On an Internal Reference for Perceptual Judgments

Dissertation
der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von
Ruben Ellinghaus
aus Bielefeld

Tübingen
2019

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen
Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation: 04.05.2020

Dekan: Prof. Dr. Wolfgang Rosenstiel

1. Berichterstatter: Prof. Dr. Rolf Ulrich

2. Berichterstatter: Dr. Karin Bausenhart

The unity of all science
consists in its methods
and not in its material

KARL PEARSON

Zusammenfassung

Im Widerspruch zu klassischen psychophysischen Modellen sind Urteile über sensorische Reize (e.g., die beurteilte Höhe oder Lautstärke eines Tons) keine direkte Funktion der entsprechenden physikalisch messbaren Reizeigenschaften (e.g., der Frequenz oder des Schalldrucks des enstprechenden Tons). Vielmehr hängen perzeptuelle Urteile in vielfältiger Weise auch von der zeitlichen Position der beurteilten Reize ab. Beispielsweise ist die in Diskriminationsaufgaben gemessene Sensitivität in der Regel höher, wenn ein konstanter Standardreiz einem variablen Vergleichsreiz vorangeht anstatt folgt (Typ B Effekt). Außerdem sind Beurteilungen sensorischer Reize häufig in Richtung vergangener Reize verzerrt (Sequenzeffekte und zentrale Tendenzen).

Derartige zeitliche Kontexteffekte sind vermutlich die Signatur eines Mechanismus, der gegenwärtige sensorische Information mit rezenter sensorischer Information integriert. Das Interne Referenz Modell (IRM, Dyjas, Bausenhart, & Ulrich, 2012) postuliert beispielsweise, dass Versuchspersonen in ihrem Wahrnehmungsurteil auf eine interne Referenz zurückgreifen. Die interne Referenz wird dem Modell zufolge als Konglomerat von Repräsentationen gegenwärtiger und vergangener Reizinstanzen kontinuierlich aktualisiert.

Ziel der vorliegenden Arbeit war es, den Geltungsbereich dieses Modells abzustecken, welches eine Reihe von zeitlichen Kontexteffekten vorherzusagen vermag, und außerdem die Natur der internen Referenz näher zu ergründen. In Studie 1 erweist sich IRM als allgemeingültiges Modell der Reizdiskrimination, da eine spezifische Vorhersage des Modells (Typ B Effekt) über verschiedenste Reizattribute und Modalitäten hinweg beobachtet werden konnte. Die Bildung einer internen Referenz erscheint somit als ein genereller Bestandteil menschlicher Wahrnehmungsleistungen. Studie 2 zeigt, dass sich Menschen in ihrer Urteilsbildung zu einem geringeren Anteil auf die Rezihistorie in Form der internen Referenz verlassen, wenn die Präsentation vergangener Reize zeitlich weiter zurückliegt. Die interne Referenz erscheint somit als perzeptuelle Gedächtnisrepräsentation, die zeitlichem Zerfall ausgesetzt ist. Studie 3 deutet darauf hin, dass Menschen in der Lage sind, separate interne Referenzen für zwei unterschiedliche und abwechselnd bearbeitete Diskriminationsaufgaben zu bilden. Den Ergebnissen zufolge isolieren und integrieren sie jedoch das aufgabenrelevante Merkmal in eine einzelne interne Referenz im Falle einer Diskriminationsaufgabe mit distinkten Reizen. Letzteres weist auf eine merkmalsbasierte Kodierung der internen Referenz hin.

Zusammengefasst etabliert die vorliegende Arbeit also das IRM als ein generelles, valides und reichhaltiges formales Modell eines Mechanismus, welcher der Integration gegenwärtiger und vergangener sensorischer Information in der menschlichen Wahrnehmung zugrunde liegen könnte. Die Funktion eines deratigen Mechnismus besteht vermutlich in der Stabilisierung einer perzeptuellen Repräsentation der Umwelt: Da die Welt über kurze Zeitdauern relativ stabil ist, mag es adaptiv sein, die jüngere Vergangenheit in die Bestimmung der Gegenwart einzubeziehen.

Abstract

Contrary to classic psychophysical models, one typically observes in psychophysical experiments that the judgment of sensory stimuli is not a direct function of the physical stimulus attributes. Rather, perceptual judgment also depends on the temporal position of a stimulus in various ways. For example, discrimination sensitivity is typically higher when a constant standard precedes rather than follows a variable comparison (Type B Effect). Furthermore, stimulus judgment is often biased towards past stimuli (sequence effects and central tendency effects). Potentially, such assimilatory history biases are the signature of a mechanism which integrates present sensory information with prior knowledge based on past sensory information in order to provide perceptual stability. For example, the Internal Reference Model (IRM, Dyjas et al., 2012) postulates that human judgment relies on an internal reference, which constitutes a continuously updating conglomerate of past and present stimulus instances.

Aim of the present work was to investigate the scope of this model and to clarify the nature of the internal reference. Study 1 establishes IRM as a general psychophysical model, since a specific prediction of IRM (Type B Effect) was observed for discrimination tasks across various modalities and stimulus attributes. The formation of an internal reference henceforth appears as a general component of human perceptual performance. Study 2 demonstrates that humans rely less on the stimulus history in the form of the internal reference when the presentation of past stimuli is temporally more distant. The internal reference thereby appears as a perceptual short term memory representation which is prone to decay over time. Study 3 suggests that humans form and maintain two separate internal references for two distinct discrimination tasks they alternate between. However they isolate and integrate the task-relevant stimulus attribute across different stimuli into a single internal reference in case of a single discrimination task with different stimuli, suggesting a feature-based coding of the internal reference.

In summary, the present work establishes IRM as a general, valid and theoretically rich formal model of a mechanism which potentially underlies the integration of present and past sensory information in order to optimize perceptual stability; since the world is relatively stable across short time intervals, it might be adaptive to rely on the recent past in order to reduce uncertainty about the presence.

Acknowledgments

I thank my supervisors and colleagues Karin M. Bausenhart and Rolf Ulrich for their continuous support and assistance, and for being role models concerning the scientific attitude. I am grateful for many inspiring discussions and conversations with every single one of my colleagues. I learned a lot from you not only professionally but also personally. Also I am indebted to all co-authors and reviewers for their contributions to the publications, and to my student assistants for helping with data collection.

Enclosed Journal Publications

Ellinghaus, R., Ulrich, R., & Bausenhart, K. M. (2018). Effects of stimulus order on comparative judgments across stimulus attributes and sensory modalities. *Journal of Experimental Psychology: Human Perception and Performance, 44*, 7–12. **(Study 1)**

Ellinghaus, R., Gick, M., Bausenhart, K. M., & Ulrich, R. (2019). Decay of internal reference information in duration discrimination: Intertrial interval modulates the Type B effect. *Quarterly Journal of Experimental Psychology, 72*, 1578–1586. **(Study 2)**

Ellinghaus, R., Giel, S., Ulrich, R., & Bausenhart, K. M. (2019). *Humans integrate duration information within and across modalities: Evidence for a supramodal internal reference of time.* Manuscript submitted to *Journal of Experimental Psychology: Learning, Memory, and Cognition.* **(Study 3)**

Contents

1 Introduction

The inquiry of human sensory perception stands a long tradition in Western intellectual history. In fact, the question how knowledge about the physical world is mediated through the senses was already investigated by natural philosophers before Plato (Read, 2015). It was not until the second half of the 19th century, however, that psychology emancipated from philosophy to investigate human perception scientifically. The potentially most important event in this regard was the publication of Gustav Theodor Fechners “Elemente der Psychophysik” (e.g., Fechner, 1860), who defined psychophysics in the foreword of his opus magnum as “an exact science of the relations between body and mind”. In fact, for Wilhelm Wundt, who founded the first laboratory for experimental psychology in Leipzig in 1879, Fechners quantitative approach to sensory perception was a main objection against Immanuel Kant’s earlier uttered objection that the human mind can in principal not be investigated scientifically (Wundt, 1874, p. 6).

How did Fechner bring the same scrutiny to a science of the mind that before was known to the physical sciences only? First, he extended on already existing influential contributions. For example, Ernst Heinrich Weber, who Fechner himself called “the father of psychophysics” (Fechner, 1860, p. VIII) had experimented on the human sense of touch before (Weber, 1834, as cited in Boring, 1929). He found that the minimal perceivable increase (i.e., the difference threshold) between two weights is a constant fraction of the lighter weight. For instance, if a 21 kg weight can just be distinguished from a 20 kg weight, a 21 g weight could just be distinguished from a 20 g weight. Fechner, after rigorous further experimentation and theoretical consideration extended Weber’ s insight into the famous Weber-Fechner-Law, stating that the perceived intensity of a stimulus increases logarithmically with the physical stimulus magnitude.

Second, in addition to this theoretical insight about the relation between stimulus and sensation, Fechner introduced various methods of threshold estimation to the field of psychophysics. Specifically, Fechner employed experimental procedures for threshold estimation in which participants had to repeatedly discriminate between two stimuli that differed alongside one physical continuum (e.g., two tones differing in loudness). Many of these methods are still being used in modern day experimental psychology. For example, in the common two-alternative forced-choice task (2AFC), originally developed by Hegelmaier (1852), a standard s with constant magnitude and a comparison c with magnitude varying from trial to trial are presented successively. The order of s and c also varies randomly from trial to trial, yielding trials with order $\langle cs \rangle$ and order $\langle sc \rangle$. Participants have to identify the larger stimulus at the end of each trial. To measure discrimination sensitivity (i.e., the difference threshold), one typically estimates the magnitude difference between s and c which enables identification of the larger stimulus with an accuracy level of 75% (Gescheider, 1997). This measure is conventionally defined as difference limen (DL; or just noticeable difference JND). Often times, also of interest is the point of subjective equality (PSE), which corresponds to the stimulus magnitude which is subjectively equal to the standard, and which is conventionally defined as the magnitude of c which is judged to be larger than s in 50 percent of cases.

In order to estimate DL and PSE, typically a psychometric function is fitted to the data obtained from a discrimination experiment (for a fitting procedure based on a maximum likelihood routine, see Bausenhart, Dyjas, Vorberg, & Ulrich, 2012). This psychometric function plots the probability that c is judged to be larger than s on the y-axis as a function of the physical magnitude of c on the x-axis. Any psychometric function monotonically increases from 0 to 1 with increasing values of c. Put differently, the larger the magnitude of c, the higher the probability for it being perceived as larger than the standard s will be, ranging from a probability of 0 corresponding to some comparably low value of c to a probability of 1 corresponding to some comparably high value of c. Figure 1 provides an example of a psychometric function for a hypothetical 2AFC discrimination experiment. Most importantly, DL and PSE are respectively captured by the slope and the location of the psychometric function. Specifically, the level of c which corresponds to a probability of 0.5 (i.e., the location of the psychometric function) for a "$c > s$" response defines the PSE. Furthermore, half the distance between the levels of c respectively corresponding to the 0.25 and 0.75 probabilities (i.e., half the interquartile range) defines DL. Thus, the steeper the psychometric function, the smaller the DL and hence the higher the participant's sensitivity.

1.1 Modeling Human Stimulus Discrimination

Investigating perceptual performance by means of discrimination tasks and threshold estimation has been central to psychophysics ever since the early days of experimental psychology. Importantly, during the 20th century, researchers were not only interested in quantifying and measuring human discrimination performance, but also became interested in understanding the mechanisms underlying this very basic human ability. Most of the models developed for understanding these mechanisms, as for example signal detection theory (Green & Swets, 1966; Macmillan & Creelman, 2005; Wickens, 2002) and other prominent psychophysical models (Luce & Galanter, 1963; Yeshurun, Carrasco, & Maloney, 2008), are so-called difference models of stimulus discrimination. These models are based on the pioneering work of Thurstone (1927a, 1927b), who postulated that participants base their judgement on the difference of the internal stimulus representations $\mathbf{D} = \mathbf{X}_1 - \mathbf{X}_2$, where $\mathbf{X}_1$ and $\mathbf{X}_2$ represent the internal magnitudes of the first and second stimulus in a trial, $\mathbf{S}_1$ and $\mathbf{S}_2$. In case $\mathbf{D}$ is larger than a fixed constant γ, the first stimulus is judged as the larger one, and in case $\mathbf{D}$ is smaller than γ, the second stimulus is judged as the larger one. The magnitude of these internal representations $\mathbf{X}_1$ and $\mathbf{X}_2$ in turn are a conglomerate function of the physical magnitude of the stimuli and the corre-

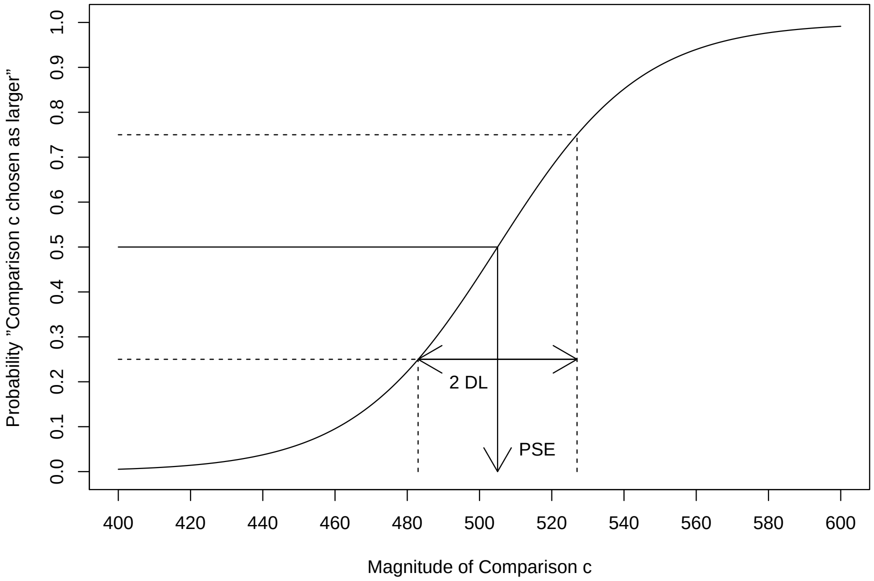

Figure 1. Hypothetical logistic psychometric function of a 2AFC duration discrimination experiment with a standard s of 500 msec. The location of the function defines the PSE, or point of subjective equality. Note that the PSE can be different from s, a phenomenon referred to as Type A effect or time order error, for example because the participant has a preference for one of the two response alternatives. Half the distance between the levels of c respectively corresponding to $x_{25\%}$ and $x_{75\%}$ (i.e., half the interquartile range) defines DL. Hence, DL is a measure of sensitivity reflected in the steepness of the psychometric function.

sponding stochastic error terms $\mathbf{E}_1$ and $\mathbf{E}_2$, modeled as random variables[1] with expected values $\mathbf{E}(\mathbf{E}_1) = \mathbf{E}(\mathbf{E}_2) = 0$. Thus, $\mathbf{S}_1$ and $\mathbf{S}_2$ are internally represented as respectively $\mathbf{X}_1$ and $\mathbf{X}_2$, with $\mathbf{X}_1 = \mathbf{S}_1 + \mathbf{E}_1$ and $\mathbf{X}_2 = \mathbf{S}_2 + \mathbf{E}_2$.

Most importantly, according to the classic difference models, perceived stimulus magnitude is essentially a mere function of physical stimulus magnitude. This is obvious, because, according to the standard difference model, except for the error term, no other value than the physical magnitude of a stimulus determines the magnitude of the internal representation of a stimulus.

---

[1] As a historical note, it should be noted that Fechner was already well aware of the random variance of the sensations caused by sensory stimuli. Concerning this matter, Fechner extended on an insight of Gauss, who found that measurement errors can be modeled as random Gaussian distributions. In detail, Fechner used this principle to infer the variances of sensations caused by sensory stimuli from the variances of stimulus judgments measured empirically. It has been argued that in doing so, Fechner was the first to uncover unobservable properties of the mind by (mathematical) deduction from observed quantities, thereby establishing psychology as a science. For a detailed discussion of this subject matter, see Link (1994).

Contrary to this basic assumption, however, it has been demonstrated that stimulus judgment also differs as a function of the relative temporal context of a stimulus. These temporal context effects can be categorized broadly into two classes. First, it has been shown that the order of the standard s and the comparison c influences discrimination performance in the 2AFC paradigm. We refer to varieties of this effect as stimulus order effects (Dyjas et al., 2012; Lapid, Ulrich, & Rammsayer, 2008; Ulrich & Vorberg, 2009). Second, it has been shown that the judged magnitude of a stimulus can be influenced by the physical magnitude of stimuli presented in previous trials of the experiment (e.g., Fischer & Whitney, 2014; Fritsche, Mostert, & de Lange, 2017). We refer to varieties of this effect as stimulus history effects.

In the following, we will inspect these temporal context effects in detail. We will then see how an extension of the classic difference models which incorporates a memory mechanism is able to account for a broad range of temporal context effects. Namely, the internal reference model (IRM, Dyjas et al., 2012; see also Lapid et al., 2008) proposes that stimulus judgment is not solely based on present stimulus representations but also on an internal reference which is a conglomerate of both previous and present stimulus instances and updates dynamically from trial to trial. In broad terms, the aim of this work is to explore the scope of the internal reference model, and to clarify the nature of the internal reference.

1.2 Stimulus Order Effects

1.2.1 Time Order Error

When experimenting on the discrimination of two successively lifted weights, Fechner (1860) observed that the first weight tended to be underestimated relative to the second weight - a phenomenon which came to be known as the negative time order error (Guilford, 1954; Hellström, 1979, 1985; Stott, 1935; Woodrow, 1935). Note that under specific circumstances, the first stimulus is overestimated relative to the second stimulus (i.e., a positive time-order error, Fechner, 1860; Köhler, 1923; Woodrow, 1933).

Although the time order has been extensively studied, its origin is still unclear. While many theories propose a response bias (i.e., a preference for one of the two response alternatives, or $\gamma \neq 0$ in terms of the standard difference model) as its cause (Yeshurun et al., 2008), other theories assume a genuine perceptual origin of the effect (Hellström, 1977).

In any case, the time order error becomes evident in a lateral shift of the psychometric function and thus, the PSE, for the two stimulus orders $\langle sc \rangle$ and $\langle cs \rangle$. That is, for trials with stimulus order $\langle sc \rangle$ the underestimation of the first stimulus implies that on average a magnitude of c smaller than s suffices in order to be perceived as equal to s. As a result, PSE will be smaller than s on $\langle sc \rangle$ trials. In contrast, for trials with stimulus order $\langle cs \rangle$ the underestimation of the first stimulus implies that on average a magnitude of c larger than s is necessary in order to be perceived as equal to s. As a result, PSE will be larger than s on $\langle cs \rangle$ trials.

1.2.2 Type B Effect

Much more essential for the present work, discrimination sensitivity as indexed by DL has also been shown to differ as a function of stimulus order in the 2AFC task. In detail, discrimination sensitivity is normally higher for trials in which the standard s precedes rather than follows the comparison c. Put differently, one typically observes a larger DL for $\langle cs \rangle$ trials than for $\langle sc \rangle$ trials. This result pattern is referred to as the negative Type B effect. Figure 2 provides an example of order-dependent psychometric functions exhibiting this effect for a hypothetical 2AFC discrimination experiment. Ulrich and Vorberg (2009) introduced the term *Type B effect* in order to distinguish this effect from the time order error, which is referred to as a *Type A effect* in their terminology. Specifically, one speaks of a negative Type B effect if DL is larger for $\langle cs \rangle$-trials than in $\langle sc \rangle$-trials, while a positive Type B effect refers to the opposite result pattern, i.e., smaller DL for $\langle cs \rangle$-trials than in $\langle sc \rangle$-trials. Almost only negative Type B effects have been reported in the literature (Dyjas et al., 2012; Lapid et al., 2008; Rammsayer, 2008; Rammsayer & Ulrich, 2012; Stott, 1935; Ulrich, 2010; Woodrow, 1935). Rare exceptional cases of positive Type B effects are limited to stimuli of short duration presented with very short interstimulus intervals (Bausenhart, Dyjas, & Ulrich, 2015, Hellström & Rammsayer, 2004, 2015).

Interestingly, and in contrast to the time-order error (Type A effect), the standard difference model cannot easily account for the Type B effect (Dyjas et al., 2012). According to these models, any subjective stimulus magnitude difference is merely a function of physical stimulus magnitudes, which certainly are independent of the temporal position or order of the stimuli. Shortly, we will see how the Type B effect can be accounted for by the Internal Reference Model (IRM, Dyjas et al., 2012; Lapid et al., 2008). Importantly, in contrast to the Type A effect, the Type B effect has almost exclusively been investigated for duration discrimination (for an excepction, see, e.g., Nachmias, 2006).

1.3 Stimulus History Effects

1.3.1 Central Tendency Effects

So far, we considered the effects that the relative temporal position of a stimulus within a trial can have on its perceived magnitude. Interestingly, not only the relative position of a stimulus within a trial, but also its relative position within the range of experienced stimuli can serve as a context which influences how the stimulus is perceived (e.g., Helson, 1964).

