphi_{d',1}(e_k)\varphi_{0,1}(e_l) + \varphi_{0,1}(e_k)\varphi_{d',1}(e_l)} \\ &= \frac{1}{1 + e^{-(e_k - e_l)d'}}. \end{aligned} \tag{3}$$

Obviously $p_{\text{cor}}$ is monotonically related to $(e_k - e_l) \cdot d' = |e_1 - e_2| \cdot d'$. For $N = 2$, the optimal indecision strategy (i.e., to decide only if $p_{\text{cor}} > 1/N + \delta$) corresponds to determining an indecision criterion, $C$: The participant selects the *don't know* button if and only if $|e_1 - e_2| \cdot d' < C$, with the relation between $C$ and $\delta$ being one-to-one on their domain of definition ($0 \le \delta < 1/2$, $0 \le C$). On the $e_1 - e_2$ axis, this strategy corresponds to deciding whether $|e_1 - e_2| < c = C/d'$, where $c$ is the *effective indecision criterion*. It depends on $d'$ for a given value of $\delta$. Figure 1A illustrates the optimal indecision strategy for the case $N = 2$, $d' = 1$, and $c = 1$.

Figure 1B shows *decider operating characteristics* (DOC) for 2AFC tasks for different values of $d'$. The probabilities that represent the coordinates can be derived from Figure 1A:

$$\begin{aligned} p(D \mid e_k = e_s) &= \frac{\int_c^\infty \varphi_{d',2}}{\int_0^\infty \varphi_{d',2}}, \\ p(D \mid e_k \neq e_s) &= \frac{\int_{-\infty}^{-c} \varphi_{d',2}}{\int_{-\infty}^{0} \varphi_{d',2}}. \end{aligned} \tag{4}$$

Although derived from symmetrical distributions, the resulting DOCs are slightly asymmetric.

Forced-choice tasks are often considered to be free of an internal response criterion. The additional response alternative *don't know* reintroduces a response criterion that can vary as a function of time or across participants. In order to compare the performance for different values of the criterion, one has to specify a quantity that serves as a performance index. In an $N$-alternative unforced-choice task, an obvious quantity is the extrapolated correct-response probability $p_{\text{ecor}} = p_{\text{cor}} + p_{\text{unsure}}/N$, where $p_{\text{cor}}$ is the correct-response probability, and $p_{\text{unsure}}$ is the probability of responding *don't know*. The last term adds what would have been contributed to $p_{\text{cor}}$ in a forced-choice task had the participant decided to throw dice instead of saying *don't know*. This extension makes $p_{\text{ecor}}$ comparable to $p_{\text{cor}}$ in a forced-choice task. Figure 1C compares psychometric functions[1] [$p_{\text{ecor}}$ as a function of $10 \cdot \log(d')$, $N = 2$] for different indecision strategies.

While in blocked designs, the participant will be able to estimate the value of $d'$ and can thus calculate $p_{\text{cor}}$ and maintain a fixed safety margin $\delta$; in randomized designs, the indecision strategy has to be based on an estimation $d'^*$ of the average of the true value of $d'$. This corresponds to maintaining a fixed value of $c$, whatever the actual value of $d'$, instead of maintaining a fixed value of $\delta$ or $C$. Figure 1C presents both types of psychometric functions. The gray lines are based on fixed values of $\delta$, whereas the black lines are based on fixed values of $c$. The leftmost curve ($c = \delta = 0$) represents the psychometric function for a forced-choice task. The psychometric functions for unforced-choice tasks for $\delta = 20\%$ (this corresponds to $C = c \cdot d' = 0.85$) and $c = 1$ differ little from that for forced-choice tasks. Larger values of $\delta$ or $c$ do alter the psychometric function to a greater, but still small, degree. Please consider that a safety margin of $\delta = 40\%$ ($C = 2.2$, rightmost gray curve) implies that the participant decides only if $p_{\text{cor}}$ is larger than 90%.

The effect of comparable shifts in the criterion of a yes/no task are illustrated in Figure 1D. Considering that a shift of 1 in the criterion $c_{\text{y/n}}$ of a yes/no task can make all the difference between a reasonable response criterion ($c_{\text{y/n}} = 1$: asymptotic level for low signal intensities 16%) and one that would make the simple up–down rule produce nonsense thresholds ($c_{\text{y/n}} = 0$: asymptotic level 50%), the small difference between an indecision criterion $c = 0$ and $c = 1$ is the more astonishing.

What underlies the relative stability of unforced-choice tasks toward variations of the response criterion? The main reason is evident from Figure 1A. The area shaded in white corresponds to the probability of the answer *don't know*. Its size does not correspond directly to a change of the performance measure $p_{\text{ecor}}$, as half of this area is counted as correct, and half of it as incorrect responses. The change of the performance measure depends not on the size of the white area, but only on its asymmetry. If the white area were symmetric—that is, if it contained as much to the left as to the right of zero—$p_{\text{ecor}}$ would be equal for forced and unforced choices. The decrease in performance results from the fact that more white area is to the right than to the left of zero, indicating that the participant gives away some of his/her possible performance by applying this response criterion. The rectangular bor-

