On estimating the difference limen in duration discrimination tasks: A comparison of the 2AFC and the reminder task

EINAT LAPID AND ROLF ULRICH
University of Tübingen, Tübingen, Germany

AND

THOMAS RAMMSAYER
University of Bern, Bern, Switzerland

This article assesses whether the two-alternative forced-choice (2AFC) and the reminder tasks (i.e., method of constant stimuli) yield identical estimates of the difference limen (DL). In a series of six experiments, participants discriminated between two duration stimuli. Experiments 1–5 employed auditory stimuli, and Experiment 6 employed visual stimuli. Experiments 1 and 2 combined each of the two tasks with an adaptive and a nonadaptive procedure for threshold estimation. Experiment 3 varied the distribution of the comparison levels, whereas Experiment 4 employed random interstimulus intervals. Experiments 5 and 6 examined the influence of the presentation order of the standard and comparison stimuli. Results indicate that both the adaptive and the nonadaptive procedures yield virtually identical DL estimates; yet, the 2AFC task produces consistently larger DLs than does the reminder task. In addition, DL increases when the standard occurs in the second rather than in the first stimulus position. In order to account for these results, we assume that participants use an internal standard instead of the actually presented standard as a reference for their judgment.

The *difference limen* (DL)—also termed *just noticeable difference* (JND)—is a traditional and important concept in psychophysical work. It provides a means for quantifying the sensitivity of a sensory system. Although signal detection theory also provides powerful tools to assess the sensitivity of a sensory system (see Macmillan & Creelman, 2005; Wickens, 2002), this theory has not replaced the concept of DL. Indeed, a 2001 special issue of *Perception & Psychophysics* edited by Klein and Macmillan covered new approaches to estimating DL.

The DL measures the amount of change in a stimulus that is required to produce a JND in the sensation of this stimulus (see, e.g., Gescheider, 1997; Guilford, 1954). For example, on each trial of a duration discrimination task, two tones are successively presented to a participant (e.g., Getty, 1975). One tone is the standard, and its duration is kept constant—say, at 500 msec—throughout the experiment. The second tone is the comparison, and the experimenter varies its duration from trial to trial. The listener may experience a difference between the two durations only when the comparison is at least 554 msec. In this case, the DL would be 54 msec—that is, equal to the duration difference between the comparison and the standard. The smaller the DL, the higher the discrimination performance.

Several psychophysical tasks are available for measuring the DL (for a review, see Gescheider, 1997). Two common tasks are the reminder task and the two-alternative forced-choice (2AFC) task (Macmillan & Creelman, 2005).¹ In either instance, the participant is presented with a standard and a comparison stimulus on each trial. Moreover, several comparison durations are employed, and in each trial, one of the comparisons is randomly chosen from this set of durations. In the reminder task, the standard is always presented first, followed by the comparison. On each trial, the magnitude of the comparison can be smaller, equal to, or larger than the magnitude of the standard (e.g., the duration of the comparison may be shorter, equal to, or longer than the duration of the standard). At the end of each trial, the participant is asked to report whether the comparison was longer or shorter than the standard by responding “longer” or “shorter,” respectively. The results generated by this reminder task can be displayed in the form of a psychometric function. This function plots the proportion of “longer” responses against comparison duration. Typically, this function resembles the shape of an ogive curve, is 0 at small comparison values, and approaches 1 for large values. A psychometric function from a certain parametric family (e.g., logistic function) is then fitted to the observed data points. Finally, the DL is es-

timated as being half the interquartile range of this fitted function—that is, $DL = (x_{.75} - x_{.25})/2$, where $x_{.25}$ and $x_{.75}$ denote the values of the comparison that yield 25% and 75% “longer” responses, respectively. The steeper this psychometric function, the smaller the DL; thus, the higher is the differential sensitivity of the observer (see Luce & Galanter, 1967). The values $x_{.25}$ and $x_{.75}$ are usually used, and the most common alternatives are $x_{.29}$ and $x_{.71}$, which arise in transformed up–down procedures (see, e.g., Bode & Carhart, 1973; Levitt, 1971).

Although the 2AFC task also employs a standard and a comparison on each trial, this task for estimating DL differs greatly from the reminder task. There are two subtle but crucial differences between these two tasks. First, with the reminder task, the standard always precedes the comparison, whereas the standard and the comparison are presented in random order across trials in 2AFC. Second, in the reminder task, the comparison can be smaller, equal to, or larger than the standard. By contrast, in the 2AFC task, the comparison is always larger than or at least equal to the standard, and the participant is asked to indicate whether the first or second stimulus is larger (e.g., in the case of a duration discrimination task, the participant indicates whether the first or second interval appeared longer). The experimenter simply notes whether the response is correct. The proportion of correct responses is displayed as a psychometric function. In contrast to the reminder task, however, this proportion increases from the chance level of .5 when the value of the comparison is equal to the standard up to 1.0 for very large comparisons. As before, a psychometric function from a parametric family is fitted to the observed data points (e.g., a scaled logistic function). On the basis of this function, one determines the difference threshold, which is usually defined as the comparison value at which the proportion of correct responses is .75 (see, e.g., McKee, Klein, & Teller, 1985). The DL is finally obtained by subtracting the value of the standard from this estimated threshold value. That is, the DL indicates the value of the comparison (i.e., $DL + \text{value of standard}$) at which the comparison is correctly detected 75% of the time.

In psychophysical research, one or the other of these two tasks is exclusively employed to measure the DL. For example, in research on time perception, DL is sometimes estimated with the 2AFC task (see, e.g., Karmarkar & Buonomano, 2003; Rammsayer & Lima, 1991; Wright, Buonomano, Mahncke, & Merzenich, 1997) and sometimes with the reminder task (e.g., Getty, 1975; Grondin, 2001; Tse, Intriligator, Rivest, & Cavanagh, 2004; Ulrich, Nitschke, & Rammsayer, 2006). This practice in research may reflect a common implicit assumption that the two tasks should result in more or less the same estimates of DL (e.g., Thompson, Schiffman, & Bobko, 1976; Wright et al., 1997).² In fact, we are not aware of an empirical study that has addressed this issue in a systematic way. Therefore, the main goal of the present study was to provide data for assessing this issue. We used a duration discrimination task for measuring the DL. Hence, in the following experiments, we employed both the reminder task and the 2AFC task in order to estimate the DL within the same group of

participants and to evaluate whether this implicit assumption holds. Moreover, in order to assess the stability of the DL obtained with each task, we tested our participants on two occasions and computed the test–retest reliability (see Linschoten, Harvey, Eller, & Jafek, 2001).

Furthermore, the experimenter not only needs to decide whether he or she will use the reminder task or the 2AFC task for measuring DL, but an additional decision also needs to be made about whether the data should be collected by means of an adaptive or a nonadaptive psychophysical procedure. With a typical nonadaptive procedure, five or more levels of the comparison value are predetermined around the threshold region and administered to the participant several times in random order. With adaptive procedures, however, levels of the comparison are not predetermined but are governed by the participant's response history. Adaptive procedures enjoy widespread use in psychophysics because they avoid trials with an inefficient placement of comparison values—that is, values that are either too small or too large (for a review, see Treutwein, 1995). Thus, in contrast to nonadaptive procedures, they rapidly extract relevant information from a psychometric function that underlies discrimination performance.

Comparisons of adaptive and nonadaptive procedures have sometimes been performed with computer simulations (see, e.g., García-Pérez & Alcalá-Quintana, 2005; Simpson, 1988; Watson & Fitzhugh, 1990). However, surprisingly little behavioral work has been directed to the question of whether adaptive and nonadaptive procedures differ in estimating DL.³ Empirical work is important to validate the knowledge that has been reached by computer simulations. Such simulations require several assumptions that may be violated in practice (see Leek, 2001, p. 1288). For example, simulations assume that the underlying psychometric function is unaffected by perceptual learning or that the perceptual outcome on a certain trial is independent from the outcome of the preceding trials.

A few empirical studies provide some clues on this issue (Brand & Hohmann, 2002; Dai, 1995; Hesse, 1986). According to these studies, an efficient method would require relatively few trials to achieve a certain level of accuracy for estimating discrimination performance. Unfortunately, these studies provide a rather inconsistent picture of results. For example, Brand and Hohmann reported that adaptive procedures are more efficient than nonadaptive ones; Hesse reported that nonadaptive procedures are more efficient, and Dai reported that both procedures are equally effective. Thus, in conclusion—and perhaps somewhat surprisingly—decisive empirical work on comparing adaptive and nonadaptive procedures is still lacking. Therefore, a secondary goal of the present study was to address this issue. Specifically, we employed test–retest reliability methodology in order to assess temporal stability of the DL estimates using the adaptive and the nonadaptive procedure.

