Attention, Perception, & Psychophysics
2009, 71 (6), 1219-1227
doi:10.3758/APP.71.6.1219

BRIEF REPORTS

Estimating the difference limen in 2AFC tasks: Pitfalls and improved estimators

ROLF ULRICH
University of Tübingen, Tübingen, Germany

AND

DIRK VORBERG
University of Münster, Münster, Germany

Discrimination performance is often assessed by measuring the difference limen (DL; or just noticeable difference) in a two-alternative forced choice (2AFC) task. Here, we show that the DL estimated from 2AFC percentage-correct data is likely to systematically under- or overestimate true discrimination performance if order effects are present. We show how pitfalls with the 2AFC task may be avoided and suggest a novel approach for analyzing 2AFC data.

The difference limen (DL; also, difference threshold or just noticeable difference) is a fundamental concept in psychophysics for assessing discrimination performance. The DL provides information on how well small stimulus differences can be detected. A prominent psychophysical task for measuring the DL is the two-alternative forced choice (2AFC) task. In a typical 2AFC task, two stimuli—a constant standard s and a variable comparison c—are presented successively on each trial, and the participant is asked to locate the temporal position (i.e., first vs. second) of the more intense (larger, louder, longer, etc.) stimulus. The temporal order of s and c varies randomly from trial to trial in 2AFC tasks.

For generating a psychometric function, the comparison c is varied while the standard s is held constant. How well the stimuli can be discriminated from each other is reflected by the steepness of the psychometric function. Thus, DL may be defined as half the difference between the stimulus values at which the comparison stimulus is judged to be larger than the standard stimulus on $100 \cdot p$ or $100 \cdot (1 - p)$ percent of the trials—that is, $\mathrm{DL} = (c_p - c_{1-p})/2$ (e.g., Luce & Galanter, 1967). Conventionally, $p = .75$ is chosen, but other percentage levels (e.g., 70% or 90%) are also used in defining DL. In 2AFC tasks, experimenters often choose only comparison stimuli c with values at least as large as s. In this case, an alternative DL definition is used—namely, $\mathrm{DL} = (c_p - c_{.5})$. For psychometric functions that are point symmetric around $c_{.5}$, these definitions are equivalent, because point symmetry implies $(c_p - c_{.5}) = (c_{.5} - c_{1-p})$.

For convenience, 2AFC responses are often scored simply as correct or incorrect, regardless of the order in which the stimulus pair was presented. Because participants will achieve 50% correct responses by sheer guessing, the convention has developed to define the DL in terms of the stimulus difference $c - s$ at which the 2AFC psychometric function equals .75—that is, halfway between chance and perfect performance. In this article, we show that this practice is valid only if certain conditions hold but is bound to lead to distorted DL estimates otherwise. Our purpose is to alert researchers to the estimation problems that arise when discrimination performance depends not on the physical difference between stimuli alone, but also on their presentation order. Note that the methodological problems we point out in the following do not invalidate the 2AFC task as such. However, they do argue against the practice of disregarding stimulus order in scoring 2AFC data.

As a case in point, consider a recent article by one of us (Lapid, Ulrich, & Rammsayer, 2008), which compared duration discrimination DLs in two different psychophysical tasks and showed superior discriminability in a 2AFC task, as compared with a task in which the standard stimulus always appeared first (*reminder task*). The authors suggested a moving average process (Morgan, Watamaniuk, & McKee, 2000; Nachmias, 2006) as an explanation of this apparent difference in discrimination sensitivity. Here, we will propose a more plausible account that attributes the divergent DL estimates to uncontrolled order effects in the 2AFC task. The purpose of our article is (1) to demonstrate why, in the 2AFC task, order effects are likely to produce misleading

R. Ulrich, ulrich@uni-tuebingen.de

estimates of DL and (2) to estimate DL—uncontaminated by order effects—from 2AFC data.

Two Classes of Order Effects, and the Objectives of This Article

In 2AFC discrimination experiments, the participant is to arrive at some comparison judgment between two stimuli. How well the comparison stimulus c can be discriminated from the standard stimulus s depends on the physical difference $c - s$; however, the temporal order of the stimuli on a trial may act as a nuisance variable that also influences the comparison judgment. In fact, Fechner (1860, pp. 87–92) already reported and discussed order effects in judgments of sequential stimuli, and such presentation order errors, more commonly known as time order errors, have been the subject of much theoretical and empirical research in classical psychophysics (see, e.g., Eisler, Eisler, & Hellström, 2008; Guilford, 1954; Hellström, 1985; Woodrow, 1935). Recently, Yeshurun, Carrasco, and Maloney (2008) have taken up this discussion; in several studies, they demonstrated strong and reliable time order effects, which pose severe problems both for traditional psychophysical measurement and for approaches based on signal detection theory (SDT).

Yeshurun et al. (2008) distinguished between two types of order effects. First, participants may prefer one response alternative to the other for some reason. For example, one of the response buttons might be easier to operate, and participants may prefer the alternative associated with the easier buttonpress. Such a preference would create a bias, leading participants to select one interval more often than the other one. In SDT, this kind of order effect is often attributed to a nonzero decision criterion (e.g., Wickens, 2002, p. 100), a notion that is related to the constant error ($CE = c_{.5} - s$) in classical psychophysics—that is, a consistent shift of the psychometric function (Wickens, 2002, p. 99). We will call this a *Type A* order effect.

Second, order effects might (also) reflect genuine sensory or perceptual interactions between successive stimuli, rather than a bias effect, a possibility much discussed in classical psychophysics; we will refer to this as the *Type B* order effect. This effect has been alluded to as a "darker possibility" by Yeshurun et al. (2008, p. 1842), who, in a reanalysis of several data sets, showed that most participants within each study displayed the same interval preference, which renders the hypothesis unlikely that these interval biases reflected just idiosyncrasies. Instead, the authors considered the possibility that sensitivity truly differs in the intervals of a 2AFC task. In classical psychophysics, this implies psychometric functions that differ in slope (and thus in DL), depending on whether the constant standard is presented first or second. In SDT, this corresponds to differential sensitivity (in terms of d') in the two temporal intervals, for which Yeshurun et al., in fact, found evidence in a modified 2AFC detection task. Similar support for the notion of a true sensitivity difference across intervals was reported by Nachmias (2006), who observed smaller DLs when c was in the second interval than when it was in the first interval. Analogous results have been reported by Hairston and Nagarajan (2007), Lapid et al. (2008), and Ulrich, Nitschke, and Rammsayer (2006).