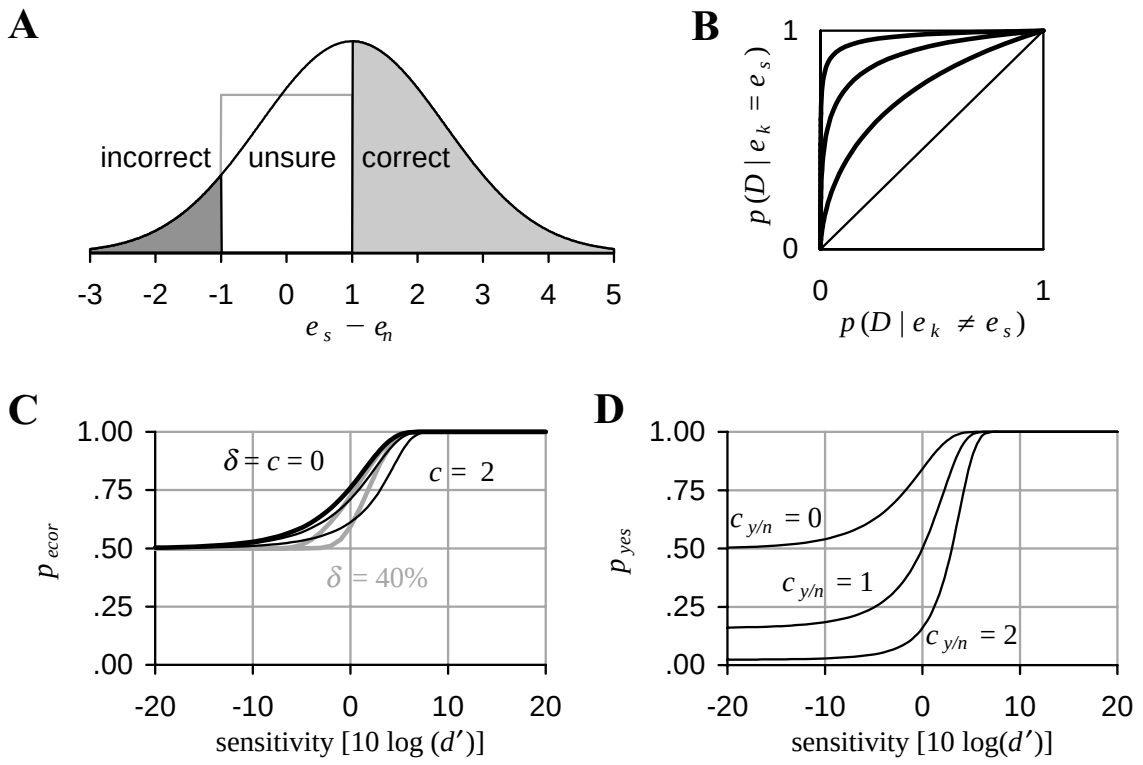


**Figure 1. Theory of indecision for two-alternative unforced-choice tasks, assuming Gaussian distributions of equal variance. (A) Distribution of the quantity $e_s - e_n$ (i.e., the difference between the decision axis event perceived in the signal stimulus and in the noise stimulus) for a signal with $d' = 1$ and an effective indecision criterion of $c = 1$. See the text for an explanation of the rectangular border shown in the *unsure* area. (B) *Decider operating characteristics* for $d' = 1$, 2, and 3. (C) Psychometric functions ($p_{ecor}$) for forced choice ($\delta = c = 0$, thick black line), for two different values of the safety margin ($\delta = 20$, 40%, thick gray lines) and for two different values of the effective indecision criterion ($c = 1$, 2, thin black lines). (D) Psychometric functions for a yes/no task, derived from a Gaussian model of signal detection, for four different values of the yes/no criterion.**

der drawn into the *unsure* area of Figure 1A shows the equivalent rectangular area of the actual *don't know* area. The excess of the right-hand (positive) part of the *unsure* area over this rectangular area corresponds to the information loss owing to the indecision criterion. If the indecision criterion $c$ is halved, this area is decreased by a factor of about four because of its triangular shape. Hence, the effect of $c$ on the performance is a second-order effect, as compared with the first-order effect produced by shifting the yes/no criterion.

### Unforced Weighted Up–Down

Adaptive staircase procedures for forced-choice tasks specify how to adapt the signal level after correct and incorrect responses (for a review, see Treutwein, 1995). Levitt (1971) has classified a wide range of adaptive staircase procedures as *transformed up–down* methods, with the signal level adapted after a certain block of responses. In contrast to these, with the WUD method (Kaernbach, 1991), the signal level is adapted after every single response. This method has proven to be superior to transformed up–down methods in simulations (Kaernbach, 1991) and also in experiments (Rammsayer, 1992). In this section, the WUD rule is modified for unforced-choice tasks.

With the WUD procedure, different step sizes are applied after correct and incorrect responses. The formula for the step size $S_{cor}$ after correct responses and $S_{incor}$ after incorrect responses is derived from the principle that the net movement should be zero when the response probability is equal to the target performance at the equilibrium point $p_{equ}$ of the adaptive run:[2]

$$S_{cor} \cdot p_{equ} + S_{incor} \cdot \left(1 - p_{equ}\right) = 0, \Leftrightarrow \frac{S_{incor}}{-S_{cor}} = \frac{p_{equ}}{1 - p_{equ}}. \quad (5)$$

For instance, for the target performance $p_{equ} = 75\%$, it follows that $S_{incor} = 3 \cdot -S_{cor}$. It is obvious that in order to approach the equilibrium point, $S_{cor}$ must be negative and $S_{incor}$ must be positive. With unforced-choice tasks, one must also specify how to adapt the signal level after *don't know* responses. One possibility is to roll a die and choose an alternative at random. That corresponds closely to what participants feel is happening when forced to choose. One

could, however, eliminate many of the stochastic features of this process by adapting the signal level in a suitable way so as to represent the average result of rolling the die again and again. Depending on the number $N$ of alternatives, the probability of making the correct response by chance, $p_{chance}$, equals $1/N$. Then, the step size $S_{unsure}$ to be applied after a *don't know* response should be

$$S_{unsure} = p_{chance} \cdot S_{cor} + \left(1 - p_{chance}\right) \cdot S_{incor}$$
$$= \frac{1}{N} S_{cor} + \frac{N-1}{N} S_{incor}. \quad (6)$$

$S_{unsure}$ can now be calculated for any desired target point of the psychometric function:

$$\frac{S_{unsure}}{-S_{cor}} = \frac{p_{equ} - 1/N}{1 - p_{equ}}. \quad (7)$$

Consider the case where the equilibrium performance $p_{equ}$ of the adaptive procedure is halfway between chance performance and perfect performance [i.e., $p_{equ} = (N + 1)/2N$]. Here we obtain

$$\frac{S_{incor}}{-S_{cor}} = \frac{N+1}{N-1}, \qquad \frac{S_{unsure}}{-S_{cor}} = 1. \quad (8)$$

Please note the close relationship between this new adaptive procedure, called UWUD henceforth, and the simple up–down rule of Békésy (with $S_{no}/-S_{yes} = 1$). The *no* button is renamed *don't know*, and the yes response is replaced by a forced-choice task. With simple up–down, a positive response by the participant is taken at face value. With UWUD, it is cross checked, and in case of a mistake, a kind of correction for guessing is introduced by taking a large upward step. Note that for large values of $N$, this is not much larger than the level adaptation following *don't know* responses. In other words, *don't know* responses are treated very similar to incorrect responses for high values of $N$. If a participant sets his/her criterion so as to avoid making any mistakes, the resulting run will resemble a simple up–down run. If the participant does not use the *don't know* response, the UWUD procedure is the same as the normal WUD procedure for forced-choice tasks.

