Atten Percept Psychophys (2015) 77:1033–1043
DOI 10.3758/s13414-015-0875-8

Effects of stimulus order on discrimination sensitivity for short and long durations

Karin M. Bausenhart · Oliver Dyjas · Rolf Ulrich

Published online: 2 April 2015

Abstract Previous studies have shown that discrimination sensitivity in 2AFC tasks depends on the presentation order of the standard and comparison stimulus. The present study examined whether this so-called Type B effect generalizes across different standard magnitudes. Therefore, Experiment 1 employed an auditory duration discrimination task with short (100 ms) and long (1,000 ms) standard durations and a constant interstimulus interval (ISI) of 1,000 ms. For both standard durations, a clear Type B effect emerged. In Experiment 2, discrimination sensitivity was assessed for short (300 ms) and long (1,000 ms) ISIs and a constant standard duration of 100 ms, in order to examine whether the Type B effect diminishes or even reverses when both stimuli are presented in rapid succession, as was suggested by previous studies. In the short, but not the long ISI condition, the Type B effect was virtually eliminated. Taken together, the present experiments suggest that the Type B effect is robust across standard magnitude, but diminishes when the time interval between both stimuli is reduced. This result pattern is discussed within the framework of the Internal Reference Model and the Sensation Weighting Model. It is also demonstrated that both models provide a quantitative account of the present results.

K. M. Bausenhart · O. Dyjas · R. Ulrich
University of Tübingen, Tübingen, Germany

K. M. Bausenhart (✉)
Department of Psychology, University of Tübingen,
Schleichstr. 4, 72076 Tübingen, Germany
e-mail: karin.bausenhart@uni-tuebingen.de

Keywords Discrimination sensitivity · Stimulus order · Type B effect · Duration discrimination · 2AFC · Psychophysics

In typical psychophysical tasks such as the two-alternative forced-choice (2AFC) task, participants are asked to discriminate between a fixed-magnitude standard stimulus s and a variable comparison stimulus c, whose magnitude can be lower, equal to or higher than the magnitude of the standard. In the temporal 2AFC task, these two stimuli are presented successively in one of two temporal orders, that is, $\langle sc \rangle$ and $\langle cs \rangle$, to balance for potential effects of stimulus order on task performance. Researchers often disregard such order effects by aggregating data across stimulus order. However, if the observed data are analyzed separately for the two stimulus orders, order effects are commonly observed. For example, the order-conditional psychophysical functions observed in a typical 2AFC task might be shifted horizontally from the point of objective equality (i.e., a Type A order effect or time-order error), such that the magnitude of the first stimulus is either judged to be higher or lower than the magnitude of the second one. Theoretically, this might be the sign of a perceptual, decisional, or response bias, and has been extensively studied (cf. Eisler, Eisler, & Hellström, 2008). More important for the purpose of the present study, however, is the so-called Type B order effect (Ulrich 2010; Ulrich & Vorberg, 2009). This effect refers to the phenomenon that the spread of the order-conditional psychometric functions may differ with regard to stimulus order. Specifically, the difference limen (DL) is typically larger and thus discrimination sensitivity lower for stimulus order $\langle cs \rangle$ than for $\langle sc \rangle$ (e.g., Lapid, Ulrich, & Rammsayer, 2008; Nachmias, 2006; Stott, 1935; Ulrich, 2010; Woodrow, 1935). This specific result pattern has been defined as a *negative* Type B effect (Dyjas & Ulrich, 2014).

A negative Type B effect was observed for both random and blocked stimulus order (Dyjas, Bausenhart, & Ulrich, 2012; Nachmias, 2006), and in different modalities including vision (Nachmias, 2006, for converging evidence, see also Patching, Englund, & Hellström, 2012), audition and vision (Grondin & McAuley, 2009; Lapid et al., 2008; Ulrich, Nitschke, & Rammsayer, 2006), and haptics (Ross & Gregory, 1964). Negative Type B effects emerge in different tasks domains (for an overview, see Dyjas et al., 2012, Table 1) including discrimination of weights (Ross & Gregory, 1964), duration (e.g., Woodrow, 1935), and visual shapes (Nachmias, 2006), even though their absence is also sometimes reported (e.g., García-Pérez & Peli, 2014, for a line bisection task). Finally, the Type B effect seems to be independent of judgment mode, since it has been demonstrated not only for comparative judgments, but also for equality judgments (e.g., Dyjas & Ulrich, 2014), and reproduction tasks (e.g., Bausenhart, Dyjas, & Ulrich 2014).

As was shown by Dyjas et al. (2012), the standard psychophysical difference model (Green & Swets, 1966; Macmillan & Creelman, 2005; Noreen, 1981; Sorkin, 1962; Thurstone, 1927a, b; Wickens, 2002), which assumes that participants base their judgments on the difference of the two stimuli's internal representations, cannot account for the Type B effect (but see, e.g., García-Pérez, 2014, for an extension of this model).[1]

Instead, Dyjas et al. (2012) suggested that participants rely on an internal reference (see also, e.g., Michels & Helson, 1954) which incorporates previous and currently available stimulus information, and thus is dynamically updated from trial to trial. According to this Internal Reference Model (IRM), the Type B effect emerges as a consequence of the dynamical updating process (Dyjas et al., 2012; Lapid et al., 2008). Specifically, IRM predicts that DL for stimulus order $\langle cs \rangle$ should be always either larger than or equal to DL for stimulus order $\langle sc \rangle$, depending on the weights assigned to previous and currently available stimulus information in the integration process. In other words, according to IRM, the negative Type B effect can either be present or absent, however, it cannot reverse. This general mechanism postulated by IRM should apply to different task domains, stimulus modalities and types of judgment, and is therefore consistent with the robust negative Type B effects reported by the studies cited above.

Likewise, the predictions of IRM should also generalize across different magnitudes of the standard. In our previous studies in the domain of duration discrimination, a standard duration of 500 msec was employed (Dyjas et al., 2012; Dyjas, Bausenhart, & Ulrich, 2014; Dyjas & Ulrich, 2014. Therefore, the goal of the present study was to examine the Type B effect across a broader range of standard durations. Interestingly, and contrary to IRM's predictions, there are some studies reporting reversed and thus *positive* Type B effects under particular circumstances (Hellström & Rammsayer, 2004; Hellström, 2003; for converging evidence, see also Hellström, 1979). Regarding duration discrimination, Hellström and Rammsayer (2004) report a positive Type B effect for very short duration stimuli, especially when presented with brief interstimulus intervals (ISI). It was suggested that processing of short durations might differ from processing of longer durations, especially when these are presented with longer ISIs. For instance, memory processes, interference between stimuli, and backward and forward masking might play a crucial role (Allan & Rousseau, 1977; Kallman & Morris, 1984; Kallman, Hirtle, & Davidson, 1986; Rammsayer & Lima, 1991). Moreover, the reversal of the Type B effect might be the sign of qualitatively different timing mechanisms operating at short and long durations (Michon, 1985; Rammsayer & Ulrich, 2011). Nevertheless, the observed reversal of the Type B effect disagrees with a large body of evidence reporting typical negative Type B effects, and contradicts IRM's predictions. Therefore, in addition to examining the Type B effect across a broader range of standard durations, a second aim of the present study is to investigate the influence of ISI on the Type B effect.