For example, it has long been known that the judged magnitude of a stimulus is often biased towards the average of the stimulus range employed in an experiment — a phenomenon referred to as the central tendency of judgment (Hollingworth, 1910): "judgments of time, weight, force, brightness, extent of movement, length, area, size of angles all show the same tendency to gravitate toward a mean magnitude" (pp 462–462). For instance Vierordt (1868) observed in a series of temporal reproduction tasks that relatively long temporal intervals

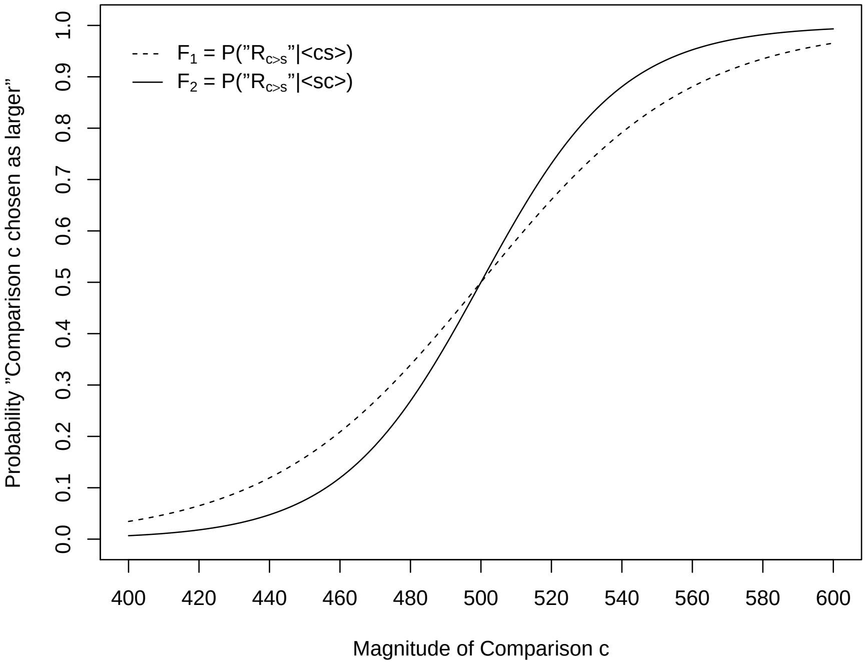

Figure 2. Hypothetical order-dependent logistic psychometric functions $F_1 = P(\text{“}R_{c>s}\text{”}|\langle cs\rangle)$ and $F_2 = P(\text{“}R_{c>s}\text{”}|\langle sc\rangle)$ of a 2AFC duration discrimination experiment with a standard s of 500 msec. Note that both functions have identical locations corresponding to an identical PSE for both stimulus orders. However, F_2 is steeper than F_1, implying a larger DL for $\langle cs\rangle$ trials than for $\langle sc\rangle$ trials and hence exhibiting a negative Type B effect.

are underestimated and relatively short temporal intervals are overestimated — reproduced durations gravitate towards the mean of the temporal intervals employed in the experiment (see also Bausenhart, Dyjas, & Ulrich, 2014). Central tendency effects are often attributed to humans exploiting statistical regularities in sensory stimulation for the sake of perceptual stability (Cicchini, Arrighi, Cecchetti, Giusti, & Burr, 2012; Jazayeri & Shadlen, 2010).

Similar evidence for effects of the whole stimulus distribution on perceptual judgments comes from a bisection task experiment by Brown, McCormack, Smith, and Stewart (2005). In this task, participants were initially presented with two standard durations, a long and a short one. Subsequently, they judged each of the successively presented comparison stimuli to be more similar to the long or to the short duration. The authors found that the bisection point (the duration that is judged equally close to the long or to the short duration) was lower for more positively skewed than for negatively skewed distributions. Henceforth, the perceived durations depended on the overall distribution of comparisons.

1.3.2 Assimilatory Sequence Effects

Similarly, the judged magnitude of a stimulus in a given experimental trial is often attracted towards a stimulus magnitude experienced on the immediately preceding trial (Garner, 1953). For example, participants in a study by Fischer and Whitney (2014) had to reproduce the orientation of a randomly orientated Gabor patch (a sinusoid grating often used in vision research) on each trial, and the reported orientations were systematically biased in the direction of the orientation presented on the preceding trial. For instance, when the Gabor on the preceding trial was oriented more clockwise than the Gabor on the present trial, participants adjusted the reproduced Gabor more clockwise than its actual orientation. This result indicates that visual perception is serially dependent in such a way that both past and present sensory information inform the representation of a stimulus.

Furthermore, Dyjas et al. (2012) found in a 2AFC duration discrimination experiment in blocks with trial order $\langle cs \rangle$, PSE to be shifted to the left (i.e., smaller than s) when c on the preceding trial was relatively long, and shifted to the right (i.e., larger than c) when c on the preceding trial was relatively short. Henceforth, c appeared relatively longer (shorter) in trial n when c in trial n - 1 was relatively long (short), indicated by the fact that a duration shorter (longer) than s was sufficient (necessary) to cause a sensation equal to s. In other words, the perceived magnitude of c in trial n gravitated towards the physical magnitude c in trial n - 1, thereby conceptually mirroring the aforementioned results obtained by Fischer and Whitney (2014).

Like central tendency effects, such assimilatory sequence effects have been documented for various tasks and domains, including both basic sensory attributes such as visual line length discrimination (Ashourian & Loewenstein, 2011), auditory frequency discrimination (Arzounian, Kerangal, & Cheveign, 2017; Chambers & Pressnitzer, 2014; Raviv, Ahissar, & Loewenstein, 2012; Raviv, Lieder, Loewenstein, & Ahissar, 2014), weight (Parducci & Marshall, 1962) and loudness (Decarlo & Cross, 1990; King & Lockhead, 1982; Ward, 1979), and also more abstract properties such as facial attractiveness (Geiselman, Haight, & Kimata, 1984; Taubert, Burg, & Alais, 2016) and judgment of price (Matthews & Stewart, 2009), as well as for different tasks including absolute identification (Donkin, Chan, & Tran, 2017; Stewart, Brown, & Chater, 2005), categorization and exemplar production (Zotov, Jones, & Mewhort, 2011), and passive viewing (Fornaciai & Park, 2018).

1.4 The Internal Reference Model

The temporal context effects outlined above demonstrate that human perception is not merely a function of the physical characteristics of the respective sensory stimuli. Rather, the temporal position of a stimulus within the experiment partly determines its perceived magnitude. Specifically, stimulus representations appear to be biased towards past stimuli, as evident in sequence effects and central tendency effects. How can we make sense of these phenomena on a theoretical level? Modern approaches to perception assume that we combine noisy sensory information with prior knowledge based on statistical regularities in the environment. In that context, perception is often regarded as a process of drawing inferences about external events and stimuli, based on both past and present internal states (e.g., Chater, Tenenbaum, & Yuille, 2006; Friston & Kiebel, 2009; von Helmholtz, 1867; Kersten, Mamassian, & Yuille, 2004; Kersten & Yuille, 2003; Knill & Richards, 1996; Petzschner, Glasauer, & Stephan, 2015; Pouget, Beck, Ma, & Latham, 2013; Yuille & Kersten, 2006). In fact, since the world is fairly stable across short temporal intervals (Dong & Atick, 1995), often it might be adaptive to predict the presence based on the past (see also Burr & Cicchini, 2014; Cicchini, Anobile, & Burr, 2014; Cicchini, Mikellidou, & Burr, 2017; Fischer & Whitney, 2014).

The details of the processes by which prior knowledge is acquired and past and present sensory information is combined in order to guide perception, however, remain to be investigated. Several researchers (e.g., Arzounian et al., 2017; Dyjas et al., 2012; Fischer & Whitney, 2014; Kalm & Norris, 2018; Raviv et al., 2014) have suggested that humans combine the memory representations of past and present stimulus representations (for an overview with an emphasis on the temporal domain, see, Bausenhart, Bratzke, & Ulrich, 2016; van Rijn, 2016). For example, according to the Internal Reference Model (IRM, Dyjas et al., 2012), participants form an internal reference which combines past and present stimulus instances (see also Durlach & Braida, 1969; Morgan, Watamaniuk, & McKee, 2000; Nachmias, 2006). Specifically, the internal reference follows a geometrically moving average and updates continuously from trial to trial. To illustrate, recall the typical 2AFC task in which the participant chooses the larger of two successively presented stimuli, $\mathbf{S}_1$ and $\mathbf{S}_2$, one of which is the standard s with constant magnitude and one of which is the comparison c with variable magnitude. According to IRM, the internal reference on trial n is a weighted sum of the first stimulus' internal representation $\mathbf{X}_{1,n}$ on the current trial n and the internal reference $\mathbf{I}_{n-1}$ from the previous trial,

$$\mathbf{I}_n = g \cdot \mathbf{I}_{n-1} + (1 - g) \cdot \mathbf{X}_{1,n} \tag{1}$$

with constant weight g, $0 \leq g < 1$. The internal reference is then used as a referent in the comparison process. That is, if $\mathbf{D}_n = \mathbf{I}_n - \mathbf{X}_{2,n} > 0$, participants judge the first stimulus to be the larger one and the second stimulus otherwise.

Interestingly, for values of $0 < g < 1$, IRM predicts that DL for stimulus order $\langle cs \rangle$ is larger than the DL for stimulus order $\langle sc \rangle$, i.e., a negative Type B effect. This holds intuitively as long as the values of c are distributed symmetrically around the standard. In that case, the internal representation of c will generally be pulled towards s on $\langle cs \rangle$ trials, because the expected value of $\mathbf{I}_n$ is equal to s. Consequently, the subjective difference $\mathbf{D}_n$ will be decreased, and discrimination sensitivity will be lower (DL will be larger) compared to $\langle sc \rangle$ trials. Note that IRM was originally developed to account for the Type B order effect observed in the auditory 2AFC duration discrimination task (Dyjas et al., 2012; Lapid et al., 2008) and an equality judgment task (Dyjas & Ulrich, 2014). Nervertheless, the model has been shown to account for a broad range of temporal context effects in the domain of time perception (Bausenhart et al., 2016), as for example the Vierordt effect (Bausenhart et al., 2014), and assimilatory sequence effects (Dyjas et al., 2012).

Put differently, IRM specifies mathematically a mechanism by which prior knowledge based on statistical regularities is combined with present sensory input in order to guide perception. However, the predictions of the model have not yet been investigated outside the temporal domain. For example, we do not know whether the formation of an internal reference is specifically prominent in the temporal domain. Furthermore, nothing is yet known about the nature of this internal reference. For instance, we do not know how stable the internal reference is and what information is coded in it.

1.5 Aim of the Present Work

The present work was dedicated to a better understanding of the internal reference as proposed by IRM. This is an important endeavor, because it might add to a better understanding of the mechanisms underlying temporal context effects. This in turn might enrich our knowledge about the processes by which present and past information are integrated in human perception to construct a model of the world. In order to achieve this goal, a series of three related studies based on 2AFC discrimination experiments investigated the Type B effect to address a) the scope of the internal reference model (Study 1) and b) the nature of the internal reference (Studies 2 and 3).

Study 1 addressed the scope of IRM by investigating the generality of the Type B effect. Specifically, we tested whether the Type B effect emerges across stimulus attributes and modalities, by carrying out an experimental series including a number of discrimination tasks (e.g., intensity, numerosity) across the visual and auditory modality. In terms of IRM, this would inform about whether the formation of an internal reference constitutes a general component of human perceptual performance rather than being restricted to the temporal domain.

Study 2 investigated the stability of the internal reference. Specifically, the literature suggested that the internal reference might be a perceptual memory representation which is prone to decay over time. Furthermore, since the Type B effect is a direct result of the internal reference, its magnitude should be reduced once the internal reference has decayed to a certain extend. Accordingly, we hypothesized that the Type B effect is reduced with a longer inter-trial-interval (ITI) compared to a shorter ITI in 2AFC duration discrimination task.

Study 3 investigated the coding of the internal reference. For instance, it is conceivable that the internal reference merely codes task-relevant stimulus information (e.g., consists of an abstract code for duration in temporal discrimination tasks), but it is also conceivable that the internal reference constitutes a comprehensive object-like representation of the stimulus carrying the task-relevant information (e.g., a representation of a sine tone with a certain duration, loudness, and so forth). In order to distinguish between various alternatives of the type of coding potentially underlying the internal reference representation, we varied task-irrelevant stimulus attributes between trials in a series of multiple 2AFC discrimination experiments. For example, in Experiment 3.1, the tones marking the temporal intervals in an auditory discrimination task either constantly stayed at a high or low pitch from trial to trial, or alternated between a high and a low pitch from trial to trial. The expectations and theoretical implications of these experiments will be explained in detail below.

2 Scope of the Internal Reference Model (Study 1)[2]

2.1 Aim of the Study

The standard difference model (Thurstone, 1927a, 1927b) provides an intuitive and straightforward account of the mechanisms underlying discrimination performance. It is therefore not surprising that most prominent psychophysical models, such as signal detection theory (Green & Swets, 1966; Macmillan & Creelman, 2005; Wickens, 2002), are still based on this model. As shown by Dyjas et al. (2012), these difference models imply that DL does not differ as a function of stimulus order.

Empirically, however, sensitivity is typically lower for $\langle cs \rangle$-trials than for $\langle sc \rangle$-trials (Dyjas et al., 2012; Lapid et al., 2008; Rammsayer, 2008; Rammsayer & Ulrich, 2012; Stott, 1935; Ulrich, 2010; Woodrow, 1935). In contrast to difference models, the internal reference model can account for this negative Type B effect. Namely, in its standard form it predicts that the Type B effect is either negative or absent (Dyjas et al., 2012).

Psychophysical models such as IRM make general predictions across domains. Accordingly, the Type B effect should emerge in discrimination tasks across various stimulus attributes and modalities. However, so far the Type B effect has almost exclusively been investigated for temporal discrimination (e.g., Dyjas et al., 2012; Dyjas, Bausenhart, & Ulrich, 2014; Hellström & Rammsayer, 2015; Grondin & McAuley, 2009; Lapid et al., 2008; Ulrich, 2010). For rare exceptions outside the temporal domain, see Nachmias (2006) and Ross and Gregory (1964), who report stimulus order dependent DL analyses for the discrimination of shapes and weights, respectively.

From a theoretical perspective it is of course important to know whether the predictions of IRM apply only to the temporal domain or rather generalize across stimulus attributes and modalities. Empirical evidence on this issue would be informative regarding the question whether reliance on an internal reference as specified by IRM is a general component of human perceptual performance rather than a specific phenomenon in the domain of temporal cognition. The latter might be expected based on the idea that the cognitive representation of temporal information might be especially fragile or transient, as there is no sensory system specifically dedicated to the processing of time. Accordingly, judgment tasks in the temporal domain might be more exposed to contextual effects (e.g., Bausenhart et al., 2016; Bruno & Cicchini, 2016; Heron et al., 2012; Jazayeri & Shadlen, 2010).

---

[2] **Ellinghaus, R.**, Ulrich, R., & Bausenhart, K. M. (2018). Effects of stimulus order on comparative judgments across stimulus attributes and sensory modalities. *Journal of Experimental Psychology: Human Perception and Performance*, 44, 7–12.

2.2 Experiments & Results

To investigate the generality of the Type B Effect systematically, 24 subjects in a within-subjects design performed 2AFC discrimination tasks for four stimulus attributes (i.e., duration, frequency, intensity, and numerosity), each in the visual and in the auditory modality. For example, for the auditory (visual) intensity discrimination task, participants had to discriminate sine tones (grey circles) differing in loudness (brightness). For the auditory (visual) duration discrimination task, participants had to discriminate sine tones (grey circles) differing in duration. For the auditory (visual) frequency discrimination task, participants had to discriminate sine tones (Gabor patches) differing in frequency. For the auditory (visual) numerosity discrimination task, participants discriminated the number of pulses (dots) in an auditory pulse train (random-dot pattern). In order to realise an efficient way of threshold estimation, an adaptive procedure was used in which the magnitude of c adapted to participants' response behavior following the weighted up-down method (Kaernbach, 1991). For each participant, DL was estimated for each discrimination task. Figure 2 of Ellinghaus, Ulrich, and Bausenhart (2018) depicts boxplots of DL as a function of stimulus order and task. Most importantly, for all tasks, median DL was larger for stimulus order $\langle cs \rangle$ than for $\langle sc \rangle$, revealing a negative Type B effect. Furthermore, the relative size of this negative Type B effect was relatively constant and thus generalizes across the investigated tasks and modalities.

2.3 Theoretical Implications

In a series of 2AFC discrimination tasks for various stimulus attributes and modalities, discrimination sensitivity in terms of DL was consistently larger for stimulus order $\langle cs \rangle$ than for order $\langle sc \rangle$. The present results thereby demonstrate that the negative Type B effect is not restricted to temporal discrimination but instead constitutes a ubiquitous phenomenon across modalities and stimulus attributes. In other words, the Type B effect appears to be the signature of a general mechanism in human perception.

According to IRM, this mechanism consists in dynamical updating of an internal reference, which conglomerates past and present stimulus representations. On a more general level, IRM is consistent with the notion that humans integrate sensory information with prior knowledge based on statistical regularities to guide perception (e.g., Friston & Kiebel, 2009; Raviv et al., 2012; Shi, Church, & Meck, 2013). As mentioned in the Introduction, such an integration mechanism might be adaptive in constructing a stable representation of the environment. Because the world is relatively stable across short temporal intervals, the recent past is a relatively good predictor of the present state of the world.

Importantly, the present results cast doubts on the validity of psychophysical models which are based on the original difference model, and thus imply that the perceived magnitude difference of two stimuli reduces to the physical magnitude difference of these stimuli. Furthermore, the present study establishes the Type B effect as a benchmark effect for future models of stimulus discrimination.

3 Nature of the Internal Reference (Study 2 and Study 3)

The results of Ellinghaus et al. (2018) suggest that the formation of an internal reference as specified by IRM is a ubiquitous and general mechanism by which past and present sensory information are combined to guide human perception. However, IRM is phrased purely abstract in mathematical terms and therefore remains agnostic about the nature of the internal reference. In order to shed light on this nature, two further studies were designed. Namely, Study 2 investigated the temporal stability of the internal reference, and Study 3 investigated what is coded in the internal reference (i.e., the representational format of the internal reference).

3.1 Temporal Stability of the Internal Reference (Study 2)[3]

3.1.1 Aim of the Study

Study 2 investigated the nature of the internal reference in terms of its temporal stability. Previous research suggested that sensory representations stored in memory decay as time passes (see Lu, Williamson, & Kaufman, 1992, for a detailed examination of this subject matter). For example, discrimination performance has been found to decrease when the temporal interval between the two stimuli to be discriminated (Inter-Stimulus-Interval, ISI) increases (e.g., Crowder, 1982; Magnussen, Thomas, & Greenlee, 1996). Similarly, when the temporal interval between trials (Inter-Trial-Interval, ITI) is increased, the influence of past stimuli on the judgment of present stimuli has been found to decrease, too (Decarlo, 1992; Jaffe-dax, Frenkel, & Ahissar, 2017). For example, Jaffe-dax et al. (2017) found that stimuli were less biased towards recent stimuli evident in both a psychophysical and an electroencephalographic measure when the ITI was increased. In the light of these findings, it seemed reasonable to assume that the internal reference constitutes a residual perceptual trace or memory representation which is prone to decay over time. Phrased in terms of IRM, participants might rely less on past stimulation in the form of an internal reference when it has already decayed to a certain extend. Since according to IRM the Type B effect is a direct result of participants relying on the internal reference, the Type B effect should be reduced in this case.

3.1.2 Experiment & Results

This latter hypothesis was tested by 24 participants fulfilling an auditory 2AFC discrimination task with a long and a short ITI condition. By the decay logic

---

[3] **Ellinghaus, R.**, Gick, M., Bausenhart, K. M., & Ulrich, R. (2019). Decay of internal reference information in duration discrimination: Intertrial interval modulates the Type B effect. *Quarterly Journal of Experimental Psychology, 72*, 1578–1586.

explained above, we expected a reduced Type B effect for the long ITI condition compared to the short ITI condition.

Discrimination sensitivity in terms of DL was estimated for each participant for each ITI condition and stimulus order. As can be seen in Figure 2 of Ellinghaus, Gick, Ulrich, and Bausenhart (2019), DL was larger for $\langle cs \rangle$-trials than for $\langle sc \rangle$-trials in both conditions for most participants, revealing the typical negative Type B effect. Most importantly, this stimulus order dependent sensitivity difference was larger for the short ITI condition than for the long ITI condition. This latter result pattern indicates a reduced negative Type B effect for the long ITI condition as compared to the short ITI condition.

In addition to this analysis, we fitted the theoretical psychophysical functions of IRM individually to each participant. The empirical and fitted psychometric functions averaged across participants for both ITI conditions are depicted in Figure 3 of Ellinghaus, Gick, et al. (2019). Consistent with our theoretical considerations, the estimated values of g were smaller for the long ITI condition than for the short ITI condition.

3.1.3 Theoretical Implications

In this 2AFC duration discrimination experiment, a meaningfully reduced Type B effect was observed when the temporal interval between trials was extended. First, these results further establish the Type B effect as a robust phenomenon by replicating earlier observations of the Type B effect (Dyjas et al., 2012; Ellinghaus et al., 2018; Dyjas & Ulrich, 2014; Bausenhart et al., 2015; Lapid et al., 2008). Furthermore, recall that according to IRM, the Type B effect is a direct result when participants partly rely on past stimulus information in the form of an internal reference. Therefore, the present observation of a reduced Type B effect for a prolonged ITI in a 2AFC duration discrimination task suggests that participants relied less on the internal reference, potentially due to decay of the corresponding memory trace during the time between trials.

In accordance with this observation, a model fit of IRM revealed a smaller estimated value of the weighting factor g for the long ITI condition than for the short ITI condition. Thus, in terms of IRM, the internal reference $\mathbf{I}_{n-1}$ is apparently integrated to a lesser extent with $\mathbf{X}_{1,n}$ when relatively more time has passed between two subsequent trials. Probably, this is so because the internal reference has already decayed to a certain extent, e.g., gotten too noisy (Ashourian & Loewenstein, 2011; Lee & Harris, 1996), after a relatively long temporal interval. Therefore, the present results suggest that the internal reference constitutes a perceptual short term memory representation, which is prone to decay in a similar fashion as perceptual memory representations in general (e.g., Magnussen et al., 1996).

On a general level, humans seem to rely less on past stimulus information when this past information is temporally more distant, and base their decisions more on current stimulus input instead.

3.2 Coding of the Internal Reference (Study 3)[4]

Study 1 and 2 established IRM as a valid and general model for perceptual prior formation in general and memory mixing in particular. However, the question still remains what exactly is mixed in memory. For example, it is conceivable that humans can generalize across stimuli varying alongside feature dimensions or modalities to form a perceptual prior which codes the task-relevant information only. Evidence in this direction stems from a study by Roach, Mcgraw, Whitaker, and Heron (2017), wherein participants had to manually reproduce the duration of stimuli presented in different modalities (visual vs. auditory) and locations (left or right to the centered fixation cross), and the durations of different stimuli were drawn from different distributions. For example, visual stimuli could range in duration around a mean of 320 msec and auditory stimuli could range around a mean of 1280 msec. Most importantly, when trials from both distributions were presented interleaved, reproductions were shifted towards the mean of the combined distribution. This result suggests that participants formed a duration prior by generalizing across stimuli that differed with respect to task-irrelevant stimulus features such as location and modality.