**Monte Carlo Simulation**

Simulations of adaptive procedures serve two purposes. The simulation data can be compared with human data. In this case, one should determine the individual psychometric functions, and the simulated procedural details should be chosen exactly as in the real experiment (see, e.g., Kollmeier, Gilkey, & Sieben, 1988). In other studies, the simulation simply serves to demonstrate some general effects (e.g., the dependence on step size or other procedural parameters). In such simulations, some arbitrary psychometric function is chosen, and the procedural details need not coincide with experimental settings (and often are published without any reference to human data; see, e.g., Kaernbach, 1991). The advantage of the second approach is that many different simulations can be carried out, varying the parameters in fine steps, which simply cannot be paralleled with behavioral data. In the present case, there is the additional difficulty of determining each individual's psychometric function because it depends on values of the response criterion under his/her control. Therefore, I decided to choose the more general approach of basing the simulations on a theoretical model instead of on individual data. In the following simulations, the performances were calculated from a Gaussian model of two-alternative unforced-choice tasks (see Figure 1). The *x*-axis (signal strength) was taken to be a logarithmic function of $d'$, and it is expressed in decibels (10 times base-10 logarithm; compare also *x*-axis of Figure 1C and note 1).

For a simulation of unforced-choice procedures, it must be specified in what way the indecision criterion varies as a function of the signal strength. If the participant had full knowledge of the effective distributions at every single trial, the optimal indecision strategy would be to decide if and only if $p_{cor} > 1/N + \delta$, where the value of $p_{cor}$ is calculated with the true value of $d'$. Let us, for illustration purposes, consider the case $N = 2$. The optimal strategy with a constant value for $\delta$ would be equivalent to maintaining a stable indecision criterion $C = c \cdot d'$ (i.e., the effective indecision criterion $c$ would have to be larger for small values of $d'$). Consider, for instance, the case of $d' = 0$: In this case, $p_{cor}$ will never be greater than $1/N$, however large $|e_1 - e_2|$ may be, and the requirement that $p_{cor} > 1/N + \delta$ is equivalent to an infinite value for the indecision criterion $c$.

If, on the other hand, the participant does not have trial-by-trial knowledge of $d'$, he/she would have to estimate an average sensitivity value $d'^*$ and base the indecision strategy on the value of $p_{cor}$ calculated with $d'^*$ instead of $d'$. In the case of $N = 2$, this is equivalent to maintaining a stable indecision criterion $c$ whatever the actual value of $d'$ may be. Given feedback, participants in adaptive procedures will probably show some but not perfect knowledge of $d'$.

The following simulations tested both types of indecision strategies. For different values of $\delta$, we simulated 100,000 runs of a UWUD procedure, simulating either an optimal indecision strategy (i.e., taking into account the true value of $d'$) or setting the internal representation $d'^*$ to be equal to one, which should represent an average estimate of $d'$ at threshold. The step width of the adaptive procedure ($S_{unsure} = -S_{cor}$) was 1 dB. Each run began at a random position within the interval [15,16] dB, where the performance was quite close to one.[3] A reversal happens when a response leading to a downward movement (correct) is followed by a response leading to an upward movement (incorrect or *don't know*), or vice versa. Each run was terminated 20 trials after the fourth reversal. The threshold estimate was the mean of these last 20 trials. The termination criterion was trials, not reversals, in order better to compare runs with different criteria that result in reversals after different average numbers of trials.

Figure 2A shows the statistical error as a function of the systematic error for the case $N = 2$. The statistical error was estimated as the standard deviation of the threshold