Experiment 1

In Experiment 1, participants performed a 2AFC duration discrimination task with standard durations of 100 and 1,000 ms presented in different blocks of trials. The ISI was kept constant at 1,000 ms. Therefore, this experiment examines whether the Type B effect generalizes to short and long standard durations.

Methods

Participants 19 women and 5 men (mean age 21.1 ± 3.4 years) volunteered in a single session in exchange for

[1] Please note that even such an extended version could not account without further assumptions for sequential effects reported in the literature (Bausenhart et al. 2014; Dyjas et al. 2012; Lages and Treisman 1998). Nevertheless, such sequential effects are consistent with the basic mechanism of the Internal Reference Model outlined in the following paragraphs.

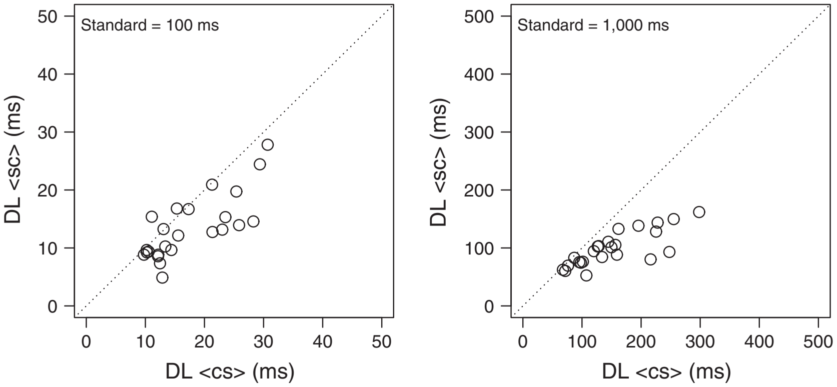

Fig. 1 Scatterplots of individual DL estimates for Experiment 1 (auditory duration discrimination with a fixed interstimulus interval of 1,000 ms). Standard durations are 100 ms (*left* panel) and 1,000 ms (*right* panel). The data points of the two replaced participants with suspiciously large DLs are not shown, because these values were considered outliers according to a predetermined three-sigma criterion. The corresponding values (x_i, y_i) of these two participants $(i = 1, 2)$ were $(x_1 = 59.5, y_1 = 212.0)$ and $(x_2 = 54.0, y_2 = 36.2)$ for the 100 ms standard and $(x_1 = 614.5, y_1 = 454.5)$ and $(x_2 = 348.8, y_2 = 464.2)$ for the 1,000 ms standard

course credit. All of them reported normal hearing and were naïve with respect to the hypotheses of the experiment. The data of two participants were replaced because their DLs in one or more conditions were larger than three standard deviations above the group mean in that condition.

Apparatus and stimuli The experiment was implemented in Matlab (The MathWorks, Inc., Version 2009a) using the Psychophysics Toolbox 3.0.8 (Brainard, 1997; Pelli, 1997). Instructions and feedback appeared on a computer screen in black ($< 1\ cd/m^2$) on a light-gray background ($49\ cd/m^2$). The "y" and "m" key of a standard German keyboard served as the left and right response key, respectively. Auditory stimuli were filled white noise intervals with rise- and fall-times of 10 ms, respectively, presented binaurally through headphones at a peak level of 65 dB SPL. A new sample of white noise was generated for each stimulus on each trial. In the *short standard condition*, the standard duration was 100 ms, and the duration of c varied from 52 to 148 ms in fixed steps of 12 ms. In the *long standard condition*, the standard duration was 1,000 ms, and the duration of c varied from 700 to 1,300 ms in fixed steps of 75 ms. Thus, there were 9 levels of c for each standard duration.

Procedure On each trial, two stimuli were presented successively separated by an offset-to-onset interstimulus interval (ISI) of 1,000 ms. For stimulus order $\langle sc \rangle$, the first stimulus was the fixed-duration standard and the second stimulus was the variable comparison stimulus. For stimulus order $\langle cs \rangle$, the order of stimuli was reversed. Stimulus order and the level of c varied randomly from trial to trial. Participants pressed the left (right) response key to indicate that the first (second) stimulus was the longer one. Following the response, either "1", "2", or "=" was displayed for 400 ms on the screen, indicating that the first or the second stimulus was the longer one or that the two stimuli were identical in duration, respectively. 1,600 ms after feedback onset, the next trial began. If the participant did not respond within 5,000 ms after the offset of the second stimulus, the trial was terminated and "zu langsam" (too slow) was displayed for 800 ms on the screen.

The short and the long standard duration were administered in separate blocks and the order of blocks was counterbalanced across participants. Each duration of c was presented 20 times for each stimulus order, such that a block consisted of 360 trials (20 repetitions × 9 levels of c × 2 stimulus orders). Participants could take a short rest after every 90 trials. At the beginning of each block, participants performed 18 practice trials (each level of c presented once for each stimulus order). Practice trials did not enter data analysis.

Design and dependent variables The dependent variables were stimulus order ($\langle sc \rangle$ vs. $\langle cs \rangle$) and standard duration (100 ms vs. 1,000 ms), thus there was a 2 × 2 within subjects design. A logistic psychometric function was fitted to the data of each participant in each condition under the constraint that the average of the psychometric functions for stimulus orders $\langle sc \rangle$ and $\langle cs \rangle$ passes through the point $(s, 0.5)$, a tautology that applies when stimulus order varies randomly and stimuli differ only in one dimension (Bausenhart, Dyjas, Vorberg, & Ulrich, 2012; García-Pérez & Alcalá-Quintana, 2010, 2011a, 2012; Ulrich & Vorberg, 2009; Ulrich 2010). From these psychometric functions, DL and PSE were calculated and submitted to separate repeated-measures analyses of variance (ANOVAs).

Results and discussion

The two scatterplots in Fig. 1 contain the estimated DLs for each participant and for the two standard durations.[2] The x-axis represents DL for stimulus order $\langle cs \rangle$ and the y-axis DL for order $\langle sc \rangle$. First, these data exhibit significant positive correlations between both estimates; $r = .77$, $t(22) = 5.7$, $p < .001$, for standard duration 100 ms and $r = .80$, $t(22) = 6.3$, $p < .001$ for 1,000 ms. Second, all but two data points are on or below the main diagonal indicating a negative Type B effect for almost all participants.