SDT analysis is usually recommended for estimating d' from 2AFC data when response bias effects exist (i.e., a Type A order effect is present). For simplicity, applications of SDT to 2AFC data often assume unbiased responding in the standard difference model, but it is well known that nonzero response bias reduces the percentage of correct responses, P_C, in the 2AFC task. Consequently, true sensitivity is *underestimated* if assessed via $d' = \sqrt{2} \cdot z(P_C)$, where $z(P_C)$ is the z value that corresponds to P_C (Green & Swets, 1973; Klein, 2001; Macmillan & Creelman, 2005; Wickens, 2002).

Treated properly, however, Type A order effects pose no serious problems to SDT if 2AFC performance is analyzed in terms of hits and false alarms, rather than percentage correct. This can be done by defining correct and incorrect *second larger* responses as hits and false alarms, respectively, which allows separating sensitivity and response bias (Green & Swets, 1973, pp. 408–411; Klein, 2001, pp. 1424–1427; Wickens, 2002, pp. 98–103). Note, however, that this corrects for Type A order effects only, and there currently exists no satisfactory model within SDT for dealing with Type B order effects (see Yeshurun et al., 2008). Moreover, it is not known whether and how Type A and Type B order effects can be distinguished from each other empirically, even when full receiver-operating characteristic (ROC) curves are traced by systematically varying the participant's response criterion.

This ROC approach of SDT is in contrast to the traditional psychophysical approach, which describes performance in terms of psychometric functions that are generated by varying the size of the comparison, c. If separate psychometric functions are obtained for the two presentation orders, it is possible to distinguish Type A from Type B errors by their slope and intercept differences (Lapid et al., 2008; Nachmias, 2006). In this article, we will present methods for doing so and will show how to obtain measures of discrimination sensitivity that are not contaminated by time order effects.

The present article has two major objectives. In the first part, we will demonstrate that the traditional 2AFC approach of estimating DL may run into fundamental problems when data are collapsed across presentation orders into a psychometric function that gives just the percentage of correct responses as a function of the comparison size. More specifically, we will show that when constant errors exist, this type of data analysis is bound to produce severely distorted DL estimates. Under most conditions, such DL estimates from the 2AFC task will be larger than the ones obtained in single-stimulus tasks, which means that true discriminability will be underestimated. To understand what causes this problem in the 2AFC task, an analysis of the psychometric functions conditional on the temporal order of the standard and the comparison stimuli is needed. The analysis of these functions is the major theoretical goal in this article.

We will follow Nachmias's (2006) suggestion of tracing separate psychometric functions for each presentation order of s and c and will introduce a maximum likelihood

procedure for estimating both CE and DL. In contrast to traditional approaches, the estimates of DL from this procedure are not distorted by constant errors. Moreover, our procedure can also be used to assess discrimination performance when order effects cannot be attributed to response bias alone but reflect a true sensitivity difference in the two intervals (Yeshurun et al., 2008).

Psychometric Functions in the 2AFC Task

Let s and c be the standard stimulus and comparison stimulus, respectively; to simplify the notation, s and c will also refer to the magnitudes of the standard and comparison. The presentation order of the two stimuli varies randomly across trials; that is, their order is either $\langle sc \rangle$ or $\langle cs \rangle$. Usually, 2AFC experiments keep s fixed and employ comparisons that are larger than or equal to s—that is, $c \geq s$. On each trial of such an experiment, the participant indicates the temporal position of the stimulus perceived as larger. Rather than scoring responses as correct or incorrect (e.g., Gescheider, 1997, pp. 146–149) and summarizing discrimination performance by a percentage-correct psychometric function, $G(c) \equiv P(\text{Correct} \,|\, s, c)$, we start by constructing separate psychometric functions from the data, one function for each order.

Let S_1 and S_2 denote the stimulus in the first or second position, respectively. Define $F_1(c) \equiv P(\text{“}S_1 > S_2\text{”} \,|\, \langle cs \rangle)$ and $F_2(c) \equiv P(\text{“}S_2 > S_1\text{”} \,|\, \langle sc \rangle)$ as the conditional probability with which the participant judges the comparison c as the larger of the two stimuli when it was presented first or second, respectively. Note that, by definition, these functions give the probabilities of correct responses for $c > s$ but of incorrect responses for $c < s$. It is helpful to note that, for $c > s$, $F_1(c)$ and $1 - F_2(c)$ correspond to the definitions of hit and false alarm rates in SDT, respectively.

We assume that both $F_1(c)$ and $F_2(c)$ increase monotonically with c, for a fixed standard s. For the moment, no parametric assumptions about their shape are required. Rather, we describe the locations and slopes of $F_1(c)$ and $F_2(c)$ by their points of subjective equality, $\text{PSE} = c_{.5}$, and their DLs (defined by $\text{DL} = c_{.75} - c_{.5}$ when c is never less than s). When needed, we index the statistics by the temporal position i ($i = 1, 2$) of c, to indicate which psychometric function they refer to. Type A and Type B order effects on judgments can now be expressed by location and slope differences, respectively, of $F_1(c)$ versus $F_2(c)$. The upper and lower panels in Figure 1 illustrate these definitions. Note that the example shows both Type A and Type B order effects, demonstrating that they can be distinguished from each other empirically in a 2AFC paradigm by separately tracing $F_1(c)$ and $F_2(c)$.

A straightforward check for Type A order effects is to compare the judgment probabilities at $c = s$—that is, when the comparison is equal to the standard in size. Obviously, on such trials, the stimulus pairs $\langle sc \rangle$ and $\langle cs \rangle$ are identical physically. Because the response alternatives exclude each other, this implies that $P(\text{“}S_1 < S_2\text{”} \,|\, \langle ss \rangle) + P(\text{“}S_2 < S_1\text{”} \,|\, \langle ss \rangle) = 1$ and, therefore, $F_1(s) = 1 - F_2(s)$. Thus, at $c = s$, the two psychometric functions are restricted to sum to 1.

Now consider the values that the order-dependent psychometric functions take on at $c = s$, and define $p \equiv F_1(s)$. What does p tell us about Type A order error? Obviously, any deviation of p from .5 indicates that the PSEs for both psychometric functions differ from the standard s, but because of $F_2(s) = 1 - p$, they do so in opposite directions. If $p > .5$, CE_1 must be negative and CE_2 positive, and vice versa if $p < .5$. This is illustrated in the upper panel of Figure 1.