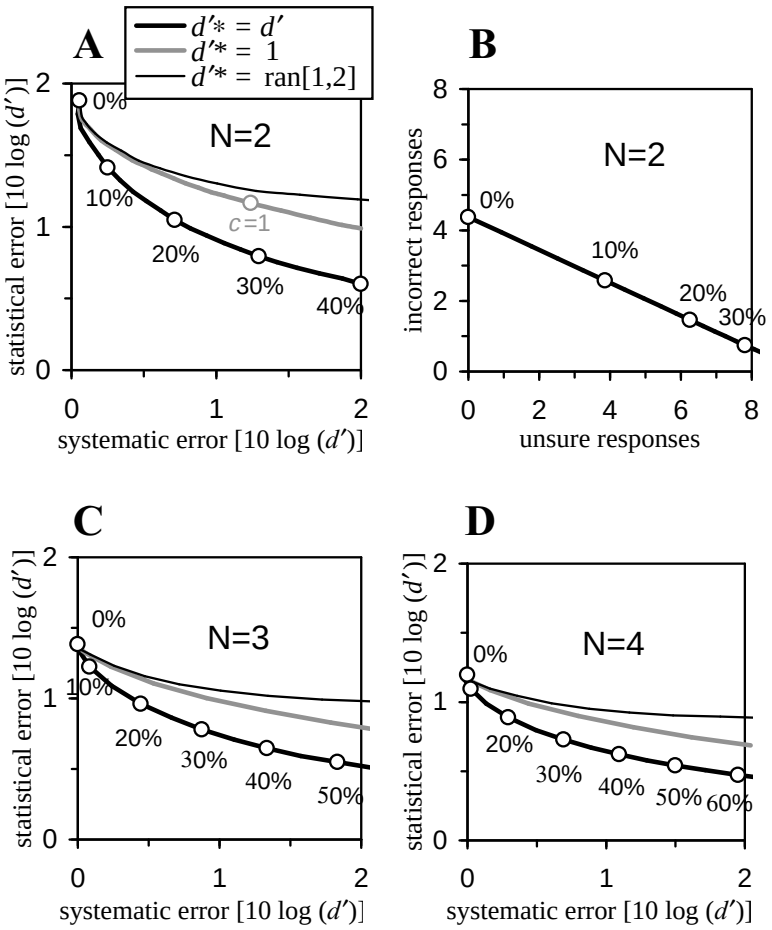


**Figure 2. Simulation results for unforced weighted up–down runs. (A) $N$ = 2: The statistical error is shown as a function of the systematic error for various values of $\delta$ (see annotations). The total error corresponds to the distance from the origin. The thick black line corresponds to the optimal indecision strategy based on a trial-by-trial knowledge of $d'$. The thick gray line was calculated with a fixed internal representation $d'^*$ of $d'$. The thin black line was calculated with the internal representation $d'^*$ taken randomly from the interval [1,2]. (B) Number of incorrect responses for the same runs as in (A) as a function of the number of unsure responses. (C, D) The same simulations as in (A) for $N$ = 3 and $N$ = 4.**

estimates resulting from these 100,000 runs. This measure reflects the statistical variations of the single threshold estimates around the average threshold estimate. The systematic error was estimated by the absolute difference between the average threshold estimate and the *true* threshold, which was assumed to be at that point where a forced-choice procedure would give 75% correct responses ($d'$ = 0.954).

The solid black line corresponds to the optimal indecision strategy, with the participant's having full trial-by-trial knowledge of $d'$ ($d'^* = d'$), for different values of $\delta$ (see annotations). The statistical error decreases as $\delta$ increases. This is a consequence of the tendency toward a deterministic behavior at low signal intensities. The systematic error increases as $\delta$ increases. This arises as follows. The forced-choice procedure, which is equivalent to unforced-choice with $\delta = 0$, will on average find the true threshold value quite precisely. As the number of *don't know* responses increases, the effective psychometric function shifts toward higher signal intensities (compare Figure 1C). The adaptive procedure finds the target performance for this shifted psychometric function and, hence, a difference (the systematic error) exists from the target performance of the original psychometric function. It is important to note, however, that the systematic error is really small for $\delta < 10\%$.

Whereas the direction of the systematic error is known, the statistical error may add itself in either direction—enlarging, reducing, or even inverting the systematic error. The average size of the resulting total error is the orthogonal sum (i.e., the square root of the sum of the squared values) of the statistical and the systematic error and equals

the distance of each data point from the origin (0,0) of the graph.[4] This distance decreases with increasing $\delta$ for moderate values of $\delta$, reaches a local minimum at $\delta = 20\%$, and then begins to increase, although remaining smaller than the value for forced choice ($\delta = 0$) up to $\delta = 36\%$. Thus, over a wide range of possible response strategies, the UWUD procedure produces fewer errors than does the normal WUD procedure. One might even argue that the statistical error is more important than the total error, since quite often it is not the absolute threshold value that is important but a change in threshold resulting from a variation of one of the experimental parameters. The statistical error with UWUD runs is smaller than that with normal WUD runs for all values of $\delta$.

The gray line shows simulation data for the case in which the participant does not have trial-by-trial knowledge of $d'$. In this case, the indecision strategy is again based on the calculation of $p_{cor}$, where the internal representation $d'^*$ was set equal to one. It should be noted that any other value $d'^{*\dagger}$ would have yielded the same curve: In the case of $N = 2$, it can be followed from Equation 3 that for any other value $d'^{*\dagger}$ there exists a value $\delta^\dagger$ that would yield the same decisions as the original parameters $d'^*$ and $\delta$. Instead of adding annotations of $\delta$ values (which would be relevant only for $d'^* = 1$), the $c = 1$ position of the curve is indicated. With a constant $d'^*$ the statistical error decreases, but not to the same degree as with the optimal indecision strategy $d'^* = d'$. This shows that a knowledge of the signal strength (i.e., of $d'$) helps the participant to decide when it would be better not to decide. There is, however, still a large range of $c$ values that lead to total errors that are smaller than for forced-choice tasks. The sweat point is at $c = 0.75$, and up to $c = 1.2$, the total error is smaller than for $c = 0$.

One important argument against reintroducing a criterion is that it may vary over the course of the experiment. The purpose of the next simulation was to determine the degree to which variability in the criterion increases the statistical error. It was assumed that the participant has no trial-by-trial knowledge of $d'$. The decision when to decide does then depend on both the internal representation $d'^*$ and the safety margin $\delta$ (or the indecision criterion $C$), and both can vary. It is, however, sufficient to vary one of them, since a change in $d'^*$ has exactly the same effect as a judiciously chosen change in $\delta$ or $C$. Whereas in the previous simulation $d'^*$ was set to one, it was now taken randomly from the interval [1,2]. This corresponds to varying the effective indecision criterion $c$ by a factor of two. The randomization was done once per run. Note that, for a given range of criteria, randomizing the criterion within runs produces less variability of the threshold estimates than does randomizing between runs, because the effective average criterion per run varies less. Hence, between-runs randomization represents the most severe test possible of the stability of the UWUD procedure against variability of the response criterion.

The thin black line in Figure 2A shows the results. In spite of the randomization of the indecision criterion, the statistical error still decreases with increasing $\delta$. For small values of $\delta$ randomization does not much affect the statistical error, whereas for high values of $\delta$, the statistical error is not so small as without randomization. The total error decreases with increasing $\delta$, up to a local minimum at $\delta = 20\%$, and then increases again. It remains smaller than the value for forced-choice, up to $\delta = 33\%$. Again, and despite the variability of the response criterion, the total error of UWUD is smaller than or comparable with that of WUD over a wide range of response criteria, and if one is interested in the statistical error, UWUD is superior for all tested values.