We used the adaptive weighted up–down procedure that was developed by Kaernbach (1991) for two reasons. First, Kaernbach's procedure provides an especially simple tracking algorithm that is easy to implement in an experiment. This facility may explain its increasing popu-

larity in psychophysical research. Second, experimental work with this procedure has often been performed on duration discrimination (see, e.g., Grondin, 1993; Grondin, Ivry, Franz, Perreault, & Metthé, 1996; Rammsayer, 1992; Rammsayer & Ulrich, 2005; Ulrich et al., 2006). In this domain of research, usually several experimental conditions are run within a single experiment, and each condition requires the estimation of DL. Therefore, it is important to know for future research whether this approach provides sufficiently stable estimates of DL.

To sum up, the major goal of this article was to assess whether DL estimates from the 2AFC task and the reminder task are of the same magnitude. As a secondary goal, each task was combined with an adaptive and nonadaptive procedure in order to assess potential differences between these two psychophysical procedures. In order to reduce the burden for the participants, we decided to decompose the complete factorial design—that is, task (reminder vs. 2AFC) $\times$ procedure (adaptive vs. nonadaptive) $\times$ order of blocks (first vs. second)—into two feasible designs. In Experiment 1, we excluded the 2AFC task from the complete design. In Experiment 2, we only excluded the nonadaptive version of the reminder task from the complete design. Thus, Experiment 1—together with Experiment 2—still enables a comparison between the 2AFC and the reminder task. Experiments 3–6 were designed to test specific hypotheses that emerged from the results of Experiments 1 and 2.

EXPERIMENT 1

Participants performed the reminder version of the duration discrimination task using filled auditory intervals. On each trial, a standard tone of 500 msec preceded a variable comparison tone that could be shorter, equal to, or longer than the standard. At the end of each trial, the participant had to indicate which of the two stimuli was longer—that is, the first or the second one. Each participant performed four consecutive blocks of trials. In the two adaptive blocks, the trial-to-trial changes of the comparison duration were governed by Kaernbach's (1991) weighted up-and-down procedure. There were two interleaved trial runs. One run estimated the 25% level of the psychometric function and the other run its 75% level, which is referred to as the *doublet procedure* (see Leek, 2001). In the two remaining nonadaptive blocks, the duration of the comparison was sampled on each trial from a fixed set of prespecified durations that were symmetrically arranged above and below the standard duration of 500 msec. Following previous suggestions (García-Pérez & Alcalá-Quintana, 2005; Hall, 1981; Leek, Hanna, & Marshall, 1992), we used a maximum-likelihood analysis to estimate DL (and also the point of subjective equality [PSE]) from the collected data in each block for each participant. Test-retest reliabilities were computed by correlating the DL estimates from the first and second block of each procedure. In a second set of analyses, we employed the Spearman-Kärber technique (J. Miller & Ulrich, 2001) in order to summarize each participant's psychometric function. In contrast to the maximum-likelihood

approach to estimation of DL, this technique does not require any assumption about the shape of the underlying psychometric function for assessing the slope of an observed psychometric function (see J. Miller & Ulrich, 2001, for a comparison of these techniques).

Method

Participants. Twenty-four female and 6 male students of the University of Tübingen (mean age: 22.7 ± 4.1 years) participated in a single experimental session that lasted approximately 60 min. They had normal hearing and were naive about the purpose of the study.

Apparatus and Stimuli. A PC controlled the presentation of the stimuli and the recording of the participants' responses. Auditory stimuli used for all tasks were temporal intervals of white noise that were generated by a SoundBlaster-compatible soundcard, and they were presented binaurally via headphones (Philips SBC Hp 200) at an intensity of 85 dB SPL.

General procedure. The time course of a single trial was identical for the adaptive and the nonadaptive procedure. A trial started with the presentation of the 500-msec standard interval. 1,000 msec after the offset of the standard, the variable comparison interval was presented. At the end of the trial, participants were asked to decide which of the two intervals was longer—the first or the second one. Participants were not informed about the existence of a standard. They pressed the left "shift" key of a computer keyboard when they judged the first interval as being longer and the right "shift" key when they judged the second interval as being longer. The experimenter emphasized accuracy over speed. Two seconds after the response, the next trial started with the presentation of the standard. The experiment consisted of four blocks, each containing 110 trials. Previous research has indicated that at least 100 trials are required for satisfactory estimation of DL (see Leek et al., 1992). Two blocks involved the adaptive procedure, and the remaining two blocks involved the nonadaptive procedure. There were six possible orders of the four blocks (i.e., AANN, ANNA, ANAN, NAAN, NANA, NNA, with N = nonadaptive, A = adaptive), and these six orders were counterbalanced across the 30 participants. In order to make sure that the participant understood the general task, a practice block was presented at the beginning of the experimental session (but not before each block). In order to not give advantage to one task over the other, the practice block consisted of 20 trials with a standard interval of 400 msec, in contrast to the experiment that engaged standards of 500 msec. In any case, no information was given on the duration of the standard.

Adaptive blocks. Durations of the comparison followed the weighted up-down rule (Kaernbach, 1991). Two step sizes were employed—10 and 30 msec (in a ratio of 1:3)—that were kept constant throughout the experiment as in previous work (e.g., Rammsayer, 1992; Ulrich et al., 2006). Two separate interleaved runs (55 trials each) of the comparison duration were employed. One run targeted at the 75% level of performance and the second run at 25%, since these two target levels are the most common for estimating DL. The starting value for the 75% run was set at 600 msec—that is, well above the standard duration. Whenever the comparison stimulus was judged as being longer, its duration was decreased by 10 msec, and the duration was increased by 30 msec when it was judged as being shorter. An exactly opposite procedure was employed for the 25% run. That is, whenever the comparison stimulus was judged as being shorter than the standard, the duration was increased by 10 msec, and it was decreased by 30 msec when judged as being longer. The starting value of the 25% run was set well below the standard at 400 msec. The two independent interleaved runs—employed in the same block—made it impossible for a participant to predict the next signal level to be presented. Figure 1 provides an illustration of these two runs.

Nonadaptive blocks. The design was identical to the one of the adaptive blocks, with the only exception being that the levels of the comparison stimuli were preselected and constant throughout a block (i.e., method of constant stimuli). Eleven levels were selected

Figure 1. One block of the adaptive procedure consisted of two independent interleaved runs (55 trials each). This figure depicts the data for one block of a real participant. The graph shows the duration of the comparison stimuli (in milliseconds) on each trial as the experiment progressed. Stimulus levels followed an adaptive rule as described in the text corresponding to the test person's responses. Filled circles represent 75% run and open circles represent 25% run. The dashed line indicates the estimated PSE ($x_{.50}$), the lower solid line the estimated $x_{.25}$, and the upper solid line the estimated $x_{.75}$.

ranging from 400 to 600 msec in constant steps of 20 msec, so that five levels were below the standard and five levels were above it. The 11th level was equal to the standard—that is, equal to 500 msec. Each level was presented to the participant 10 times in random order, resulting in a total of 110 trials.

Estimation of DL. The technique for estimating DL was identical for adaptive and nonadaptive blocks and followed previous research (e.g., García-Pérez & Alcalá-Quintana, 2005; Leek et al., 1992). All 110 trials of a single block were always used to generate the full psychometric function. A logistic psychometric function

$$\Psi(x) = \frac{1}{1 + \exp[-(x - a)/b]} \quad (1)$$

was used to compute the maximum likelihood estimates of DL and the PSE, where x denotes the length of the comparison duration. The PSE is equal to a , and DL is equal to $b \cdot \log(.75/.25)$ (see Bush, 1967). This function specifies the probability $\Psi(x)$ of the response “comparison longer than standard” for each level x of the comparison. The function $\Psi(x)$ ranges from 0 for extremely brief comparison intervals to 1 for extremely long comparison durations (Figure 2). In this and the following experiments, obtained estimates were checked by plotting the data together with the corresponding psychometric function in order to see whether the fit was reasonable.

Results and Discussion

A separate two-way ANOVA with factors procedure (adaptive vs. nonadaptive) and block (first vs. second) was performed for DL and PSE. Figure 3 depicts the result for DL. The overall mean DL ($\pm SE$) was 32.2 ± 1.2 msec. DL estimates did not differ significantly between the two procedures [$F(1,29) = 2.43$, $MS_e = 54.5$, $p = .13$, $\eta^2 = .08$]; mean DL was 31.1 ± 1.9 msec and 33.2 ± 1.7 msec for the adaptive and nonadaptive procedures, respectively. This result suggests that both procedures give about the same

estimates of DL. Although performance slightly improved with practice, the main effect of block did not reach statistical significance [$F(1,29) = 2.73$, $MS_e = 80.0$, $p = .11$, $\eta^2 = .09$]. The mean DL was 33.5 ± 1.67 msec and 30.8 ± 1.93 msec for the first and second block, respectively. The interaction of both factors was not significant [$F(1,29) = 0.02$, $MS_e = 35.2$, $p = .88$, $\eta^2 = .00$]. The obtained mean Weber fractions were 6.6% for the nonadaptive procedure and 6.2% for the adaptive. Almost identical fractions have been reported for duration discrimination with auditory stimuli in previous studies (Getty, 1975; Grondin, 1993; Rammsayer & Ulrich, 2005). The present results support the idea that the means of collecting the data have very little effect on the DL estimates.

