EVALUATION OF THE EQUIVALENCE OF TWO DIFFERENT METHODS OF THRESHOLD ESTIMATIONS WITH ADAPTIVE PSYCHOPHYSICAL PROCEDURES

Amélie Pfäffli & Thomas H. Rammsayer
Institute of Psychology, University of Bern, CH-3000 Bern 9, Switzerland and Distance Learning University Switzerland, CH-3900 Brig

Abstract

Because of its higher efficiency, adaptive psychophysical procedures are becoming increasing popular compared to traditional non-adaptive procedures. The present study evaluates the equivalence of two different methods for threshold estimation based on single trials and so-called mid-run estimates, respectively. For this purpose, 286 participants performed four psychophysical temporal tasks: duration discrimination of filled and empty intervals in the range of milliseconds, duration discrimination of filled intervals in the range of seconds, and rhythm perception. All tasks employed an adaptive rule based on the weighted-up-down procedure (Kaernbach, 1991). As a psychophysical indicator of performance, 75%-difference thresholds were determined on the basis of both single trials and mid-run estimates. As a measure of congruence, Pearson correlations were computed across individual threshold estimates obtained with both methods. Correlation coefficients ranging from r_{xy} = .95 to r_{xy} = .98 suggest a highly reliable level of equivalence for both methods of threshold estimation.

Adaptive methodologies are widely used in different fields of psychological research. In adaptive testing, the item characteristics of each test item are determined by the responses that have been given on the last item or sequence of items. The decision of which item will be presented next is made through adaptive algorithms. This allows a very quick assessment of the measured criterion, as the number of presented test items is reduced: only items which offer maximal information about the skill level of any subject are presented. This characteristic of adaptive testing constitutes the most obvious advantage of this measurement technique: it allows measuring a criterion with relatively few items compared to non-adaptive procedures but without cutting back on item difficulties or on the efficiency of the test in general. Other advantages are the increase of test security as the presented items depend on individual skill levels and therefore vary from subject to subject as well as the small sample reliability of adaptive testing, which has been pointed out by Levitt (1971).

There are also two major disadvantages of adaptive methodologies: First, the development of appropriate items is more complex than in non-adaptive testing and, secondly, computer-based testing it often required because complex algorithms are necessary to identify items with appropriate item difficulties.

In sum, adaptive procedures can be more efficient than non-adaptive, classical testing procedures. Especially research on sensation and perception relies highly on adaptive methodologies. In psychophysics, an adaptive procedure can be described as “one in which the stimulus level of any one trial is determined by the preceding stimuli and responses” (Levitt, 1971, p. 1279). This makes adaptive measurements highly efficient for psychophysiological research, because the number of presented stimuli can be reduced without making the measurement less reliable.

Because changes of item characteristics are associated with changes in the ability to give correct answers, measures of performance on the items or tasks presented constitute a psychometric function (Leek, 2001). This means, that there is a clear relationship between the difficulty of tasks presented and the probability for a correct answer. The psychometric function contains all relevant data for psychophysics by displaying the stimulus strength on the abscissa and the subject's response on the ordinate (Klein, 2001).

During the last century, there has been constant development and refinement on adaptive procedures to gain highly efficient and reliable measurement techniques (Leek, 2001). In her description of adaptive procedures in psychophysical research, Leek describes three categories of procedures, which have been widely used in this research area during the last decades: Parameter estimation by sequential testing (PEST), maximum-likelihood adaptive procedures, and staircase procedures. This article will focus only on staircase procedures, which can be broadly described as follows: They normally use the previous response(s) within an adaptive track to select the next placement of trial and then provide a threshold estimate which normally varies between $x_{.50}$ and $x_{.75}$ of the individual psychometric function depending on the adaptive rule that has been chosen. The threshold estimate itself is "a measure of location of the function along the stimulus axis" (Leek, 2001, p 1279). It is determined as a level of detection (or discrimination) performance. For instance, $x_{.50}$ means, that at this point of the psychometric function the track targets the stimulus level for which the probability of a correct or positive response equals the probability of a negative or incorrect response (Leek, 2001).