Figure 2B shows the error rate as a function of the unsure responses during the last 20 trials of each run. The number of errors decreases monotonically with increasing $\delta$, and at the same time, the number of *don't know* responses increases. For $\delta = 15\%$, the fraction of errors is halved, as compared with forced choice ($c = 0$). Especially in situations with feedback, this may be expected to improve the participants' comfort because it reduces the amount of negative feedback.

The reversal number during the last 20 trials increases with increasing $\delta$ from about 7.1 ($\delta = 0\%$) to 8.9 ($\delta = 10\%$) and reaches, finally, 10.5 ($\delta = 49\%$). This is evidence of the transition from normal WUD to simple up–down, which is known to produce more reversals. The effect, although small, may influence the run length (in reversals) set by the experimenter. The effect of the step size on the reversal number cannot be assessed from the present simulation data because the step size was not varied.

Figures 2C and 2D show the result of simulations of UWUD runs for $N$-alternative tasks with $N = 3$ and $N = 4$. Again, Equations 1 and 2 were used to calculate the probability $p_{cor}$ for a correct response, with the virtual participant refusing to choose among the $N$ alternatives if $p_{cor} < 1/N + \delta$. Data sets with optimal indecision strategies ($d'^* = d'$) based on trial-by-trial knowledge of $d'$, with indecision strategies based on a single fixed internal representation $d'^* = 1$, and with a randomized internal representation $d'^*$ taken from the interval [1,2] are presented in the same way as in Figure 2A. All simulation results show clearly that there is a reduction of the statistical error and the total error owing to the inclusion of the *don't know* response type. The advantage is, however, not as dramatic as for $N = 2$. Again, the optimal indecision strategies produce smaller statistical errors for comparable systematic errors than do the suboptimal strategies based on fixed values of $d'^*$. With increasing $N$, the statistical error gets smaller, as would be expected. The value of $\delta$ at the sweat point with the minimal total error is about 20% for all values of $N$.

**Behavioral Data**

Simulation data may prove the apparent superiority of a method only to discover that it, in fact, fails in real experiments. The likely reason for this is that real participants behave differently than simulated participants. They may have different psychometric functions, they may like or dislike certain aspects of a method, and certain aspects of

the procedures, such as the systematic or apparently random succession of signal intensities, may help or hamper their ability to concentrate on their task. Therefore, the real test of a psychophysical procedure is comparing it with established procedures in a behavioral experiment.

Six paid psychology students (4 females, 2 males; age range, 20–22) without any prior experience in adaptive psychophysical procedures participated in the experiment. Two performed in 120 runs and 4 in 360 runs of three different procedures run in cyclic order: simple up–down, weighted up–down, and unforced weighted up–down. The task was to detect a brief sinusoid of 1000 Hz with a Gaussian envelope ($\sigma$ = 150 msec) centered in 800 msec of white noise with 100-msec ramps. For simple up–down runs, one interval contained the sinusoid, and the participant was asked whether he/she had heard it. For WUD and UWUD, two intervals, separated by 200 msec, were presented; just one contained the sinusoid, and the other consisted of noise alone. The runs began at a random position between 12 and 24 dB, where 0 dB was defined as equal rms power within a third-octave band centered around 1000 Hz. The initial step size was 4 dB. The step size was halved after the second and the fourth reversals so that the final step size was 1 dB. A run lasted until 16 reversals had occurred. With the two-interval procedures, feedback was given of incorrect responses.

Given the similarity of the procedures, special care was taken to ensure that the participants recognized the existence of the additional response possibility with unforced-choice tasks. In particular, prior to each unforced-choice run, the participant was required to press the *don't know* button to initiate the run. After each unforced-choice run,