Thus, judgments must have been subject to Type A order error when p is found to deviate systematically from .5, which is easy to test statistically by a binomial or chi-square test. Note that this check (which does not require estimating PSEs) is neutral with respect to Type B order error; it is conceivable that presentation order induces a relative shift of the psychometric functions without affecting their slopes. Slope differences can be revealed by separately estimating DLs for $F_1(c)$ and $F_2(c)$, which requires no assumption of zero constant errors. However, nonzero CEs signaled by $p \neq .5$ should alert us to problems in estimating DL from probability correct data.

For probability-correct 2AFC psychometric functions, the DL is usually defined as $\text{DL}_G = c_{.75} - s$, where $c_{.75}$ is the value of the comparison stimulus that gives $G(c) = P(\text{Correct} \,|\, s, c) = .75$—that is, the stimulus level halfway between chance guessing and perfect responding. This DL definition is problematic, however, unless $\text{CE} = 0$. To anticipate, the problem is not caused by replacing $c_{.5}$ by s; just the opposite: Every probability-correct psychometric function must obey $c_{.5} = s$, whether constant errors exist or not! This fact implies that crucial information about constant errors is lost by collapsing data across stimulus orders. To show this, we relate $G(c)$ to the order-dependent functions $F_1(c)$ and $F_2(c)$ and demonstrate that DL_G is likely to underestimate true discriminability if CE differs from zero.

Linking $G(c)$ to $F_1(c)$ and $F_2(c)$

We begin by decomposing $G(c)$ into the two psychometric functions conditional on presentation order. If stimuli are presented equally often in the orders $\langle sc \rangle$ or $\langle cs \rangle$, the probability of correctly identifying the comparison c, $s \leq c$, equals

$$
\begin{aligned}
G(c) &= P(\text{Correct} \,|\, s, c) && (1)\\
&= P(\text{Correct} \,|\, \langle cs \rangle) \cdot .5 \\
&\quad + P(\text{Correct} \,|\, \langle sc \rangle) \cdot .5 && (2)\\
&= P(\text{“}S_1 > S_2\text{”} \,|\, \langle cs \rangle) \cdot .5 \\
&\quad + P(\text{“}S_2 > S_1\text{”} \,|\, \langle sc \rangle) \cdot .5 && (3)\\
&= F_1(c) \cdot .5 + F_2(c) \cdot .5 && (4)\\
&= [F_1(c) + F_2(c)] \cdot .5. && (5)
\end{aligned}
$$

By analogous reasoning, we find for $c \leq s$:

$$
\begin{aligned}
G(c) &= P(\text{Correct} \,|\, s, c) && (6)\\
&= [1 - F_1(c)] \cdot .5 + [1 - F_2(c)] \cdot .5 && (7)\\
&= 1 - [F_1(c) + F_2(c)] \cdot .5. && (8)
\end{aligned}
$$