A similar analysis of the PSE again yielded no significant main effect of procedure [$F(1,29) = 0.2$, $MS_e = 252.7$, $p = .97$, $\eta^2 = .01$]. The overall mean PSE was 502.2 ± 1.7 msec and was virtually identical for both procedures—that is, 502.1 msec and 502.3 msec for the nonadaptive and adaptive procedures, respectively. Neither the main effect of block [$F(1,29) = 3.5$, $MS_e = 246.5$, $p = .07$, $\eta^2 = .11$] nor the interaction of both factors [$F(1,29) = 0.2$, $MS_e = 174.2$, $p = .681$, $\eta^2 = .01$] became statistically significant.

In a second set of analyses, the nonparametric Spearman–Kärber technique was used to summarize each participant's psychometric function. This technique treats a psychometric function as a cumulative probability function (see, e.g., Finney, 1952; Trevan, 1927) and summarizes it in terms of its moments (i.e., mean, standard deviation [SD], and higher moments). We used the computer

Figure 2. Psychometric logistic function fitted to the data depicted in Figure 1. The data themselves are represented by filled circles. This function has a PSE of 536 msec and a DL of 19 msec.

program PMETRIC (J. Miller & Ulrich, 2004) to compute the mean, the *SD*, and the skewness of each participant's psychometric function. The mean (like the PSE) measures the central tendency of a psychometric function, and the *SD* (like the DL) measures its steepness and thus a participant's discrimination performance.⁴ We also employed a measure of skewness to assess whether the observed psychometric functions would deviate meaningfully from symmetry. As before, each summary statistic was submitted to a separate ANOVA. First, the overall mean was 502 ± 1.63 msec, and thus virtually identical to the overall PSE. The average mean tended to be slightly larger in the first block than in the second block [$F(1,29) = 3.1$, $MS_e = 231.0$, $p = .088$, $\eta^2 = .10$] (i.e., 504.7 msec vs. 499.9 msec). No other effects were significant ($F < 1$). Second, the overall *SD* was 46.5 ± 1.2 msec. The *SD* decreased somewhat—yet significantly—with practice indicating a slightly improved discrimination performance in the second half of the experiment [$F(1,29) = 5.6$, $MS_e = 81.0$, $p = .025$, $\eta^2 = .16$] (i.e., 48.4 ± 1.5 msec vs. 44.5 ± 1.7 msec). There was no significant main effect of procedure [$F(1,29) = 2.8$, $MS_e = 57.0$, $p = .106$, $\eta^2 = .09$] or a significant interaction [$F(1,29) = 2.8$, $MS_e = 57.0$, $p = .106$, $\eta^2 = .09$]. Finally, the coefficient of skewness (i.e., the third central moment divided by *SD*; see Evans, Hastings, & Peacock, 2000) was computed for each single psychometric function. The average coefficient was 0.238 and was significantly larger than zero [$F(1,29) = 36.6$, $MS_e = 0.186$, $p < .001$, $\eta^2 = .56$]. Although this value indicates a small positive skewness, the observed functions were virtually symmetrical.⁵ There were no further significant effects on this coefficient.

We also assessed whether the parametric and nonparametric estimates are correlated as one would expect if both index the same concept. In agreement with this expecta-

tion, the mean obtained from the Spearman–Kärber technique was highly correlated across participants with the parametric PSE estimate—that is, the product moment correlation ranged from .95 to .99 across the different conditions. In addition, the *SD* of the Spearman–Kärber technique and the DL estimate were also highly correlated (.973 to .997).

Thus, in summary, the data of Experiment 1 show that the adaptive and the nonadaptive procedures produce virtually identical DL and PSE results. The same conclusion ap-

Figure 3. The graph includes the results of Experiments 1 and 2 and thus shows mean DL as a function of task and procedure. The x-axis shows the different procedures grouped according to task, whereas the y-axis shows DL (msec). Black bars represent the adaptive procedure; light gray bars represent the nonadaptive procedure, and the dark gray bar represents the additionally included control condition of Experiment 2.

Figure 4. Illustration of the 2AFC adaptive task. Comparison duration is plotted as a function of trial number. Data were generated by a participant.

plies to the summary statistics from the Spearman–Kärber technique. In addition, the estimates of both analyses agree surprisingly well, although the underlying assumptions and the computational steps of both techniques differ greatly. (The results on test–retest reliability will be presented and discussed together with those of Experiment 2.)

EXPERIMENT 2

This experiment was similar to Experiment 1, except that the DL was estimated by means of the 2AFC task. We also included the adaptive version of the reminder task in the design of Experiment 2 in order to enable a comparison between the 2AFC and the reminder task. DL was estimated by a standard parametric approach and again by the nonparametric Spearman–Kärber technique (Ulrich & Miller, 2004).

Method

Participants. A new group of 30 volunteers (9 males, 21 females) ranging in age from 20 to 44 years (mean age: 27.8 ± 6.3 years) was recruited for this experiment. All had normal hearing and were naive about the purpose of the study. Two participants were replaced due to DL estimates that were 2 *SDs* above the mean of the entire sample.

Apparatus, Stimuli, and Procedure. The apparatus and the auditory stimuli were identical to those employed in Experiment 1. On a single trial of the 2AFC task, the 500-msec standard occurred either first or second. In contrast to Experiment 1, the comparison was always longer than the standard, and the presentation order of the standard and comparison varied randomly from trial to trial. Participants were asked to indicate the longer interval. An experimental session comprised six blocks: two adaptive 2AFC blocks (A), two nonadaptive 2AFC blocks (B), and two reminder adaptive blocks (C). There were six possible orders of the four blocks (i.e., ABCABC, ACBACB, BACBAC, BCABCA, CABACB, and CBACBA), and these orders were counterbalanced across the 30 participants.

Adaptive 2AFC blocks. In the adaptive 2AFC condition, the comparison duration followed the same adaptive rule as that in Experiment 1, with the same step sizes of 10 msec and 30 msec

converging to a stimulus level with 75% correct responses. The initial duration of the comparison was 600 msec. In contrast to Experiment 1, only one run of 100 trials was employed. Whenever the comparison duration reached the standard duration of 500 msec, the comparison duration was changed to its previous level for the next trial in order to keep the comparison duration longer than the standard duration. Figure 4 provides an example of such a run. In order to estimate DL, the data of such a run were first ordered according to the obtained levels of the comparison from short to long durations. Then, the corresponding percentage of correct responses was computed for each level. As in Experiment 1, we adopted the standard definition of the threshold as the value yielding 75%-correct performance in the 2AFC task. We again employed a standard parametric technique for estimating this threshold (see Ulrich & Miller, 2004). Specifically, the following 2AFC logistic function was fitted to the observed data points

$$\Psi(x) = 0.5 + \frac{0.5}{1 + \exp[-(x - a)/b]} \quad (2)$$

in order to estimate the parameters *a* and *b*. Here, *x* represents the difference between the comparison and the standard duration, and $\Psi(x)$ denotes the probability of a correct response at difference *x* (see Ulrich & Miller, 2004). In contrast to Equation 1, the parameter *a* now represents the threshold value (i.e., DL) instead of the PSE; that is, $x_{.75} = a$, and *b* > 0 a scale parameter. Figure 5 shows the fitted function for the data shown in Figure 4.

Nonadaptive 2AFC blocks. The estimation of DL was conducted in the same manner as that for the 2AFC adaptive condition, with the only exception being that the comparison durations were preselected and kept constant throughout the experiment. Specifically, the following 10 comparison durations were used: 515, 530, 545, 560, 575, 590, 605, 620, 635, and 650 msec. These durations were presented in random order across the trials in a single 2AFC block.

Reminder-adaptive blocks. This condition was identical to the one applied in Experiment 1.