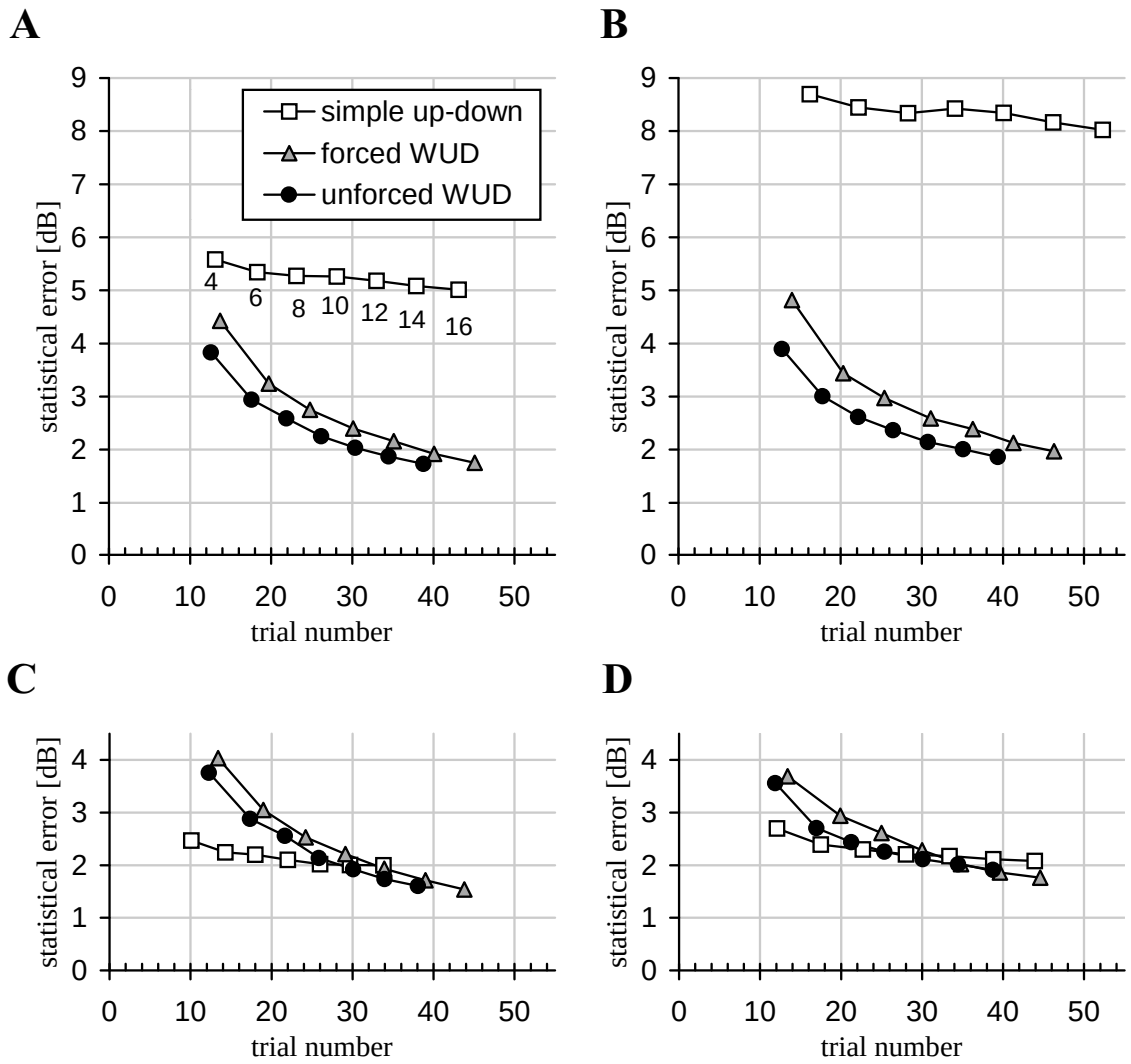


**Figure 3. Results of a behavioral comparison of the simple up–down, forced, and unforced weighted up–down procedures. (A) Statistical error averaged across all 6 participants, as a function of the average run length in trials. (B, C) The same for two clusters of participants: those that could not cope with the simple up–down procedure (B), and those that could (C). (D) Data of 4 participants after a substantial amount of training.**

the total number of incorrect and unsure responses was provided, and when the ratio of these two quantities exceeded 2.5 in either direction, the participant was encouraged to make either greater or less use of the *don't know* button. Our intention was to aid the participants to maintain stable criteria. Such a message occurred on only 6% of the runs, 5% due to too few unsure responses, and 1% due to too many of them.

One threshold estimate was calculated by averaging all intensities following the second reversal, up to the intensity following the 4th reversal.[5] Six further estimates were based on all intensities following the 4th reversal, up to the intensity following the 6th, 8th, . . ., 16th reversals. So each single run gave rise to seven different threshold estimates. Additionally, the number of trials needed to reach the 4th, 6th, . . . 16th reversals was determined.

Figure 3A shows the statistical errors (i.e., the standard deviation of the threshold estimates) calculated over the first 120 runs (40 runs per method), as a function of the average trial number. It is given separately for the three different adaptive procedures and for the seven different run lengths. As is clear, both weighted up–down procedures proved to be far more efficient than the simple up–down procedure, reaching smaller statistical errors for the same trial numbers. UWUD is a little bit more efficient than WUD.

The poor results for the simple up–down deserve further analysis. From the single-participant data, it became obvious that the 6 participants could be clustered into two groups of 3 each, with one group having had problems in achieving stable threshold estimates with the simple up–down method (Cluster A), and the other group coping well with this method (Cluster B).

Figure 3B shows the data from Cluster A. The difference in efficiency between UWUD and WUD, as well as the poor performance for simple up–down, is most prominent for Cluster A. Figure 3C shows the data averaged over Cluster B. In this case, the simple up–down procedure generated threshold estimates that are comparable to those generated by the two weighted up–down procedures. For short runs, simple up–down was sometimes even superior to weighted up–down. It is interesting to notice that, for both clusters, the variability of the threshold estimates did not decrease for long runs as much as it did for the weighted up–down procedures. This could have resulted from the participants' inability to respond stochastically independent from their earlier responses. The participants inferred that the stimulus was always present and that its intensity was reduced after a *yes* response and was increased after a *no* response. Stochastic response behavior should have appeared to them as illogical and inconsistent. A possible strategy was for the participants to respond alternately with *yes* and *no*, once the threshold region had been reached. This would then prevent further reduction of the threshold variability.

To analyze how further training might modify these results, 4 participants (2 from each cluster) were asked to continue until they had reached 360 runs (120 per method). Figure 3D shows the statistical errors calculated over the last 180 runs. The earlier difficulties with the simple up–down procedure disappeared, and the data of Figure 3D look very similar to those of Figure 3C. Again, for short runs, simple up–down produced fewer statistical errors than did the weighted up–down procedures, but it cannot be reduced further by using long runs. Comparing the two weighted up–down procedures, the unforced version shows a slight advantage in efficiency.

Some people may view the systematic error as an important quantity by itself, whereas others may not. At least its variations are important in that they determine the interparticipant statistical error. To have evaluated the systematic error directly, the psychometric function would have had to be determined for each participant. In order to avoid temporal effects, this should have been done interleaved with the adaptive runs. This was not done. Instead, the systematic error was estimated by comparing the thresholds determined using the different adaptive procedures. From all tested procedures, the only criterion-free one was the WUD procedure. Therefore, all data are referenced to the results of this procedure. Table 1 shows the mean and the standard deviations (across participants) of the threshold estimates for simple up–down and UWUD, with WUD serving as reference. Four rows correspond to the data presented in Figures 3A–D. The data on the systematic errors (Table 1) parallel those on the statistical errors