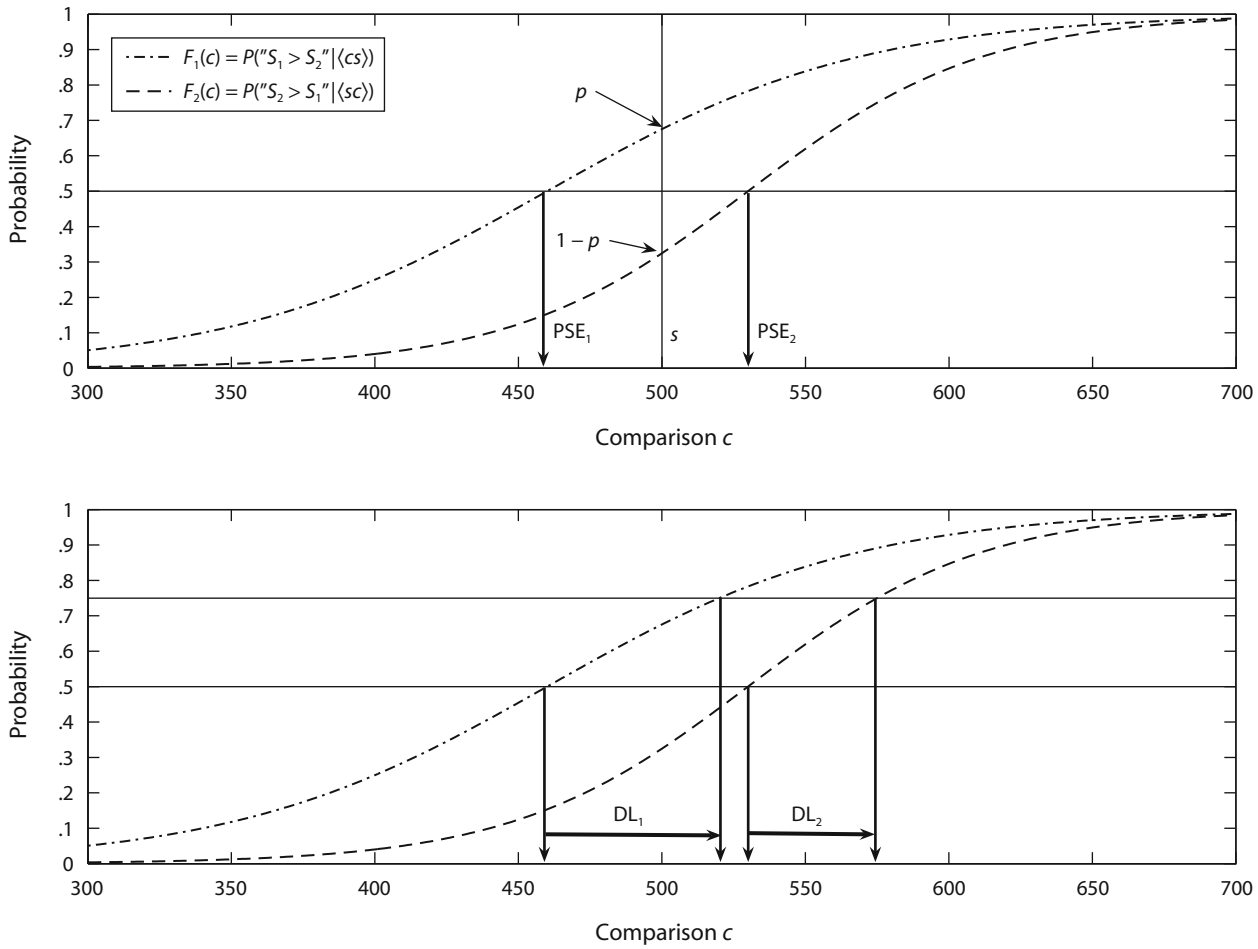

Figure 1. A hypothetical example illustrating the order-dependent psychometric functions $F_1(c) = P("S_1 > S_2" | \langle cs \rangle)$ and $F_2(c) = P("S_2 > S_1" | \langle sc \rangle)$. The size of the standard is $s = 500$. Upper panel: Point of subjective equality (PSE) for each of the two psychometric functions; they are $PSE_1 = 460$ and $PSE_2 = 530$ for $F_1(c)$ and $F_2(c)$, respectively. The constant errors, $CE = PSE - s$, associated with these functions thus are $CE_1 = 460 - 500 = -40$ and $CE_2 = 530 - 500 = 30$. The intersections of the functions $F_1(c)$ and $F_2(c)$ with the vertical at $c = 500$ yield $F_1(500) = p = .68$ and $F_2(500) = 1 - p = .32$, which implies that PSE_1 lies below the standard s and PSE_2 above. By the deviation of p and, thus, of $1 - p$ from .5, the psychometric functions are subject to Type A order error. Lower panel: The difference limens $DL_1 = 60$ and $DL_2 = 45$ for $F_1(c)$ and $F_2(c)$, respectively. This DL difference indicates superior sensitivity when c is in the second, as compared with the first, position—that is, a Type B order effect. Thus, this example exhibits both types of order effects.

Figure 2 illustrates the relationship between $G(c)$, $F_1(c)$, and $F_2(c)$.

From Equation 5, two important properties can be inferred that must hold for probability-correct psychometric functions $G(c)$ in general: First, for $c = s$, $G(c)$ equals .5, because

$$G(s) = F_1(s) \cdot .5 + F_2(s) \cdot .5 = p \cdot .5 + (1 - p) \cdot .5 = .5. \quad (9)$$

Note that this holds even when discrimination performance is subject to constant errors—that is, if both $F_1(s) \neq .5$ and $F_2(s) \neq .5$. Second, because $F_1(c)$ and $F_2(c)$ are monotonically increasing with the comparison, c, Equations 5 and 8 imply that $G(c)$ grows from .5 toward 1 as $|c - s|$ increases from 0 to infinity—that is, the further away the comparison is from the standard. Figure 2 also illustrates these properties.[1]

With these results, we are now in a position to understand why assessing the DL of the probability-correct psychometric function $G(s)$ does not reveal the true discriminability. Let DL_G be the difference between the stimulus levels at which c is judged correctly with probabilities of .75 and .5, respectively (Figure 2). Because $G(s) = .5$, which, as we have just seen, holds in general, it is correct to simplify to $DL_G = c_{.75} - s$, the definition conventionally employed for the 2AFC task (e.g., Gescheider, 1997). This DL_G generally reflects neither DL_1 nor DL_2, and it will generally also not be equal to the average of these two DLs. True sensitivity may be either under- or overestimated, depending on the relative magnitudes of Type A and Type B order effects. The deeper problem, however, is that by averaging across presentation orders, probability-correct 2AFC psychometric functions provide no clues on whether Type A order errors exist or not and, thus, offer no way to correct for them if necessary.

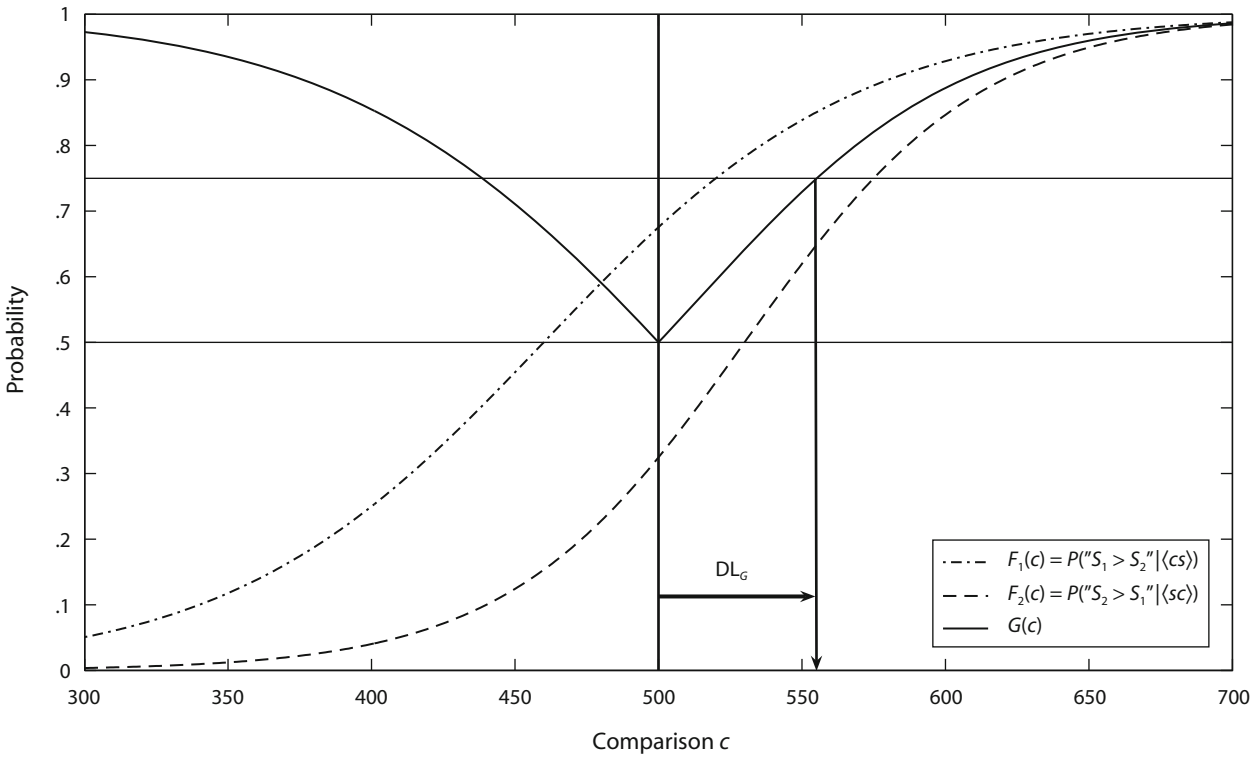

Figure 2. An example illustrating how the probability-correct function $P(\text{Correct}\,|\,s, c) = G(c)$ relates to the order-dependent functions $F_1(c)$ and $F_2(c)$. $F_1(c)$ and $F_2(c)$ are the same as in Figure 1. The solid line gives $G(c) = [F_1(c) + F_2(c)] \cdot .5$ for $c > s$, and $G(c) = 1 - [F_1(c) + F_2(c)] \cdot .5$ for $c < s$. For $G(c)$, the difference limen equals $DL_G = c_{.75} - s = 555 - 500 = 55$. This value exceeds the average of $DL_1 = 60$ and $DL_2 = 45$, which illustrates the fact that discriminability assessed from 2AFC correct data will be underestimated when a Type A order effect exists.