In 1991, Kaernbach introduced the weighted up-down method (WUDM) as a simple staircase procedure for adaptive testing. The WUDM can be seen as a modification of Derman's (1957) version of the simple up-down procedure. As with the simple up-down procedure, the WUDM follows the general principle of decreasing the signal level after a correct response and increasing the signal level after an incorrect response. However, unlike the simple up-down procedure, the WUDM allows a different step size for the upward steps (S_{up}) than for the downward steps (S_{down}). According to Kaernbach (1991), "the equilibrium condition for convergence point X_p is

$$S_{up}\, p = S_{down}\, (1 - p).$$

For $x_{.75}$, it follows that $S_{up}/S_{down} = 1/3$. The rule for a convergence to the $x_{.75}$ point would thus read: Decrease the Level 1 step after each correct response, and increase it 3 steps after each incorrect response" (p. 227).

The adoption of an adaptive algorithm as a rule for placing observations (for instance, Kaernbach's version of the weighted up-down method which places observation around the $x_{.75}$ point) leads to data that rise and fall in difficulty depending on either positive or negative responses of any given subject. These increments (either increasing or decreasing) are referred to as *steps,* while a series of steps in one direction (i.e., only positive or negative responses, respectively) is defined as a *run* (Levitt, 1971).

After the placing of observations by means of an adaptive algorithm, the data have to get analyzed. There are different methods which can be used for threshold estimation, whereas differences between these methods mainly concern the number of trials being involved in the estimation process. In the present article, the focus will be set on two methods: Mid-run-estimates (Wetherill, 1963) and an estimate based on single trials.

The method of mid-run-estimates averages the peaks and valleys of every second run to provide an estimation of performance (Levitt, 1971), so the midpoint of every second run is used for the estimation procedure. According to Levitt (1971), the method of mid-run estimates has shown to be a simple but robust procedure with excellent precision. A

noteworthy disadvantage of this method is the small number of data points being used for threshold estimation, i.e., only the midpoints of every *second* run. This can reduce the reliability of the measurement, especially when the number of stimulus presentations is rather small.

The method of threshold estimation on the basis of single trials is an attempt to overcome the disadvantage of mid-run estimates as, in this method, every single data point is used for threshold estimation, i.e., not the midpoints of every second run are being involved in the estimation process but the mean across all single trials. This leads to a larger number of data points as a basis of threshold estimation compared to the method of mid-run-estimates. Because results can be biased due to inadequate initial step sizes at the start of the experiment, different techniques for preventing such a bias are available (Levitt, 1971). In the present study, only data points after an initial stabilization phase of 12 trials were used for threshold estimation.

To date, threshold estimates based on single trials have not been compared to those derived from mid-run estimates. The aim of the present study, therefore, was to evaluate the equivalence of threshold estimates obtained by means of mid-run estimates as well as those based on single trials. For this purpose, both methods of threshold estimation were compared by means of four psychophysical timing tasks, namely duration discrimination of filled and empty intervals in the range of milliseconds, duration discrimination of filled intervals in the range of seconds, and rhythm perception. As a measure of equivalence, Pearson correlations between both methods of threshold estimation were computed. Furthermore, absolute values of threshold estimates resulting from both methods were compared by means of *t* tests.

Method

Participants

Participants were 134 male and 152 female volunteers ranging in age from 15 to 51 years (mean ± standard deviation of age: 25.3 ± 6.3 years). All participants had normal hearing and normal or corrected-to-normal sight.

Experimental Tasks

Duration discrimination tasks

Because interval timing may be influenced by type of interval (filled vs. empty) and base duration, the duration discrimination task consisted of one block of filled and one block of empty intervals with a base duration of 50 ms each, as well as one block of filled intervals with a base duration of 1000 ms. Furthermore, when participants are asked to compare time intervals, many of them adopt a counting strategy. Since explicit counting becomes a useful timing strategy for intervals longer than approximately 1200 ms (Grondin, Meilleur-Wells, & Lachance, 1999), the "long" base duration was chosen not to exceed this critical value.