An ANOVA on DL confirmed the latter impression. Specifically, DL was larger for stimulus order $\langle cs \rangle$ (85 ms) than for $\langle sc \rangle$ (56 ms), $F(1, 23) = 37.8$, $p < .001$, $\eta_p^2 = .62$, that is, a typical negative Type B effect emerged (cf. Fig. 2, top panel). DL was larger for the long standard (126 ms) than for the short standard (15 ms), $F(1, 23) = 167.4$, $p < .001$, $\eta_p^2 = 0.88$. Perhaps unsurprisingly, the magnitude of this negative Type B effect increased numerically with standard duration, $F(1, 23) = 31.4$, $p < .001$, $\eta_p^2 = .58$. In order to compare the magnitude of the Type B effect across the different standard durations, Weber Fractions (WF) were computed as DL / standard duration. WF was slightly larger for the short (0.15) than the long (0.13) standard duration, $F(1, 23) = 9.9$, $p < .01$, $\eta_p^2 = .30$. Also, larger WFs were observed for stimulus order $\langle cs \rangle$ (0.16) than for stimulus order $\langle sc \rangle$ (0.12), $F(1, 23) = 42.0$, $p < .001$, $\eta_p^2 = .65$, reflecting the Type B effect. Crucially, this effect was not modulated by standard duration, $F(1, 23) = 1.6$, $p = .22$, $\eta_p^2 = .07$ (cf. Fig. 2, middle panel). Thus, the magnitude of the negative Type B effect is comparable across standard duration, and most clearly, it was not reversed for the short duration.

As one expects, PSE was larger for the long standard than for the short standard, $F(1, 23) = 112, 181.0$, $p < .001$, $\eta_p^2 = 1.0$. Neither the effect of stimulus order, $F(1, 23) = 1.1$, $p = .30$, nor the interaction ($F < 1$) were significant (cf. Fig. 2, lower panel). The overall pattern of results is thus consistent with our previous research (Dyjas et al., 2012, 2014) showing that stimulus order exerts dissociable effects on DL and PSE.

Experiment 2

In Experiment 1, a typical negative Type B effect emerged for short as well as long standard durations. Previous research had demonstrated a reversal of the Type B effect at short standard durations, especially when paired with short ISIs (Hellström, 2003; Hellström & Rammsayer, 2004). Therefore, Experiment 2 examined whether the Type B effect would be modulated by shortening ISI from 1,000 ms to 300 ms. For example, within IRM it seems plausible that the updating of the internal reference with the information from the first stimulus takes some time and therefore cannot be accomplished within a short ISI, before the first stimulus representation is masked by the representation of the second stimulus. Consequently, the Type B effect would be diminished in the short ISI condition. As outlined above, however, no reversal of the Type B effect would be implied by IRM.

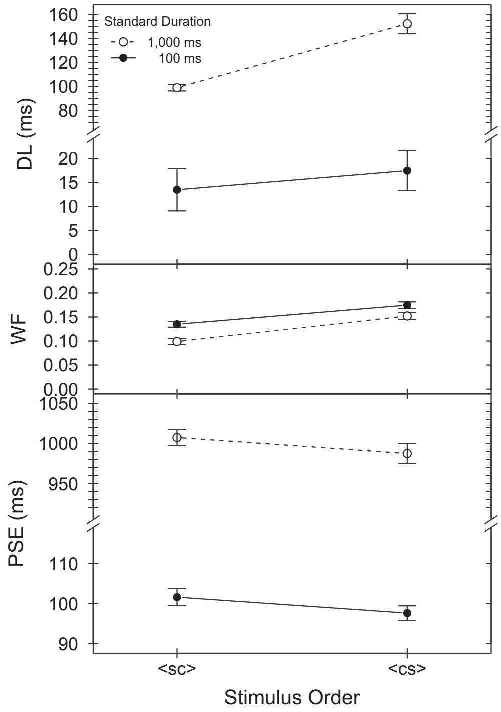

Fig. 2 Results of Experiment 1 (auditory duration discrimination with a fixed interstimulus interval of 1,000 ms). Mean difference limen (DL, *top* panel), mean Weber Fraction (WF, *middle* panel), and mean point of subjective equality (PSE, *lower* panel) $\pm 1 \cdot SE$ as a function of stimulus order and standard duration. The standard error SE of the mean was calculated according to Cousineau (2005) for a within-subjects design. Please note that axis breaks and scaling discontinuities for mean DL and PSE were employed to provide suitable scales for both standard durations and to grant comparability with the results of Experiment 2 displayed in Fig. 4

Methods

Participants A new sample of 20 women and 4 men (mean age: 25.2 ± 8.8 years) participated in exchange for course credit. They reported normal hearing and were naïve with

[2] We thank Miguel A. García-Pérez for suggesting this scatterplot analysis.

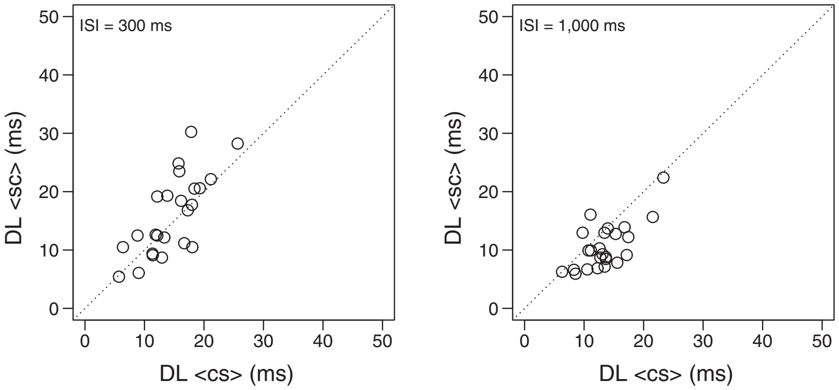

Fig. 3 Scatter plots of individual DL estimates for Experiment 2 (auditory duration discrimination with a fixed standard duration of 100 ms). Short ISI (*left* panel) and long ISI (*right* panel). The data points of one replaced participant with suspiciously large DLs are not shown, because these values were considered outliers according to a predetermined three-sigma criterion. The corresponding values (x, y) of this participant were $(x = 65.6, y = 89.3)$ for the short ISI and $(x = 65.6, y = 59.8)$ for the long ISI

respect to the hypotheses. None of them had participated in the previous experiment. The data of one participant were replaced because DLs in all conditions were larger than three standard deviations above the corresponding group means.

Apparatus, stimuli, procedure, and design These were identical to the ones used in Experiment 1, except for the following changes. First, only the short standard duration was employed. Second, ISI was 300 ms (short ISI) in one block of trials and 1,000 ms (long ISI) in another block of trials; the order of blocks was counterbalanced across participants. Thus, there was a 2 (stimulus order $\langle sc \rangle$ vs. $\langle cs \rangle$) × 2 (ISI: 300 ms vs. 1,000 ms) within-subjects design.