If averaging across presentation orders is the problem, the obvious solution is to assess CE and DL separately for $F_1(c)$ and $F_2(c)$. Before we show how to do so, we will illustrate the foregoing arguments and conclusions by graphical examples. In particular, we will briefly demonstrate that Type A order effects will produce an underestimation of true discrimination sensitivity—that is, DL_G will be larger than the DLs associated with $F_1(c)$ and $F_2(c)$. The order-dependent psychometric functions depicted in the three panels of Figure 3 have an identical slope ($DL_1 = DL_2 = 50$) but exhibit a Type A order effect, the size of which is varied across the three panels. In panel A, $F_1(c)$ and $F_2(c)$ do not exhibit such an effect ($CE_1 = CE_2 = 0$ and, thus, $p = .5$). Only this condition reflects DL_G's true sensitivity. Panels B and C display a moderate and a strong Type A order effect, respectively. As can be seen in these two cases, DL_G overestimates the DL of $F_1(c)$ and $F_2(c)$ and, thus, underestimates true discrimination sensitivity. Note that the amount of overestimation would not depend on the sign of CE_1 but on its absolute size only—that is, $|CE_1|$.

In summary then, the preceding analysis shows that DL_G generally does not accurately reflect discrimination ability but is contaminated by order effects. First, DL_G generally underestimates true sensitivity if only Type A order effects exist. In the extreme, the bias introduced by Type A effects may be such that DL_G equals $DL_1 + DL_2$; that is, the slope of $G(c)$ may be only half of that of either $F_1(c)$ or $F_2(c)$, as can be shown by approximating $F_1(c)$ and $F_2(c)$ by linear functions with equal slopes. Second, Type B order effects seem to *reduce* the bias in DL_G that is caused by Type A error. Under the linear approximation assumption, it can be shown that DL_G cannot exceed the average slope $(DL_1 + DL_2)/2$ when there is Type B error only (i.e., $p = .5$). Third, DL_G accurately reflects true discrimination performance only if neither a Type A nor a Type B order effect exists. This summary of our theoretical analyses forms the basis of the recommendations given in the following section.

Recommendations for Measuring DL in 2AFC Tasks

DL_G does not truly reflect discrimination performance. The usefulness of the 2AFC task can be restored, however, by data analysis methods that take order effects into account. Our suggested methods follow Nachmias's (2006) proposal of separately analyzing the order-dependent psychometric functions in 2AFC tasks.

An obvious solution of the problems sketched above is to estimate DL and CE for each order-dependent psychometric function separately and then to combine these measures appropriately, rather than trying to estimate them from the percentage-correct data. One obstacle to doing so is that, traditionally in 2AFC experiments, only