Results and Discussion

DL results. Figure 3 depicts the results of mean DL. As we had hoped, the mean DL estimates from the reminder adaptive blocks were virtually identical to those

Figure 5. A typical psychometric function from the 2AFC task. The graph shows the percentage of correct responses as a function of comparison duration. Filled circles represent the actually visited comparisons duration (in milliseconds) and their corresponding probabilities of correct responses as generated by a real participant. The solid line represents the best fitted psychometric function. Notice that chance level is now equal to .5.

of the previous experiment, suggesting that the estimates were quite robust. A separate two-way ANOVA with the factors condition (2AFC adaptive vs. nonadaptive 2AFC vs. reminder adaptive) and block (first vs. second) yielded a reliable main effect of condition [$F(2,58) = 32.7$, $MS_e = 384.0$, $p < .001$, $\eta^2 = .53$]. Mean DL was 59.3 ± 3.8 msec, 56.5 ± 3.5 msec, and 32.9 ± 1.5 msec for 2AFC-adaptive, 2AFC-nonadaptive, and reminder-adaptive blocks, respectively. A Scheffé test revealed a significant DL difference between the reminder-adaptive procedure and each of the two 2AFC procedures, but no significant DL difference between the estimates of the two 2AFC procedures. We also employed the Spearman-Kärber technique to compute nonparametric DL estimates for the two 2AFC procedures (see Ulrich & Miller, 2004) and found almost the same results as those for the parametric estimates. In fact, the parametric and nonparametric DL estimates were again highly correlated ($r = .956-.984$). Rather surprisingly, the present results show that the DL estimates that were obtained with the 2AFC task were almost twice as large as those obtained with the reminder task. Consistent with Experiment 1, however, both the adaptive and the nonadaptive approach yielded virtually identical DL estimates in the 2AFC condition.⁶

Test-retest reliabilities. We used the data that were obtained in the two blocks of each task to compute the test-retest reliability of DL using the Pearson correlation coefficient (as performed by Linschoten et al., 2001). We did so separately for the 2AFC data in Experiment 2 and also for data of the reminder paradigm in Experiments 1 and 2. More specifically, we correlated for each single condition the parametric DL estimates that were obtained in the first and second block across all participants. All tasks exhib-

ited modestly strong but highly significant correlations (Table 1). Test-retest reliabilities of similar magnitude were reported by Linschoten et al. (2001) for a different discrimination task. Test-retest reliabilities were also computed for the nonparametric estimates, and these results were virtually identical to those of the parametric estimates.

EXPERIMENT 3

In the previous experiment, the DL estimates from the 2AFC task were almost twice as large as those from the reminder task. One may attribute this discrepancy to the different ranges of comparison durations that were employed by each task. In the reminder task, the comparison durations were shorter and longer than the standard duration. In the 2AFC task, however, the comparison durations were only longer than the standard duration. There is, however, evidence that the estimation of DL is sensitive to the dis-

Table 1
Test-Retest Correlations of DL As a Function of Task and Procedure Separately for Parametric Estimates (Probit Analysis) and Nonparametric Estimates (Spearman-Kärber) in Experiments 1 and 2

Procedure	Task	
	2AFC	Reminder
Adaptive		
Parametric	.75*	.63*
Nonparametric	.71*	.62*
Nonadaptive		
Parametric	.68*	.70*
Nonparametric	.76*	.73*

* $p < .001$.

tribution of the comparison durations in comparison with the standard duration (Grondin, Ouellet, & Roussel, 2001), indicating smaller DLs when the comparison durations are shorter than the standard duration and larger DLs when the comparisons are longer than the standard. Hence, one might assume that the 2AFC task yields larger DLs than does the reminder task.⁷ In order to test this hypothesis, only shorter comparison durations than the standard duration were employed in the 2AFC condition of this experiment. With this procedure, we expected smaller DLs in the 2AFC condition than in the reminder condition.

Method

Participants. A new group of 20 volunteers—12 females and 8 males (mean age: 30.3 ± 9.3 years)—participated in this experiment. They had normal hearing and were naive about the purpose of the study.

Apparatus, Stimuli, and Procedure. The apparatus and the auditory stimuli were identical to those employed in Experiments 1 and 2. A single session comprised one 2AFC block and one reminder block. The order of these two blocks was counterbalanced across participants. The 2AFC task was identical to the nonadaptive version of this task applied in Experiment 2, and the reminder task was identical to the nonadaptive version of this task in Experiment 1. The method for determination of DL in the 2AFC task was analogous to the one of Experiment 2.

Results and Discussion

Contrary to the prediction of the proposed hypothesis, mean DL was again significantly larger for the 2AFC task (47.2 ± 3.7 msec) than for the reminder task (35.4 ± 2.6 msec) [$t(19) = 3.29, p = .002, \eta^2 = .36$]. A one-sided t test for independent samples failed to reveal a significant difference between mean DL for the 2AFC task in this experiment and the corresponding mean DL of Experiment 2 [$t(48) = 1.36, p = .090$]. In addition, mean DL for the reminder task did not significantly differ from the corresponding one of Experiment 2 [$t(48) = 0.51, p = .308$]. In summary, the observed discrepancy between the DL values obtained with the 2AFC and the reminder tasks in Experiment 2 cannot be attributed to the idea that the comparison durations were always longer than the standard duration. The present data suggest rather that the 2AFC task yields consistently larger DLs than the reminder task, irrespective of whether the comparisons are shorter or longer than the standard duration.

EXPERIMENT 4

Another possible reason of the discrepant results might be the different temporal pattern of the two tasks. Note that in the reminder task, the duration of the first stimulus is always the same, whereas in the 2AFC task, the duration of the first stimulus varies from trial to trial. One may therefore assume that this temporal variability of the first interval increases the overall temporal uncertainty when the second stimulus will be presented. It is well documented that an increase of temporal uncertainty impairs perceptual discrimination not only for nontemporal features (see, e.g., Rolke & Hofmann, 2007) but also for temporal features (Grondin & Rammsayer, 2003). According

to this explanation, the constant interstimulus interval (ISI) of 1,000 msec between the first and second interval would facilitate discrimination in the reminder but not in the 2AFC task. Thus, we supposed that a random ISI in the reminder task might impair discrimination performance and consequently yield DLs similar to those of 2AFC task with a constant ISI. In order to evaluate this hypothesis, Experiment 4 combined each of the two tasks with a random and a constant ISI.

Method

Participants. A new group of 24 volunteers—16 females and 8 males (mean age: 21.0 ± 0.4 years)—participated in this experiment. They had normal hearing and were naive about the purpose of this study.

Apparatus, Stimuli, and Procedure. The apparatus and the auditory stimuli were identical to those employed in Experiments 1, 2, and 3. A single session comprised two 2AFC blocks and two reminder blocks. Each block combined one task with either a fixed ISI of 1,000 msec or a random ISI that followed a normal distribution with mean of 500 msec and an SD of 100 msec. In the random condition, we employed a mean ISI of 500 msec rather than of 1,000 msec in order to omit exceedingly long ISIs. The order of these four blocks was counterbalanced across participants. The 2AFC task was identical to the nonadaptive version of this task that was applied in Experiment 1, and the reminder task was identical to the nonadaptive version of this task in Experiments 2 and 3.

Results and Discussion

Figure 6 depicts mean DL. A two-way ANOVA with factors task (reminder vs. 2AFC) and ISI (fixed vs. random) again revealed reliably larger DLs for the 2AFC (57.0 ± 4.1 msec) than for the reminder task (38.3 ± 2.7 msec) [$F(1,23) = 49.1, MS_e = 170.4, p < .001, \eta^2 = .68$]. As one might expect from research on temporal preparation (see, e.g., Rolke & Hofmann, 2007), discrimination performance was worse in the random (51.8 ± 10.6 msec) than in the constant (43.5 ± 8.9 msec) ISI condition [$F(1,23) = 6.9, MS_e = 242.6, p = .015, \eta^2 = .23$]. The significant interaction of both factors indicates that performance on the 2AFC task worsened when the ISI varied from trial to trial rather than when it was kept constant across trials [$F(1,23) = 6.4, MS_e = 121.8, p = .019, \eta^2 = .22$]. In fact, separate t tests indicated that this interaction effect emerged from the 2AFC task [$t(23) = 2.98, p = .007, \eta^2 = .28$] and not from the reminder task [$t(23) = 0.91, p = .37$].

The results of this experiment strengthen the notion that the 2AFC task produces generally enlarged DLs. In contrast to our expectations and somewhat surprisingly, however, the random ISI did not worsen discrimination performance in the reminder task. This finding rejects our conjecture that assumes that the reminder task in Experiments 1 and 2 benefited from temporal certainty. This finding, however, puts forward the idea that the superior discrimination performance in the reminder task stems from the continuous presentation of the standard in the first position.

EXPERIMENT 5

Experiment 5 examined this alternative explanation by manipulating the presentation order of the standard and the comparison in the reminder task. This order could be

Figure 6. Mean DL as a function of task and ISI condition for Experiment 4.

either standard in the first position or in the second position. If the order of these two stimuli matters, then one should observe worse performance under the standard-second condition than under the standard-first condition. As a matter of fact, experimental evidence suggests that the order of these two stimuli has an effect on both temporal (Marchman, 1969; McGavren, 1965; Rammsayer & Wittkowski, 1990; see Ulrich et al., 2006; Van Allen, Benton, & Gordon, 1966) and spatial (Nachmias, 2006) discriminations. Experiment 5 also included the 2AFC task as a control condition.