Results and discussion

Figure 3 depicts the individual DLs for short and long ISIs. First, as in Experiment 1, these data show positive correlations; $r = .73$, $t(22) = 5.1$, $p < .001$ for short ISI, and $r = .68$, $t(22) = 4.4$, $p < .001$ for long ISI. Second, visual inspection suggests that only the data points for the long ISI show a systematic negative Type B effect, whereas the data points for the short ISI rather scatter around the identity line.

This subjective impression was strengthened by ANOVA. In particular, overall there was no main effect of stimulus order on DL, $F(1, 23) = 1.8$, $p = .19$, $\eta_p^2 = .07$ (cf. Fig. 4, top panel). DL was larger for the short ISI (15.2 ms) than for the long ISI (12.0 ms), $F(1, 23) = 14.3$, $p < .001$, $\eta_p^2 = .38$. There was an interaction of stimulus order and ISI, $F(1, 23) = 12.1$, $p < .01$, $\eta_p^2 = .34$. Specifically, there was a typical negative Type B effect in the long ISI condition, $t(23) = 4.4$, $p < .001$, replicating the result of Experiment 1. However, this effect vanished for the short ISI condition, $t(23) = 1.4$, $p = .16$. Accordingly, the temporal interval between the two successive stimuli in the 2AFC task modulates the magnitude of the Type B effect. This is consistent with the idea that the integration of the first stimulus into the internal reference is hampered when the second stimulus follows the first one closely in time.

ISI did not affect PSE, $F(1, 23) = 1.3$, $p = .26$ (cf. Fig. 4, lower panel). However, PSE was larger for stimulus order $\langle sc \rangle$ (104 ms) than for order $\langle cs \rangle$ (95 ms), $F(1, 23) = 30.5$, $p < .001$, $\eta_p^2 = .57$. This effect suggests that the magnitude of the first stimulus (s in the $\langle sc \rangle$ condition, and c in the $\langle cs \rangle$ condition) is overestimated as compared to the magnitude of the second stimulus. This Type A effect corresponds to a positive time-order error, which is often observed for rather short duration stimuli

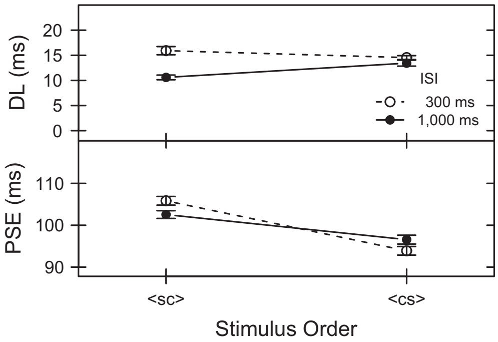

Fig. 4 Results of Experiment 2 (auditory duration discrimination with a fixed standard duration of 100 ms). Mean difference limen (DL, *top* panel) and mean point of subjective equality (PSE, *lower* panel) $\pm 1 \cdot SE$ as a function of stimulus order and ISI. The standard error SE of the mean was calculated according to Cousineau (2005) for a within-subjects design

(e.g., Allan, 1977). This effect, however, was modulated by ISI, $F(1, 23) = 6.4$, $p < .05$, $\eta_p^2 = .22$, such that in the short ISI condition, the overestimation of the first stimulus compared to the second one was even larger than in the long ISI condition. Again, this is consistent with the findings of previous studies (e.g., Hellström & Rammsayer, 2004).

General discussion

The major aim of the present study was to investigate whether the Type B effect is modulated by the magnitude of the standard stimulus. Experiment 1 demonstrated higher duration discrimination sensitivity for stimulus order $\langle sc \rangle$ than for stimulus order $\langle cs \rangle$, independent of whether relatively short (100 ms) or long standard durations (1,000 ms) were employed. Thus, the finding of a negative Type B effect generalizes from a standard duration of 500 ms (Dyjas et al., 2012, 2014; Dyjas & Ulrich, 2014) to longer as well as shorter standard durations. Nevertheless, in Experiment 2, this negative Type B effect diminished when both stimuli were separated by a brief ISI (300 ms), rather than a longer one (1,000 ms). Yet, there was no reversal of the Type B effect. Accordingly, these results fit within the scope of IRM, which predicts that discrimination sensitivity for trials with stimulus order $\langle sc \rangle$ should be either higher than or equal to sensitivity for trials with stimulus order $\langle cs \rangle$ (as shown in the Appendix, IRM also provides a quantitative account of the present data). Therefore, in the domain of auditory duration discrimination, IRM seems to apply to a relatively broad range of standard durations and ISIs.

Within IRM, the absence of the Type B effect for the brief ISI may be attributed to a lack of integration of the current stimulus representation with information from previous trials. This seems plausible under the assumption that the integration does not proceed in a completely automatic but rather in a more controlled and maybe time-consuming fashion. Previous evidence from a cueing study suggests that the integration process is indeed under cognitive control (Dyjas et al., 2014). Clearly, further studies are required to substantiate this speculation, for example by manipulating the time available for the integration process more directly.

As outlined in the Introduction, there is previous evidence for a reversal of the Type B effect in duration discrimination when an even shorter standard duration (50 ms) than in the present experiments was employed, especially with relatively short ISIs ($\leq$ 300 ms, Hellström & Rammsayer, 2004). In addition, the procedure to measure discrimination performance in this study differed in several ways from the one of the present experiments. Specifically, ISI was manipulated between participants and an adaptive testing scheme was administered to assess performance for the two stimulus orders. While stimulus order was randomly intermixed between trials (as in the present experiments), performance at the 25th and 75th percentile was assessed in separate blocks of trials rather than within the same block of trials. Under such conditions, estimates of DL might be influenced by shifts of the underlying psychometric functions between blocks (cf. Ulrich & Vorberg, 2009). These methodological differences thus hamper a direct comparison between the present study and the one by Hellström and Rammsayer (2004).

It should be noted that converging evidence for positive Type B effects has sometimes also been found in other task domains (as loudness discrimination and line length) under certain stimulation conditions and using different assessment methods (Hellström, 1979, 2003). Therefore, of course, we cannot refute the findings of a positive Type B effect under specific conditions. Although such a reversed Type B effect could not be explained by IRM, the present results nonetheless show that IRM is applicable to a wide range of standard durations and ISIs. A more general framework, such as *Sensation Weighting* (SW; Hellström, 1979, 1985), would be needed to account for any reversal of the Type B effect.[3] According to this framework, the internal representations of both stimuli in a discrimination task would be weighted differentially, with the assigned weights depending on the ISI duration. Accordingly, if the second stimulus receives a larger weight than the first stimulus, the SW framework implies a negative Type B effect, whereas a reversed effect is implied when the first stimulus receives a larger weight than the second stimulus (please refer to the Appendix for a quantitative account of the present results based on this framework). In any case, neither the typical negative Type B effect nor a reversal of this effect can be explained by psychophysical accounts based on the standard difference model (Thurstone, 1927a, b).