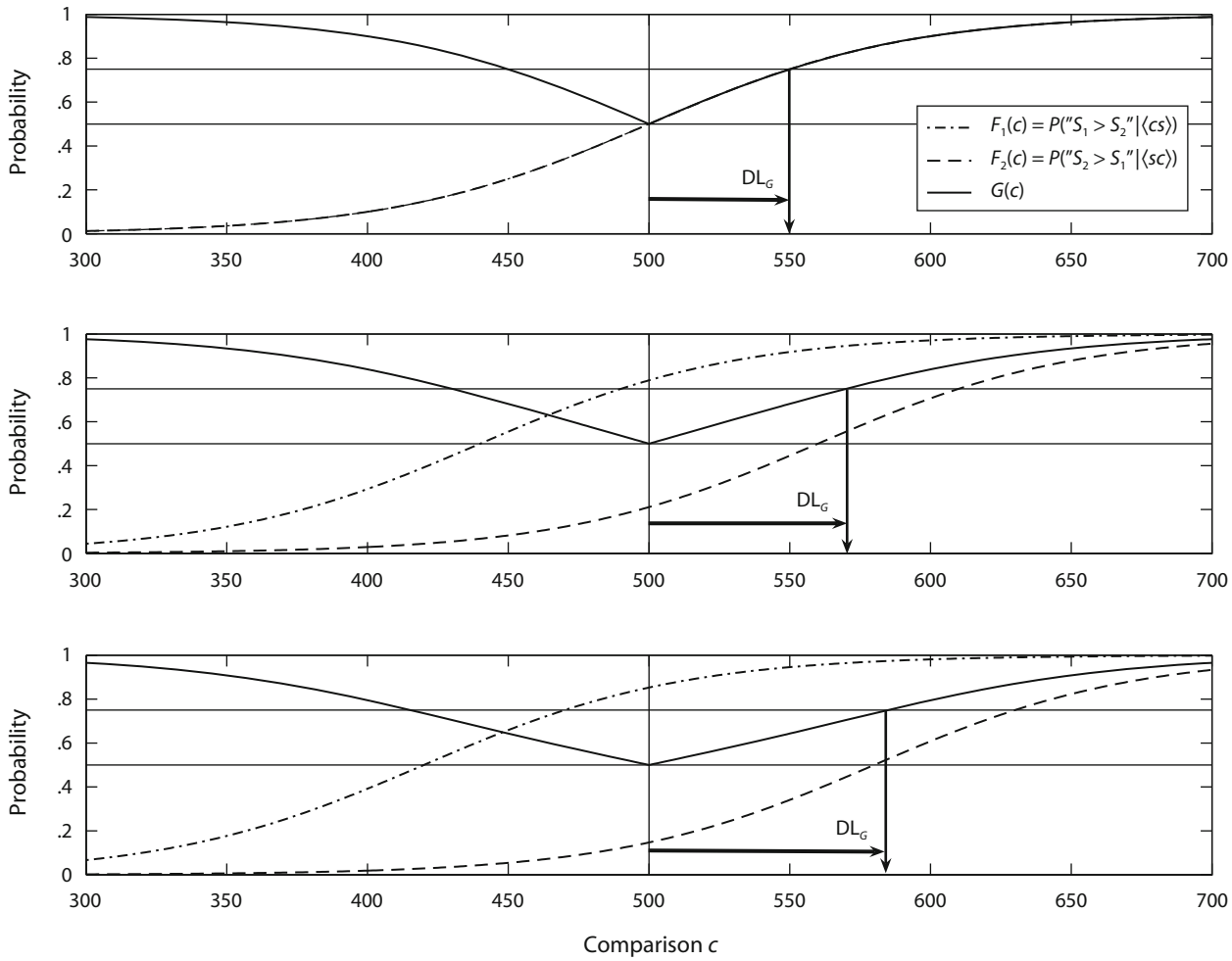

Figure 3. Illustration of how the size of a Type A order effect influences the difference limen, DL_G, associated with the probability-correct psychometric function $G(c)$. In all three panels, the difference limen of $F_1(c)$ and $F_2(c)$ is $DL_1 = DL_2 = 50$. Thus, these functions do not display a Type B order effect. Upper panel: There is no Type A order effect present—that is, $CE_1 = 0$ ($CE_2 = 0$). Thus, $DL_G = DL_1 = DL_2 = 50$. Middle panel: A moderate Type A order effect—that is, $CE_1 = -60$ ($CE_2 = 60$)—decreases the steepness of $G(c)$ and thus increases $DL_G = 70$. Lower panel: A large Type A order effect—that is, $CE_1 = -80$ ($CE_2 = 80$)—increases $DL_G = 84$ even more.

stimulus pairs are presented in which the comparison c is never less than s. This creates difficulties in estimating these statistics, because only a part of each psychometric function is covered. There are two ways of overcoming this problem: (1) employing the full stimulus range with $c < s$ and $c > s$, or (2) extrapolating by fitting a parametric model of the psychometric functions within the limited range, $c > s$. We will illustrate both approaches.

Extending the traditional procedure. Consider a 2AFC experiment in which stimulus pairs with $c < s$ and $c > s$ are presented. This allows assessing both psychometric functions $F_1(c)$ and $F_2(c)$ and, thus, estimating the respective DLs and CEs by traditional methods—for example, by reading off $c_{.25}$, $c_{.5}$, and $c_{.75}$ from the smoothed empirical functions by interpolation. This provides direct information on the prevalence and extent of Type A and Type B order errors. The question then arises as to whether and how to combine these measures across presentation orders. If Type A effects are dominant (i.e., $|CE_1| > 0$ and $|CE_2| > 0$), whereas Type B effects are small (i.e., no essential difference between DL_1 and DL_2), a reasonable sensitivity index for practical work may be obtained by averaging these measures as $DL_{2AFC} = (DL_1 + DL_2)/2$. If Type B effects are dominant, however—that is, if presentation order strongly affects the slopes of the psychometric functions—this may provide important information on the underlying mechanisms, which makes averaging questionable. Nevertheless, the DL averaged across orders provides a less biased estimate of true discrimination sensitivity than does DL_G estimated from the percentage-correct data if Type A effects are negligible. Let us illustrate these recommendations by studying an explicit model of the order-dependent psychometric functions.

In principle, any psychometric function that ranges from 0 to 1 qualifies as a model for $F_1(c)$ and $F_2(c)$. A function that lends itself to explicit analysis is the logistic function (see Bush, 1967). Thus, we assume that each psychometric function follows a logistic, with means and

Table 1
Illustrative Data: Response Frequencies *A*, *B*, *C*, and *D*, Along With Estimates of $F_1(c)$, $F_2(c)$, and $G(c)$, As a Function of Comparison Level c_i

	Comparison Level c_i							
	500	520	540	560	580	600	620	640
A_i	47	67	55	60	62	75	74	80
B_i	53	33	45	40	38	25	26	20
C_i	48	62	75	87	86	98	99	99
D_i	52	38	25	13	14	2	1	1
$\hat{F}_1(c_i)$	.470	.670	.550	.600	.620	.750	.740	.800
$\hat{F}_2(c_i)$	.480	.620	.750	.870	.860	.980	.990	.990
$\hat{G}_1(c_i)$	.475	.645	.650	.735	.740	.865	.865	.895

Note—These data are from a duration discrimination experiment. The first author served as a participant in five sessions. This experiment was virtually identical to the nonadaptive 2AFC condition in Experiment 2 of Lapid, Ulrich, and Rammsayer (2008).

slopes that may depend on presentation order ($a_1 = \mathrm{PSE}_1$ and $a_s = \mathrm{PSE}_2$) and slopes (i.e., b_1 and b_2); that is,

$$F_1(c) = \frac{1}{1 + \exp\left[-\left(c - a_1\right) / b_1\right]} \tag{10}$$

$$F_2(c) = \frac{1}{1 + \exp\left[-\left(c - a_2\right) / b_2\right]} = 1 - \frac{1}{1 + \exp\left[\left(c - a_2\right) / b_2\right]}. \tag{11}$$

For logistic functions, the DL obeys $\mathrm{DL}_i = b_i \cdot \ln(3)$ (see Bush, 1967).

Remember that, at $c = s$, these functions are constrained by $F_1(s) + F_2(s) = 1$ (see Figure 2). The two functions sum to 1 at $c = s$ if

$$-\frac{\left(s - a_1\right)}{b_1} = \frac{\left(s - a_2\right)}{b_2} \tag{12}$$

or, equivalently, if

$$a_2 = \frac{b_2}{b_1}\left(s - a_1\right) + s \tag{13}$$

holds. Therefore, a_2 is determined by b_1, b_2, a_1, and s, which implies that there are only three free parameters to be determined. (We will show below how to estimate them from the data by a maximum likelihood method.)

Defining, as suggested above, $\mathrm{DL}_{2\mathrm{AFC}}$ by the average inverse slope of $F_1(c)$ and $F_2(c)$, we obtain

$$\mathrm{DL}_{2\mathrm{AFC}} = \frac{\left(b_1 + b_2\right)}{2} \cdot \ln(3), \tag{14}$$

a measure that both is uncontaminated by Type A error and, moreover, reduces to the common inverse slope when Type B error vanishes. This parametric approach also enables the estimation of the constant errors associated with $F_1(c)$ and $F_2(c)$—that is, $\mathrm{CE}_1 = a_1 - s$ and $\mathrm{CE}_2 = a_2 - s$, respectively. If there is no Type A error, the true constant errors should be equal to zero.

Fitting a parametric model to the data. Assuming a parametric model for the order-dependent psychometric functions, it is possible to derive parameter estimates uncontaminated by order errors even if the data come from the traditional 2AFC experiment, in which the comparison c is never smaller than the standard s. As before, we estimate parameters and derive appropriate measures of DL by keeping trials with s in the first position separate from those with s in the second position. If wanted, the estimates of DL_1 and DL_2 can be combined as suggested by Equation 14.