Stimuli. Filled intervals were white-noise bursts from a computer-controlled sound generator (Phylab Model 1), presented binaurally through headphones (Vivanco SR85) at an intensity of 67 dB. The empty intervals were marked by onset and offset clicks 3 ms in duration, with an intensity of 88 dB. These intensity levels were chosen on the basis of the results of a prior pilot experiment in which 12 subjects were asked to adjust the loudness of a 3-ms click until it matched that of a 1000-ms white-noise signal.

Procedure. The order of blocks was counterbalanced across participants. Each block consisted of 32 trials, and each trial consisted of one standard interval (= base duration) and one comparison interval. The duration of the comparison interval varied according to an

adaptive rule (Kaernbach, 1991) to estimate $x_{.75}$ of the individual psychometric function, that is, the comparison interval at which the response "longer" was given with a probability of 0.75. Threshold estimation was performed twice: based on single trials and mid-run estimates, respectively. To reduce estimation bias, the first two runs were not considered for threshold estimation based on mid-run-estimates. For single trials, estimation of difference-threshold values was based on Trials 13-32. This resulted in 20 data points for single trials and a mean number of 9.2 data points for mid-run-estimates of the duration discrimination task with filled intervals in the range of milliseconds, 10.3 data points for the duration discrimination task with empty intervals in the range of milliseconds, and 8.0 data points for the duration discrimination task in the range of seconds, respectively.

Within each experimental block, the order of presentation for the standard interval and the comparison interval was randomized and balanced with each interval being presented first in 50% of the trials. Trials were randomly interleaved within a block. Within each trial, the two intervals were presented with an interstimulus interval of 900 ms. The participant's task was to decide which of the two intervals was longer and to indicate his or her decision by pressing one of two designated response keys. After each response, visual feedback ("+", i.e., correct; "–", i.e., false) was displayed on the computer screen. The next trial started 900 ms after the feedback.

Rhythm perception task

Stimuli. The stimuli consisted of 3-ms clicks presented binaurally through headphones at an intensity of 88 dB.

Procedure. Participants were presented with auditory rhythmic patterns, each consisting of a sequence of six 3-ms clicks marking five beat-to-beat intervals. Four of these intervals (Interval 1, 2, 3, and 5) were of a constant duration of 150 ms, while one interval (Interval 4) was variable (150 ms+x). The magnitude of x changed from trial to trial depending on the participant's previous response according to the weighted up–down procedure (Kaernbach, 1991) which converged on a probability of hits of 0.75. Correct responding resulted in a decrease of x and incorrect responses made the task easier by increasing the value of x. Thus, the weighted up–down procedure was used to determine the 75% threshold as an indicator of performance on rhythm perception. Again, estimation of differences thresholds was performed twice: based on single trials and mid-run estimates, respectively. To reduce estimation bias, the first two runs were not considered for threshold estimation. An experimental block consisted of 32 trials.

The participant's task was to decide whether the presented rhythmic pattern was perceived as "regular" (i.e., all beat-to-beat intervals appeared to be of the same duration) or "irregular" (i.e., one beat-to-beat interval was perceived as deviant). Participants indicated their decision by pressing one of two designated response keys. No feedback was given, as there were no perfectly isochronous ("regular") patterns presented.

As for the duration discrimination tasks, estimation bias was reduced by excluding Runs 1 and 2 for mid-run-estimates and Trials 1-12 for single trials from the estimation process. This resulted in a mean number of 13.4 data points for mid-run-estimates and 20 data points for single trials, respectively.

Results

Table 1 shows Pearson correlations between both methods of threshold estimation as a measure of equivalence. Correlations ranging from r_{xy} = .95 to r_{xy} = .98 indicate very high consistency between the two methods. In addition, single trials and mid-run-estimates correlate much higher within the same task than between tasks.

Table 1
Pearson correlations of threshold estimates based on mid-run-estimates (MRE) and single trials (ST) for the four psychophysical timing tasks.