On the contrary, however, results reported by Rhodes, Seth, and Roseboom (2018) suggest that participants rather form separate priors for different stimuli, henceforth potentially coding a composite of task-relevant and task-irrelevant features. In detail, participants in this latter study had to indicate whether a stimulus appeared late or early relatively to what they expected based on a series of three stimuli separated by a constant inter stimulus interval. Participants‘ responses were biased towards the inducing stimuli, indicating that they formed a perceptual prior. Interestingly, auditory stimuli were biased towards a prior induced by auditory stimuli and visual stimuli were biased towards a prior induced by visual stimuli. The same pattern was observed when stimuli did not vary between modalities but within the auditory modality, i.e., white noise stimuli were biased towards white noise stimuli and pure tones were biased towards pure tones. Importantly, however, the differentiation was not observed for stimuli varying between low and high pitch. This result pattern indicates that participants formed specific perceptual priors based on the task-irrelevant characteristics of the stimuli which marked the task-relevant durations, at least as long as stimulus types were sufficiently distinct. In case stimuli varied only in pitch, participants apparently generalized across stimulus distributions.

Reconciling these findings in the context of IRM, at least two possibilities concerning the coding of the internal reference come to mind. On the one hand, it appears that participants might isolate and extract the task-relevant sensory feature into the internal reference, as for example the duration of the presented stimuli when performing a duration discrimination task. One may in this case speak of a feature-based coding of the internal reference, because

---

[4] **Ellinghaus, R.**, Giel, S., Ulrich, R., & Bausenhart, K. M. (2019). *Humans integrate duration information within and across modalities: Evidence for a supramodal internal reference of time.* Manuscript submitted to *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

only the task-relevant feature of the sensory stimuli enters the internal reference representation. In that case, the internal reference might for example consist of an abstract code for duration in any duration discrimination task, independent of whether the interval lengths to be discriminated are marked by, for example, low frequency or high frequency tones or by, for example, visual or auditory stimuli. In this case, the internal reference would correspond to a common perceptual prior for the different stimulus types. On the other hand, it also seems plausible that participants might form separate internal references for different stimuli. One may in this case speak of a stimulus-based coding of the internal reference, because the internal reference would consist in a composite representation including both task-relevant and task-irrelevant features. In that case, the internal reference representation might for example consist in a composite representation of a sine tone with a certain duration, a certain frequency and a certain loudness in an auditory duration discrimination task.

For the present study, four 2AFC discrimination experiments were carried out in order to investigate the coding of the internal reference. The first two experiments focused on differentiating between the two aforementioned different types of coding potentially underlying the internal reference, i.e., feature-based coding vs. stimulus-based coding. For that matter, stimuli either stayed constant between trials or varied trial-by-trial along a certain task-irrelevant stimulus dimension. Note that in case humans form separate stimulus-based internal references, these separate internal references would have to be stored in memory for a relatively longer temporal interval in the variable condition before the same stimulus is repeated and the internal reference can be updated. As a consequence, one should observe a reduced Type B effect in the variable condition compared to the constant condition due to relatively stronger memory decay of the internal reference.

A third experiment investigated a third variant of coding, namely multiple-feature-coding (see below). Finally, a fourth experiment differentiated between abstract feature-based coding and sensory feature-based coding. The hypotheses and the logic of these experiments will be outlined below in the corresponding sections 3.3 and 3.4.

3.2.1 Feature-Based Coding vs. Stimulus-Based Coding (Experiments 3.1 and 3.2)

In Experiment 3.1, 48 participants performed a 2AFC duration discrimination task wherein the frequency of the sine tones to be discriminated either stayed constant at a high pitch or low pitch from trial to trial (constant condition) or alternated between a high and a low pitch between trials (variable condition). Averaged across conditions, a typical negative Type B effect was observed, again indicating that participants relied on an internal reference integrating past and present stimulus instances when comparing the stimuli. Most importantly, as can be seen in Figure 1 of Ellinghaus, Giel, Ulrich, and Bausenhart (2019), the magnitude of the Type B effect did not meaningfully differ between the variable and the constant condition. This result renders unlikely that participants formed separate internal references for the high frequency and the low frequency tones in the alternating condition. In the case of separate internal references, each of the internal references associated with respectively one of the two frequencies most likely would have decayed to a substantial degree when the corresponding frequency is repeated after two trials, that is, after an effective ITI of 6600 msec, hence resulting in a reduced Type B effect (cf. Ellinghaus, Gick, et al., 2019). Rather, the observed result pattern is consistent with the study by Roach et al. (2017) reporting that participants can integrate temporal information for prior formation across stimuli varying alongside task-irrelevant attributes. Specifically, in the context of IRM, the observed result pattern is in line with the idea that participants can isolate and extract the task-relevant temporal information into the internal reference when performing the discrimination task.

Experiment 3.2 was a conceptual replication of Experiment 3.1. Again, 24 participants completed an auditory 2AFC duration discrimination task. The frequency of the sine tones either varied randomly from trial to trial between 15 different frequencies (variable condition) or stayed again constant during blocks at the midpoint pitch of these 15 sine tones (constant condition). Conceptually mirroring the main result of Experiment 3.1, a reliable negative Type B effect similar in magnitude within the constant and the variable condition was observed. This result pattern renders highly unlikely that participants formed separate (stimulus-based) internal references, for two reasons. First, one can speculate that maintaining 15 separate internal references most likely causes a memory load exceeding short term memory capacity. Second, each internal reference most certainly would have decayed before it could be updated after 14 trials on average. Thus, the present results strengthen the conclusion drawn from Experiment 3.1, namely that participants form a single internal reference encoding the task-relevant duration information irrespective of other task-irrelevant attributes. Thus, in sum, Experiment 3.1 and 3.2 suggest a feature-based coding rather than a stimulus-based coding of the internal reference.

3.2.2 Single-Feature Coding vs. Multiple-Feature Coding (Experiment 3.3)

As explained in the previous section, the data of Experiment 3.1 and 3.2 are inconsistent with the hypothesis that participants formed separate (stimulus based) internal references. However, we cannot distinguish whether participants integrated only the task-relevant feature or also task-irrelevant features into a single internal reference. Note that features such as frequency and loudness are quantitative in nature and thus can theoretically be integrated in the same manner as duration. Empirically, this has been demonstrated by the observed Type B effects for frequency and loudness discrimination tasks in Ellinghaus et al. (2018) Henceforth, we cannot exclude that the internal reference in Experiment 3.1 and 3.2 consisted in a composite representation coding a conglomerate of these various integrated feature values. In the following, we refer to this hypothesis as *multiple-feature-coding*, as apposed to the aforementioned *single-feature coding*.

Furthermore, recall that the task-irrelevant stimulus attribute varied in Experiment 3.1 and 3.2 was frequency, and Rhodes et al. (2018) argued that temporal intervals marked by stimuli of different frequencies may be too similar to stimulate the formation of different perceptual priors. Thus, it is conceivable that participants form separate internal references for different stimuli as long as these stimuli are sufficiently distinct.

In order to explore these variants of coding (i.e., multiple-feature coding and separate internal references in case of sufficient distinctness), a third 2AFC duration discrimination experiment employing higher distinctness of the stimuli regarding task-irrelevant features was carried out with 24 participants. Specifically, in Experiment 3.3, the stimuli which marked the temporal intervals either varied between the auditory and the visual modality (variable condition) from trial to trial or constantly stayed in the visual or auditory modality from trial to trial (constant condition). Note that integration of task-irrelevant features across these distinct stimuli (i.e., multiple-feature coding) seems very implausible in this case due to the qualitative differences associated with different modalities. Henceforth, a Type B effect of similar size in the variable and the constant condition would suggest that participants isolated and extracted the task-relevant duration information into a single internal reference. On the contrary, a reduced Type B effect should be observed for the variable condition compared to the constant condition in case participants formed separate internal references for the auditory and visual stimuli, respectively.

Again conceptually mirroring the main result pattern of Experiment 3.1 and Experiment 3.2, a reliable negative Type B effect was observed, and this effect did not meaningfully differ between the constant and the variable condition. Henceforth, it appears that even if the stimuli are highly distinct (i.e., presented in different modalities), participants seem to isolate and extract the task-relevant duration information into one single internal reference (single-feature coding).

3.2.3 Separate Internal References for Distinct Discrimination Tasks (Experiment 3.4)

In this last 2AFC discrimination experiment, we aimed to assess the Type B effect and thus the formation of the internal reference across stimuli that vary along both task-relevant and task-irrelevant features in an intensity discrimination task. Specifically, participants either constantly discriminated the loudness of two tones or the brightness of two light circles in each trial or they alternated between these auditory and visual intensity discrimination tasks from trial to trial. On the one hand, since intensity is a stimulus dimension arranged alongside a quantitative continuum, participants might be able to integrate it into a single internal reference across these distinct tasks and stimulus types, in line with the idea of common magnitude-based representation of the stimulus dimension intensity (e.g., Walsh, 2003). On the other hand, the distinctness of the two stimulus types is even greater than in Experiment 3.3 and hence might be sufficient to cause the formation of separate internal references (cf. Rhodes et al., 2018). In that case, a reduced Type B effect should be observed due to temporal decay of the internal reference, as outlined above. Replicating all previous experiments, the typical negative Type B effect was observed averaged across conditions. Most importantly, however, and in contrast to the previous experiments of this study, the magnitude of the Type B effect now differed between the constant and the variable condition, that is, the magnitude of the Type B effect was meaningfully reduced for the variable condition. This result is consistent with the hypothesis that participants formed separate internal references for the two tasks/stimulus types. Other than might have been predicted based on common magnitude theories (e.g., Walsh, 2003), the data render unlikely that participants can integrate and store the intensity information across the distinct stimuli as a general and abstract code. Rather, variation of both task-relevant and task-irrelevant features rendered stimulus types sufficiently distinct to enable the formation of separate (stimulus-based) internal references, which were prone to decay while the respective other stimulus type was processed.

3.2.4 Theoretical Implications

To summarize, this study contained a series of four 2AFC discrimination experiments. These were designed to investigate the coding of the internal reference by employing discrimination tasks in which stimulus types either stayed constant between trials or varied between trials. In Experiments 3.1 - 3.3 the stimuli marking the temporal intervals in a duration discrimination task either varied with regard to a task-irrelevant stimulus attribute between trials or stayed constant regarding all task-irrelevant stimulus attributes from trial to trial. Specifically, in Experiments 3.1 and Experiment 3.2, stimuli varied in frequency and in Experiment 3.3 stimuli varied in modality (visual vs. auditory) between the trials of a duration discrimination task. In all of these first three experiments, the Type B effect was not meaningfully reduced when stimuli varied between trials compared to when stimuli stayed constant across trials. Contrary to what might have been expected based on the observation of separate perceptual priors reported by Rhodes et al. (2018), this result indicates that participants did not form separate internal references for the different stimuli. In that case, the internal reference associated with a specific frequency or modality most likely would have decayed to a substantial degree before repetition of the respective stimulus type, implying a reduced Type B effect (cf. Ellinghaus et al., 2018). However, our observation is coherent with Roach et al. (2017) reporting that participants can generalize across stimuli varying alongside task-irrelevant attributes for the formation of a single perceptual prior. Specifically, in the context of IRM, participants potentially isolated and extracted the duration information into the internal reference when performing the discrimination tasks. Therefore, Experiment 3.1 - 3.3 suggest that the coding of the internal reference is feature-based (coding only the task-relevant stimulus feature) rather than stimulus-based (encoding a composite representation of both task-relevant and task-irrelevant features). With specific regards to duration perception, these results correspond well with theories suggesting that time is represented as an abstract amodal code (Wearden, Todd, & Jones, 2006) or a supramodal auditory code (Bratzke, Seifried, & Ulrich, 2012).

Finally, in Experiment 3.4, both task-irrelevant and task-relevant features were varied in a non-temporal task. In this experiment, participants alternated between an auditory intensity (i.e., loudness) and a visual intensity (i.e., brightness) discrimination task alternated from trial to trial in the variable condition. Interestingly, in this case, the Type B effect was meaningfully reduced compared to when participants performed either the loudness or the brightness discrimination task in isolation (constant condition). Following the same temporal decay logic as in the previous experiments, this result renders unlikely that participants were able to integrate the intensity information across the distinct stimuli to maybe store it in the form of a general magnitude-based representation, since in this case no reduction of the Type B effect should have been observed. Such a common code as suggested by common coding theories (e.g., Walsh, 2003) arguing for a common representational system for quantitative stimulus dimensions such as space, time, number, and intensity, thus does not seem to form the basis of the internal reference. Rather, it appears that participants formed separate internal references for the two distinct stimulus types or discrimination tasks.

Taken together, humans seem to integrate the task-relevant information into a single internal reference across stimuli differing with respect to task-irrelevant features such as frequency or modality (Experiments 3.1 - 3.3). However, in case stimuli differ with respect to both task-relevant and task-irrelevant features and hence the distinct stimulus types are associated with different discrimination tasks, humans seem to form separate internal references for the distinct stimulus types (Experiment 3.4). This latter interpretation is also coherent with Rhodes et al. (2018) arguing that separate perceptual priors are only formed when stimulus types are sufficiently distinct. Importantly, the present study also establishes the Type B phenomenon as a useful tool to tackle these specific research questions about perceptual prior formation.

4 General Discussion

Before we turn to a general discussion of the reported studies, let us first summarize the preceding chapters. Contrary to influential psychophysical models (Green & Swets, 1966; Macmillan & Creelman, 2005; Wickens, 2002; Luce & Galanter, 1963; Yeshurun et al., 2008), the judged magnitude of a stimulus (e.g., the perceived pitch of a sine tone) is not merely a function of the physical stimulus magnitude (e.g., the frequency of the tone). Rather, stimulus judgment also varies with the relative temporal position of a stimulus. For example, one often observes in psychophysical experiments that sensory stimuli are perceptually attracted towards immediately preceding stimuli (sequence effects, e.g., Fischer & Whitney, 2014; Garner, 1953) or towards the mean of the distribution of the stimulus magnitudes presented in the experiment (central tendency effects, e.g., Hollingworth, 1910; Bausenhart et al., 2014) Also, discrimination sensitivity in terms of DL has been found to depend on the order of the standard s and the comparison c in the 2AFC task, an intriguing phenomenon referred to as the Type B effect (Lapid et al., 2008; Dyjas et al., 2012; Ulrich & Vorberg, 2009). In detail, DL is typically lower (i.e., sensitivity is higher) for $\langle sc \rangle$ trials compared to $\langle cs \rangle$ trials. Potentially, such temporal context effects are a signature of a mechanism which integrates present sensory information with prior knowledge based on past sensory information. For example, IRM (Dyjas et al., 2012) postulates that human judgment relies on an internal reference, which constitutes a conglomerate of past and present stimulus instances and updates continuously from trial to trial.

Importantly, IRM can account for a broad range of temporal context effects, including sequence effects, central tendency effects, and the negative Type B effect. The aim of the present work was to investigate the scope of this model and to clarify the nature of the internal reference. In Study 1 it was shown that although IRM was originally developed to account for the effects of stimulus order on duration discrimination performance (Dyjas et al., 2012; Lapid et al., 2008), the Type B effect is not restricted to the domain of temporal perception, but instead reflects a general phenomenon across modalities and stimulus attributes. Henceforth, the formation of an internal reference as specified by IRM appears to be a ubiquitous phenomenon of human perception.

Study 2 and Study 3 investigated the nature of the internal reference. Specifically, in Study 2 it was shown that the magnitude of the Type B effect decreases when the temporal interval between trials increases. Viewed through the goggles of IRM, this result pattern suggests that humans rely less on previous stimulus information when its initial presentation is temporally more distant, potentially due to partial temporal decay of the internal reference. This is in line with the idea the internal reference constitutes a perceptual short term memory representation which is prone to decay over time, similar to perceptual short term memory representations in general (e.g., Lu et al., 1992; Magnussen et al., 1996).

Finally, in Study 3 it was shown that the coding of the internal reference might be feature-based (i.e., representing only the task-relevant stimulus attribute) rather than stimulus-based (i.e., consisting in a composite representation of the multiple features constituting the stimuli). Specifically, the results of Experiments 3.1 - 3.3 suggested that participants seemingly isolated and extracted the task-relevant feature (i.e., duration) into one single internal reference when stimuli varied with respect to task irrelevant features such as frequency (Experiment 3.1 and 3.2) or modality (Experiment 3.3). Furthermore, the results of Experiment 3.4 suggested that participants formed separate internal references for the distinct stimuli associated with different discrimination tasks (i.e., a loudness discrimination task and a brightness discrimination task participants alternated between). This latter interpretation is coherent with Rhodes et al. (2018) arguing that separate perceptual priors are only formed when stimulus types are sufficiently distinct.

Taken together, the present work further establishes IRM as an empirically valid formal model of a mechanism potentially underlying temporal context effects. Accordingly, the fomation of an internal reference as specified by IRM turns out to be a plausible candidate mechanism underlying the combination of present sensory information with prior knowledge based on past sensory experience in perception. Furthermore, although IRM itself is phrased purely abstract and thus agnostic about various psychological aspects of the mechanism underlying the formation of the internal reference, the studies of the present work enrich our understanding of these very psychological aspects, specifically with regard to the temporal stability and the coding of the internal reference representation. Importantly, the present work also establishes the Type B effect as a useful empirical tool to tackle specific research questions about perceptual prior formation in general and internal reference formation in particular.

4.1 Putting IRM into Context

4.1.1 Potential Function of the Internal Reference

Many recently developed related models (e.g., Arzounian et al., 2017; Kalm & Norris, 2018; Raviv et al., 2012; Shi et al., 2013) share with IRM the basic assumption that the representation of a currently judged stimulus is biased towards or mixed with memory representations of previously presented stimuli. From a theoretical perspective, it is interesting to speculate about the potential function of such a mechanism. In that regard, it has been suggested that reliance on the immediate past serves the function of providing a stable representation in the presence of a noisy and uncertain environment. As the world is fairly stable across short temporal intervals, humans potentially effectively reduce uncertainty about the present by relying on the recent past (Chater et al., 2006; Pouget et al., 2013). Consistent with this assumption, predictive coding theories of perception and cognition assume that humans constantly predict forthcoming events based on internal models of the world (Friston & Kiebel, 2009). Interestingly, recycling of previous sensory information might also be efficient in the sense of saving processing resources and thus reducing cortical processing (Manassi, Liberman, Kosovicheva, Zhang, & Whitney, 2018).

It should be noted, however, that reliance on the past for the sake of perceptual stability obviously stands in conflict with another functional requirement of the perceptual system, namely to be adaptive and to be sensitive to change. Put differently, in case one relies too heavily on the past in order to determine the present, changes in the environment might go unnoticed or be falsely overwritten by outdated information from memory. On the other hand, in case one relies too little on the past in order to determine the present, a valid and efficient source of information might be disregarded, potentially resulting in unstable or noisy perceptual representations. On a general level, then, it seems that the perceptual system needs to balance out these two opposing needs for stability and flexibility/adaptability.

Consistent with these considerations, human reliance on perceptual priors has been found to be conditional on the specific processing demands and stimulus characteristics of a given situation. For example, the relative weighting of past stimulus information has been found to depend on stimulus reliability. For instance, Cicchini, Mikellidou, and Burr (2018) varied stimulus reliability in an orientation reproduction task by varying the spatial frequency and orientation of Gabor patches. In this case, assimilatory sequence effects were strongest when stimulus reliability was low (e.g., low spatial frequency oblique Gabors). This result pattern conceptually extends to the domain of numerosity judgments, were low numerosities, which are represented more reliably, are known to exhibit less assimilatory serial dependencies (Cicchini et al., 2014). Thus, humans potentially in general rely on past stimulus information to a larger extent when current stimulus information is less reliable and uncertain, presumably because the need for compensation is higher in this case (see also Kersten et al., 2004).

Coherently, the weighting factor g which modulates the relative weighting of past and present stimulus information in the context of IRM, has also been found to vary flexibly with the strength or certainty of stimulus information. For example, (Dyjas et al., 2014) observed that the magnitude of the Type B effect (and hence the weighting factor g) decreases when participants were informed about the order of s and c in a 2AFC discrimination task. Thus, in case a valid cue about stimulus position increased the informational value of the comparison stimulus, participants in this study relied less on remote stimulus information in the form of the internal reference. Interestingly, Study 2 of the present work (Ellinghaus, Gick, et al., 2019) suggests that participants rely less on the internal reference when the representation has already decayed (e.g., the representation of the internal reference has gotten too noisy) to a substantial degree after a longer compared to a shorter ITI. Here, humans seem to rely less on previous information when it has become more uncertain compared to present information.

Taken together, it appears that humans may flexibly adjust the relative weighting of past and present information depending on the respective strength or noisiness of past and present representations. One could speculate that humans integrate past and present information in an optimal fashion to maximize perceptual stability, depending on the respective reliability of the two information sources. It might be interesting to test this in an orthogonal design were past and present stimulus reliability are varied independently. For example, reliability of past stimuli could be lowered by introducing various lengths of ITI (cf., Ellinghaus, Gick, et al., 2019), and reliability of present stimuli could be lowered by adding various levels of external noise. Assimilatory history effects such as sequence effects or the Type B effect should be the strongest (weakest) were the reliability of present stimuli is the lowest (highest) and the reliability of past stimuli is the highest (lowest).

4.1.2 The Role of Stimulus Similarity: Distinct Priors/References for Distinct Stimuli?

Interestingly, assimilatory history effects have also been found to be dependent on the proximity of past and present stimuli. As mentioned above, participants in the study by Fischer and Whitney (2014) had to reproduce the orientation of a randomly orientated Gabor on each trial, and the reported orientations were systematically biased in the direction of the orientation presented on the preceding trial. For instance, when the Gabor on the preceding trial was oriented more clockwise than the Gabor on the present trial, participants adjusted the reproduced Gabor more clockwise than its actual orientation. Interestingly, these positive sequential dependencies were stronger for spatially and temporally proximal stimuli, which led these authors to postulate a spatiotemporal continuity field wherein which stimulus information is integrated over time. Similarly, stimuli have been found to be biased towards similar past stimuli more strongly than towards more distinct past stimuli. For example, in the aforementioned studies by Cicchini et al. (2018) and Cicchini et al. (2014), sequential effects were positively modulated by stimulus similarity. A positive relation between stimulus similarity and assimilation has also been reported by several recent studies (Liberman, Fischer, & Whitney, 2014; Liberman, Zhang, & Whitney, 2016; Xia, Leib, & Whitney, 2016).