We illustrate the recommended procedure by applying it to data from a duration discrimination study that one of us (R.U.) ran with himself as participant (Table 1). The data (response frequencies) in this table are reported according to the presentation order of s and c. There were eight comparison stimuli, c_i ($i = 1, \ldots, 8$), all equal to or larger than the standard s. Each comparison level was presented 200 times, 100 times in the first stimulus position and 100 times in the second position. The upper half of the table shows the response frequencies as a function of response category, c_i, and presentation order, summarizing the response frequencies

$$A_i = N(\text{“}S_1 > S_2\text{”} \mid \langle c_i s \rangle) \tag{15}$$

$$B_i = N(\text{“}S_2 > S_1\text{”} \mid \langle c_i s \rangle) \tag{16}$$

$$C_i = N(\text{“}S_2 > S_1\text{”} \mid \langle s c_i \rangle) \tag{17}$$

and

$$D_i = N(\text{“}S_1 > S_2\text{”} \mid \langle s c_i \rangle). \tag{18}$$

In the lower half of the table, the estimates of $F_1(c)$, $F_2(c)$, and $G(c)$ are given, which were computed from these response frequencies.

We used the maximum likelihood method to estimate the parameters b_1, b_2, and a_1 of the logistic model presented in the preceding section. Assuming stochastic independence of responses, the likelihood function L for the 2AFC task is

$$L\left(A, B, C, D \mid b_1, b_2, a_1\right) = \prod_{i=1}^{8} F_1(c)^{A_i} \cdot \left[1 - F_1(c)\right]^{B_i} \cdot F_2(c)^{C_i} \cdot \left[1 - F_2(c)\right]^{D_i}. \tag{19}$$

Maximizing this function with a numerical procedure yielded the estimates $\hat{b}_1 = 121.1$ msec, $\hat{b}_2 = 32.0$ msec, and $\hat{a}_1 = 490.8$ msec for the data in Table 1. Therefore, the estimated DL for $F_1(c)$ and $F_2(c)$ are $\mathrm{DL}_1 = 121.1 \cdot \ln(3) = 133.0$ msec and $\mathrm{DL}_2 = 32.0 \cdot \ln(3) = 35.2$ msec, respectively, and their average is equal to $\mathrm{DL}_{2\mathrm{AFC}} = 84.1$ msec. Note that responding was strongly affected by Type B error, as evidenced by the huge difference between DL_1 and DL_2. By Equation 13, the estimate of a_2 was calculated, which amounts to $\hat{a}_2 = 502.4$ msec. Therefore, the constant errors are $\mathrm{CE}_1 = 490.8 - 500 = -9.2$ msec and $\mathrm{CE}_2 = 502.4 - 500 = 2.4$ msec. These errors suggest that the true constant error is close to zero, indicating only a slight Type A order effect, if any.

The predicted psychometric functions are in fair agreement with the data (Figure 4). Assessing the fit by a minimum chi-square test yields $\chi^2(13) = 17.3, p = .19$, which indicates that the model fits the data moderately well. A likelihood ratio test [$\chi^2(1) = 52.9, p < .01$] reveals that $F_2(c)$ is significantly steeper than $F_1(c)$. This finding is

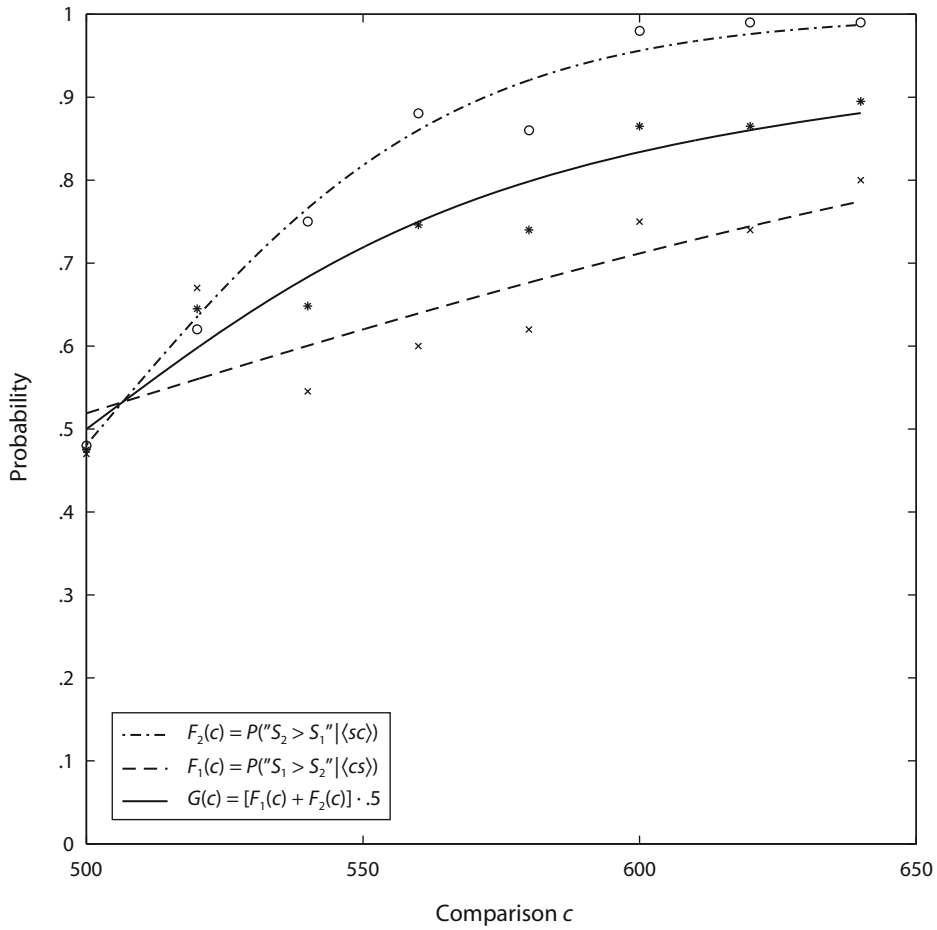

Figure 4. Predicted functions $P("S_1 > S_2" | \langle cs \rangle) = F_1(c)$ and $P("S_2 > S_1" | \langle sc \rangle) = F_2(c)$ for the data in Table 1 and the predicted function $G(c)$.

consistent with data reported by Hairston and Nagarajan (2007), Lapid et al. (2008), Nachmias (2006), and Ulrich et al. (2006), who observed better discrimination performance when the comparison *c* was presented in the second rather than in the first stimulus position. The traditional estimate of DL_G from $G(c) = .5 \cdot [F_1(c) + F_2(c)]$ turned out to be 60.5 msec, which, as was predicted, is less than the average DL of the two functions—that is, $DL_{2AFC} = 84.1$ msec.