**Table 1**
**Average Threshold Estimates (in Decibels) Obtained With Two Procedures, With the Values Obtained With the Third One, the Criterion-Free Forced Weighted Up–Down Procedure, Serving as Reference**

| Simple Up–Down Minus Forced Weighted Up–Down | Unforced Weighted Up–Down Minus Forced Weighted Up–Down | $F$ test, Significant if $f > F(n-1, n-1)$ | $\alpha$ |
|---|---|---|---|
| First 120 runs (40 per method) for all participants ($n = 6$) | | | |
| –3.4 ± 5.5 | 0.03 ± 0.5 | $F(5,5) = 10.97$, $f = 11.26$** | .01 |
| First 120 runs for Cluster A (bad simple up–down performers, $n = 3$) | | | |
| –6.9 ± 6.1 | –0.05 ± 0.2 | $F(2,2) = 19.00$, $f = 26.27$* | .05 |
| First 120 runs for Cluster B (good simple up–down performers, $n = 3$) | | | |
| 0.1 ± 0.8 | 0.1 ± 0.7 | $F(2,2) = 19.00$, $f = 1.18$ | .05 |
| Last 180 runs for remaining participants ($n = 4$) | | | |
| –0.3 ± 2.4 | 0.09 ± 0.3 | $F(3,3) = 9.28$, $f = 9.21$ | .05 |

(Figure 3A–D): Especially during the first 120 runs, Cluster A shows, on average, large systematic errors and a large variation of these systematic errors for the simple up–down procedure. The participants of Cluster B show comparable systematic errors for simple up–down and UWUD. After additional training (last line of Table 1), the averages of the systematic errors for these two procedures approach each other, whereas the variability of the systematic error for simple up–down data remains high. Due to the small number of participants in the last phase of the experiment ($n = 4$), the difference of the variance of the systematic error across trained participants for the two procedures misses significance by a narrow margin (see $F$ test in Table 1). Especially in comparing different groups of participants, the simple up–down procedure could introduce systematic errors. In contrast, the systematic errors of the UWUD procedure are rather small, indicating a low value of the indecision criterion employed by the participants.

In order to determine the indecision strategy applied by the participants, the data were compared with the simulation data. Figure 4A shows the distribution of the numbers of incorrect and unsure responses for runs that simulated as precisely as possible the experimental details (step size reduction, termination after 16 reversals). The indecision strategy ($\delta = 10\%$, $d'^* = 1 \Rightarrow c = 0.4$) was chosen in order to mimic the experimental data shown in Figure 4B. The two solid lines in Figure 4B represent the boundaries beyond which participants were encouraged to make greater or less use of the *don't know* button. Ninety-four percent of the runs lay inside these boundaries, with the main part

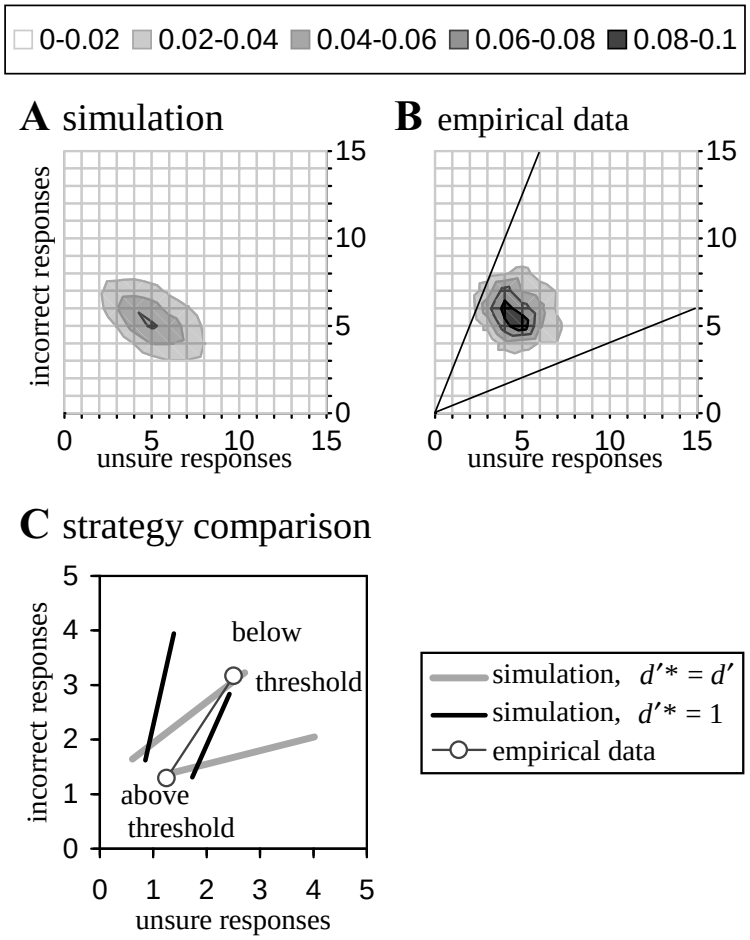


**Figure 4. Distribution of runs as a function of the number of unsure and incorrect responses. (A) Simulation results for $\delta = 10\%$ and $d'^* = 1$ ($c = 0.4$). (B) Behavioral data. The solid lines represent the borderlines beyond which a specific encouragement was given. (C) Comparison of simulation results and empirical data. Here, the response types were evaluated separately for the below-threshold and above-threshold parts of the runs. Thick gray lines show simulation results for optimal indecision strategies (left line: $\delta = 5\%$, right line: $\delta = 10\%$) and thick black lines for strategies using an average value for $d'^*$ (same values for $\delta$, $d'^* = 1$). The experimental results are shown with a thin black line with circles.**

of the distribution concentrated much more tightly than the total area defined by the boundaries. This means that, even without any attempt to stabilize the criterion, the participants must have chosen their response criteria in a stable and reproducible manner. The distribution of the experimental runs is constrained to a smaller region than that of the simulated runs, indicating that the participants monitored their response behavior and avoided excesses of either response type by adjusting their response criteria. If, for instance, in a certain run, a large number of incorrect responses occurred, the participant enlarged the indecision criterion and so was more prone to press the *don't know* button. The value of the indecision criterion is rather small, and at these values, the simulations predict a rather small systematic error (see Figure 2A), quite in line with the data presented in Table 1.