Method

Participants. A new group of 24 volunteers participated in this experiment (16 females and 8 males; mean age: 21.4 ± 4.5 years). They had normal hearing and were naive about the purpose of this study. One participant was replaced because of an inability to comply with the task; that is, his DL was about three times larger than the mean of the entire sample.

Apparatus, Stimuli, and Procedure. These were identical to those of the previous experiments using the nonadaptive procedure. In the standard-second condition of the reminder task, however, the order of the standard and the comparison was reversed. Each participant performed three blocks. A single block of trials was used to measure DL in each condition (i.e., standard-first reminder, standard-second reminder, and 2AFC). The order of these three conditions was counterbalanced across participants.

Results and Discussion

A one-way ANOVA revealed a significant difference between the three conditions [$F(2,46) = 13.9$, $MS_e = 170.4$, $p < .001$, $\eta^2 = .38$]. Mean DL was 43.3 ± 3.9 msec for the regular condition, 64.4 ± 6.2 msec for the irregular condition, and 68.6 ± 5.6 msec for the 2AFC condition. A Scheffé test ($\alpha = .05$; critical difference = 12.9 msec) revealed a significant DL difference only between the standard-first and the two other conditions. This pattern of results supports the above idea that the order of the standard and the comparison in the reminder task has a strong effect on discrimination performance, and that the standard in the first position facilitates the discrimination process. Converging evidence for this standard position

effect has been reported by Ulrich et al. (2006)—although they did not manipulate standard position within a single experiment—and by Marchman (1969), McGavren (1965), Rammsayer and Wittkowski (1990), and Van Allen et al. (1966), who separated trials according to standard position after running the experiment. Interestingly, this finding has recently also been reported for a spatial discrimination task (Nachmias, 2006). In fact, this effect may suggest an explanation for the performance difference between the regular reminder task and the 2AFC task.

EXPERIMENT 6

All previous experiments revealed discrepant DLs between the 2AFC and the regular reminder tasks. All of them, however, employed auditory stimuli. An important question, therefore, is whether this discrepancy is restricted only to the auditory modality or whether it would generalize across modalities. In order to address this question, we replicated Experiment 5 with visual duration stimuli.

Method

Participants. A new group of 24 volunteers—18 females and 6 males (mean age: 25.7 ± 5.8 years)—participated in this experiment. They had normal vision and were naive about the purpose of the study.

Apparatus, Stimuli, and Procedure. This experiment was identical to Experiment 5, with the exception of the sensory modality. This experiment used the visual modality employing duration discrimination of light provided by a green light emitting diode (LED; diameter 0.48°, viewing distance 60 cm, luminance 48 cd/m²). The LED was attached to the center of the computer screen, and the background of the screen was black. Standard duration was again 500 msec. The comparison durations were 300, 340, 380, 420, 460, 500, 540, 580, 620, 640, and 700 msec for the standard-first and the standard-second conditions of the reminder task, and they were 530, 560, 590, 620, 650, 680, 710, 740, 770, and 800 msec for the 2AFC task. These durations were selected in such a way so that this task was about as equally difficult as it was in the previous experiments. Participants pressed the left “shift” key when the first light appeared longer than the second one, and they pressed the right “shift” key when the second light appeared longer than the first one.

Results and Discussion

A one-way ANOVA again revealed a highly significant effect of condition [$F(2,46) = 11.4$, $MS_e = 114.3$, $p < .0001$, $\eta^2 = .33$]. The mean DL was 77.1 ± 1.6 msec for the standard-first condition, 138.6 ± 3.8 msec for the standard-second condition, and 146.2 ± 3.3 msec for the 2AFC condition. A Scheffé test ($\alpha = .05$, critical difference = 41.2 msec) indicated a significant DL difference only between the standard-first and the two other conditions. These results clearly show that the discrepant DL estimates between the 2AFC and the standard-first reminder task are not restricted to the auditory modality. Furthermore—and consistent with the previous experiment—the position of the standard duration in the reminder task again had a strong effect on DL, being reliably larger when the standard appeared after the comparison.

GENERAL DISCUSSION

The present work compared DL estimates produced by the 2AFC task and by the reminder task. The basic question

was whether both tasks yield DLs of similar magnitude and about the same reliability. Therefore, we combined each of the two tasks with an adaptive and a nonadaptive procedure. A duration discrimination task was employed to address this question empirically.

Main Findings

As was discussed in the introduction, both the reminder task and the 2AFC task are used to measure the discrimination performance of a participant. Each of the two tasks can provide an estimate of the DL. Therefore, in practice, the reminder task and also the 2AFC task have been used to estimate DL. There is no *a priori* theoretical reason why DL estimates derived from the two tasks should systematically differ. In fact, both tasks are usually believed to produce the same results (see, e.g., Thompson et al., 1976; Wright et al., 1997). The results of the main Experiments 1 and 2 and also those of the additional Experiments 3–6, however, revealed a meaningful DL difference between the two tasks (for a summary, see Table 2). More specifically, the DL estimated by the 2AFC task turned out to be almost twice as large as the DL estimated by the reminder task, and this discrepancy was observed for both visual and auditory conditions. This difference cannot be attributed to potential biases in the way DL is estimated, as can be confirmed by computer simulations (see Appendix A). The results of Experiments 5 and 6 strongly indicate that this difference is related to the position of the standard stimulus.

These discrepant DL estimates are of great concern, because they produce a misleading picture when the results of several studies are compared without considering the task for estimating DL (see, e.g., Thompson et al., 1976; Wright et al., 1997). For example, in the domain of duration discrimination, several studies that examined the Weber fraction have employed one or the other task for estimating DL. According to our results—and consistent with other research—estimates of the Weber fraction seem to be larger when employing a 2AFC task (e.g., Grondin, 1993; Rammsayer & Lima, 1991; Wright et al., 1997) rather than a reminder task (e.g., Getty, 1975; Rammsayer & Ulrich, 2005). Therefore, the results of the present study strongly suggest that valid comparisons can only be reached if the task for estimating the DL is taken into account.

A supplementary topic of our study addressed the two standard psychophysical procedures for data collection—that is, the adaptive versus the nonadaptive procedure. The two procedures yielded virtually identical DL results. In addition, the test–retest reliabilities of the two procedures were also similar in magnitude, suggesting that both procedures yield equally stable DL estimates. This finding clearly supports the argument that the two procedures are equally efficient (see, e.g., Dai, 1995; Simpson, 1988). Finally, we also assessed quantification of discrimination performance by means of the Spearman–Kärber method. The present study reinforces previous results from computer simulations (J. Miller & Ulrich, 2001; Ulrich & Miller, 2004) showing high correlations between the estimates provided by the traditional probit analysis and by the Spearman–Kärber method.

A Possible Account: Formation of an Internal Standard

Next in this discussion, we provide a possible explanation of why DL is larger in the 2AFC than in the reminder task and why there is a positional effect of the standard in the reminder task. This explanation rests on the assumption that participants integrate stimulus information across trials in order to form an internal standard instead of employing the stimulus information just of the current trial as most psychophysical theories claim (see, e.g., Green & Swets, 1966; Thurstone, 1927). For example, these theories usually assume that the first stimulus in a trial is stored in memory and that the participant compares the second stimulus of the same trial with the memory trace of the first stimulus. If this assumption were true, however, then it is difficult to see how classical approaches in psychophysics can account for present findings.

The notion of an internal standard receives support from more recent work (Lages & Treisman, 1998; Morgan, Watamaniuk, & McKee, 2000; Nachmias, 2006; Viemeister, 1970) and also from classical psychophysical work (see, e.g., Woodworth & Schlosberg, 1954). These authors pointed out the favorable use of an internal standard—also termed an *implicit* standard in contrast to the explicit standard that is actually presented in a single trial—as a reference for comparison. All of these studies have revealed that discrimination performance usually does not worsen when the standard is omitted in the reminder task.