Similarly, Study 3 of the present work suggests that participants separately aggregate stimulus information from different stimulus types into separate internal references, as long as stimulus types are sufficiently distinct. For example, the results of Experiment 3.4 suggested that participants formed separate internal references for the brightness and the loudness task they alternated between. In case stimuli differed in pitch or in modality between trials of a duration discrimination task, participants rather integrated the task-relevant information across these distinct stimulus types (Experiments 3.1-3.3). Accordingly, the picture that arises is that humans integrate information across moderately distinct stimuli, but separately aggregate stimulus information from highly distinct stimulus types or different tasks. From a functional perspective, this seems to be adaptive intuitively, since stimuli of the same type (i.e., sharing many features or belonging to the same task) or in spatiotemporal proximity are potentially more likely to be related and/or to stay constant across time compared to stimuli which either do not share many features, require different judgment or actions, or are spatially or temporally distant.

Another factor related to separate prior formation might be distinctness regarding motor outputs. In the temporal reproduction study by Roach et al. (2017), multiple separate duration priors were formed when the respective stimulus distributions were coupled to different motor responses, such as different forms of duration judgment or different motor systems. In case distributions differed only by stimulus characteristics such as spatial location, temporal sequence, or sensory modality, participants formed a single prior that generalized across distributions. Similarly, Nagai, Suzuki, Miyazaki, and Kitazawa (2012) reported that participants formed a general prior in a temporal order judgment task when stimulus distributions were separated by color, however multiple priors were formed when the two colors also served as a cue for the participants to respectively shift their gaze up or down. Finally, Gekas, Chalk, Seitz, and Seriès (2013) demonstrated that participants can learn separate motion direction priors for distributions separated by color but only when one of the distributions was a vertically flipped image of the other one.

Apparently, humans in some cases differentiate between, and in other cases integrate across stimulus distributions when forming priors, while it is not yet clear what separates the former from the latter cases. One can speculate stimuli need to differ from each other to a certain extent before multiple priors are formed. As suggested by Roach et al. (2017), differences in motor outputs might be a crucial factor in this regard. Intuitively, this makes sense since humans (and other intelligent agents) need to perceptually differentiate between objects that need to be acted upon differently in order to enable goal-directed behavior.

Yet it should be noted that in some studies (e.g. Gekas et al., 2013; Kerrigan & Adams, 2013; Rhodes et al., 2018; Ellinghaus, Giel, et al., 2019) mere perceptual differences were decisive as to whether participants generalized or separated distributions for prior formation. Clearly, more research is needed to further investigate the boundary conditions of generalization vs. separation in prior acquisition and the question how stimulus distinctness influences perceptual prior formation is at best partly solved at present. In any case, however, the Type B effect and IRM offer a useful theoretical and empirical framework for studying specific research questions devoted to this issue.

For example, the influence of distinct motor outputs and perceptual differences on internal reference formation may be studied by varying these factors independently in a 2AFC discrimination experiment. For example, participants could give responses alternatingly with the left or right hand, or they could alternate between hand and foot responses. Additionally, stimuli could vary regarding task-irrelevant attributes such as modality or frequency, akin to Study 3 of the present work. In such a design, it might be interesting to investigate under which particular combinations of perceptual and motor differences participants would form separate internal references rather than integrate across distinct stimulus types and/or motor output systems. Such a study might especially help to close a research gap between the studies on the Type B effect on the one hand and assimilatory biases such as central tendencies and sequence effects on the other hand, thereby informing about the relation between these phenomena.

4.1.3 Adaptive vs. Repulsive History Effects and the Functional Locus of the Type B Effect

Interestingly, stimuli are in some circumstances contrasted away rather than assimilated towards previous stimulus instances. A prime example in this regard might be the well-known tilt-aftereffect (Gibson & Radner, 1937), where the perceived orientation of a grating (e.g., a Gabor) is contrasted away from a previously seen grating. For instance, a vertical grating may appear slightly tilted clockwise after being presented with a grating tilted counter-clockwise. It is generally assumed that the mechanisms underlying such repulsive effects serve the requirement of the human perceptual system to be adaptive and sensitive to change (e.g., Clifford, Wyatt, Arnold, Smith, & Wenderoth, 2001; Chopin & Mamassian, 2012; Müller, Metha, Krauskopf, & Lennie, 1999). Another example of repulsive history effects stems from absolute identification tasks, where participants have to identify a stimulus from a set of items. In these tasks one typically observes assimilation to the response of the preceding trial but contrast to the response of earlier trials (e.g., Mori & Ward, 1995).

Recently, there has been an ongoing debate about the functional loci of attractive and repulsive sequence effects, i.e., whether they respectively arise late or early in the processing stream, ranging from early sensory to post-decisional stages of processing (e.g., Fornaciai & Park, 2018; Fritsche et al., 2017; Frund, Wichmann, & Macke, 2014; St.John-Saaltink, Kok, Lau, & de Lange, 2016). Within this line of research, it has been suggested that assimilatory history biases emerge in decisional or motor stages while repulsive effects emerge earlier in perceptual stages (Braun, Urai, & Donner, 2018; Fritsche et al., 2017; Pascucci, Mancuso, Santandrea, Libera, & Plomp, 2017). However, perception-based assimilatory effects have also been demonstrated (Fornaciai & Park, 2018; St.John-Saaltink et al., 2016). More specifically, Gekas, McDermott, and Mamassian (2019) observed different directions of perceptual serial effects at different timescales. These authors investigated the influence of stimulus history on the responses given in an orientation judgment task. Importantly, participants in this design had to attend but not to react to every stimulus presented, thereby minimizing response or motor biases. Interestingly, the immediate past exerted a positive influence on present stimulus judgments, while the recent past exerted a negative influence and the more distant past exerted again a positive influence. In sum, the literature suggests that while response-based history effects are generally positive, perceptual history effects may be both attractive and repulsive in nature.

Connecting the present work to this debate, it should be noted that the Type B effect cannot easily be explained as a response-driven phenomenon, since response history does not systematically vary between $\langle cs \rangle$ and $\langle sc \rangle$ trials. Rather, the present work suggests that the Type B effect is based on assimilation of stimulus representations held in short term memory towards persisting memory traces of previously stored stimulus instances. Consistent with this presumption, the Type B effect is absent for brief (100 msec) stimuli separated by brief (300 msec) ISIs (Bausenhart et al., 2015), potentially because there is not enough time for the representation of the first stimulus to be integrated with representations of earlier stimuli. Some studies even report evidence for positive Type B effects for very brief stimuli (Hellström & Rammsayer, 2004, 2015). Coherently, Fritsche et al. (2017) observed that attractive serial biases occur only when judgment occurs some time after stimulus offset, whereas a repulsive bias occurs in case a judgment is made at the moment of perception, providing further evidence that stimulus representations are being assimilated towards memory traces of past stimuli while being held in short term memory. Furthermore, Bliss, Sun, and D'Esposito (2017) varied the delay between stimulus and response in a spatial delayed response task. Conceptually mirroring the aforementioned study by Fritsche et al. (2017), they observed that responses given some time after stimulus offset but not responses given directly after stimulus offset showed evidence of positive serial dependence, strengthening the short term memory interpretation of positive assimilatory effects. Nevertheless, Manassi et al. (2018) demonstrated that attractive serial dependencies can be present at the moment of perception.

In any case, it should be noted that IRM in its original formulation can not account for repulsive history effects, since IRM implies history biases to be either positive or absent. Importantly, however, observations of repulsive history effects do by no means invalidate IRM, since there is no obvious reason to assume that the mechanisms which are respectively responsible for repulsive and assimilatory history effects cannot in principle be effective at the same time. In fact, is has been suggested that attractive effects stem from higher-level attention-based processing (Bliss et al., 2017; Fritsche et al., 2017; Fritsche & de Lange, 2019; Dyjas et al., 2014), whereas repulsion effects stem from lower level adaptation (Kohn, 2007), and both types of mechanisms might run in parallel to balance out the need for perceptual stability on the one hand and change sensitivity on the other hand (Czoschke, Fischer, Beitner, Kaiser, & Bledowski, 2019; Fornaciai & Park, 2019).

4.1.4 Interrelations with Short Term Memory Research

Interestingly, sequence effects resembling psychophysical assimilation effects such as the Type B effect have been reported for paradigms designed to study short term or working memory. For example, the likelihood of a probe item to be incorrectly identified as part of a current array is higher when that particular item was part of an array presented on the previous trial (Makovski & Jiang, 2008). Similarly, the target location of a saccade can be biased to the location of a target in the previous trial (Papadimitriou, Ferdoash, & Snyder, 2019). Typically, such effects are explained as a form of positive interference of the representations in short term or working memory. As Kiyonaga, Scimeca, Bliss, and Whitney (2017) point out, positive sequence effects observed in both working memory and perceptual judgment tasks might be commonly attributed to memory traces persisting across trials. In detail, these authors adhere to the fact that both working memory tasks and psychophysical judgment tasks require subjects to make a response based on the features of a stimulus some time after stimulus offset, and hence representations of current stimuli might be attracted to previous stimuli while they are being stored in short term memory until the judgment is being made. This view is coherent with the aforementioned studies demonstrating that the magnitude of positive history effects correlates positively with the temporal interval between stimulus offset and stimulus judgment (Bliss et al., 2017; Gekas et al., 2019). In sum, assimilatory dependencies in perceptual judgment and interference effects in short term or working memory, although studied within separate lines of research, seem to be related, and memory mixing models such as IRM might to some extent offer a general theoretical framework under which these phenomena can be reconciled.

4.2 Conclusion

The present thesis aimed at a better understanding of the internal reference model as an explanation for temporal context effects. The resulting contribution might be subsumed under two interrelated aspects. First, it was shown that although IRM was originally developed to account for order and sequence effects in duration discrimination, the model generalizes across stimulus attributes and modalities. Accordingly, the formation of an internal reference for perceptual judgments turns out to be a general component of human perceptual performance. Second, the present work informs about psychological aspects of the model which it is agnostic about in its original form, such as the representational nature of the internal reference in terms of its coding and its temporal stability.

On a more general level, by enriching our understanding of the formation of internal references as a general perceptual process, the present work adds to a timely discussion about the question how present and recent sensory information is integrated in human perception in order to ensure perceptual stability. In doing so, two separate lines of research are bridged, namely research on the stimulus order effects originating in traditional psychophysics on the one hand and (Dyjas et al., 2012; Fechner, 1860; Ulrich & Vorberg, 2009) and research on perceptual priors and sequential dependencies originating in computational biology or neuroscience on the other hand (Akrami, Kopec, Diamond, & Brody, 2018; Chater et al., 2006; Fischer & Whitney, 2014; Friston & Kiebel, 2009; Pouget et al., 2013). Although many issues about the integration of past and previous sensory information in human perception remain unsettled, a picture arises in which various disciplines including psychology and neuroscience each applying their own methods can collectively contribute to an understanding of this subject matter. Recently, theorists have argued that such an integrative perspective is beneficial regarding research on perception and cognition in general (Krakauer, Ghazanfar, Gomez-Marin, MacIver, & Poeppel, 2017; Kelber, Evangelisti, & Ellinghaus, 2020). These theoretical considerations draw heavily on the heritage of famous vision scientist David Marr, who argued that mental processes are best understood if tackled at the cognitive level via behavioral data and computational modeling first before questions about the neuronal realisations of these processes may be addressed. This view might be best illustrated in his famous quote: “Trying to understand perception by studying only neurons is like studying bird flight by studying only feathers, it simply cannot be done” (Marr, 1982, p. 27). In that regard, the present work may set an impetus for future investigation, since it provides an example how knowledge about perceptual processes can be acquired by mathematical modeling and behavioral data alone, independent of asking how these mechanisms might eventually be realised by neuronal processes (see also Ulrich, 2009). Finally, it may also be noteworthy that the classic psychophysical methods such as the 2AFC paradigm, which were introduced by G. T. Fechner and hence are as old as the scientific inquiry of the human mind itself, are still valuable for addressing timely research questions today. This might speak for the enduring quality and importance of these research tools.

References

Akrami, A., Kopec, C. D., Diamond, M. E., & Brody, C. D. (2018). Posterior parietal cortex represents sensory history and mediates its effects on behaviour. *Nature*, *554*, 368.

Arzounian, D., Kerangal, M. D., & Cheveign, A. D. (2017). Sequential dependencies in pitch judgments. *The Journal of the Acoustical Society of America*, *142*, 3047–3057.

Ashourian, P., & Loewenstein, Y. (2011). Bayesian inference underlies the contraction bias in delayed comparison tasks. *PloS one*, *6*, e19551.

Bausenhart, K. M., Bratzke, D., & Ulrich, R. (2016). Formation and representation of temporal reference information. *Current Opinion in Behavioral Sciences*, *8*, 46–52.

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.

Bausenhart, K. M., Dyjas, O., & Ulrich, R. (2015). Effects of stimulus order on discrimination sensitivity for short and long durations. *Attention, Perception, & Psychophysics*, *77*, 1033–1043.

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.

Bliss, D. P., Sun, J. J., & D'Esposito, M. (2017). Serial dependence is absent at the time of perception but increases in visual working memory. *Scientific Reports*, *7*, 14739.

Boring, E. G. (1929). *A history of experimental psychology.* New York: London: The Century CO.

Bratzke, D., Seifried, T., & Ulrich, R. (2012). Perceptual learning in temporal discrimination : asymmetric cross-modal transfer from audition to vision. *Experimental Brain Research*, *221*, 205–210.

Braun, A., Urai, A. E., & Donner, T. H. (2018). Adaptive history biases result from confidence-weighted accumulation of past choices. *Journal of Neuroscience*, *38*, 2418–2429.

Brown, G. D. A., McCormack, T., Smith, M., & Stewart, N. (2005). Identification and bisection of temporal durations and tone frequencies: Common models for temporal and nontemporal stimuli. *Journal of Experimental Psychology: Human Perception and Performance*, *31*, 919–938.

Bruno, A., & Cicchini, G. M. (2016). Multiple channels of visual time perception. *Current Opinion in Behavioral Sciences*, *8*, 131–139.

Burr, D., & Cicchini, G. M. (2014). Vision : Efficient Adaptive Coding. *Current Biology*, *24*, R1096–R1098.

Chambers, C., & Pressnitzer, D. (2014). Perceptual hysteresis in the judgment of auditory pitch shift. *Attention, Perception, & Psychophysics*, *76*, 1271–1279.

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, *10*, 287–291.

Chopin, A., & Mamassian, P. (2012). Predictive properties of visual adaptation. *Current biology*, *22*, 622–626.

Cicchini, G. M., Anobile, G., & Burr, D. C. (2014). Compressive mapping of number to space reflects dynamic encoding mechanisms, not static logarithmic transform. *Proceedings of the National Academy of Sciences*, 201402785.

Cicchini, G. M., Arrighi, R., Cecchetti, L., Giusti, M., & Burr, D. C. (2012). Optimal Encoding of Interval Timing in Expert Percussionists. *The Journal of Neuroscience*, *32*, 1056–1060.

Cicchini, G. M., Mikellidou, K., & Burr, D. (2017). Serial dependencies act directly on perception. *Journal of Vision*, *17*, 1–9.

Cicchini, G. M., Mikellidou, K., & Burr, D. C. (2018). The functional role of serial dependence. *Proceedings of the Royal Society B*, *285*, 20181722.

Clifford, C. W., Wyatt, A. M., Arnold, D. H., Smith, S. T., & Wenderoth, P. (2001). Orthogonal adaptation improves orientation discrimination. *Vision research*, *41*, 151–159.

Crowder, R. G. (1982). Decay of auditory memory in vowel discrimination. *Journal of Experimental Psychology: Learning Memory and Cognition*, *8*, 153–162.

Czoschke, S., Fischer, C., Beitner, J., Kaiser, J., & Bledowski, C. (2019). Two types of serial dependence in visual working memory. *British Journal of Psychology*, *110*, 256–267.

Decarlo, L. T. (1992). Intertrial interval and sequential effects in magnitude scaling. *Journal of Experimental Psychology: Human Perception and Performance*, *18*, 1080–1088.

Decarlo, L. T., & Cross, D. V. (1990). Sequential effects in magnitude scaling: Models and theory. *Journal of Experimental Psychology: General*, *119*, 375–396.

Dong, D. W., & Atick, J. J. (1995). Statistics of natural time-varying images. *Network : Computation in Neural Systems Statistics of natural time-varying images*, *6*, 345–358.

Donkin, C., Chan, V., & Tran, S. (2017). The effect of blocking inter-trial interval on sequential effects in absolute identification. *The Quarterly Journal of Experimental Psychology*, *68*, 129–143.

Durlach, N. I., & Braida, L. D. (1969). Intensity perception. I. Preliminary theory of intensity resolution. *The Journal of the Acoustical Society of America*, *46*, 372–383.

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.

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

Ellinghaus, R., Gick, M., Ulrich, R., & Bausenhart, K. M. (2019). Decay of internal reference information in duration discrimination : Intertrial interval modulates the Type B effect. *Quarterly Journal of Experimental Psychology*, *72*, 1578–1586.

Ellinghaus, R., Giel, S., Ulrich, R., & Bausenhart, K. M. (2019). *Humans integrate duration information within and across modalities: Evidence for a supramodal internal reference of time.* (Mansucript submitted for publication)

Ellinghaus, R., Ulrich, R., & Bausenhart, K. M. (2018). Effects of stimulus order on comparative judgments across stimulus attributes and sensory modalities. *Journal of Experimental Psychology: Human Perception and Performance*, *44*, 7–12.

Fechner, G. T. (1860). *Elemente der Psychophysik* [Elements of psychophysics]. Leipzig, Germany: Breitkopf und Härtel.

Fischer, J., & Whitney, D. (2014). Serial dependence in visual perception. *Nature Neuroscience*, *17*, 738–743.

Fornaciai, M., & Park, J. (2018). Attractive serial dependence in the absence of an explicit task. *Psychological Science*, *29*, 437–446.

Fornaciai, M., & Park, J. (2019). Spontaneous repulsive adaptation in the absence of attractive serial dependence. *Journal of vision*, *19*, 21–21.

Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society*, *364*, 1211–1221.

Fritsche, M., & de Lange, F. P. (2019). The role of feature-based attention in visual serial dependence. *bioRxiv*, 584789.

Fritsche, M., Mostert, P., & de Lange, F. P. (2017). Opposite effects of recent history on perception and decision. *Current Biology*, *27*, 590–595.

Frund, I., Wichmann, F. A., & Macke, J. H. (2014). Quantifying the effect of intertrial dependence on perceptual decisions. *Journal of Vision*, *14*, 9–9.

Garner, W. R. (1953). An informational analysis of absolute judgements of loudness. *Journal of Experimental Psychology*, *46*, 373–380.

Geiselman, R. E., Haight, N. A., & Kimata, L. G. (1984). Context effects on the perceived physical attractiveness of faces. *Journal of Experimental Social Psychology*, *20*, 409–424.

Gekas, N., Chalk, M., Seitz, A. R., & Seriès, P. (2013). Complexity and specificity of experimentally induced expectations in motion perception. *BMC Neuroscience*, *14*, P355.

Gekas, N., McDermott, K. C., & Mamassian, P. (2019). Disambiguating serial effects of multiple timescales. *Journal of vision*, *19*, 24–24.

Gescheider, G. A. (1997). *Psychophysics: The fundamentals* (3rd ed.). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Gibson, J. J., & Radner, M. (1937). Adaptation, after-effect and contrast in the perception of tilted lines. i. quantitative studies. *Journal of Experimental Psychology*, *20*, 453.

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.

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

Hegelmaier, F. (1852). Ueber das Gedächtniss für Linear-Anschauungen [On memory for visually perceived lines]. *Archiv für physiologische Heilkunde*, *11*, 844–853.

Hellström, Å. (1977). Time errors are perceptual. *Psychological Research*, *39*, 345–388.

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., & 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.

Hellström, Å., & Rammsayer, T. H. (2015). Time-order errors and standard-position effects in duration discrimination: An experimental study and an analysis by the sensation-weighting model. *Attention, Perception, & Psychophysics*, *77*, 2409–2423.

Helson, H. (1964). *Adaptation-level theory.* New York: Harper & Row.

Heron, J., Aaen-Stockdale, C., Hotchkiss, J., Roach, N. W., McGraw, P. V., & Whitaker, D. (2012). Duration channels mediate human time perception. *Proceedings of the Royal Society B: Biological Sciences*, *279*, 690–698.

Hollingworth, H. (1910). The central tendency of judgment. *The Journal of Philosophy, Psychology and Scientific Methods*, *7*, 461–469.

Jaffe-dax, S., Frenkel, O., & Ahissar, M. (2017). Dyslexics ' faster decay of implicit memory for sounds and words is manifested in their shorter neural adaptation. *eLife*, *6*, 1–19.

Jazayeri, M., & Shadlen, M. N. (2010). Temporal context calibrates interval timing. *Nature Neuroscience*, *13*, 1020–1026.

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

Kalm, K., & Norris, D. (2018). Visual recency bias is explained by a mixture model of internal representations. *Journal of vision*, *18*, 1–1.

Kelber, P., Evangelisti, T. H., & Ellinghaus, R. (2020). *Zwischen Reduktion und Integration: Interdisziplinäre Theoriebildung in der Wahrnehmungsforschung.* (Mansucript in preparation)

Kerrigan, I. S., & Adams, W. J. (2013). Learning different light prior distributions for different contexts. *Cognition*, *127*, 99–104.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as bayesian inference. *Annual Review of Psychology*, *55*, 271-304.

Kersten, D., & Yuille, A. (2003). Bayesian models of object perception. *Current Opinion in Neurobiology*, *13*, 150-158.

King, M. C., & Lockhead, G. R. (1982). Response scales and sequential effects in judgment. *Perception & Psychophysics*, *30*, 599–603.

Kiyonaga, A., Scimeca, J. M., Bliss, D. P., & Whitney, D. (2017). Serial dependence across perception, attention, and memory. *Trends in Cognitive Sciences*, *21*, 493–497.

Knill, D. C., & Richards, W. (Eds.). (1996). *Perception as bayesian inference*. New York, NY, USA: Cambridge University Press.

Köhler, W. (1923). Zur Theorie des Sukzessivvergleichs und der Zeitfehler. *Psychological Research*, *4*, 115–175.