In summary, the results of this study demonstrate that the negative Type B effect is robust across different stimulus magnitudes. In general, the widespread presence of Type B order effects provides a benchmark for the formulation and advancement of psychophysical theories of stimulus discrimination.

Author Note We thank Å. Hellström and M. A. García-Pérez for helpful comments and Matthias Karlbauer for assistance in data collection. Correspondence concerning this article should be addressed to Karin M. Bausenhart, Cognition and Perception, Department of Psychology, University of Tübingen, Schleichstr. 4, 72076 Tübingen, Germany. E-Mail may be sent to karin.bausenhart@uni-tuebingen.de. This research was supported by the DFG grant UL 116/13-1.

[3] Actually, IRM can be regarded as a restricted case of the SW framework, however, also going beyond SW by adding an explicit formulation of the dynamic formation of the internal reference.

Appendix: Quantitative accounts of the present data

The Type B effect provides an important benchmark for evaluating the predictions of models of discrimination processes. Therefore, it was the primary goal of the present work to assess the robustness of the negative Type B effect. For example, the Internal Reference Model (IRM; Dyjas et al., 2012; Lapid et al., 2008) is consistent with a negative Type B effect but not with a positive one, whereas the basic Sensation Weighting Model (SWM; or Weighted Difference Model, e.g., Hellström, 1979, 1985, 2003) would be consistent with not only negative but also with positive Type B effects. Thus, according to Popper's theory of science (cf. Glöckner & Betsch, 2011), SWM involves less empirical content than IRM. This appendix goes beyond the primary goal of the present work and examines whether IRM and SWM would also provide adequate quantitative accounts of the psychometric functions observed in the two experiments reported in the main text.

Internal reference model

IRM's predicted psychometric functions for comparative judgments with random stimulus orders $\langle sc\rangle$ and $\langle cs\rangle$ were derived in Dyjas and Ulrich (2014). In order to keep things simple, Dyjas and Ulrich derived these predictions under the simplifying assumption that participants judge the first stimulus to be larger than the second stimulus, if $\mathbf{D}_n > 0$ rather than $\mathbf{D}_n > \gamma$, where $\mathbf{D}_n = \mathbf{I}_n - \mathbf{X}_{2,n}$ is the difference between the internal representation of the internal standard $\mathbf{I}_n$ and the second stimulus $\mathbf{X}_{2,n}$ on trial n, and where γ is a bias parameter. Dyjas and Ulrich (2014, p. 1127) explicitly noted that $\gamma \neq 0$ would also entail the prediction of the Type A effect — a common assumption of difference models. Including γ as a further parameter within IRM's framework would not alter IRM's prediction regarding the Type B effect but also allow IRM to account for Type A effects.

Thus amending the predictions of IRM such that γ may differ from zero, yields the following predicted psychometric functions (see Appendix B in Dyjas & Ulrich, 2014, especially Equations 23 and 26, p. 1147), that is, judging the comparison duration c larger than the standard duration s

$$P(\text{“}\, c > s \,\text{”}|\langle sc\rangle) = \Phi\left[\frac{\gamma + (c-s)}{\kappa}\right] \tag{1}$$

for stimulus order $\langle sc\rangle$ and

$$P(\text{“}\, c > s \,\text{”}|\langle cs\rangle) = \Phi\left[\frac{-\gamma + (c-s)\cdot(1-g)}{\kappa}\right] \tag{2}$$

for stimulus order $\langle cs\rangle$ with

$$\kappa = \sqrt{\frac{2\cdot\sigma^2}{1+g} + \frac{g^2\cdot(1-g)\cdot\sigma_c^2}{2\cdot(1+g)}}. \tag{3}$$

The parameter σ denotes the variability (noise) of the internal stimulus representation and this parameter is expected to increase with standard duration according to Weber's law. The constant g, $0 \leq g < 1$, denotes the weight for updating the internal reference in the current trial. The constant σ_c is the standard deviation of the employed comparison durations and thus determined by the experimental setting.

It can be seen that for $g = 0$, the standard difference model is implied for $\gamma = 0$ and the difference model with bias (cf. García-Pérez & Alcalá-Quintana, 2011b) is implied for $\gamma \neq 0$. Specifically, for $g = 0$, Eqs. 1 and 2 simplify to

$$P(\text{“}\, c > s \,\text{”}|\langle sc\rangle) = \Phi\left[\frac{\gamma + (c-s)}{\sqrt{2}\cdot\sigma}\right] \tag{4}$$

and

$$P(\text{“}\, c > s \,\text{”}|\langle cs\rangle) = \Phi\left[\frac{-\gamma + (c-s)}{\sqrt{2}\cdot\sigma}\right]. \tag{5}$$

In addition, it can be shown that Eqs. 1 and 2 satisfy the constraint

$$P(\text{“}\, c > s \,\text{”}|\langle sc\rangle) + P(\text{“}\, c > s \,\text{”}|\langle cs\rangle) = 1 \tag{6}$$

for $s = c$ (see Ulrich & Vorberg, 2009).

The predicted psychometric functions (i.e., Eqs. 1 and 2) were fitted to the observed psychometric functions by minimizing the root mean squared error (RMSE). Specifically, these functions were fitted to each participant's data and then these predicted functions were averaged across all participants. Figures 5 and 6 depict these predicted average functions along with the average relative response frequencies for Experiments 1 and 2, respectively. These figures show that IRM provides reasonable fits and also accounts for the negative Type B effect observed in these data.

Furthermore, Table 1 contains the average estimated parameters and the average RMSE along with the corresponding standard errors. First, it can be seen that the average weighting constant g remains remarkably stable across the two standard durations in Experiment 1, that is, g does not significantly differ between the two conditions, $t(23) = 0.1$, $p = .94$. Second, as one should expect according to Weber's law, the internal noise (σ) increases with standard duration in Experiment 1. Third, the interstimulus interval (ISI) between the standard and comparison stimulus in Experiment 2 affects g, $t(23) = 4.0$, $p < .001$. For ISI = 1,000 ms, g is quite similar to the estimates of g obtained in Experiment 1. For ISI = 300 ms, however, g is close to zero — within the framework of IRM this would indicate that this interstimulus interval is too short to enable an updating of the internal reference. Here, it should be noted that the estimated values of g must be somewhat biased towards