For example, Viemeister (1970) employed a single-stimulus task. His participants were asked to rate the intensity of an auditory stimulus on a 4-point rating scale. In one condition, an intensity cue preceded the stimulus, and in another condition, there was no cue. Interestingly, the discrimination performance did not differ between the two conditions. On the basis of this finding, Viemeister concluded that participants ignore the cue but compare the intensity of the stimulus against a stored reference. Furthermore, in a hyperacuity study, Morgan et al. (2000) employed the reminder and the single-stimulus task to assess performance using line separation discrimination. Discrimination performance was virtually identical in both tasks. In line with

Table 2
Observed Mean DL ($\pm$ Standard Error of the Mean) in
Milliseconds for Each Experiment

	Reminder/ Adaptive	Reminder/ Nonadaptive	2AFC/ Adaptive	2AFC/ Nonadaptive
Experiment 1	31.1 $\pm$ 1.9	33.2 $\pm$ 1.7		
Experiment 2	32.9 $\pm$ 1.5		59.3 $\pm$ 3.8	56.5 $\pm$ 3.5
Replication ^a				44.6 $\pm$ 3.5
Experiment 3 ^b		35.4 $\pm$ 2.6		47.2 $\pm$ 3.7
Experiment 4 ^c		38.2 $\pm$ 2.7		57.0 $\pm$ 4.1
Experiment 5 ^d		43.3 $\pm$ 3.9 ^f		68.6 $\pm$ 5.6
		64.4 $\pm$ 6.2 ^g		
Experiment 6 ^e		77 $\pm$ 1.6 ^f		146 $\pm$ 3.25
		143 $\pm$ 3.8 ^g		

^aThis replication is reported in note 6. ^bIn this 2AFC task, the comparison was always shorter than the standard. ^cExperiment 4 employed random ISIs. ^dExperiment 5 manipulated the position of the standard. ^eExperiment 6 used visual instead of auditory duration stimuli. ^fIn this condition, the standard appeared in the first position. ^gIn this condition, the standard appeared in the second position.

Viemeister, this result led these authors to conclude that participants employ an internal representation of the standard to judge the size of the comparison. They also conducted a simulation that indicates that such an internal representation is built up by sampling stimulus information over as many as 20 trials. In fact, they suggested that an internal standard is more stable—that is, has smaller variance than the sensory representation of the explicit standard—and this might actually explain why participants use an internal standard rather than rely on an explicit standard. Finally, Lages and Treisman (1998) also suggested that participants establish an internal standard, although they focused their analysis on the PSE rather than on the DL.

More recently, Nachmias (2006) assessed discrimination of simple visual patterns with the reminder task and with the single-stimulus task, which omits a preceding standard. He reported almost identical discrimination performance for the two tasks. In addition, he also found that the position of the standard matters. As in our study, when the standard was presented in the first position, performance was better than when the standard occurred in the second position. A similar positional effect has been observed in some previous timing studies (Marchman, 1969; McGavren, 1965; Rammsayer & Wittkowski, 1990; Ulrich et al., 2006; Van Allen et al., 1966). In these studies, participants also tended to give more correct responses for pairs of a fixed standard and a variable comparison interval when the presentation order was standard–comparison than when it was comparison–standard.

Incorporating a recent theoretical idea of Nachmias (2006), we suggest a simple account of how such an internal standard—that is, I —might be formed. Accordingly, I combines information A from previous trials with the internal representation X_1 of the first stimulus in the present trial. In addition, we proceed from the idea that X_1 is noisy; that is, it may be randomly smaller or larger than the corresponding physical duration of the first stimulus. According to this account, participants keep a moving average of the internal representations that precede the current trial; this moving average is denoted by A . For example, I might be a weighted combination of A and X_1 ; that is, $I = g \cdot A + (1 - g) \cdot X_1$ with $0 < g < 1$. Participants judge the second stimulus as being longer than the first one, if the internal representation X_2 —which is also noisy—of the second interval is larger than the updated value of I .

In order to evaluate the prediction of Nachmias's (2006) model, we implemented it in a computer simulation. In agreement with the data of Experiments 5 and 6, the model predicts a larger DL when the comparison appeared in the first rather than in the second stimulus position (Appendix B provides a more formal analysis of this prediction). Most importantly, this mechanism can also be applied to the 2AFC task. In agreement with the results of the present experiments, our simulations clearly indicate a larger DL for the 2AFC than for the regular reminder task.⁸ In addition, our simulations reveal that this prediction holds irrespective of whether the standard in the 2AFC task is always smaller or always larger than the comparison. Thus, this model provides a promising research perspective for

future modeling of discrimination behavior, because it suggests a plausible mechanism for how the internal standard I may evolve.

Conclusion

In summary, then, both the 2AFC and the reminder task are widely in use in psychophysics. It is a common belief that estimates produced by these two tasks are equivalent. The present results, however, do not support this notion—at least not for duration discrimination tasks. DL estimates obtained with the 2AFC task were reliably larger than those obtained from the reminder task. Whether this discrepancy only applies to duration discrimination needs to be addressed in future work. The finding that a standard position effect can also occur in spatial tasks (Nachmias, 2006) suggests that the present results also apply to other nontemporal tasks. In fact, the proposed models in this article do not differentiate between temporal and nontemporal information processing.

In this article, we provided a novel model that is based on ideas put forward by Nachmias (2006) and Morgan et al. (2000). This model incorporates the notion of a stable internal criterion that includes sensory information not only from the current trial, but also integrates information across preceding trials. The underlying mechanism implies that discrimination performance is sensitive to the position of the standard. Specifically, discrimination performance worsens whenever the standard stimulus is presented in the second stimulus position within a given trial. The model might well account for other findings reported in the psychophysical literature—for example, the finding that discrimination performance increases when the standard is repeatedly presented before a comparison is required (see, e.g., Drake & Botte, 1993; Ivry & Hazeltine, 1995; Schulze, 1989). Elaborations of the model might possibly also account for the important finding that discrimination performance is strongly affected by the number of comparison intervals (N. S. Miller & McAuley, 2005). The model might also explain the finding that performance is better when the standard duration is fixed rather than varied across trials (i.e., roving standard). It may also account for the worsened performance when the range of standard durations is increased (McAuley & Jones, 2003; N. S. Miller & McAuley, 2005). Future research is necessary to evaluate the prediction of this novel model in these related domains of temporal information processing.

AUTHOR NOTE

We thank Tanja Seifried for her help in programming the experiments and Karin Bausenhardt for helpful discussions. We appreciate the constructive comments by Neil A. Macmillan, Devin McAuley, and an anonymous reviewer. This work was supported by Deutsche Forschungsgemeinschaft UL 116/8-3. Correspondence concerning this article should be sent to E. Lapid, Cognitive and Biological Psychology, University of Tübingen, Friedrichstr. 21, 72072 Tübingen, Germany (e-mail: einat.lapid@uni-tuebingen.de).

REFERENCES

- BODE, D., & CARHART, R. (1973). Measurements of articulation functions using adaptive test procedures. *IEEE Transactions on Audio & Electroacoustics*, **AU-21**, 196-201.