Kohn, A. (2007). Visual adaptation: physiology, mechanisms, and functional benefits. *Journal of neurophysiology*, *97*, 3155–3164.

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., & Poeppel, D. (2017). Neuroscience needs behavior: correcting a reductionist bias. *Neuron*, *93*, 480–490.

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.

Lee, B., & Harris, J. (1996). Contrast transfer characteristics of visual short-term memory. *Vision Research*, *36*, 2159–2166.

Liberman, A., Fischer, J., & Whitney, D. (2014). Serial dependence in the perception of faces. *Current Biology*, *24*, 2569–2574.

Liberman, A., Zhang, K., & Whitney, D. (2016). Serial dependence promotes object stability during occlusion. *Journal of vision*, *16*, 16–16.

Link, S. W. (1994). Rediscovering the past: Gustav Fechner and signal detection theory. *Psychological Science*, *5*, 335–340.

Lu, Z., Williamson, S. J., & Kaufman, L. (1992). Behavioral lifetime of human auditory sensory memory predicted by physiological measures. *Science*, *258*, 1668–1670.

Luce, R. D., & Galanter, E. (1963). Discrimination. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. I, pp. 191–243). New York: John Wiley & Sons.

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

Magnussen, S., Thomas, J. P., & Greenlee, M. W. (1996). Parallel processing in visual short-term memory. *Journal of Experimental Psychology: Human Perception and Performance*, *22*, 202–212.

Makovski, T., & Jiang, Y. V. (2008). Proactive interference from items previously stored in visual working memory. *Memory & Cognition*, *36*, 43–52.

Manassi, M., Liberman, A., Kosovicheva, A., Zhang, K., & Whitney, D. (2018). Serial dependence in position occurs at the time of perception. *Psychonomic Bulletin & Review*, 2245–2253.

Marr, D. (1982). Vision: A computational investigation into the human representation and processing of visual information.

Matthews, W. J., & Stewart, N. (2009). Psychophysics and the judgment of price: Judging complex objects on a non-physical dimension elicits sequential effects like those in perceptual tasks. *Judgment and Decision Making*, *4*, 64–81.

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.

Mori, S., & Ward, L. M. (1995). Pure feedback effects in absolute identification. *Perception & Psychophysics*, *57*, 1065–1079.

Müller, J. R., Metha, A. B., Krauskopf, J., & Lennie, P. (1999). Rapid adaptation in visual cortex to the structure of images. *Science*, *285*, 1405–1408.

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

Nagai, Y., Suzuki, M., Miyazaki, M., & Kitazawa, S. (2012). Acquisition of multiple prior distributions in tactile temporal order judgment. *Frontiers in Neuroscience*, *3*, 1–7.

Papadimitriou, C., Ferdoash, A., & Snyder, L. H. (2019). Ghosts in the machine : memory interference from the previous trial. *American Journal of Physiology-Heart and Circulatory Physiology*, 567–577.

Parducci, A., & Marshall, L. M. (1962). Assimilation vs. contrast in the anchoring of perceptual judgments of weight. *Journal of Experimental Psychology*, *63*, 426–437.

Pascucci, D., Mancuso, G., Santandrea, E., Libera, C. D., & Plomp, G. (2017). Laws of concatenated perception : Vision goes for novelty, Decisions for perseverance. *bioRxiv*, 1–45.

Petzschner, F. H., Glasauer, S., & Stephan, K. E. (2015). A bayesian perspective on magnitude estimation. *Trends in Cognitive Sciences*, 1–9.

Pouget, A., Beck, J. M., Ma, W. J., & Latham, P. E. (2013). Probabilistic brains: knowns and unknowns. *Nature neuroscience*, *16*, 1170.

Rammsayer, T. H. (2008). Neuropharmacological approaches to human timing. In S. Grondin (Ed.), *Psychology of time* (pp. 295–320). Bingley: Emerald Group.

Rammsayer, T. H., & Ulrich, R. (2012). The greater temporal acuity in the reminder task than in the 2AFC task is independent of standard duration and sensory modality. *Canadian Journal of Experimental Psychology*, *66*, 26–31.

Raviv, O., Ahissar, M., & Loewenstein, Y. (2012). How recent history affects perception: The normative approach and its heuristic approximation. *PLoS Computational Biology*, *8*, e1002731.

Raviv, O., Lieder, I., Loewenstein, Y., & Ahissar, M. (2014). Contradictory behavioral biases result from the influence of past stimuli on perception. *PLoS Computational Biology*, *10*, e1003948.

Read, J. C. A. (2015). The place of human psychophysics in modern neuroscience. *Neuroscience*, *296*, 116–129.

Rhodes, D., Seth, A. K., & Roseboom, W. (2018). Multiple duration priors within and across the senses. *bioRxiv*, 467027.

Roach, N. W., Mcgraw, P. V., Whitaker, D. J., & Heron, J. (2017). Generalization of prior information for rapid Bayesian time estimation. *Proceedings of the*

National Academy of Sciences, *114*, 412–417.

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.

Shi, Z., Church, R. M., & Meck, W. H. (2013). Bayesian optimization of time perception. *Trends in Cognitive Sciences*, *17*, 556–564.

Stewart, N., Brown, G. D. A., & Chater, N. (2005). Absolute identification by relative judgment. *Psychological Review*, *112*, 881–911.

St.John-Saaltink, E., Kok, P., Lau, H. C., & de Lange, F. P. (2016). Serial dependence in perceptual decisions is reflected in activity patterns in primary visual cortex. *Journal of Neuroscience*, *36*, 6186–6192.

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

Taubert, J., Burg, E. V. D., & Alais, D. (2016). Love at second sight: Sequential dependence of facial attractiveness in an on-line dating paradigm. *Scientific Reports*, *6*, 22740.

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. (2009). Uncovering unobservable cognitive mechanisms: The contribution of mathematical models. In F. Rösler, C. Ranganath, B. Röder, & R. H. Kluwe (Eds.), *Neuroimaging of human memory: Linking cognitive processes to neural systems* (pp. 25–41). New York: Oxford University Press.

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.

van Rijn, H. (2016). Accounting for memory mechanisms in interval timing: A review. *Current Opinion in Behavioral Sciences*, *8*, 245 - 249.

Vierordt, K. (1868). *Der Zeitsinn nach Versuchen* [The time sense according to experiments]. Tübingen, Germany: Verlag der H. Laupp'schen Buchhandlung.

von Helmholtz, H. (1867). *Handbuch der physiologischen Optik.*

Walsh, V. (2003). A theory of magnitude: Common cortical metrics of time, space and quantity. *Trends in Cognitive Sciences*, *7*, 483-488.

Ward, L. M. (1979). Stimulus information and sequential dependencies in magnitude estimation and cross-modality matching. *Journal for Experimental Psychology: Human Perception and Performance*, *5*, 444–459.

Wearden, J. H., Todd, N. P. M., & Jones, L. A. (2006). When do auditory / visual differences in duration judgements occur? *The Quarterly Journal of Experimental Psychology*, *59*, 1709–1724.

Weber, E. H. (1834). *De pulsu, resorptione, auditu et tactu: annotationes anatomicae et physiologicae, auctore.* prostat apud CF Koehler.

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

Woodrow, H. (1933). Weight discrimination with a varying standard. *The American Journal of Psychology*, *45*(3).

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

Wundt, W. (1874). *Grundzüge der physiologischen Psychologie.* Leipzig: Engelmann.

Xia, Y., Leib, A. Y., & Whitney, D. (2016). Serial dependence in the perception of attractiveness. *Journal of Vision*, *16*, 28.

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.

Yuille, A., & Kersten, D. (2006). Vision as bayesian inference : analysis by synthesis? *Trends in Cognitive Sciences*, *10*, 301 − 308.

Zotov, V., Jones, M. N., & Mewhort, D. J. K. (2011). Contrast and assimilation in categorization and exemplar production. *Attention, Perception, & Psychophysics*, *73*, 621–639.

Appendix A Study 1

Ellinghaus, R., Ulrich, R., & Bausenhart, K. M. (2018). Effects of stimulus order on comparative judgments across stimulus attributes and sensory modalities. *Journal of Experimental Psychology: Human Perception and Performance*, 44, 7–12.

Journal of Experimental Psychology:
Human Perception and Performance
2018, Vol. 44, No. 1, 7–12

0096-1523/18/$12.00 http://dx.doi.org/10.1037/xhp0000495

OBSERVATION

Effects of Stimulus Order on Comparative Judgments Across Stimulus Attributes and Sensory Modalities

Ruben Ellinghaus, Rolf Ulrich, and Karin M. Bausenhart
University of Tübingen

In psychophysical experiments, participants are often asked to compare the magnitude of a constant standard against the magnitude of a variable comparison. According to prominent models of stimulus discrimination, discrimination sensitivity should depend only on the physical magnitude difference between these two stimuli but not on the order of their presentation. However, previous experiments on auditory duration discrimination have shown that discrimination sensitivity is higher when the standard precedes rather than follows the comparison. It is presently unclear whether this Type B effect emerges only for duration discrimination or generalizes across modalities and stimulus attributes. Therefore, we conducted a study in which participants performed several discrimination tasks for various stimulus attributes (i.e., duration, frequency, intensity, and numerosity), each in the visual and in the auditory modality. In all cases, discrimination sensitivity was higher when the standard preceded rather than followed the comparison. This result indicates that the Type B effect is not restricted to the domain of temporal cognition but rather reflects a general phenomenon across a range of domains and modalities. The outcome of the present experiment is consistent with the internal reference model according to which the Type B effect is a consequence of a dynamically updated internal reference, which is used in the comparison process. Alternatively, a weighted difference model with a larger weight for the second stimulus position than for the first stimulus position can also account for this result.

Public Significance Statement
Discriminating between stimuli is a fundamental human ability that is relevant in almost every aspect of everyday life (e.g., speech perception, food intake, music perception). Understanding the basic mechanisms involved in human discrimination is therefore of utmost relevance. The present research challenges standard theories of human discrimination and stresses the importance of memory processes.

Keywords: comparative judgments, discrimination sensitivity, psychophysical models, Type B effect

The fundamental ability to compare and discriminate between stimuli has been investigated ever since experimental psychology was established (e.g., Fechner, 1860). Until today, stimulus discrimination is studied with psychophysical procedures in which participants repeatedly compare and discriminate between stimuli that differ in physical magnitude (e.g., two tones differing in loudness). For example, in the two-alternative, forced-choice task (2AFC, Hegelmaier, 1852), a standard s with constant magnitude and a comparison c with magnitude varying from trial to trial are presented successively and participants identify the larger stimulus. The temporal order of s and c varies randomly between trials, yielding trials with order $\langle sc \rangle$ or $\langle cs \rangle$. To measure discrimination sensitivity, one typically estimates the magnitude difference between s and c, which enables identification of the larger stimulus with an accuracy level of 75% (Gescheider, 1997). This measure is conventionally defined as difference limen (*DL*; or just noticeable difference).

Interestingly, *DL* often differs between $\langle sc \rangle$ and $\langle cs \rangle$ trials, a phenomenon referred to as the Type B effect (Ulrich & Vorberg, 2009).[1] If sensitivity is lower for $\langle cs \rangle$ trials than for $\langle sc \rangle$ trials, the Type B effect is called negative and positive if sensitivity is higher

---

Ruben Ellinghaus, Rolf Ulrich, and Karin M. Bausenhart, Cognition and Perception, Department of Psychology, University of Tübingen.

This research was supported by DFG grant BA 4110/5-2.

We thank Soraja Fejzuli and Sophie Giel for assistance in testing participants.

Correspondence concerning this article should be addressed to Ruben Ellinghaus, Department of Psychology, University of Tübingen, Schleichstrasse 4, 72076 Tübingen, Germany. E-mail: ruben.ellinghaus@uni-tuebingen.de

[1] Ulrich and Vorberg (2009) introduced this notation to distinguish the Type B effect from the Type A effect, which refers to a horizontal shift of the psychometric functions for orders $\langle sc \rangle$ and $\langle cs \rangle$. Thus, the Type A effect refers to order-dependent differences in judged magnitude (e.g., because of a response bias or a time-order error, Fechner, 1860), whereas the Type B effect reflects order-dependent differences in sensitivity.

in ⟨*cs*⟩ trials than in ⟨*sc*⟩ trials (Dyjas & Ulrich, 2014). Type B effects reported in the literature, however, are almost exclusively negative (Dyjas, Bausenhart, & Ulrich, 2012; Lapid, Ulrich, & Rammsayer, 2008; Rammsayer, 2008; Rammsayer & Ulrich, 2012; Stott, 1935; Ulrich, 2010; Woodrow, 1935). Positive Type B effects have rarely been reported in the case of rather short-duration stimuli, especially when presented with very short interstimulus intervals (Hellström & Rammsayer, 2004, 2015; but see Bausenhart, Dyjas, & Ulrich, 2015).

Considering this effect as robust, comprehensive models of human stimulus discrimination performance would ideally be able to account for it. However, standard psychophysical models such as signal detection theory (Green & Swets, 1966; Macmillan & Creelman, 2005; Wickens, 2002) and other prominent models (Luce & Galanter, 1963; Yeshurun, Carrasco, & Maloney, 2008) cannot predict this phenomenon (cf. Dyjas, Bausenhart, & Ulrich, 2012). In brief, all of these models follow the pioneering work of Thurstone (1927a, 1927b), according to which participants base their judgment on the difference of the internal stimulus representations $D = X_1 - X_2$, whereby X_1 and X_2 represent the internal magnitudes of the first and second stimulus in a trial. The first (second) stimulus is judged as the larger one for D larger (smaller) than a fixed constant γ (with $\gamma \neq 0$, indicating a response bias).

Extensions of these classical difference models can account for the Type B effect. For example, the Sensation Weighting Model (SWM, Hellström, 1979, 1985) explains the Type B effect by postulating that X_1 and X_2 are weighted differently by weighting factors w_1 and w_2. For $w_2 > w_1$, the SWM predicts a negative Type B effect, and for $w_1 > w_2$, it predicts a positive Type B effect (see Bausenhart, Dyjas, & Ulrich, 2015; Dyjas & Ulrich, 2014, for a mathematical analysis of the SWM). Furthermore, the Internal Reference Model (IRM, Dyjas et al., 2012; Lapid, Ulrich, & Rammsayer, 2008) replaces X_1 in the decision process by an internal reference I (Durlach & Braida, 1969; Helson, 1947, 1964; Morgan, Watamaniuk, & McKee, 2000; Nachmias, 2006), which incorporates previous and current stimulus instances and provides a mechanism how I is updated from trial to trial. It has been shown that IRM predicts that the Type B effect should either be negative or absent (Dyjas, Bausenhart, & Ulrich, 2012). Intuitively, this holds because in ⟨*cs*⟩ trials, the variable *c* is merged into I, which effectively attenuates D and thus impairs discrimination performance. In ⟨*sc*⟩ trials, however, the constant *s* is merged into I, whereas the information from the variable *c* enters fully into D, enabling higher discrimination performance than for ⟨*cs*⟩ trials (cf. Bausenhart, Bratzke, & Ulrich, 2016, Figure 1).

Because these models generally apply to any stimulus discrimination task, the Type B effect should emerge across various task domains and sensory modalities. To date, the Type B effect has been mainly investigated for temporal discrimination (e.g., Dyjas, Bausenhart, & Ulrich, 2012; Dyjas, Bausenhart, & Ulrich, 2014; Hellström & Rammsayer, 2015; Grondin & McAuley, 2009; Lapid et al., 2008; Ulrich, 2010; but see García-Pérez & Alcalá-Quintana, 2010) and only sporadically been reported for other domains such as discrimination of shapes (Nachmias, 2006) and weights (Ross & Gregory, 1964).

One may thus argue that the Type B effect is especially prominent in the temporal domain. Specifically, because there is no dedicated sensory system for time (cf. Wearden, 2016), the representation of temporal information might be unstable and thus particularly susceptible to contextual factors (e.g., Bausenhart, Bratzke, & Ulrich, 2016; Bruno & Cicchini, 2016; Heron et al., 2012; Jazayeri & Shadlen, 2010). Accordingly, the formation of an internal reference may help to stabilize the representation of temporal information. By contrast, if the Type B effect is observed consistently across different tasks and sensory modalities, this would further stress its theoretical relevance for psychophysical models. Therefore, in the present study, we took a systematic approach to investigate 2AFC discrimination performance for four stimulus attributes (i.e., duration, frequency, intensity, and numerosity), each in the visual and in the auditory modality.

Method

Participants

Fifteen female and nine male volunteers (mean age 26 ± 9.2 years) participated in the study.[2] All participants provided written consent and received 10 €/h or course credit as reimbursement. Three participants were replaced because estimated *DL*s were outside the predefined three-sigma range.

General Procedure

Participants performed adaptive 2AFC discrimination tasks for four stimulus attributes (i.e., duration, frequency, intensity, and numerosity), each in the visual and the auditory modality. Testing was distributed over four separate sessions of two tasks, each within a maximum time span of two weeks. Task order was counterbalanced between participants with a balanced Latin square. In each task, participants had to discriminate two successively presented stimuli (constant standard *s* and variable comparison *c*), differing in magnitude on the respective stimulus attribute. The magnitude of *c* adapted to participants' response behavior following the weighted up-down method (Kaernbach, 1991). Two start values for *c* (one above and one below *s*) combined with two stimulus orders (⟨*sc*⟩ and ⟨*cs*⟩) yielded four trial runs. These were presented randomly interleaved and consisted of 50 trials each (see Figure 1). The 200 trials of each task were divided into eight blocks of 25 trials each, with a self-terminated break after each block. At the beginning of each task, 25 practice trials were administered but discarded from data analysis.

The same adaptive rules were used during practice and experimental blocks. Specifically, for the upper (lower) runs, the magnitude of *c* was decreased (increased) by fixed step size Δ whenever *c* was judged as being larger (smaller) than *s* and increased (decreased) by $3 \cdot \Delta$ when *c* was judged as being smaller (larger) than *s*. Starting values were reset after practice. The lower and upper runs targeted the 25th and 75th percentile of the psychometric function, respectively. For each run the reversal points were

---

[2] From the results of Dyjas, Bausenhart, & Ulrich (2012), we estimated an effect size of the Type B effect of $d = 0.7$. From this, one computes a statistical power of $1 - \beta = .91$ for a paired *t* test with $n = 24$ participants for detecting a Type B effect with $\alpha = .05$ (two-sided test).

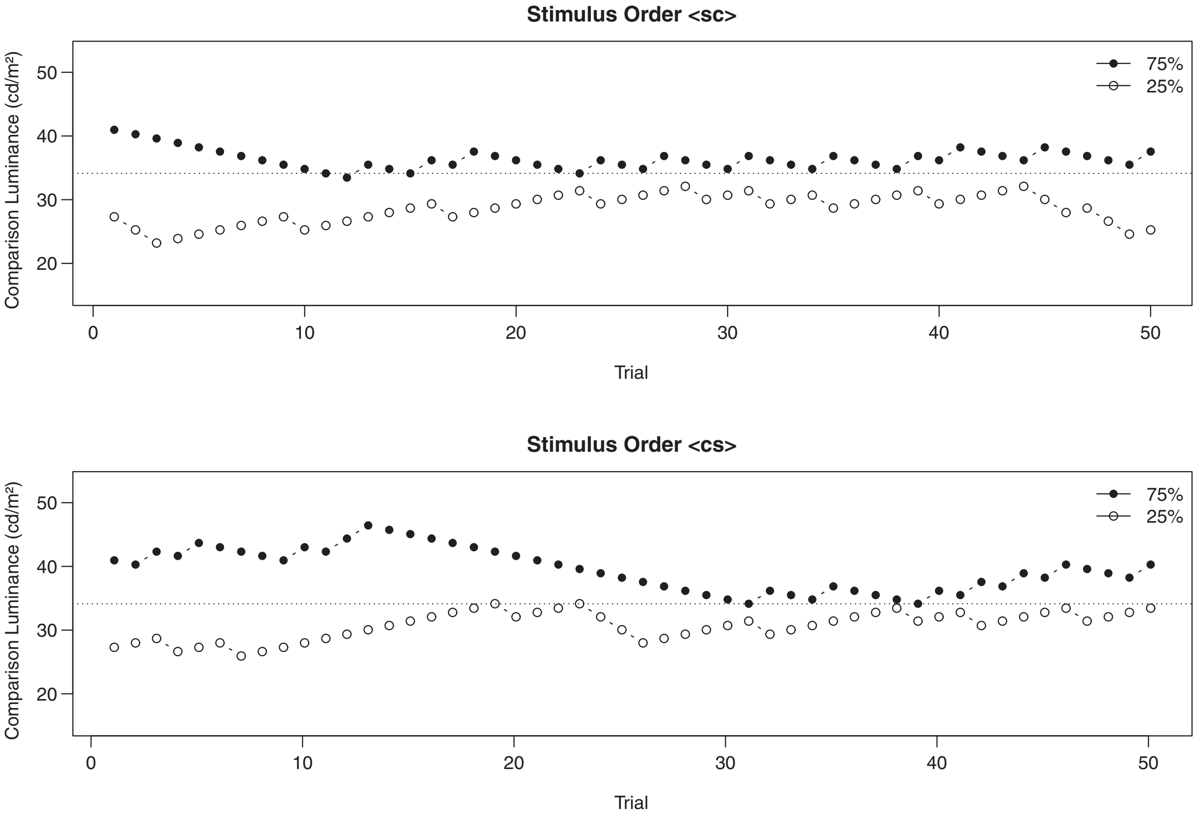

Figure 1. Example data of a single participant illustrating the adaptive runs of the weighted up-down procedure for visual intensity discrimination and for both stimulus orders ⟨*sc*⟩ (upper panel) and ⟨*cs*⟩ (lower panel). The trajectory with closed circles represents a 75% run and the one with open circles a 25% run. For the upper panel, the midrun estimate $x_{25\%}$ and $x_{75\%}$ for the 25% and 75% run are 28.92 cd/m^2 and 35.89 cd/m^2, respectively. Hence, the estimated difference limen $DL_{\langle sc \rangle}$ is computed as $DL_{\langle sc \rangle} = (35.89 - 28.92)/2 = 3.49$ cd/m^2. For the lower panel, an analogous calculation yields the estimate $DL_{\langle cs \rangle} = 4.13$ cd/m^2.

determined and averaged, yielding midrun estimates $x_{25\%}$ and $x_{75\%}$ for the lower and upper run, respectively.[3] *DL* was defined as $DL = (x_{75\%} - x_{25\%})/2$.