	ST Task 1	ST Task 2	ST Task 3	ST Task 4
MRE Task 1	**.973*****			
MRE Task 2	.301***	**.980*****		
MRE Task 3	.142*	.258***	**.974*****	
MRE Task 4	.117*	.269***	.156**	**.951*****

Task 1: Duration discrimination with filled intervals in the range of milliseconds
Task 2: Duration discrimination with empty intervals in the range of milliseconds
Task 3: Duration discrimination with filled intervals in the range of seconds
Task 4: Rhythm perception

***$p \leq .001$, **$p \leq .01$, *$p \leq .05$ (2-tailed)

To compare absolute values of threshold estimates based on single trials and mid-run-estimates, paired *t* tests were computed. The results of these analyses are shown in Table 2. Differences between threshold estimates obtained by the single-trial method and by mid-run-estimates reached statistical significance for the following tasks: duration discrimination with filled intervals in the range of milliseconds [$t(285) = -4.29$, $p < .001$, $d = -.03$], duration discrimination with filled intervals in the range of seconds [$t(282) = -3.09$, $p < .0.1$, $d = -.05$], and rhythm perception [$t(285) = 12.38$, $p < .001$, $d = .31$]. While effect sizes for the three duration discrimination tasks were extremely small, rhythm perception yielded a moderate effect size. Percent differences between threshold estimates derived from both methods were -3.9%, 0.9%, -3.4% and 10.8% for Tasks 1-4.

Discussion

A possible disadvantage of the well-known method of mid-run-estimates is the relatively small number of data points available for threshold estimation. As an alternative method, threshold estimation based on single trials was compared with the former one and tested for equivalence by means of four different adaptive timing tasks. Correlations between both methods indicate a strong association and high consistency for each of the four

Table 2
Means (M) and standard errors of means (SEM) as well as *t* values (*t*) and effect sizes (*d*) for threshold estimates based on single trials and mid-run-estimates for the four psycho-physical timing tasks

	Single Trials		Mid-run-estimates		*t*	*p*	*d*
	M	SEM	M	SEM			
Task 1	10.3	.41	10.7	.37	-4.29	.001	-.03
Task 2	21.0	.76	20.8	.69	1.43	.15	.01
Task 3	157.5	6.48	162.9	7.24	-3.09	.002	-.05
Task 4	55.4	1.30	49.4	.97	12.38	.001	.31

Abbreviations: see Table 1

psychophysical tasks used in the present study. This finding endorses the view that both methods yield virtually identical threshold estimates. It should be noted, however, that *t* tests revealed statistically significant differences in absolute threshold values obtained with both methods for three of the four timing tasks. Considering this result, two points need to be discussed further: The direction of the differences in threshold values and the inspection of effect size estimates.

First, the direction of statistically significant differences in duration discrimination and rhythm perception does not suggest systematic under- or overestimation for one of the two methods. For threshold estimation obtained by mid-run-estimates, threshold values were significantly smaller for the rhythm perception task, but reliably larger for the duration discrimination tasks with filled intervals in the range of milliseconds and seconds. Thus, differences in threshold estimates obtained by the two methods seem to be rather task-inherent than being dependent on the method used for threshold estimation.

Secondly, the inspection of effect size estimates suggests that the revealed differences in the statistically significant *t* tests of the two duration discrimination tasks are of only negligible practical importance as the effect sizes turned out to be rather small. A special case, however, represents the rhythm perception task as this task yielded a considerably higher effect size estimate. The difference threshold obtained by mid-run-estimates was notably smaller compared to the one based on single trials. To this point, it remains unclear what factors account for the discrepancies found for the rhythm perception task. Therefore, further investigations are certainly indicated.

In sum, threshold estimation by means of mid-run-estimates and based on single trials appear to constitute equivalent alternatives for quantification of performance on adaptive duration discrimination tasks. Nevertheless, observed differences in absolute magnitude of threshold values obtained by mid-run estimates and the single-trial method are still to be accounted for.

References

Grondin, S., Meilleur-Wells, G., & Lachance, R. (1999). When to start explicit counting in time-intervals discrimination task: A critical point in the timing process of humans. *Journal of Experimental Psychology. Human Perception and Performance, 25,* 993−1004.

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

Klein, S. A. (2001). Measuring, estimating, and understanding the psychometric function: A commentary. *Perception & Psychophysics, 63,* 1421-1455.

Leek, M. R. (2001). Adaptive procedures in psychophysical research. *Perception & Psychophysics, 63,* 1279-1292.

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

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