- BRAND, T., & HOHMANN, V. (2002). An adaptive procedure for categorical loudness scaling. *Journal of the Acoustical Society of America*, **112**, 1597-1604.
- BUSH, R. R. (1967). Estimation and evaluation. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (2nd ed., Vol. 1, pp. 429-469). New York: Wiley.
- DAI, H. (1995). On measuring psychometric functions: A comparison of the constant-stimulus and adaptive up-down methods. *Journal of the Acoustical Society of America*, **98**, 3135-3139.
- DRAKE, C., & BOTTE, M.-C. (1993). Tempo sensitivity in auditory sequences: Evidence for a multiple-look model. *Perception & Psychophysics*, **54**, 277-286.
- EVANS, M., HASTINGS, N., & PEACOCK, B. (2000). *Statistical distributions* (3rd ed.). New York: Wiley.
- FINNEY, D. J. (1952). *Probit analysis* (2nd ed.). Cambridge: Cambridge University Press.
- GARCÍA-PÉREZ, M. A., & ALCALÁ-QUINTANA, R. (2005). Sampling plans for fitting the psychometric function. *Spanish Journal of Psychology*, **8**, 256-289.
- GESCHIEDER, G. A. (1997). *Psychophysics: The fundamentals* (3rd ed.). Mahwah, NJ: Erlbaum.
- GETTY, D. J. (1975). Discrimination of short temporal intervals: A comparison of two models. *Perception & Psychophysics*, **18**, 1-8.
- GREEN, D. M., & SWETS, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.
- GRONDIN, S. (1993). Duration discrimination of empty and filled intervals marked by auditory and visual signals. *Perception & Psychophysics*, **54**, 383-394.
- GRONDIN, S. (2001). Discriminating time intervals presented in sequences marked by visual signals. *Perception & Psychophysics*, **63**, 1214-1228.
- GRONDIN, S., IVRY, R. B., FRANZ, E., PERREAULT, L., & METTHÉ, L. (1996). Markers' influence on the duration discrimination of intermodal intervals. *Perception & Psychophysics*, **58**, 424-433.
- GRONDIN, S., OUELLET, B., & ROUSSEL, M. E. (2001). About optimal timing and stability of Weber fraction for duration discrimination. *Acoustical Science & Technology*, **22**, 370-372.
- GRONDIN, S., & RAMMSAYER, T. (2003). Variable foreperiods and temporal discrimination. *Quarterly Journal of Experimental Psychology*, **56A**, 731-765.
- GUILFORD, J. P. (1954). *Psychometric methods*. New York: McGraw-Hill.
- HALL, J. L. (1981). Hybrid adaptive procedure for estimation of psychometric functions. *Journal of the Acoustical Society of America*, **69**, 1763-1769.
- HESSE, A. (1986). Comparison of several psychophysical procedures with respect to threshold estimates, reproducibility and efficiency. *Acustica*, **59**, 263-273.
- IVRY, R., & HAZELTINE, R. E. (1995). The perception and production of temporal intervals across a range of durations: Evidence for a common timing mechanism. *Journal of Experimental Psychology: Human Perception & Performance*, **21**, 1-12.
- KAERNBACH, C. (1991). Simple adaptive testing with the weighted up-down method. *Perception & Psychophysics*, **49**, 227-229.
- KARMAKAR, U. R., & BUONOMANO, D. V. (2003). Temporal specificity of perceptual learning in an auditory discrimination task. *Learning & Memory*, **10**, 141-147.
- KLEIN, S. A., & MACMILLAN, N. A. (Eds.) (2001). Psychometric functions and adaptive methods [Special issue]. *Perception & Psychophysics*, **63**(8).
- LAGES, M., & TREISMAN, M. (1998). Spatial frequency discrimination: Visual long-term memory or criterion setting? *Vision Research*, **38**, 557-572.
- LEEK, M. R. (2001). Adaptive procedures in psychophysical research. *Perception & Psychophysics*, **63**, 1279-1292.
- LEEK, M. R., HANNA, T. E., & MARSHALL, L. (1992). Estimation of psychometric functions from adaptive tracking procedures. *Perception & Psychophysics*, **51**, 247-256.
- LEVITT, H. (1971). Transformed up-down methods in psychoacoustics. *Journal of the Acoustical Society of America*, **49**, 467-477.
- LINSCHOTEN, M. R., HARVEY, L. O., JR., ELLER, P. M., & JAFEK, B. W. (2001). Fast and accurate measurement of taste and smell thresholds using a maximum-likelihood adaptive staircase procedure. *Perception & Psychophysics*, **63**, 1330-1347.
- LUCE, R. D., & GALANTER, E. (1967). Discrimination. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (2nd ed., Vol. 1, pp. 191-244). New York: Wiley.
- MACMILLAN, N. A., & CREELMAN, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, NJ: Erlbaum.
- MARCHMAN, J. N. (1969). Discrimination of brief temporal durations. *Psychological Record*, **19**, 83-92.
- MCAULEY, J. D., & JONES, M. R. (2003). Modeling effects of rhythmic context on perceived duration: A comparison of interval and entrainment approaches to short-interval timing. *Journal of Experimental Psychology: Human Perception & Performance*, **29**, 1102-1125.
- MCGAVREN, M. (1965). Memory of brief auditory durations in comparison discriminations. *Psychological Record*, **15**, 249-260.
- McKEE, S. P., KLEIN, S. A., & TELLER, D. Y. (1985). Statistical properties of forced-choice psychometric functions: Implications of probit analysis. *Perception & Psychophysics*, **37**, 286-298.
- MILLER, J., & ULRICH, R. (2001). On the analysis of psychometric functions: The Spearman-Kärber method. *Perception & Psychophysics*, **63**, 1399-1420.
- MILLER, J., & ULRICH, R. (2004). A computer program for Spearman-Kärber and probit analysis of psychometric function data. *Behavior Research Methods, Instruments, & Computers*, **36**, 11-16.
- MILLER, N. S., & MCAULEY, J. D. (2005). Tempo sensitivity in isochronous tone sequences: The multiple-look model revisited. *Perception & Psychophysics*, **67**, 1150-1160.
- MORGAN, M. J., WATAMANIUK, S. N., & McKEE, S. P. (2000). The use of an implicit standard for measuring discrimination thresholds. *Vision Research*, **40**, 2341-2349.
- NACHMIAS, J. (2006). The role of virtual standards in visual discrimination. *Vision Research*, **46**, 2456-2464.
- 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.
- RAMMSAYER, T. H., & LIMA, S. D. (1991). Duration discrimination of filled and empty auditory intervals: Cognitive and perceptual factors. *Perception & Psychophysics*, **50**, 565-574.
- RAMMSAYER, T. [H.], & ULRICH, R. (2005). No evidence for qualitative differences in the processing of short and long temporal intervals. *Acta Psychologica*, **120**, 141-171.
- RAMMSAYER, T. H., & WITTKOWSKI, K. M. (1990). Zeitfehler und Positionseffekt des Standardreizes bei der Diskrimination kurzer Zeitdauern [Time-order error and position effect of the standard stimulus in the discrimination of short durations]. *Archiv für Psychologie*, **142**, 81-89.
- ROLKE, B., & HOFMANN, P. (2007). Temporal uncertainty degrades perceptual processing. *Psychonomic Bulletin & Review*, **14**, 522-526.
- SCHULZE, H. H. (1989). The perception of temporal deviations in isochronic patterns. *Perception & Psychophysics*, **45**, 291-296.
- SIMPSON, W. A. (1988). The method of constant stimuli is efficient. *Perception & Psychophysics*, **44**, 433-436.
- THOMPSON, J. G., SCHIFFMAN, H. R., & BOBKO, D. J. (1976). The discrimination of brief temporal intervals. *Acta Psychologica*, **40**, 489-493.
- THURSTONE, L. L. (1927). A law of comparative judgment. *Psychological Review*, **34**, 273-286.
- TREUTWEIN, B. (1995). Adaptive psychophysical procedures. *Vision Research*, **35**, 2503-2522.
- TREVAN, J. W. (1927). The error of determination of toxicity. *Proceedings of the Royal Society B*, **101**, 483-514.
- TSE, P. U., INTRILIGATOR, J., RIVEST, J., & CAVANAGH, P. (2004). Attention and the subjective expansion of time. *Perception & Psychophysics*, **66**, 1171-1189.
- ULRICH, R., & MILLER, J. (2004). Threshold estimation in two-alternative forced-choice (2AFC) tasks: The Spearman-Kärber method. *Perception & Psychophysics*, **66**, 517-533.
- ULRICH, R., NITSCHKE, J., & RAMMSAYER, T. (2006). Crossmodal temporal discrimination: Assessing the predictions of a general pacemaker-counter model. *Perception & Psychophysics*, **68**, 1140-1152.
- VAN ALLEN, M. W., BENTON, A. L., & GORDON, M. C. (1966). Temporal discrimination in brain-damaged patients. *Neuropsychologia*, **4**, 159-167.

- VIEMEISTER, N. F. (1970). Intensity discrimination: Performance in three paradigms. *Perception & Psychophysics*, **8**, 417-419.
- WATSON, B. A., & FITZHUGH, A. (1990). The method of constant stimuli is inefficient. *Perception & Psychophysics*, **47**, 87-91.
- WICKENS, T. D. (2002). *Elementary signal detection theory*. Oxford: Oxford University Press.
- WOODWORTH, R. S., & SCHLOSBERG, H. (1954). *Experimental psychology* (3rd ed.). London: Methuen.
- WRIGHT, B. A., BUONOMANO, D. V., MAHNCKE, H. W., & MERZENICH, M. M. (1997). Learning and generalization of auditory temporal-interval discrimination in humans. *Journal of Neuroscience*, **17**, 3956-3963.

NOTES

1. The reminder task has been traditionally called the *method of constant stimuli*. The term may create some confusion. Note that it originates from the fact that an experimenter determines the set of comparison stimuli before running an experiment and uses this set throughout the experiment (Woodworth & Schlosberg, 1954, p. 200). Unfortunately, this definition also applies to the 2AFC method when the set of stimuli is predetermined. For this reason, we do not employ the term *constant stimuli* in the present article. Instead, we will use the term *nonadaptive* when the set of stimuli is predetermined and the term *adaptive* when the stimulus set is determined by the response history of a participant during an experiment. These latter terms, however, will be introduced later.