Apparatus

Stimuli were generated and presented on a Mac Pro 3.1 using the Matlab PsychToolbox extensions (Brainard, 1997; Pelli, 1997). Participants were seated in a dimly lit and sound-attenuated booth, placed in front of a gamma-corrected 21-inch CRT monitor (1,024 × 768 pixels, 150 Hz) at a viewing distance of approximately 35 cm. Auditory stimuli were presented binaurally through headphones. Responses of first stimulus larger and second stimulus larger were collected through the y and m keys of a QWERTZ keyboard, respectively.

Discrimination Tasks and Stimuli

Each trial of each task started with a fixation point presented at the screen center. After 1,000 msec, the first stimulus was presented. After an interstimulus interval of 1,000 msec, the second stimulus was presented. Stimuli lasted for 500 msec (except for the duration discrimination task, see below). Participants were instructed to indicate which stimulus (first or second) had the larger magnitude. No feedback was given. After an intertrial interval of 1,000 msec, the next trial began.

Standard magnitudes, start values, and step sizes for each task are given in Table 1. Pilot testing and reference values from the psychophysical literature were used to determine these parameters.

---

[3] According to an alternative stopping rule for the up-down method, each trial run stops after a fixed number of reversals instead of a fixed number of trials (cf. Lu & Dosher, 2014). Therefore, a difference in *DL* between stimulus orders ⟨*sc*⟩ and ⟨*cs*⟩ might in principle be attributed to different numbers of reversals. However, additional analysis revealed that the number of reversals did not meaningfully differ between stimulus orders, thereby ruling out this possible confound.

Table 1
Experimental Parameters of the Up-and-Down Procedure for the Eight Discrimination Tasks

Discrimination task	Standard	Lower start value	Upper start value	Step size Δ
Duration				
Auditory	500 msec	400 msec	600 msec	10 msec
Visual	500 msec	300 msec	700 msec	20 msec
Frequency				
Auditory	500 Hz	495.69 Hz	504.35 Hz	1 cent
Visual	2 c/deg	1.6 c/deg	2.4 c/deg	.02 c/deg
Intensity				
Auditory	50 dB	40 dB	60 dB	1 dB
Visual	34.1 cd/m^2	27.1 cd/m^2	41.6 cd/m^2	.7 cd/m^2
Numerosity				
Auditory	10 pulses	5 pulses	15 pulses	1 pulse
Visual	30 dots	20 dots	40 dots	1 dot

Duration discrimination. In the auditory modality, participants discriminated 500-Hz sine tones (ramped 5 msec on- and offsets). The tones differed in duration and had a constant intensity level of 65 dB. In the visual modality, stimuli were gray disks with a diameter of 6.5° and a luminance of 34.1 cd/m^2, centrally presented against a black background.

Frequency discrimination. In the auditory modality, participants discriminated sine tones differing in frequency with a constant intensity level of 65 dB and ramped 5 msec on- and offsets. In the visual modality, participants discriminated Gabor patterns, which differed in spatial frequency and were presented centrally against a gray background within a rectangular window (256 × 265 pixels). *SD* of the Gaussian filter was 10% of the width of that area, and phase was kept constant for all stimuli.

Intensity discrimination. In the auditory modality, participants discriminated 500-Hz sine tones (ramped 5 msec on- and offsets), which differed in sound intensity. In the visual modality, participants discriminated centrally presented gray circles with a diameter of 6.5°, which differed in their levels of luminance. All stimuli were presented against a black background.

Numerosity discrimination. In the auditory modality, participants discriminated the number of pulses in an auditory pulse train. A single pulse train consisted of a square wave tone, which switched at randomly chosen time points between 0 dB and 60 dB, creating a hearing impression like /tr/—/tr/-/tr/–/tr/-/tr/—/tr/. In the visual modality, participants discriminated random-dot patterns differing in dot number. Specifically, dot patterns consisted of white dots (65 cd/m^2, 0.26° × 0.26°) presented at randomly chosen locations against a black background within a centrally presented rectangular region (300 × 500 pixels).

Results

Figure 2 depicts boxplots adjusted for skewness (Hubert & Vandervieren, 2008) of *DL* as a function of stimulus order and task.[4] These plots indicate some mild outliers and one severe outlier in the auditory numerosity discrimination task. Nevertheless, we included these values because they were within the predefined three-sigma range. For all tasks, median *DL* was larger for stimulus order $\langle cs \rangle$ than for $\langle sc \rangle$, revealing a negative Type B effect. This impression was confirmed by Wilcoxon signed-rank tests, which produced statistical significance for all tasks except visual frequency and auditory numerosity, which yielded trends (see Table 2). Moreover, positive correlations among the *DL*s for both stimulus orders pointed to a common component in discrimination performance (see Table 2).

We also conducted a within-subjects analysis of variance with factors *sensory modality* and *stimulus attribute* on the relative increase of *DL*, computed as $(DL_{\langle cs \rangle} - DL_{\langle sc \rangle})/DL_{\langle sc \rangle} \cdot 100\%$ (Table 2; cf. Dyjas, & Ulrich, 2012). If the Type B effect is the sign of a general discrimination process, this measure should not meaningfully vary across stimulus attributes and sensory modalities. Consistent with this assumption, neither the main effects of sensory modality, $F(1, 23) = 0.01$, $p = .916$, and stimulus attribute, $F(3, 69) = 1.22$, $p = .308$, nor their interaction, $F(3, 69) = 1.91$, $p = .137$, was significant.[5] The negative Type B effect was evident in the significant intercept, $F(1, 23) = 32.84$, $p < .001$, $\eta_p^2 = 0.59$. In sum, a negative Type B effect of approximately the same relative magnitude was observed in all experimental conditions.

Discussion

The present study examined the influence of stimulus order on discrimination sensitivity in standard 2AFC tasks. For various stimulus attributes and modalities, sensitivity was higher when a constant standard preceded rather than followed a variable comparison. The relative size of this negative Type B effect was relatively constant across discrimination tasks. Thus, the present results strengthen the view that the negative Type B effect is not restricted to temporal discrimination but instead constitutes an ubiquitous and robust phenomenon.

This outcome is theoretically important because prominent Thurstonian models cannot account for the Type B effect. According to these difference models, participants compare the internally represented stimulus magnitudes of the two presented stimuli within a given trial. Consequently, discriminability of two successively presented stimuli should depend on their physical difference but not on their presentation order. Our results clearly speak against this notion, thereby casting doubt on the validity of these models and hence of the conclusions drawn from them.

A theoretically important question is why the Type B effect occurs. SWM (Hellström, 1979, 1985) explains the Type B effect by postulating that the two stimulus positions are weighted differentially. Specifically, SWM entails a negative Type B effect if the weight w_2 associated with the second stimulus is larger than the weight w_1 associated with the first stimulus. However, SWM does not specify why a certain stimulus position is weighted more strongly because the weighting can be inferred only post hoc from the data. In contrast, IRM (Dyjas, Bausenhart, & Ulrich, 2012; Lapid et al., 2008; for a review see Bausenhart, Bratzke, & Ulrich, 2016) builds on the idea that participants compare the internal representation of the second presented stimulus against an internal

---

[4] A normal Q-Q plot suggested that the estimated *DL*s were not normally distributed but rather exhibited a systematic skew to the right. We thank an anonymous reviewer for suggesting this analysis.

[5] The calculated power for a medium effect (i.e., $f^2 = 0.15$) of sensory modality, stimulus attribute, and their interaction is .78, .46, and .78, respectively ($\alpha = .05$ and $n = 24$).

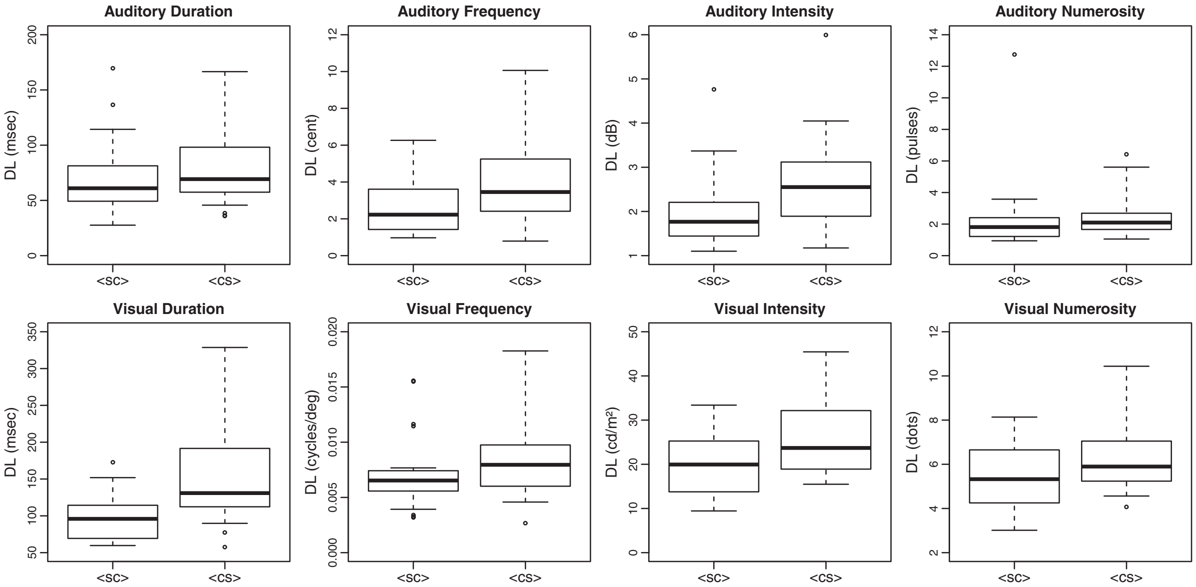

Figure 2. Adjusted boxplot of *DL* as a function of stimulus order, sensory modality, and stimulus attribute. The middle line of each box represents the median. The lower and upper end points of the box represent Q_1 (i.e., 25th) and Q_3 (i.e., 75th) percentile of the data, respectively. The whiskers define the lower and upper boundaries of the data. The length of the whiskers were adjusted for the skewed *DL* distribution according to the exponential model proposed by Hubert and Vandervieren (2008). All boxplots were computed with the R-function *adjbox*.

reference, which is exclusively based on past and present stimulus instances. Consequently, IRM makes the strong prediction of a negative Type B effect, as was observed in the present study. In general, the idea of an internal reference relates to the notion that humans rely on recent sensory stimulation to increase the signal-to-noise ratio in perception (e.g., Raviv, Ahissar, & Loewenstein, 2012; Raviv, Lieder, Loewenstein, & Ahissar, 2014; Shi, Church, & Meck, 2013).

Table 2
Summary of Basic Statistical Results

Task	W^a	p	r_s	DL increase^b
Duration				
Auditory	152	.034	.66	25.1 (8.3)
Visual	26	<.001	.48	58.6 (10.3)
Frequency				
Auditory	58	<.001	.81	52.8 (12.5)
Visual	184	.101	.25	37.3 (14.7)
Intensity				
Auditory	96	.003	.31	48.2 (14.5)
Visual	98	.003	.38	36.7 (12.3)
Numerosity				
Auditory	178	.084	.73	30.4 (11.6)
Visual	146	.027	.23	26.2 (8.0)

Note. Column r_s gives Spearman's rank correlation between the individual estimates of $DL_{\langle cs\rangle}$ and $DL_{\langle sc\rangle}$. The last column shows the negative Type B effect as a relative increase of *DL* between stimulus orders $\langle sc\rangle$ and $\langle cs\rangle$.
[a] Test statistic of the Wilcoxon signed-rank test. [b] This increase in *DL* is defined as $\frac{DL_{\langle cs\rangle}-DL_{\langle sc\rangle}}{DL_{\langle sc\rangle}}\cdot 100\%$. The numbers in parentheses give the standard error of mean.

The present results do not invalidate the 2AFC task for measuring discrimination performance. However, researchers should not collapse both trial orders to derive *DL* but instead estimate *DL* separately for each order and average the two *DL*s to index discrimination sensitivity (for a detailed discussion of this issue, see Ulrich & Vorberg, 2009). R and Matlab code for such order-conditional analyses have been provided by Bausenhart, Dyjas, Vorberg, and Ulrich (2012).

To conclude, a negative Type B effect was observed not only for duration discrimination but also in other domains such as numerosity, intensity, and frequency in both the visual and auditory modality. This indicates that the Type B effect is the signature of a general mechanism. However, the exact mechanism causing the Type B effect is still under debate. Candidate explanations stress past sensory information (e.g., IRM) or differential weighting of the two stimulus positions (e.g., SWM). In any case, however, the Type B effect appears to be a general and robust phenomenon, and we propose that it should be considered a benchmark effect for future models of stimulus discrimination.

References

Bausenhart, K. M., Bratzke, D., & Ulrich, R. (2016). Formation and representation of temporal reference information. *Current Opinion in Behavioral Sciences, 8,* 46–52.

Bausenhart, K. M., Dyjas, O., & Ulrich, R. (2015). Effects of stimulus order on discrimination sensitivity for short and long durations. *Attention, Perception, & Psychophysics, 77,* 1033–1043.

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.

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

Bruno, A., & Cicchini, G. M. (2016). Multiple channels of visual time perception. *Current Opinion in Behavioral Sciences, 8,* 131–139.

Durlach, N. I., & Braida, L. D. (1969). Intensity perception. I. Preliminary theory of intensity resolution. *Journal of the Acoustical Society of America, 46,* 372–383.

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.

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.

Fechner, G. T. (1860). *Elemente der Psychophysik* [Elements of psychophysics]. Leipzig, Germany: Breitkopf und Härtel.

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.

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

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (revised ed.). Los Altos, CA: Peninsula Publishing, reprinted edition 1988.

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

Hegelmaier, F. (1852). Ueber das Gedächtnis für Linear-anschauungen [On memory for visually perceived lines]. *Archiv für Physiologische Heilkunde, 11,* 844–853.

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., & 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.

Hellström, Å., & Rammsayer, T. H. (2015). Time-order errors and standard-position effects in duration discrimination: An experimental study and an analysis by the sensation-weighting model. *Attention, Perception, & Psychophysics, 77,* 2409–2423.

Helson, H. (1947). Adaptation-level as frame of reference for prediction of psychophysical data. *American Journal of Psychology, 60,* 1–29.

Helson, H. (1964). *Adaptation-level theory*. New York: Harper & Row.

Heron, J., Aaen-Stockdale, C., Hotchkiss, J., Roach, N. W., McGraw, P. V., & Whitaker, D. (2012). Duration channels mediate human time perception. *Proceedings of the Royal Society B: Biological Sciences, 279,* 690–698.

Hubert, M., & Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. *Computational Statistics and Data Analysis, 52,* 5186–5201.

Jazayeri, M., & Shadlen, M. N. (2010). Temporal context calibrates interval timing. *Nature Neuroscience, 13,* 1020–1026.

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

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.

Lu, Z.-L., & Dosher, B. (2014). *Visual psychophysics*. Cambridge, MA: The MIT Press.

Luce, R. D., & Galanter, E. (1963). Discrimination. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. I, pp. 191–243). 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.

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

Rammsayer, T. H. (2008). Neuropharmacological approaches to human timing. In S. Grondin (Ed.), *Psychology of time* (pp. 295–320). Bingley, United Kingdom: Emerald Group.

Rammsayer, T. H., & Ulrich, R. (2012). The greater temporal acuity in the reminder task than in the 2AFC task is independent of standard duration and sensory modality. *Canadian Journal of Experimental Psychology, 66,* 26–31.

Raviv, O., Ahissar, M., & Loewenstein, Y. (2012). How recent history affects perception: The normative approach and its heuristic approximation. *PLoS Computational Biology, 8,* e1002731.

Raviv, O., Lieder, I., Loewenstein, Y., & Ahissar, M. (2014). Contradictory behavioral biases result from the influence of past stimuli on perception. *PLoS Computational Biology, 10,* e1003948.

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.

Shi, Z., Church, R. M., & Meck, W. H. (2013). Bayesian optimization of time perception. *Trends in Cognitive Sciences, 17,* 556–564.

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.

Wearden, J. (2016). *The psychology of time perception*. London, England: MacMillan.

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

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.

Received March 13, 2017
Revision received August 28, 2017
Accepted September 4, 2017 ■

Appendix B Study 2

Ellinghaus, R., Gick, M., Bausenhart, K. M., & Ulrich, R. (2019). Decay of internal reference information in duration discrimination: Intertrial interval modulates the Type B effect. *Quarterly Journal of Experimental Psychology, 72*, 1578–1586.

Original Article

QJEP

Quarterly Journal of Experimental Psychology
1–9

DOI: 10.1177/1747021818808187
qjep.sagepub.com
SAGE

Decay of internal reference information in duration discrimination: Intertrial interval modulates the Type B effect

Ruben Ellinghaus, Mareike Gick, Rolf Ulrich and Karin M Bausenhart

Abstract

Psychophysical evidence suggests that human perception of a stimulus is assimilated towards previous stimuli. The internal reference model (IRM) explains such assimilation through an internal reference (IR), which integrates past and present stimulus representations and thus might be conceived as a form of perceptual memory. In this study, we investigated whether the IR decays with time, as previously shown for perceptual memory representations in general. One specific prediction of IRM is higher discrimination sensitivity when a constant standard precedes rather than follows a variable comparison in a two-alternative forced-choice (2AFC) discrimination task. Furthermore, the magnitude of this so-called negative Type B effect should decrease with decreasing weighting of past stimulus information in the integration process. Therefore, decay of the IR should result in a reduced Type B effect. To examine this prediction, we carried out a 2AFC duration discrimination experiment with a short (1,600 ms) and a long (3,200 ms) intertrial interval (ITI). As expected, a reduced negative Type B effect was observed at the long compared with the short ITI, consistent with the idea that humans rely on the immediate past when evaluating current sensory input, however, less so when the IR incorporating the perceptual short-term memory representation of these past stimuli has already decayed.

Keywords

Stimulus discrimination; assimilation; internal reference model; Type B effect; perceptual memory

Received: 10 April 2018; revised: 14 August 2018; accepted: 27 August 2018

The human ability to compare and discriminate between stimuli is of fundamental importance and has been investigated ever since the dawn of experimental psychology (e.g., Fechner, 1860; Hegelmaier, 1852). Until today, stimulus discrimination and its underlying mechanisms are typically studied with psychophysical procedures in which participants compare and discriminate stimuli that differ in physical magnitude (e.g., two tones differing in loudness or duration). Most theories of human stimulus discrimination which model the cognitive processes underlying discrimination performance are based on the Thurstonian assumption that participants rely solely on the difference of the internal representations of the two stimuli to be compared (Green & Swets, 1966; Luce & Galanter, 1963; Macmillan & Creelman, 2005; Thurstone, 1927a, 1927b; Wickens, 2002; Yeshurun, Carrasco, & Maloney, 2008). Accordingly, any judgement should only depend on the physical magnitude of the currently presented stimuli and be independent of previous stimuli.

However, it has been demonstrated experimentally that this independence assumption does not hold; judgement on a current trial can be influenced by stimuli presented on previous trials (e.g., Helson, 1964). For example, the judged magnitude of a stimulus is often biased towards the average of the stimulus range employed in the experiment—a phenomenon referred to as the central tendency of judgement (Hollingworth, 1910). Similarly, the judged magnitude of a stimulus in a given experimental trial is often pulled towards the magnitude of a stimulus presented on the immediately preceding trial (Garner, 1953). Such assimilation effects are ubiquitous in the perceptual judgement literature and have been documented for various tasks including auditory duration discrimination (Dyjas, Bausenhart, & Ulrich, 2012), auditory frequency discrimination (Arzounian, Kerangal, & Cheveign, 2017; Chambers

Department of Psychology, University of Tübingen, Tübingen, Germany

Corresponding author:
Ruben Ellinghaus, Department of Psychology, University of Tübingen, Schleichstr. 4, 72076 Tübingen, Germany.
Email: ruben.ellinghaus@uni-tuebingen.de

& Pressnitzer, 2014; Raviv, Ahissar, & Loewenstein, 2012; Raviv, Lieder, Loewenstein, & Ahissar, 2014), visual line length discrimination (Ashourian & Loewenstein, 2011), and judgements of visual orientation (Fischer & Whitney, 2014), weight (Parducci & Marshall, 1962) and loudness (Decarlo & Cross, 1990; King & Lockhead, 1982; Ward, 1979). Interestingly, assimilation extends also to more complex stimulus attributes such as facial attractiveness (Geiselman, Haight, & Kimata, 1984; Taubert, Burg, & Alais, 2016) and also non-physical attributes such as judgement of price (Matthews & Stewart, 2009). Furthermore, assimilation has been observed not only for relative judgement tasks but also in absolute identification (Donkin, Chan, & Tran, 2017; Stewart, Brown, & Chater, 2005), as well as for categorisation and exemplar production (Zotov, Jones, & Mewhort, 2011), and in a passive viewing paradigm (Fornaciai & Park, 2018).

Several researchers (e.g., Arzounian et al., 2017; Dyjas et al., 2012; Fischer & Whitney, 2014; Raviv et al., 2014) have suggested that such assimilation effects emerge as a consequence of mixing the memory representations of previously presented stimuli with the representations of the current stimuli (for overviews in the domain of temporal discrimination, see Bausenhart, Bratzke, & Ulrich, 2016; van Rijn, 2016). For example, the internal reference model (IRM; Dyjas et al., 2012) provides a quantitative account of how incoming stimulus information is continuously updated from trial to trial and integrated into an internal reference representation (see also Durlach & Braida, 1969; Morgan, Watamaniuk, & McKee, 2000; Nachmias, 2006). Originally, this model has been developed to account for stimulus order effects on discrimination sensitivity in the classical two-alternative forced-choice (2AFC) task (Hegelmaier, 1852). In the temporal 2AFC task, sometimes also referred to as two-interval forced-choice (2IFC) task, participants are asked to select the larger of two successively presented stimuli: one of which is a standard s with constant magnitude and the other is a comparison c with magnitude varying from trial to trial. The order of s and c also varies randomly from trial to trial, thus yielding trials with stimulus orders $\langle sc\rangle$ and $\langle cs\rangle$. According to the IRM, participants complete this task in the current trial n by comparing the internal representation $\mathbf{X}_{2,n}$ of the second stimulus in this trial against the current internal reference $\mathbf{I}_n$, which is a conglomerate of previous and current stimulus instances and updates continuously from trial to trial:

$$\mathbf{D}_n = \mathbf{I}_n - \mathbf{X}_{2,n} \tag{1}$$

The internal reference $\mathbf{I}_n = g \cdot \mathbf{I}_{n-1} + (1\text{–}\mathbf{g}) \cdot \mathbf{X}_{1,n}$ on trial n is a weighted sum of the first stimulus' internal representation $\mathbf{X}_{1,n}$ on the current trial n and the internal reference $\mathbf{I}_{n-1}$ from the previous trial $n-1$ with constant weight g, $0 \le g < 1$. If $\mathbf{D}_n = \mathbf{I}_n - \mathbf{X}_{2,n} > 0$, participants judge the first stimulus to be the larger one and the second stimulus otherwise.