Our numerical example demonstrates that Type A and Type B order effects can be disentangled from each other even in the traditional 2AFC task if responses are analyzed separately for each presentation order. Whether the DLs of the two functions should be averaged to obtain an overall index of discriminability performance must be decided as the case arises (but see our recommendations in the preceding section). In any case, we recommend always analyzing the data conditionally on presentation order and calculating separate estimates of DLs and CEs for each order.

Conclusion

In psychophysical research, the .75 threshold of correctly responding in a 2AFC task is often measured to index a participant's discrimination performance. Under plausible and nonrestrictive assumptions, we have shown that this measure—that is, DL_G—does not correctly reflect true discrimination performance when order effects are present. One such effect is the Type A order effect, as reflected in a nonzero CE. The more CE deviates from zero, the more inflated becomes DL_G. This hampers the interpretation of results from 2AFC tasks—for example, when experimenters compare estimates of DL_G across experimental conditions that differ in CE. Thus, a difference in DL_G between such conditions might just reflect a response bias, rather than a true difference in discriminability. Furthermore, Type B order errors also hamper the interpretation of DL_G. According to this error, discrimination performance depends on the presentation order of *s* and *c*. In this case, DL_G is not equal to the average DLs from the two psychometric functions but tends to underestimate it. In brief, the occurrence of a pure Type A order error inflates DL_G, whereas the occurrence of a pure Type B order error tends to deflate DL_G.

Following Nachmias (2006), we suggested an alternative approach for analyzing 2AFC data and provided mathematical tools with which to accomplish this goal. In particular, we recommended that researchers should not

average the two psychometric functions $P(\text{“}S_2 > S_1\text{”} \mid \langle sc \rangle)$ and $P(\text{“}S_1 > S_2\text{”} \mid \langle cs \rangle)$ in order to compute $G(c)$ for estimating DL. Instead, a DL may be computed for each of these two functions, and the two DLs may be averaged to assess the overall discrimination performance in the 2AFC task. We introduced a maximum likelihood procedure for computing this average DL. This approach will assess more accurately true discrimination performance in a 2AFC task, whether or not the data are subject to order effects.

Although SDT offers an approach to correct d' for Type A order errors in 2AFC tasks (see Klein, 2001, pp. 1424–1427), we are not aware of an analogous approach that permits correcting for Type B order effects. Thus, to develop models within the signal detection approach that account for both Type A and Type B order effects in a satisfactory way remains a challenge for future research.

AUTHOR NOTE

This work was supported by the Deutsche Forschungsgemeinschaft (Grant Ul 116/8-3). We thank Stanley A. Klein, Jeff Miller, Tanja Seifried, and an anonymous reviewer for constructive comments. Correspondence concerning this article should be addressed to R. Ulrich, Department of Cognitive and Biological Psychology, Psychological Institute, University of Tübingen, Friedrichstr. 21, 72072 Tübingen, Germany (e-mail: ulrich@uni-tuebingen.de).

REFERENCES

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.

EISLER, H., EISLER, A. D., & HELLSTRÖM, Å. (2008). Psychophysical issues in the study of time perception. In S. Grondin (Ed.), *Psychology of time* (pp. 75-109). Bingely, U.K.: Emerald.

FECHNER, G. T. (1860). *Elemente der Psychophysik* (Erster Teil) [Elements of Psychophysics (Vol. 1)]. Leipzig, Germany: Breitkopf & Härtel. [Citation is based on the third unchanged edition, which was published in 1907 by Breitkopf & Härtel]

GESCHEIDER, G. A. (1997). *Psychophysics: The fundamentals* (3rd ed.). Hillsdale, NJ: Erlbaum.

GREEN, D. M., & SWETS, J. A. (1973). *Signal detection theory and psychophysics*. New York: Wiley. [1988 reprint edition published by Peninsula Publishing, Los Altos, CA]

GUILFORD, J. P. (1954). *Psychometric methods* (2nd ed.). New York: McGraw-Hill.

HAIRSTON, I. S., & NAGARAJAN, S. S. (2007). Neural mechanisms of the time-order error: An MEG study. *Journal of Cognitive Neuroscience*, **19**, 1163-1174.

HARVEY, L. O. J. (1986). Efficient estimation of sensory thresholds. *Behavior Research Methods, Instruments, & Computers*, **18**, 623-632.

HELLSTRÖM, Å. (1985). The time-order error and its relatives: Mirrors of cognitive processing in comparing. *Psychological Bulletin*, **97**, 35-61.

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

LAPID, E., ULRICH, R., & RAMMSAYER, T. (2008). On estimating the difference limen in duration discrimination tasks: A comparison of the 2AFC and the reminder task. *Perception & Psychophysics*, **70**, 291-305.

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.

MORGAN, M. J., WATAMANIUK, S. N. J., & 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.

SIMPSON, W. A. (1989). The step method: A new adaptive psychophysical procedure. *Perception & Psychophysics*, **45**, 572-576.

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.

WICHMANN, F. A., & HILL, N. J. (2001). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics*, **63**, 1293-1313.

WICKENS, T. D. (2002). *Elementary signal detection theory*. New York: Oxford University Press.

WOODROW, H. (1935). The effect of practice on time-order errors in the comparison of temporal intervals. *Psychological Review*, **42**, 127-152.

YESHURUN, Y., CARRASCO, M., & MALONEY, L. T. (2008). Bias and sensitivity in two-interval forced choice procedures: Tests of the difference model. *Vision Research*, **48**, 1837-1851.

NOTE

1. It is often assumed in psychophysical work that $G(c)$ can be scaled as $G(c) = .5 + .5 \cdot P(c)$, where $P(c)$ represents a cumulative probability function (e.g., Harvey, 1986; Simpson, 1989; Ulrich & Miller, 2004; Wichmann & Hill, 2001). Interestingly, Equation 2 does not advocate such a scaling assumption. In fact, if $P(s) > 0$, the scaling assumption holds that $G(s) > .5$, which contradicts the first property of $G(c)$. This might bias results and increase sampling variability of threshold estimates. In addition and on the contrary to what has been claimed previously (e.g., Ulrich & Miller, 2004), $P(c)$ corresponds to $F_2(c)$ only under certain assumptions, as will become evident in the text that follows.

(Manuscript received March 1, 2008;
revision accepted for publication March 22, 2009.)