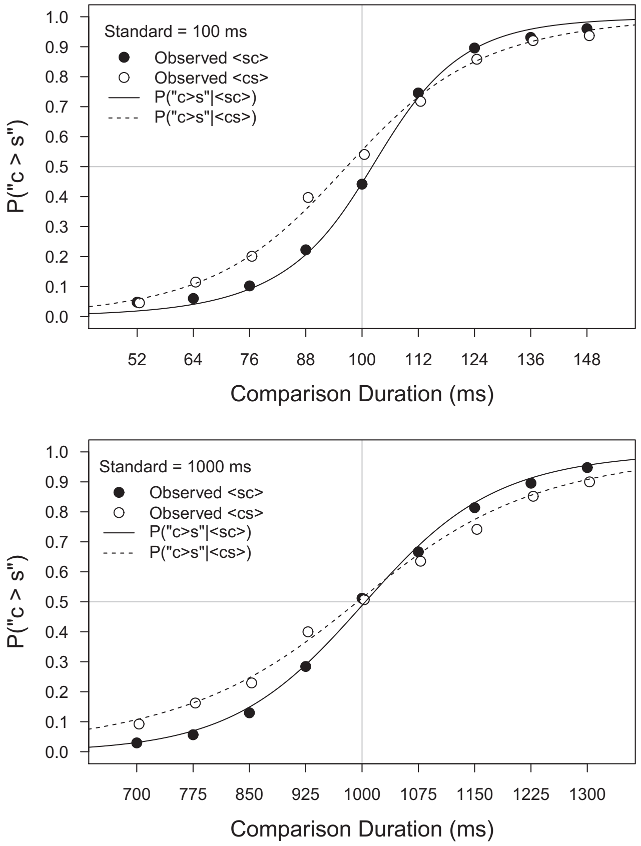

Fig. 5 Average of the individual predicted psychometric functions of the internal reference model (IRM) for stimulus orders $\langle sc \rangle$ and $\langle cs \rangle$ and for the data of Experiment 1. The x-axis represents the duration of the comparison c and the y-axis represents the probability of judging the comparison duration c longer than the standard duration s. The single data points depict observed average relative response frequencies. *Upper* panel: Standard duration $s = 100$ ms. *Lower* panel: Standard duration $s = 1,000$ ms

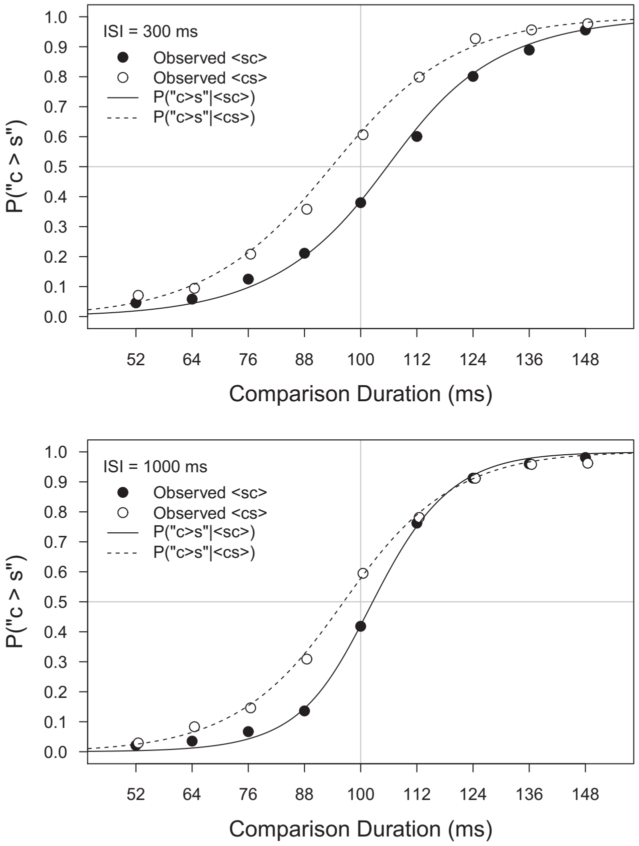

Fig. 6 Average of the individual predicted psychometric functions of the internal reference model (IRM) for stimulus orders $\langle sc \rangle$ and $\langle cs \rangle$ and for the data of Experiment 2. The x-axis represents the duration of the comparison c and the y-axis represents the probability of judging the comparison duration c longer than the standard duration s. The single data points depict observed average relative response frequencies. *Upper* panel: Interstimulus Interval ISI = 300 ms. *Lower* panel: Interstimulus Interval ISI = 1,000 ms

positive values, even if the true underlying value of g would be equal to zero. This is due to the sampling error associated with each observed psychometric function. Because g is restricted to $0 \leq g < 1$, estimates of g for participants that exhibit positive Type B effects (be it by random or by systematic deviation) must be equal to 0; whereas estimates of g for participants which exhibit negative Type B effects (again by random or systematic deviation) will be larger than 0. Thus, the average of g across participants must be larger than 0.

Therefore, we fitted Eqs. 1 to 2 also to the observed psychometric functions averaged across participants, since sampling error will be reduced in these average psychometric functions. In this case, the estimate of g is virtually equal to zero in the 300 ms ISI condition of Experiment 2 (cf. Table 2).[4]

Basic sensation weighting model (SWM)

Dyjas and Ulrich (2014, pp. 1143–1144) considered a special case of SWM that like IRM involves three free model

[4] Another approach to deal with the potential overestimation of the g parameter would be fitting an extended version of IRM which allows for g to take negative values, $-1 < g < 1$ (cf. Dyjas et al., 2012, p. 1832). Such an extension might reflect a negative weighting of prior information (e.g., contrast effects). In this case, IRM would also entail a positive Type B effect. Following this approach, we fitted parameters g, σ, and γ to the data of each observer. In the 300 ms ISI condition of Experiment 2, on average, estimates of g were indeed slightly negative ($g = -0.080$, $SD = 0.368$). However, these estimates did not differ meaningfully from zero, $t(23) = 1.1$, $p = .30$, and only tended to be smaller than the corresponding estimates summarized in Table 1, $t(23) = 2.0$, $p = .06$. For all other conditions, again, g was well above zero (all $ps < .01$), even though average estimates were numerically (yet not significantly) smaller (Experiment 1: $g = 0.252$, $SD = 0.293$ for the 100 ms and $g = 0.283$, $SD = 0.172$ for the 1,000 ms standard duration, Experiment 2: $g = 0.205$, $SD = 0.301$ for the 1,000 ms ISI) than the corresponding ones summarized in Table 1, all $ps > .05$. Thus, this analysis lends further support to our main conclusion of a consistently negative Type B effect, except for the short ISI condition of Experiment 2, where the Type B effect seems to be absent.