Optimal indecision strategies based on trial-by-trial knowledge of $d'$ show a different pattern of the use of the *don't know* response than do indecision strategies based on a fixed internal representation $d'^*$. Figure 4C shows the number of incorrect responses as a function of the number of the don't know responses for the above-threshold and the below-threshold parts of adaptive runs. The two thick gray lines show simulation data for optimal indecision strategies ($d'^* = d'$). Each line corresponds to a set of data for a certain value of $\delta$ (5% or 10%), separated in responses given below the threshold (upper right-hand point) and above the threshold (lower left-hand point). The two thick black lines show the same data in the case of a fixed internal representation $d'^* = 1$. Obviously the optimal indecider makes more use of the *don't know* response in the below-threshold region, and less use of it in the above-threshold region. The thin black line with circles shows the experimental data. They are close to the simulation data for a fixed internal representation $d'^*$, but with a slightly smaller slope. This could be interpreted as demonstrating that the participants could adapt their indecision strategy to the actual signal strength to a certain degree.

Introspective observations of the participants indicated that they felt most comfortable in the unforced-choice tasks. Under forced-choice tasks, they complained about the need to make a choice even when they felt they did not perceive anything. With yes/no tasks, they complained about "not knowing where to stop" (i.e., about the need to maintain a stable decision criterion). Unforced-choice tasks gave them the response possibilities they needed: the possibility when they did not perceive anything of saying so, and the possibility to control themselves if they thought they had got the signal.

## Conclusions

In summary, it was demonstrated both in Monte Carlo simulations and in a behavioral study that adaptive procedures can be based on unforced-choice tasks without losing reliability. And a slight gain in efficiency is actually achieved. A major argument for using unforced-choice tasks is the resulting greater comfort of the participants. This fact is especially important for experiments with naive and inexperienced participants and/or participants in clinical settings. Experienced participants yield reliable results in yes/no tasks once they have learned to maintain a stable yes/no criterion, but that does not reduce the variability due to different criteria among participants.

Normal WUD runs produce slightly more reversals in a fixed number of trials than do transformed up–down runs (about 20%; see Kaernbach, 1991, Figure 2), and UWUD runs add another 20% to the reversal count. In order to obtain a comparable run length, it may be advisable to increase the number of reversals requested for run termination.

The behavioral data presented in Figure 4B show that the participants did not excessively use of the *don't know* response. On the contrary, the results might be improved by directly encouraging greater use of the *don't know* button (e.g., by suggesting this possibility at every incorrect response). The simulation data presented in Figure 2 seem to recommend values for $c$ around 0.65, roughly twice the values actually chosen by the participants. Were they more optimal, the *don't know* response would occur more than twice as often as incorrect responses.

While the simulations tested also $N$-alternative tasks for $N > 2$, the behavioral test was done with two-alternative tasks only. Given the similarity of the outcome of the simulations for $N = 2$ and $N > 2$, it can be assumed that UWUD is behaviorally superior to WUD also for $N$-alternative tasks with more than two alternatives.

### REFERENCES

García-Pérez, M. A. (1998). Forced-choice staircases with fixed step sizes: Asymptotic and small sample properties. *Vision Research*, **38**, 1861-1881.

Green, D. M., & Swets, J. A. (1974). *Signal detection theory and psychophysics.* New York: Krieger.

Kaernbach, C. (1991). Simple adaptive testing with the weighted up–down method. *Perception & Psychophysics*, **49**, 227-229.

Kollmeier, B., Gilkey, R. H., & Sieben, U. (1988). Adaptive staircase techniques in psychoacoustics: A comparison between theoretical results and empirical data. *Journal of the Acoustical Society of America*, **83**, 1852-1862.

Levitt, H. (1971). Transformed up–down methods in psychophysics. *Journal of the Acoustical Society of America*, **49**, 467-477.

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide*. New York: Cambridge University Press.

Neutzler, F. (1999). *Sequentielle Schätzverfahren für Wahrnehmungsschwellen* [Sequential estimators for perception thresholds]. Unpublished master's thesis, Leipzig University.

Rammsayer, T. H. (1992). An experimental comparison of the weighted up–down method and the transformed up–down method. *Bulletin of the Psychonomic Society*, **30**, 425-427.

Treutwein, B. (1995). Adaptive psychophysical procedures. *Vision Research*, **35**, 2503-2522.

### NOTES

1. Psychometric functions resulting from experimental data are often shown as a logarithmic function of intensity, allowing to assess performance and $d'$ for a given intensity. For small intensities it is reasonable to assume a linear relation between log(intensity) and log($d'$). For modeling purposes, it is then sufficient and more general to plot the psychometric function as a logarithmic function of $d'$ instead of intensity.

2. Please note that in Kaernbach (1991) there is a typographical error in Equation 1 and in the following line: The subscripts *up* and *down* were mistakenly interchanged. Also in Kaernbach, the step sizes were taken

as absolute values, whereas in the present paper, they are considered to be signed quantities.

3. Starting at higher values, say 21.3, would lead to quasi-deterministic run-downs, passing the [15,16] interval, in this case at 15.3. It is thus sufficient to randomize across the range of one step size.

4. This definition of the total error is equivalent to referring the variance to the *true* value instead of to the mean. Let $\sigma^2_{\text{stat}}$ denote the mean of the quadratic deviations of the threshold estimates from the mean estimate: $\sigma^2_{\text{stat}} = <(x - <x>)^2>$. Let $\sigma^2_{\text{sys}}$ denote the squared difference between the mean $<x>$ and the true value $x_t$. Then the squared total error $\sigma^2_{\text{tot}} = \sigma^2_{\text{sys}} + \sigma^2_{\text{stat}}$ is equal to $<(x - x_t)^2>$.

5. Unpublished simulations by Neutzler (1999) tested the best way to analyze data obtained by the simple up–down method. They exhibited a slight advantage for averaging all intensities, as compared with averaging only the intensities of the reversal points. Moreover, if the discard was $N_d$ and the total reversal number was $N_t$, it proved advantageous to begin averaging at the intensity following reversal $N_d$ (i.e., excluding the intensity at reversal point $N_d$ from the analysis) and to include the intensity following to reversal $N_t$ (i.e., the one that would have been tested next) into the analysis.

(Manuscript received August 12, 1999;
revision accepted for publication February 13, 2001.)