Importantly, IRM does not only account for trial-by-trial assimilation effects (Dyjas et al., 2012) and the central tendency of judgement (Bausenhart, Dyjas, & Ulrich, 2014) as outlined above, but also makes the counterintuitive prediction that discrimination sensitivity differs between stimulus orders in the 2AFC task. Specifically, for values of $0 < g < 1$, IRM predicts that the difference limen (DL) for stimulus order $\langle cs\rangle$ is larger than the DL for stimulus order $\langle sc\rangle$, a phenomenon referred to as the negative Type B effect (Ulrich & Vorberg, 2009).[1] This holds because the expected value of $\mathbf{I}_n$ is equal to s, as long as the values of c are distributed symmetrically around the standard. Therefore, the internal representation of c will generally be pulled towards s on $\langle cs\rangle$ trials, which in turn decreases the subjective difference $\mathbf{D}_n$ and thus lowers discrimination sensitivity (i.e., increases DL). Hence, according to IRM, the negative Type B effect is a direct consequence of the suggested trial-by-trial assimilation mechanism (Dyjas et al., 2012; Lapid, Ulrich, & Rammsayer, 2008). In fact, negative Type B effects have been repeatedly demonstrated, not only for 2AFC discrimination but also for equality judgements (Dyjas & Ulrich, 2014) and across several stimulus attributes and modalities (e.g., Ellinghaus, Ulrich, & Bausenhart, 2018; Thönes, Von Castell, Iflinger, & Oberfeld, 2018).

Although the IRM offers a plausible account of the negative Type B effect and related phenomena, the specific characteristics of the internal reference representation have yet to be investigated (Bausenhart et al., 2016). One of these characteristics concerns the temporal stability of the internal reference. In fact, one might conceive the internal reference as a residual perceptual trace or memory representation of prior stimulation. Importantly, memory representations may fade over time and/or be prone to retroactive interference from similar stimuli (e.g., Ashourian & Loewenstein, 2011; Magnussen & Greenlee, 1999; Schab & Crowder, 1988; Wearden & Ferrara, 1993). Decay of perceptual memory is typically demonstrated by showing that discrimination performance decreases as the interstimulus interval (ISI) between s and c increases. For example, Crowder (1982) found that auditory discrimination sensitivity in a vowel discrimination task decreases monotonically up to an asymptotic level with increasing ISI, suggesting that the memory representation of the first presented stimulus fades as time passes. In the visual modality, similar results have, for example, been obtained for contrast discrimination (Magnussen, Thomas, & Greenlee, 1996). Interestingly, also the typical negative Type B effect diminishes or may be even inverted (Bausenhart, Dyjas, & Ulrich, 2015; Hellström & Rammsayer, 2004), at least for duration discrimination of brief stimuli, when separated by a brief ISI. This may indicate that an unbiased, immediate sensory representation of a brief first stimulus in the 2AFC task can only be maintained very briefly and thus be directly compared with the second interval when ISI is relatively short. For longer ISIs, the internal representation must be retained in a perceptual short-term store, which might render it prone to integration with residual memory traces of previously stored stimulus instances, thereby evoking the negative Type B effect (for an overview on temporal processing differences within different timescales, see, for example, Wittmann, 2013).

By a similar logic, the time course of the decay of memory representations of recent stimuli has been studied by varying the intertrial interval (ITI). For example, Decarlo (1992) demonstrated that the stimulus magnitude on the previous trial in an auditory magnitude discrimination task exerted less influence on the response in the current trial when the ITI was increased. More recently, Jaffe-Dax, Frenkel, and Ahissar (2017) found weaker assimilation towards preceding stimuli that was evident in both psychophysical and electroencephalography (EEG) measures as a result of an increased ITI. Hence, empirical results are consistent with the idea that implicit memory representations of previously presented stimuli decay as time unfolds (see Lu, Williamson, & Kaufman, 1992, for a thorough discussion of the decay of sensory memory representations).

Reconciling these findings with the updating mechanism specified by IRM, we conjectured that prolonging ITI in a 2AFC discrimination task should result in a decay of the internal reference representation $\mathbf{I}_{n-1}$. Consequently, if this representation is less available for integration with current information, it should receive less weight in the integration process. In terms of IRM, this corresponds to a reduction in the weighting factor g, which in turn implies a reduction in the negative Type B effect (cf. Dyjas et al., 2012, Figure 2). In this study, we tested this hypothesis by conducting a 2AFC duration discrimination experiment with varying ITI. Specifically, we assessed DL for both stimulus orders $\langle sc \rangle$ and $\langle cs \rangle$ with either short (1,600 ms) or long (3,200 ms) ITI. The reduction in the negative Type B effect, as predicted by IRM, should then become evident in an interaction of stimulus order and ITI length.[2]

Methods

Participants

In total, 18 female and 6 male volunteers (mean age $25.4 \pm 4.3\,\text{years}$) participated in the study. All participants provided informed written consent and were reimbursed with €8/h or course credit. Every participant reported normal hearing. Data were screened for outliers according to a predefined exclusion criterion, that is, we assessed whether each participant's parameter estimates exceeded the sample's three-sigma range within any of the experimental conditions. The data of two participants were replaced due to this criterion.[3]

Apparatus and stimuli

The experiment was written in MATLAB (The MathWorks, Inc., Natick, MA, USA) using the Psychophysics Toolbox extensions (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007; Pelli, 1997). Instructions and feedback were presented on a computer screen in white letters on a black background. A white dot in the centre of the screen (diameter 1 mm) served as fixation point. Auditory stimuli were 800-Hz sine tones (ramped 5 ms on- and offsets), presented binaurally through headphones with an intensity of 65 dB (A). The duration of the standard s was kept constant at 500 ms. The duration of the comparison c ranged from 400 to 600 ms in steps of 20 ms, that is, 400 ms, 420 ms, 440 ms, etc. The level of $c = 500$ ms was omitted because neither of the two response alternatives (first or second stimulus longer) is objectively correct when s and c are physically equal. This yielded a total of 10 levels of duration for c. The "y" and "m" keys of a standard QWERTZ keyboard served as response keys.

Design and procedure

In each trial, a fixation cross was presented at the screen centre and remained on the screen during the whole trial until the response. The two auditory stimuli, s and c, were presented successively, separated by an ISI of 1,000 ms. The order of these two stimuli varied randomly from trial to trial, that is, in half of the trials, s preceded c (order $\langle sc \rangle$), and the other half, c preceded s (order $\langle cs \rangle$). Participants then indicated whether the first or the second stimulus was longer by pressing "y" or "m," respectively. Immediately after the response, feedback was given for 400 ms by a centrally presented "+" or "–" after a correct or incorrect response, respectively. Then, the fixation point reappeared and remained on the screen for an ITI of 1,600 ms (short ITI condition) or 3,200 ms (long ITI condition), before the next trial started. One half of the experiment consisted of the short ITI condition and the other half consisted of the long ITI condition, with the order of these conditions counterbalanced between participants. In each half of the experiment, that is, for each ITI length, participants completed one practice block and 16 experimental blocks with 20 trials (i.e., 10 levels of $c \times 2$ stimulus orders), presented in random order. A feedback indicating the percentage of correct responses was presented after each block. Practice trials did not enter data analysis.

Results

Logistic psychometric functions were fitted to the data of each condition and each participant. The psychometric functions for the two stimulus orders in each ITI condition were constrained such that their average passes through the point (s, 0.5) (for a rationale of this constraint, see Bausenhart, Dyjas, Vorberg, & Ulrich, 2012; Ulrich, 2010; Ulrich & Vorberg, 2009). DL and PSE were derived from these psychometric functions as measures of discrimination sensitivity and judged duration.

The scatterplot in Figure 1 depicts individual (dots) and average (squares) DL estimates for stimulus order $\langle cs \rangle$

on the y-axis plotted against the DL estimates for stimulus order $\langle sc \rangle$ on the x-axis, as a function of ITI condition. For both ITI conditions, all but a few data points lie on or above the main diagonal, illustrating a negative Type B effect for most participants. Importantly, data points for the short ITI condition (filled symbols) are, on average, shifted upwards relative to the long ITI (empty symbols) condition. This latter result pattern indicates a stronger negative Type B effect for the short ITI condition than for the long ITI condition, consistent with our hypothesis.

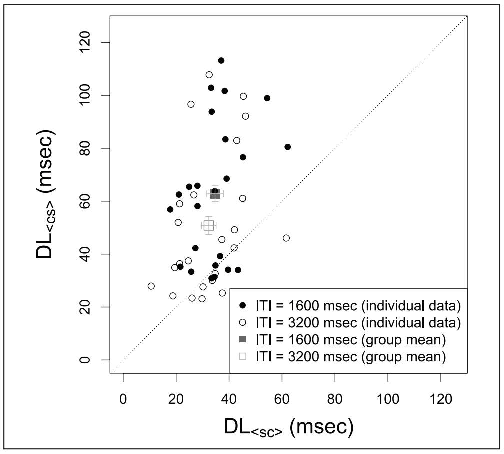

Figure 1. Scatterplot of individual DL estimates (points) and group means (squares) ±1 standard error (cf. Morey, 2008) for both stimulus orders as a function of ITI length.

The left panel of Figure 2 depicts mean DL as a function of stimulus order and ITI. For statistical analysis, a 2 (stimulus order) × 2 (ITI length) repeated-measures analysis of variance (ANOVA) was conducted. The main effect of stimulus order was statistically significant, $F(1, 23) = 22.35, p < .001, \eta_p^2 = 0.49$, reflecting larger average DL in $\langle cs \rangle$ trials (57 ms, $SD = 27$ ms) compared with $\langle sc \rangle$ trials (34 ms, $SD = 11$ ms). This finding reveals a negative Type B effect. The main effect of ITI length was significant, too, $F(1, 23) = 20.07, p < .001, \eta_p^2 = 0.47$, reflecting a larger average DL in the short ITI condition (49 ms, $SD = 24$ ms) compared with the long ITI condition (42 ms, $SD = 22$ ms). Most importantly, the interaction of stimulus order and ITI length on DL was also significant, $F(1, 23) = 13.10, p = .001, \eta_p^2 = 0.36$, reflecting a smaller DL increase in the long ITI condition (18 ms) than in the short ITI condition (28 ms), hence confirming our expectation of a decreased negative Type B effect with longer ITI.

PSE differed only negligibly between ITI conditions, that is, 501 ms ($SD = 16$ ms) in the short ITI condition versus 501 ms ($SD = 18$ ms) in the long ITI condition (see right panel of Figure 2). Neither the main effects of stimulus order nor ITI length were significant, $F(1, 23) = 0.51, p = .482$ and $F(1, 23) = 0.15, p = .702$, respectively. The interaction of both factors was insignificant, too, $F(1, 23) = 2.34, p = .140$.

Model fit

IRM was fitted individually to the data of each participant according to the procedure outlined in Bausenhart et al. (2015). Specifically, IRMs' theoretical psychometric

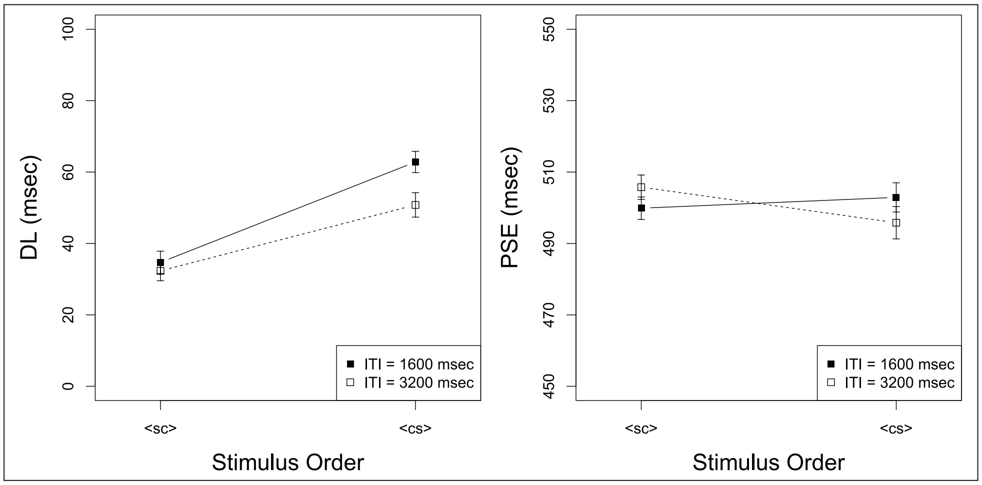

Figure 2. Mean DL (left panel) and PSE (right panel) as a function of stimulus order and ITI. Error bars reflect ±1 within-subjects standard error of the mean according to a suggestion of Morey (2008).

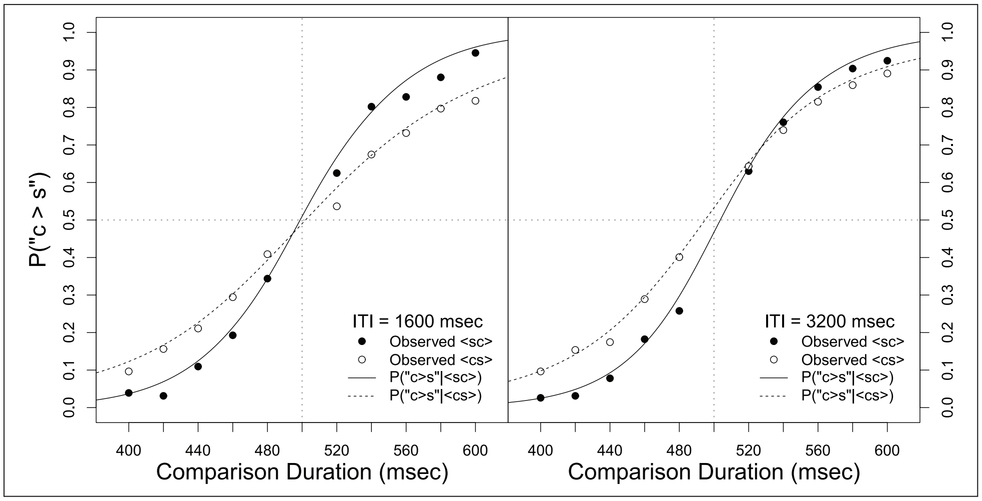

Figure 3. Fitted psychometric functions of the internal reference model (IRM) for stimulus orders $\langle sc\rangle$ and $\langle cs\rangle$ for the short ITI (left panel) and the long ITI (right panel). The *x*-axis represents the duration of the comparison c and the *y*-axis represents the probability of judging the comparison duration c longer than the standard duration s. Data points and psychometric functions depict the average relative response frequencies and the average of fitted psychometric functions across participants, respectively.

functions for a 2AFC task with random stimulus order, depicting the probability of judging the comparison duration c larger than the standard duration s, are (cf. Dyjas & Ulrich, 2014)

$$P(''c > s'' \mid \langle sc\rangle) = \Phi\left[\frac{\gamma + (c-s)}{\kappa}\right] \tag{2}$$

for stimulus order $\langle sc\rangle$ and

$$P(''c > s'' \mid \langle cs\rangle) = \Phi\left[\frac{-\gamma + (c-s)\cdot(1-g)}{\kappa}\right] \tag{3}$$

for stimulus order $\langle cs\rangle$, with

$$\kappa = \sqrt{\frac{2\cdot\sigma^2}{1+g} + \frac{g^2\cdot(1-g)\cdot\sigma_c^2}{2\cdot(1+g)}} \tag{4}$$

The parameter σ reflects the variability (noise) of the perceptual stimulus representations, γ is a bias parameter (e.g., indicating a preference for one of the two response options). σ_c is a constant referring to the standard deviation of the comparison durations employed in the experiment. Most important to the present investigation is the weighting parameter g, which reflects the relative weighting of past and present stimulus information. Larger values of g correspond to stronger weighting of past relative to present stimulus information. These functions were fitted to each observed individual psychometric function by minimising the root mean squared error (RMSE). Figure 3 depicts these models fits, and the average estimated parameters along with the average RMSEs are given in Table 1. Consistent with the theoretical considerations outlined in the Introduction, the average estimated g is smaller in the long than in the short ITI condition, as confirmed by a one-sided paired-samples *t* test, $t(23) = 2.01, p = .028$.

Discussion

Psychophysical research indicates that humans' stimulus discrimination not only relies on present stimulus information but also depends on the history of stimulation. For example, the IRM postulates that participants compare stimuli against a dynamically updating internal reference which represents a conglomerate of past and current stimulus information (Dyjas et al., 2012). This study examined the temporal stability of this internal reference representation. Specifically, in the light of earlier psychophysical research (e.g., Ashourian & Loewenstein, 2011; Magnussen & Greenlee, 1999; Schab & Crowder, 1988; Wearden & Ferrara, 1993) it seemed plausible that the internal reference constitutes short-term memory representation that may fade over time. Accordingly, the prediction was made that a relatively long temporal interval between trials in a 2AFC task should lead to decay of the representation that incorporates stimulus information from previous trials, and consequently to a decreased negative Type B effect, compared with a relatively short interval.

Consistent with this assumption, the observed *DL* difference between $\langle sc\rangle$ and $\langle cs\rangle$ trials was smaller in the

Table 1. Means and standard errors of estimated model parameters of the internal reference model.

Intertrial interval (ITI)	g	σ	γ	RMSE
1,600 ms	0.35 (0.07)	40.0 (2.8)	0.3 (2.5)	0.091 (0.004)
3,200 ms	0.23 (0.08)	34.4 (3.0)	−5.2 (2.6)	0.081 (0.004)

Note. Functions were fitted to the observed individual psychometric functions. The parameters σ and γ are measured in milliseconds. The root mean squared error (RMSE) indicates the goodness of fit.
When fitting IRM to individual psychometric functions, the parameter g was constrained to a range of $-1 < g < 1$, to avoid systematic bias or overestimation for participants exhibiting a numerically positive Type B effect (cf. Bausenhart, Dyjas, & Ulrich, 2015).

long ITI condition than in the short ITI condition, that is, the negative Type B effect was reduced for the long compared with the short ITI. Notably, prolonging the ITI did not just evoke a general decrease in discrimination sensitivity irrespective of stimulus order, but specifically led to a modulation of discrimination sensitivity that was dependent on stimulus order. Consistent with this result, a model fit of IRM to the data of the short and long ITI condition indicates a reduction in the parameter g, which regulates the relative contribution of prior and current stimulus information in judgement. Hence, looking at the present results through the goggles of IRM, participants might give less weight to past stimulus information when it has already become too noisy (Ashourian & Loewenstein, 2011; Lee & Harris, 1996) after a relatively long temporal retention interval.

Interestingly, decay of sensory memory representations occurs for some but not all stimulus attributes (for a review, see Magnussen & Greenlee, 1999). For example, Greenlee, Magnussen, and Thomas (1991) argued that memory decay occurs for *prothetic* perceptual continua but not for *metathetic* ones (in their terminology, *intensive* and *extensive*, respectively). According to Stevens (1957) and Stevens and Galanter (1957), one refers to a prothetic continuum when a change in a stimulus dimension produces a quantitative change in sensation (e.g., loudness, brightness, heaviness, contrast). Conversely, when a change produces a qualitative difference in sensation, one speaks of a metathetic continuum (e.g., colour, pitch, spatial position). The difference between these two dimensions has also been linked to different types of neural codes representing perceptual memories of metathetic and prothetic attributes (see Stevens, 1957; Stevens & Galanter, 1957). These authors have conjectured that a prothetic continuum is associated with additive neuronal activity, whereas a metathetic one emerges form substitutive neuronal activity (e.g., activity associated with different populations of receptors).

With regard to this distinction, it is particularly interesting that the present results suggest decay of the duration information stored in the internal reference. On one hand, this finding is consistent with the hypothesis that duration belongs the class of prothetic continua (Stevens & Galanter, 1957) and thus is susceptible to memory decay as suggested by Greenlee et al. (1991). On the other hand, as time is not a physical stimulus (e.g., Donahoe & Burgos, 1999), it is difficult to unambiguously classify duration as either a metathetic or prothetic stimulus attribute. In this case, it is also conceivable that not the duration information per se was subject to decay in this study, but other prothetic aspects of the stimulation incorporated in the internal reference representation. For example, the internal reference in the present duration discrimination task may have, besides temporal information, contained nontemporal prothetic stimulus aspects (e.g., loudness), which carry the temporal information and which are susceptible to decay. In other words, the representational format of the internal reference might not be completely abstract, but rather stimulus-based.

This study investigated the temporal stability of the internal reference through its supposed decay with the passage of time. A related issue worthwhile of investigation may be its susceptibility to interfering stimulation (e.g., Ashourian & Loewenstein, 2011; Huang & Sekuler, 2010; Magnussen & Greenlee, 1999). For example, we are currently investigating whether and how concurrent stimuli presented during the ITI further distort the integration of prior and current stimulus information. Varying nontemporal and temporal characteristics of the distorting information might prove useful to inform about the properties and representational format, of the internal reference, in particular, and of duration information, in general.

Various formal models closely related to IRM have been put forth (e.g., Arzounian et al., 2017; Raviv et al., 2012; Shi, Church, & Meck, 2013), all of which share the core assumption that the representation of a currently judged stimulus is biased towards or mixed with a memory representation of previously presented stimuli. In general, such a mechanism might serve the function of providing a stable representation of a constantly changing and noisy environment (see Bausenhart et al., 2016; van Rijn, 2016, for reviews with an emphasis on the temporal domain). Put differently, humans seem to reduce uncertainty about the present by relying on their memory of the immediate past, as the past generally predicts the future fairly well, because the world is relatively stable across short time periods (see also Fischer & Whitney, 2014; St John-Saaltink, Kok, Lau, & de Lange, 2016).