Table 1 Means and their standard errors (in italics) of the estimated model parameters of the Internal Reference Model (IRM) and of the basic Sensation Weighting Model (SWM). Parameters were fitted to the individual observed psychometric functions and then averaged across participants

	IRM				*SWM*			
Experiment and condition	g	σ	γ	RMSE	σ_1	σ_2	γ^*	RMSE
Experiment 1, $s = 100$ ms	0.292	13.4	−1.6	0.070	16.3	13.8	−18.3	0.070
	0.044	*1.4*	*1.2*	*0.004*	*1.4*	*1.3*	*3.9*	*0.004*
Experiment 1, $s = 1,000$ ms	0.287	120.5	−8.0	0.077	143.4	117.7	−180.9	0.077
	0.033	*8.6*	*8.8*	*0.004*	*11.0*	*8.0*	*24.8*	*0.004*
Experiment 2, $ISI = 300$ ms	0.051	15.6	−5.6	0.072	15.2	16.0	−1.6	0.067
	0.025	*1.3*	*1.0*	*0.005*	*1.1*	*1.4*	*4.0*	*0.005*
Experiment 2, $ISI = 1,000$ ms	0.253	10.1	−2.8	0.061	12.4	10.8	−16.0	0.061
	0.045	*1.0*	*0.8*	*0.004*	*0.8*	*0.8*	*3.9*	*0.004*

The parameters σ, γ, σ_1, σ_2, γ^* are measured in milliseconds. The root mean squared error (RMSE) was averaged across the individual RMSEs

parameters. Like IRM, this special case assumes that the internal representations of stimuli are noisy. In contrast to IRM, however, it is assumed that the amount of noise can differ between the first and second stimulus in a trial. For example, is seems plausible that the memory trace of the first stimulus fades over time and thus at the time of comparison, the internal representation $\mathbf{X}_1$ of the first stimulus is noisier than the internal representation $\mathbf{X}_2$ of the second stimulus. In order to compensate for this effect and to optimize discrimination performance, the weights w_1, $0 < w_1 < 1$, and $w_2 = 1 - w_1$ of the weighted difference

$$\mathbf{D}_n = w_1 \cdot \mathbf{X}_1 - w_2 \cdot \mathbf{X}_2 + u \tag{7}$$

need to be adjusted accordingly; u is a constant and represents a bias parameter. The variance of $\mathbf{D}_n$ is minimized and thus discrimination performance maximized, if

$$w_1 = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} \tag{8}$$

$$w_2 = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2} \tag{9}$$

where σ_1 and σ_2 are the standard deviations (i.e., amount of internal noise) associated with $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively (see Dyjas & Ulrich, 2014, p. 1143). Evoking the common assumption that $\mathbf{X}_1$ and $\mathbf{X}_2$ are normally distributed and independent and their expected values are equal to the physical magnitudes of the stimuli, then

$$P(\text{“}\, c > s \,\text{”}|\langle sc \rangle) = P(\mathbf{D}_n \leq \gamma | \langle sc \rangle) \tag{10}$$

$$= \Phi\left[\frac{\gamma - [w_1 \cdot s - w_2 \cdot c + u]}{\sqrt{w_1^2 \cdot \sigma_1^2 + w_2^2 \cdot \sigma_2^2}}\right] \tag{11}$$

$$= \Phi\left[\frac{\gamma^* - [w_1 \cdot s - w_2 \cdot c]}{\sqrt{w_1^2 \cdot \sigma_1^2 + w_2^2 \cdot \sigma_2^2}}\right] \tag{12}$$

with $\gamma^* = \gamma - u$. Likewise, for stimulus order $\langle cs \rangle$ one obtains the predicted psychometric function

$$P(\text{“}\, c > s \,\text{”}|\langle cs \rangle) = P(\mathbf{D}_n \geq \gamma | \langle cs \rangle) \tag{13}$$

$$= \Phi\left[\frac{-\gamma^* - [w_2 \cdot s - w_1 \cdot c]}{\sqrt{w_1^2 \cdot \sigma_1^2 + w_2^2 \cdot \sigma_2^2}}\right]. \tag{14}$$

Table 2 Estimated model parameters of the Internal Reference Model (IRM) and of the basic Sensation Weighting Model (SWM). Parameters were fitted to the observed psychometric functions averaged across participants

	IRM				*SWM*			
Experiment and condition	g	σ	γ	RMSE	σ_1	σ_2	γ^*	RMSE
Experiment 1, $s = 100$ ms	0.304	14.7	−2.0	0.025	17.4	14.5	−19.1	0.025
Experiment 1, $s = 1,000$ ms	0.308	123.3	−4.8	0.017	143.7	119.5	−184.9	0.017
Experiment 2, $ISI = 300$ ms	0.000	16.3	−5.9	0.021	16.0	16.6	0.6	0.020
Experiment 2, $ISI = 1,000$ ms	0.278	11.1	−2.6	0.022	13.1	11.1	−17.7	0.022

The parameters σ, γ, σ_1, σ_2, γ^* are measured in milliseconds. The root mean squared error (RMSE) measures the goodness of fit

It is easy to verify that for $\sigma_1 = \sigma_2$, SWM also includes the standard difference model (for $\gamma^* = 0$) and the difference model with bias (for $\gamma^* \neq 0$, cf. García-Pérez & Alcalá-Quintana, 2011b) as special cases. Also, SWM obeys the constraint embodied in Eq. 6.

Equations 12 and 14 were also fitted to the psychometric functions obtained in Experiments 1 and 2. The estimated parameters σ_1, σ_2, and γ^* are given in Table 1 for the average fits based on individual psychometric functions, and in Table 2 for the model fit based on the observed psychometric functions averaged across observers. Consistent with the assumption above, σ_1 is estimated to be larger than σ_2 (Experiment 1, standard = 100 ms: $t(23) = 4.5$, $p < .001$; standard = 1,000 ms: $t(23) = 6.0$, $p < .001$; Experiment 2, ISI = 1,000 ms: $t(23) = 3.8$, $p = .001$). However, for ISI = 300 ms, the estimates of σ_1 and σ_2 are almost identical, $t(23) = 1.3$, $p = .20$. Within the framework of this specific version of SWM, a natural interpretation is that a stable memory representation of $\mathbf{X}_1$ can be maintained over a relatively short interval (i.e., ISI = 300 ms) but not over a longer interval (i.e., ISI = 1,000 ms). Hence, this special case of SWM presents a plausible alternative quantitative account of the observed data.

In general, the quality of the fits, as indicated by the RMSEs, is virtually identical to the ones of IRM. The only exception can be found in the 300 ms ISI condition of Experiment 2, in which SWM provides a somewhat better fit for the individual observed psychometric functions (cf. Table 1), because it also accounts for the participants which exhibit a numerically positive Type B effect. As indicated by the group-level analyses presented above (cf. Table 2 and Footnote 4) and in the results section of Experiment 2, however, these values seem to reflect sampling error associated with the individual psychometric functions rather than a systematically higher discrimination sensitivity in $\langle cs \rangle$ than in $\langle sc \rangle$ trials.

In general, the present data therefore are well explained both by IRM and by SWM. However, it should be noted that IRM and SWM are not mutually exclusive, that is, they are not logical alternatives in the sense that when one is true the other must be false. Thus, these two models might even be merged into a single and more comprehensive account of human stimulus discrimination processes.

References

Allan, L. G. (1977). The time-order error in judgments of duration. *Canadian Journal of Psychology*, *31*, 24–31.

Allan, L. G., & Rousseau, R. (1977). Backward masking in judgments of duration. *Perception & Psychophysics*, *21*, 482–486.

Bausenhart, K. M., Dyjas, O., Vorberg, D., & Ulrich, R. (2012). Estimating discrimination performance in two-alternative forced-choice tasks: Routines for MATLAB and R. *Behavior Research Methods*, *44*, 1157–1174.