2. Theory of signal detection (SDT) might lead one to expect better discrimination performance in the 2AFC task than in the reminder task. We must stress, however, that the reminder task as well as the 2AFC task that are employed in the domain of SDT research differ from the ones used in this article. In the SDT domain, the stimulus sequences in each trial are either $\langle S, C \rangle$ or $\langle C, S \rangle$ in the 2AFC task, and either $\langle S, S \rangle$ or $\langle S, C \rangle$ in the reminder task, where S denotes the standard duration (e.g., 500 msec) and C the duration of the comparison (e.g., 600 msec). In this case, SDT predicts better performance (i.e., larger d') in the 2AFC task than in this type of reminder task employed in SDT studies (see Macmillan & Creelman, 2005). The reason for the higher task performance in the 2AFC than in the reminder task is the difference in stimulus magnitudes between the first and second stimulus. In the 2AFC task, this difference is either $S - C$ (e.g., -100 msec) in $\langle S, C \rangle$ trials and $C - S$ (e.g., 100 msec) in $\langle C, S \rangle$ trials. In the reminder task, however, this difference is either $S - S$ (0 msec) in $\langle S, S \rangle$ trials and $S - C$ (e.g., -100 msec) in $\langle S, C \rangle$ trials. Because the size of the two differences in the 2AFC task is twice as large as that in the reminder task, SDT predicts better discrimination performance in the 2AFC task. In the present experiments, however, we did not use this type of reminder task

that originated in the SDT domain. In our case, not only one comparison C but several comparison stimuli are used to assess DL. In this kind of task, it is more appropriate to employ the classical psychophysical approach. Indeed, as shown in the General Discussion, this approach predicts identical DLs.

3. There are some studies, however, that employed a single, fixed stimulus level within a block of trials, which can be regarded as a special case of the method of constant stimuli. These studies indicate lower discrimination performance with fixed level than with adaptive procedures (for a review, see Leek, 2001).

4. Note that the DL of a logistic psychometric function is related to the SD of this function as the following: $SD = 1.65 \cdot DL$. Therefore, the overall mean DL of 32.2 msec is associated with an overall SD of 53.1 msec.

5. For comparison, the coefficient of skewness of a 30-step gamma distribution is equal to 0.365, although this distribution nearly resembles a normal distribution according to the central limit theorem. It is therefore justified to use a symmetrical psychometric function (such as the logistic distribution) for probit analysis for the present data. In fact—and not surprisingly—the goodness of fit for the logistic psychometric function was satisfactory in all cases.

6. We reran the present experiment with another sample. In order to simplify the design, we omitted the reminder task, therefore making the design of this replication similar to that of Experiment 1. As before, there was no main effect of procedure [$F(1,23) = 0.5$, $MS_e = 183.8$, $p = .467$, $\eta^2 = .01$], no main effect of block [$F(1,23) = 0.2$, $MS_e = 252.4$, $p = .630$, $\eta^2 = .02$], and the interaction of both factors was again not significant [$F(1,23) = 1.0$, $MS_e = 334.5$, $p = .320$, $\eta^2 = .04$]. Mean DL was 43.6 ± 5.0 msec and 45.6 ± 3.8 msec for the nonadaptive and adaptive procedure, respectively. Hence, even when omitting the reminder task from the design, the pattern of results remains the same.

7. We thank Jeff Miller for proposing this hypothesis.

8. These simulations assumed that the internal representations of the standard and the comparison (i.e., X_S and X_C) are noisy—that is, normally distributed with a mean equal to the physical duration of the stimuli and an SD of 50 msec. Furthermore, the standard and comparison durations were identical to those used in the experiments. The moving average A was computed across 20 trials that preceded the current trial. We varied the weight g in separate simulations from 0 to 1. For example, for the reminder task with $g = 0.3$, DL was around 40 msec and 58 msec when the standard was at the first and the second position, respectively. Increasing g enhanced predicted performance when the standard was in the first position, but worsened performance when the standard was in the second position. The DL for the 2AFC task was approximately 54 msec, irrespective of whether the standard duration was always smaller or larger than the comparison durations.

APPENDIX A

Monte Carlo Simulation

The purpose of the following simulations was to rule out the possibility that the discrepancy of the DL estimates between the 2AFC and the reminder tasks is due to the estimation procedure applied. Four programs—each matching exactly a single combination of task (2AFC vs. reminder) and procedure (adaptive vs. nonadaptive)—were written to simulate the responses of a virtual participant according to psychometric functions (1) and (2). These data were used to estimate the DL with the routines that were also used to estimate the DL for the real participants in the present experiments. All other aspects (e.g., stimulus levels, number of trials) were exactly matched to the real experiments. The parameters for the underlying psychometric function were $DL = 50$ msec for both tasks, and $PSE = 500$ msec for the reminder task. Each program simulated 10,000 virtual participants. Table A1 summarizes the results of these simulations. Specifically, it shows the overall mean DL and the standard deviation of the estimates. The last column shows the percentage of cases in which the estimation program did not converge. Generally, the procedures did satisfactorily recover the expected value of $DL = 50$ msec. The results of the simulations do not indicate that the estimation procedure is a possible source of the discrepant DL from the 2AFC and the reminder tasks. The second parameter that was estimated from the simulation results was the PSE. For the adaptive as well as for the nonadaptive procedure, the targeted value of $PSE = 500$ msec was obtained; thus, the simulation results suggest no bias in estimating the PSE (Table A2).

Table A1
DL Simulation Results: Mean Estimated DL and
the Standard Deviation of the DL Estimates
As a Function of Task and Procedure

Task/Procedure	DL (msec)	SD (msec)	Percentage of Nonconvergence
Reminder/adaptive	43.8	8.7	0
Reminder/nonadaptive	50.5	10.5	0
2AFC/adaptive	50.1	9.6	0.02
2AFC/nonadaptive	48.9	13.8	0.02

Table A2
PSE Simulation Results for the Reminder Paradigm: Mean
Estimated PSE and the Standard Deviation of the PSE
Estimates As a Function of Procedure

Task/Procedure	PSE (msec)	SD (msec)
Reminder/adaptive	499.95	12.76
Reminder/nonadaptive	500.06	10.63

APPENDIX B

Moving Average Model and the Positional Effect of the Standard

This appendix proves that the moving average model predicts a smaller DL when the standard is in the first rather than in the second stimulus position. In order to simplify things, we will focus on the reminder task in this appendix and only address the standard position effect (a formal analysis of the 2AFC task is complex; thus, it is beyond the purpose of this article). Note that the second interval is perceived as being longer than the first one if the event $X_2 > I$ occurs—that is, when the internal representation X_2 of the second stimulus is larger than the internal standard I . Thus, the probability of the response $S_2 > S_1$ (i.e., the second interval appears longer than the first interval) is given by

$$\begin{aligned}
 P\{S_2 > S_1' | t_C\} &= P\{X_2 > I | t_C\} \\
 &= 1 - P\{X_2 - I \leq 0 | t_C\} \\
 &= 1 - P\{X_2 - g \cdot A - (1 - g) \cdot X_1 \leq 0 | t_C\} \\
 &= 1 - \Phi \left[\frac{0 - E[X_2 - g \cdot A - (1 - g) \cdot X_1]}{\sigma_*} \right] \\
 &= \Phi \left[\frac{E[X_2 - g \cdot A - (1 - g) \cdot X_1]}{\sigma_*} \right].
 \end{aligned} \tag{B1}$$

APPENDIX B (Continued)

Note that $E[X_2 - g \cdot A - (1 - g) \cdot X_1]$ and σ_* denote the mean and the standard deviation of the random variable $D = X_2 - g \cdot A - (1 - g) \cdot X_1$, respectively. The variable D represents the internal difference between the second stimulus and the internal standard. One can assume that the standard deviation σ_* does not depend on the order of the standard and the comparison; however, the mean of D does depend on this order.

Specifically, if the standard is in the first position, the expected value of D is

$$\begin{aligned} E[D | \langle S, C \rangle] &= E[X_2 - g \cdot A - (1 - g) \cdot X_1] \\ &= E[X_2] - g \cdot E[A] - (1 - g) \cdot E[X_1] \\ &= t_C - g \cdot t_S - (1 - g) \cdot t_S \\ &= t_C - t_S. \end{aligned} \quad (\text{B2})$$

Inserting (B2) into (B1) yields the psychometric function for the case that the standard occurs always in the first position

$$P\{S_2 > S_1' | \langle S, C \rangle, t_C\} = \Phi \left[\frac{t_C - t_S}{\sigma_*} \right]. \quad (\text{B3})$$

The DL associated with (B3) is equal to $0.67 \cdot \sigma_*$.

A similar reasoning shows that when the standard occurs in the second position, the corresponding psychometric function is given by

$$\begin{aligned} P\{S_2 > S_1' | \langle C, S \rangle, t_C\} &= \Phi \left[\frac{(1 - g) \cdot (t_S - t_C)}{\sigma_*} \right] \\ &= \Phi \left[\frac{\frac{t_S - t_C}{\sigma_*}}{1 - g} \right]. \end{aligned} \quad (\text{B4})$$

Therefore, the DL associated with (B4) must be equal to $0.67 \cdot \sigma_*/(1 - g)$. Note that this DL must be larger than the DL associated with (B3), which completes the proof. This prediction follows from the fact that mean D is smaller when the standard is presented in the second rather than in the first stimulus position. This reduction in mean D diminishes the perceptible difference between the standard and the comparison and, as a result, lowers discrimination performance.