Bausenhart, K. M., Dyjas, O., & Ulrich, R. (2014). Temporal reproductions are influenced by an internal reference: Explaining the Vierordt effect. *Acta Psychologica*, *147*, 60–67.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433–436.

Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson's method. *Tutorials in Quantitative Methods for Psychology*, *1*, 42–45.

Dyjas, O., & Ulrich, R. (2014). Effects of stimulus order on discrimination processes in comparative and equality judgements: Data and models. *Quarterly Journal of Experimental Psychology*, *67*, 1121–1150.

Dyjas, O., Bausenhart, K. M., & Ulrich, R. (2012). Trial-by-trial updating of an internal reference in discrimination tasks: Evidence from effects of stimulus order and trial sequence. *Attention, Perception, & Psychophysics*, *74*, 1819–1841.

Dyjas, O., Bausenhart, K. M., & Ulrich, R. (2014). Effects of stimulus order on duration discrimination sensitivity are under attentional control. *Journal of Experimental Psychology: Human Perception and Performance*, *40*, 292–307.

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). Bingley: Emerald.

García-Pérez, M. A. (2014). Does time ever fly or slow down? The difficult interpretation of psychophysical data on time perception. *Frontiers in Human Neuroscience*, *8*, 415. doi:10.3389/fnhum.2014.00415

García-Pérez, M. A., & Alcalá-Quintana, R. (2010). Reminder and 2AFC tasks provide similar estimates of the difference limen: A re-analysis of the data from Lapid, Ulrich, & Rammsayer (2008) and a discussion of Ulrich & Vorberg (2009). *Attention, Perception & Psychophysics*, *72*, 1155–1178.

García-Pérez, M. A., & Alcalá-Quintana, R. (2011a). Improving the estimation of psychometric functions in 2AFC discrimination tasks. *Frontiers in Psychology*, *2*, 96. doi:10.3389/fpsyg.2011.00096

García-Pérez, M. A., & Alcalá-Quintana, R. (2011b). Interval bias in 2AFC detection tasks: Sorting out the artifacts. *Attention, Perception & Psychophysics*, *73*, 2332–2352.

García-Pérez, M. A., & Alcalá-Quintana, R. (2012). Correction to "reminder and 2AFC tasks provide similar estimates of the difference limen: A re-analysis of the data from Lapid, Ulrich, & Rammsayer (2008) and a discussion of Ulrich & Vorberg (2009)". *Attention, Perception, & Psychophysics*, *74*, 489–492.

García-Pérez, M. A., & Peli, E. (2014). The bisection point across variants of the task. *Attention, Perception & Psychophysics*, *76*, 1671–1697. doi:10.3758/s13414-014-0672-9

Glöckner, A., & Betsch, T. (2011). The empirical content of theories in judgment and decision making: Shortcomings and remedies. *Judgment and Decision Making*, *6*, 711–721.

Green, D. M., & Swets, J. A. (1966). *Signal Detection Theory and Psychophysics* (Rev. ed.) Los Altos, CA: Peninsula Publishing, reprinted Edition 1988.

Grondin, S., & McAuley, J. D. (2009). Duration discrimination in crossmodal sequences. *Perception*, *38*, 1542–1559.

Hellström, Å. (1979). Time errors and differential sensation weighting. *Journal of Experimental Psychology: Human Perception and Performance*, *5*, 460–477.

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

Hellström, Å. (2003). Comparison is not just subtraction: Effects of time- and space-order on subjective stimulus difference. *Perception & Psychophysics*, *65*, 1161–1177.

Hellström, Å., & Rammsayer, T. H. (2004). Effects of time-order, interstimulus interval, and feedback in duration discrimination of noise bursts in the 50- and 1000-ms ranges. *Acta Psychologica*, *116*, 1–20.

Kallman, H. J., Hirtle, S. C., & Davidson, D. (1986). Recognition masking of auditory duration. *Perception & Psychophysics*, *40*, 45–52.

Kallman, H. J., & Morris, M. D. (1984). Duration perception and auditory masking. *Annals of the New York Academy of Sciences*, *423*, 608–609.

Lages, M., & Treisman, M. (1998). Spatial frequency discrimination: Visual long-term memory or criterion setting? *Vision Research*, *38*, 557–572.

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.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.) Mahwah, New Jersey: Lawrence Erlbaum Associates.

Michels, W. C., & Helson, H. (1954). A quantitative theory of time-order effects. *The American Journal of Psychology*, *67*, 327–334.

Michon, J. A. (1985). The compleat time experiencer. In J. A. Michon, & J. L. Jackson (Eds.), *Time, mind, and behavior* (pp. 21–52). Berlin: Springer.

Nachmias, J. (2006). The role of virtual standards in visual discrimination. *Vision Research*, *46*, 2456–2464.

Noreen, D. L. (1981). Optimal decision rules for some common psychophysical paradigms. In S. Grossberg (Ed.), *Mathematical psychology and psychophysiology – Proceedings of the SIAM-AMS*, (Vol. 13, pp. 237–279). Providence, RI: American Mathematical Society.

Patching, G. R., Englund, M. P., & Hellström, Å. (2012). Time- and space-order effects in timed discrimination of brightness and size of paired visual stimuli. *Journal of Experimental Psychology: Human Perception and Performance*, *38*, 915–940.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*, 437–442.

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. (2011). Elaborative rehearsal of nontemporal information interferes with temporal processing of durations in the range of seconds but not milliseconds. *Acta Psychologica*, *137*, 127–133.

Ross, H. E., & Gregory, R. L. (1964). Is the Weber fraction a function of physical or perceived input? *Quarterly Journal of Experimental Psychology*, *16*, 116–122.

Sorkin, R. (1962). Extension of the theory of signal detectability to matching procedures in psychoacoustics. *Journal of the Acoustical Society of America*, *34*, 1745–1751.

Stott, L. H. (1935). Time-order errors in the discrimination of short tonal durations. *Journal of Experimental Psychology*, *18*, 741–766.

Thurstone, L. L. (1927a). A law of comparative judgment. *Psychological Review*, *34*, 273–286.

Thurstone, L. L. (1927b). Psychophysical analysis. *American Journal of Psychology*, *38*, 368–389.

Ulrich, R. (2010). DLs in reminder and 2AFC tasks: Data and models. *Attention, Perception, & Psychophysics*, *72*, 1179–1198.

Ulrich, R., & Vorberg, D. (2009). Estimating the difference limen in 2AFC tasks: Pitfalls and improved estimators. *Attention, Perception & Psychophysics*, *71*, 1219–1227.

Ulrich, R., Nitschke, J., & Rammsayer, T. (2006). Crossmodal temporal discrimination: Assessing the predictions of a general pacemaker-counter model. *Perception & Psychophysics*, *68*, 1140–1152.

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

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