Although this study supports the existing picture that assimilation towards past stimuli is a general and ubiquitous mechanism in stimulus judgement, it should be noted that under certain circumstances present stimuli are contrasted away rather than assimilated towards recent stimuli. For example, in absolute identification, where participants have to identify a stimulus from a set of items, a routinely observed finding is that there is assimilation to the response of the preceding trial but contrast to the response of earlier trials (Mori & Ward, 1995). It has been suggested that the presence of feedback might play a crucial role in regulating the dynamics of assimilation and contrast (Ward & Lockhead, 1971). Regarding this issue, it should be noted that we conducted experiments with feedback indicating which stimulus was the longer one (Bausenhart et al., 2012; Dyjas et al., 2012; Dyjas & Ulrich, 2014), whether the response was correct or not (Dyjas, Bausenhart, & Ulrich, 2014, present study), and without feedback (Ellinghaus et al., 2018). Interestingly, reliable Type B effects were observed in all cases, suggesting that the emergence of the Type B effect does not crucially depend on the presence or absence of a certain type of feedback. Nevertheless, a within-subjects comparison of different feedback conditions might be informative concerning the influence of feedback on the Type B effect and its underlying mechanism.

An interesting recent development in the investigation of sequential effects is the attempt to dissociate whether they are based on previous stimuli, previous responses, or both. Within this line of research, it has been suggested that sequential effects may emerge in different stages of processing, ranging from early sensory to decisional stages of processing (e.g., Fornaciai & Park, 2018; Fritsche, Mostert, & de Lange, 2017; Frund, Wichmann, & Macke, 2014; St John-Saaltink et al., 2016). Some researchers argue that stimulus history might evoke contrastive effects emerging in early perceptual stages (e.g., repulsive after-effects), whereas assimilatory biases are caused by response history (Fritsche et al., 2017; Pascucci, Mancuso, Santandrea, Libera, & Plomp, 2017), whereas other studies also provide evidence for perception-based assimilatory effects (Fornaciai & Park, 2018; St John-Saaltink et al., 2016).

Regarding this debate, it is interesting to note that the Type B effect, that is, a sensitivity difference between $\langle sc \rangle$ and $\langle cs \rangle$ trials, cannot easily be explained as a response-driven phenomenon, as response history does not systematically differ between stimulus orders. Moreover, the absence of a Type B effect for brief (100 ms) stimuli separated by brief (300 ms) ISIs (Bausenhart et al., 2015) suggests that the Type B phenomenon does not emerge immediately during stimulus encoding, but rather unfolds when stimulus representations have to be maintained in short-term memory for a certain duration, for example, until a comparative decision with a subsequently presented second stimulus can be made. Thus, the Type B effect might be based on assimilation towards persisting memory traces of previously stored stimulus instances within perceptual short-term memory. Of course, this tentative psychological interpretation of the sequential updating mechanism mathematically specified by IRM needs to be substantiated in upcoming research. In summary, however, this study supports the notion that stimulus order effects on discrimination sensitivity in the 2AFC task result from a dynamically updated internal reference which is prone to decay over time.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

This research was supported by the German Research Foundation, DFG grant BA 4110/5-2.

Notes

1. One distinguishes between positive and negative Type B effects (Dyjas & Ulrich, 2014). A positive Type B effect means that sensitivity is lower for $\langle sc \rangle$ trials, whereas a negative Type B effect means that sensitivity is lower for $\langle cs \rangle$ trials. The majority of Type B effects reported in the literature, however, are negative (e.g., Ellinghaus, Ulrich, & Bausenhart, 2018, but see Hellström & Rammsayer, 2015, for an example of a positive Type B effect).
2. We conducted a pre-study with $n = 24$ participants. This pilot experiment contained slightly more trials (360 per ITI duration), no trial-by-trial feedback, and less breaks than the present experiment. In this study, no significant Type B effect could be observed, $F(1, 23) = 0.09$, $p < .770$, although the statistical power was high, that is, close to 90% (see also Ellinghaus et al., 2018). Consequently, there was also no modulation of the Type B effect by ITI duration, $F(1, 23) = 0.75$, $p = .396$ and also no main effect of ITI, $F(1, 23) = 0.16$, $p = .696$. We believe that the null result regarding the Type B effect reflects a Type II error. Actually, the failure to observe a significant effect in a series of replications is quite likely (see Francis, 2012). In any case, absence of the Type B effect is within the scope of IRM (i.e., $g = 0$). Yet, the presence of a reliable Type B effect is a prerequisite to derive any valid conclusion on the role of the ITI manipulation, which is at the core of this study. Therefore, given the lack of Type B effect, the results of the pilot experiment are uninformative regarding the present experimental question. Nevertheless, we think it is important to mention this failure of replication of the Type B effect to avoid publication bias (e.g., Ulrich, Miller, & Erdfelder, 2018).
3. We routinely reject participants with extreme parameter estimates as these may indicate that participants did not adhere to the experimental instructions or were unable to perform the task adequately. However, an analysis of variance (ANOVA) on DLs including all tested participants yielded virtually the same results as reported in the "Results" section. That is, the main effect of stimulus order, the main effect of ITI, and the interaction effect all were significant, $F(1, 25) = 24.91$, $p < .001$; $F(1, 25) = 11.92$, $p = .002$; and $F(1, 25) = 13.29$, $p = .001$, respectively.

References

Arzounian, D., Kerangal, M. D., & Cheveign, A. D. (2017). Sequential dependencies in pitch judgments. *The Journal of the Acoustical Society of America*, *142*, 3047-3057.

Ashourian, P., & Loewenstein, Y. (2011). Bayesian inference underlies the contraction bias in delayed comparison tasks. *PLoS ONE*, *6*, e19551.

Bausenhart, K. M., Bratzke, D., & Ulrich, R. (2016). Formation and representation of temporal reference information. *Current Opinion in Behavioral Sciences*, *8*, 46–52.

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.

Bausenhart, K. M., Dyjas, O., & Ulrich, R. (2015). Effects of stimulus order on discrimination sensitivity for short and long durations. *Attention, Perception, & Psychophysics*, *77*, 1033-1043.

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.

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

Chambers, C., & Pressnitzer, D. (2014). Perceptual hysteresis in the judgment of auditory pitch shift. *Attention, Perception, & Psychophysics*, *76*, 1271–1279.

Crowder, R. G. (1982). Decay of auditory memory in vowel discrimination. *Journal of Experimental Psychology: Learning Memory and Cognition*, *8*, 153–162.

Decarlo, L. T. (1992). Intertrial interval and sequential effects in magnitude scaling. *Journal of Experimental Psychology: Human Perception and Performance*, *18*, 1080–1088.

Decarlo, L. T., & Cross, D. V. (1990). Sequential effects in magnitude scaling: Models and theory. *Journal of Experimental Psychology: General*, *119*, 375–396.

Donahoe, J. W., & Burgos, J. E. (1999). Timing without a timer. *Journal of the Experimental Analysis of Behavior*, *71*, 257–263.

Donkin, C., Chan, V., & Tran, S. (2017). The effect of blocking inter-trial interval on sequential effects in absolute identification. *The Quarterly Journal of Experimental Psychology*, *68*, 129–143.

Durlach, N. I., & Braida, L. D. (1969). Intensity perception. I. Preliminary theory of intensity resolution. *The Journal of the Acoustical Society of America*, *46*, 372–383.

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.

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

Ellinghaus, R., Ulrich, R., & Bausenhart, K. M. (2018). Effects of stimulus order on comparative judgments across stimulus attributes and sensory modalities. *Journal of Experimental Psychology: Human Perception and Performance*, *44*, 7–12.

Fechner, G. T. (1860). *Elemente der Psychophysik* [Elements of psychophysics]. Leipzig, Germany: Breitkopf und Härtel.

Fischer, J., & Whitney, D. (2014). Serial dependence in visual perception. *Nature Neuroscience*, *17*, 738–743.

Fornaciai, M., & Park, J. (2018). Attractive serial dependence in the absence of an explicit task. *Psychological Science*, *29*, 437–446.

Francis, G. (2012). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin and Review*, *19*, 975–991.

Fritsche, M., Mostert, P., & de Lange, F. P. (2017). Opposite effects of recent history on perception and decision. *Current Biology*, *27*, 590–595.

Frund, I., Wichmann, F. A., & Macke, J. H. (2014). Quantifying the effect of intertrial dependence on perceptual decisions. *Journal of Vision*, *14*, 9.

Garner, W. R. (1953). An informational analysis of absolute judgements of loudness. *Journal of Experimental Psychology*, *46*, 373–380.

Geiselman, R. E., Haight, N. A., & Kimata, L. G. (1984). Context effects on the perceived physical attractiveness of faces. *Journal of Experimental Social Psychology*, *20*, 409-424.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (Rev. ed.). Los Altos, CA: Peninsula Publishing (reprinted edition 1988).

Greenlee, M. W., Magnussen, S., & Thomas, J. P. (1991). Different neural codes for spatial frequency and contrast. In A. Valberg & B. B. Lee (Eds.), *From pigments to perception: Advances in understanding visual processes* (pp. 451–452). New York, NY: Plenum Press.

Hegelmaier, F. (1852). Ueber das Gedächtniss für Linear-Anschauungen [On memory for visually perceived lines]. *Archiv für physiologische Heilkunde*, *11*, 844–853.

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.

Hellström, Å, & Rammsayer, T. H. (2015). Time-order errors and standard-position effects in duration discrimination: An experimental study and an analysis by the sensation-weighting model. *Attention, Perception, & Psychophysics*, *77*, 2409–2423.

Helson, H. (1964). *Adaptation-level theory*. New York, NY: Harper & Row.

Hollingworth, H. (1910). The central tendency of judgment. *The Journal of Philosophy, Psychology and Scientific Methods*, *7*, 461–469.

Huang, J., & Sekuler, R. (2010). Distortions in recall from visual memory: Two classes of attractors at work. *Journal of Vision*, *10*, 24.1–24.27.

Jaffe-Dax, S., Frenkel, O., & Ahissar, M. (2017). Dyslexics' faster decay of implicit memory for sounds and words is manifested in their shorter neural adaptation. *eLife*, *6*, 1–19.

King, M. C., & Lockhead, G. R. (1982). Response scales and sequential effects in judgment. *Perception & Psychophysics*, *30*, 599–603.

Kleiner, M., Brainard, D., & Pelli, D. G. (2007). What's new in Psychtoolbox-3? *Perception*, *36*, 1.

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.

Lee, B., & Harris, J. (1996). Contrast transfer characteristics of visual short-term memory. *Vision Research*, *36*, 2159–2166.

Lu, Z., Williamson, S. J., & Kaufman, L. (1992). Behavioral lifetime of human auditory sensory memory predicted by physiological measures. *Science*, *258*, 1668–1670.

Luce, R. D., & Galanter, E. (1963). Discrimination. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. *I*, pp. 191–243). New York, NY: John Wiley & Sons.

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

Magnussen, S., & Greenlee, M. W. (1999). The psychophysics of perceptual memory. *Psychological Research*, *62*, 81–92.

Magnussen, S., Thomas, J. P., & Greenlee, M. W. (1996). Parallel processing in visual short-term memory. *Journal of Experimental Psychology: Human Perception and Performance*, *22*, 202–212.

Matthews, W. J., & Stewart, N. (2009). Psychophysics and the judgment of price: Judging complex objects on a non-physical dimension elicits sequential effects like those in perceptual tasks. *Judgment and Decision Making*, *4*, 64-81.

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology*, *4*, 61–64.

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.

Mori, S., & Ward, L. M. (1995). Pure feedback effects in absolute identification. *Perception & Psychophysics*, *57*, 1065–1079.

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

Parducci, A., & Marshall, L. M. (1962). Assimilation vs. contrast in the anchoring of perceptual judgments of weight. *Journal of Experimental Psychology*, *63*, 426-437.

Pascucci, D., Mancuso, G., Santandrea, E., Libera, C. D., & Plomp, G. (2017). Laws of concatenated perception: Vision goes for novelty, Decisions for perseverance. *bioRxiv*. Advance online publication. doi:10.1101/229187

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

Raviv, O., Ahissar, M., & Loewenstein, Y. (2012). How recent history affects perception: The normative approach and its heuristic approximation. *PLoS Computational Biology*, *8*, e1002731.

Raviv, O., Lieder, I., Loewenstein, Y., & Ahissar, M. (2014). Contradictory behavioral biases result from the influence of past stimuli on perception. *PLoS Computational Biology*, *10*, e1003948.

Schab, F. R., & Crowder, R. G. (1988). The role of succession in temporal cognition: Is the time-order error a recency effect of memory? *Perception & Psychophysics*, *44*, 233–242.

Shi, Z., Church, R. M., & Meck, W. H. (2013). Bayesian optimization of time perception. *Trends in Cognitive Sciences*, *17*, 556–564.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review*, *64*, 153–181.

Stevens, S. S., & Galanter, E. H. (1957). Ratio scales and category scales for a dozen perceptual continua. *Journal of Experimental Psychology*, *54*, 377–411.

Stewart, N., Brown, G. D. A., & Chater, N. (2005). Absolute identification by relative judgment. *Psychological Review*, *112*, 881–911.

St John-Saaltink, E., Kok, P., Lau, H. C., & de Lange, F. P. (2016). Serial dependence in perceptual decisions is reflected in activity patterns in primary visual cortex. *Journal of Neuroscience*, *36*, 6186–6192.

Taubert, J., Burg, E. V. D., & Alais, D. (2016). Love at second sight: Sequential dependence of facial attractiveness in an on-line dating paradigm. *Scientific Reports*, *6*, 22740.

Thönes, S., Von Castell, C., Iflinger, J., & Oberfeld, D. (2018). Color and time perception: Evidence for temporal overestimation of blue stimuli. *Scientific Reports*, *8*, 1688.

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., Miller, J., & Erdfelder, E. (2018). Effect size estimation from t-statistics in the presence of publication bias. *Zeitschrift für Psychologie*, *226*, 56–80.

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

van Rijn, H. (2016). Accounting for memory mechanisms in interval timing: A review. *Current Opinion in Behavioral Sciences*, *8*, 245–249.

Ward, L. M. (1979). Stimulus information and sequential dependencies in magnitude estimation and cross-modality matching. *Journal of Experimental Psychology: Human Perception and Performance*, *5*, 444–459.

Ward, L. M., & Lockhead, G. R. (1971). Response system processes in absolute judgment. *Perception & Psychophysics*, *9*, 73–78.

Wearden, J. H., & Ferrara, A. (1993). Subjective shortening in humans' memory for stimulus duration. *The Quarterly Journal of Experimental Psychology*, *46*, 163–186.

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

Wittmann, M. (2013). The inner sense of time: How the brain creates a representation of duration. *Nature Reviews Neuroscience*, *14*, 217–223.

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.

Zotov, V., Jones, M. N., & Mewhort, D. J. K. (2011). Contrast and assimilation in categorization and exemplar production. *Attention, Perception, & Psychophysics*, *73*, 621–639.

Appendix C Study 3

Ellinghaus, R., Giel, S., Ulrich, R., & Bausenhart, K. M. (2019). *Humans integrate duration information within and across modalities: Evidence for a supramodal internal reference of time.* Manuscript submitted to *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

Humans Integrate Duration Information Within and Across Modalities: Evidence for a Supramodal Internal Reference of Time

Ruben Ellinghaus[1], Sophie Giel[1], Rolf Ulrich[1], Karin M. Bausenhart[1]
[1] University of Tübingen

Author Note

Address correspondence to Ruben Ellinghaus, Department of Psychology, University of Tübingen, Schleichstr. 4, 72076 Tübingen, Germany, Electronic mail may be sent to karin.bausenhart@uni-tuebingen.de.

Abstract

Perception is not only driven by current stimulation but also by previous sensory experience, which may serve as a perceptual prior for stimulus processing. A possible mechanism underlying this phenomenon is formalized in the internal reference model (IRM), which assumes that humans rely on an internal reference that updates continuously by integrating past and present stimulus representations. As a direct consequence of this process, discrimination sensitivity is higher when a constant standard precedes rather than follows a variable comparison in two-alternative forced choice tasks. The present study exploited this Type B effect, in order to examine whether an internal reference can be formed across stimuli varying within and across modalities. In a series of four experiments, task-irrelevant and/or task-relevant features either remained constant or varied from trial to trial. In duration discrimination, the Type B effect as a proxy of perceptual prior formation was not affected by variation of pitch or modality. However, in intensity discrimination, it was considerably reduced when stimulus modality and thus also the task-relevant feature (i.e., brightness and loudness) alternated rather than remained constant. These results indicate that temporal information can be isolated from stimulus-specific features and integrated across subsequent trials into a perceptual prior for stimulus judgment, whereas intensity information from different modalities seems to hamper this process. In general, our findings contribute to a growing body of research aimed at understanding the formation and content of perceptual priors.

Keywords: time perception, stimulus discrimination, internal reference, Type B effect, stimulus history

Humans Integrate Duration Information Within and Across Modalities: Evidence for a Supramodal Internal Reference of Time

A main endeavor of psychophysical research is to determine the relation between sensory stimuli and the sensations they elicit (e.g., Fechner, 1860). An intriguing finding hereby is that sensory experience and judgment ist not only driven by present but also by past stimulation (e.g., Helson, 1964; Kiyonaga, Scimeca, Bliss, & Whitney, 2017). For instance, the judged magnitude of a current stimulus is often attracted towards a recent stimulus (e.g., Fischer & Whitney, 2014; Garner, 1953) or towards the average of all stimuli presented in an experiment (e.g., Hollingworth, 1910). Such assimilation effects towards the history of stimulation have been documented across a range of different modalities, tasks, and domains, as for example judgments of orientation (Fischer & Whitney, 2014), numerosity (Fornaciai & Park, 2018), position (Manassi, Liberman, Chaney, & Whitney, 2017), motion (Alais, Leung, & Burg, 2017), summary scene statistics (Manassi et al., 2017), facial attractiveness (Xia, Leib, & Whitney, 2016), ambiguous figures (Maloney, Martello, Sahm, & Spillmann, 2005), and, most important for the present purposes, duration (Gu & Meck, 2011; Jazayeri & Shadlen, 2010; Maaß, Schlichting, & van Rijn, 2019).

On a functional level, these effects might reflect a mechanism which is adaptive for constructing a stable internal representation of the environment. Specifically, the memory of the immediate past is usually a valid predictor of the present, because the world is fairly stable across short temporal intervals (Dong & Atick, 1995). Thus, humans may effectively reduce uncertainty about the present by relying — at least to a certain extent — on the past (Burr & Cicchini, 2014; Cicchini, Anobile, & Burr, 2014; Cicchini, Mikellidou, & Burr, 2017; Fischer & Whitney, 2014). Generally, one may therefore conceive perception as a process of drawing inferences about external events and stimuli based on both past and present internal states, that is, as predictive coding (e.g., von Helmholtz, 1867; Friston & Kiebel, 2009; Kersten, Mamassian, & Yuille, 2004; Kersten & Yuille, 2003; Knill & Richards, 1996; Yuille & Kersten, 2006).

Memory-Mixing as a source of the central tendency, sequential effects, and stimulus-order effects

An emerging question in this regard concerns the mechanisms and principles that guide this inferential process. Various researchers have suggested that past and present stimulus information is combined by mixing the memory representations of previously presented stimuli with the representations of current stimuli (e.g., Arzounian, Kerangal, & Cheveign, 2017; Dyjas, Bausenhart, & Ulrich, 2012; Fischer & Whitney, 2014; Jazayeri & Shadlen, 2010; Penney, Gibbon, & Meck, 2000; Raviv, Lieder, Loewenstein, & Ahissar, 2014; Shi, Church, & Meck, 2013). This memory mixing may proceed such that each currently encountered stimulus representation is combined with a memory prior representing the whole distribution of previously presented stimuli (e.g., as in the Bayesian integration account brought forward by Jazayeri & Shadlen, 2010), or in a more dynamic fashion such that a single "internal reference" representation is continuously updated with each encountered stimulus and thus changes on a trial-by-trial basis (e.g., Burr & Cicchini, 2014; Bliss, Sun, & D'Esposito, 2017; Dyjas et al., 2012). Such memory-mixing models account for a variety of well-established findings, such as the central tendency of judgements, biased judgements resulting from uneven stimulus distributions, and (at least in the case of dynamic accounts) also for sequential trial-by-trial modulations in perceptual judgements.

Recently, another empirical finding, the so called Type-B effect, has also been attributed to the build-up of an internal representation based on both past and current stimulation. This effect, which can be observed in Two-Alternative Forced-Choice (2AFC) tasks, holds that the sensitivity for discriminating between two stimuli depends on their order of presentation.[1] Typically, sensitivity for the discrimination between a constant standard stimulus s and a variable comparison stimulus c is lower when s precedes (i.e., stimulus order $\langle sc \rangle$) rather than follows c (i.e., stimulus order $\langle cs \rangle$). This specific result pattern can be observed when stimulus order is constant within a block of trials as well as when it varies

---

[1] The Type-B effect refers to a genuine difference in the discrimination threshold and should thus not be confused with Type-A order effects (time-order errors) that may reflect response biases or under-/overestimation of one of two successively presented stimuli (Ulrich & Vorberg, 2009).

unpredictably from trial to trial (Dyjas et al., 2012). Moreover, it has been observed for various stimulus features and modalities (Ellinghaus, Ulrich, & Bausenhart, 2018; Thönes, Von Castell, Iflinger, & Oberfeld, 2018) and different judgment modes as stimulus comparison and equality judgments (Dyjas & Ulrich, 2014).

Interestingly, this rather peculiar phenomenon follows directly from a model that posits an internal reference which combines past and present stimulus instances and updates continuously from trial to trial (see also Durlach & Braida, 1969; Morgan, Watamaniuk, & McKee, 2000; Nachmias, 2006). Specifically, the Internal Reference Model (IRM, Dyjas et al., 2012; Lapid, Ulrich, & Rammsayer, 2008) states that participants choose the larger (i.e., brighter, louder, etc.) of two successively presented stimuli in a current trial by comparing the internal representation of the second stimulus in this trial against an internal reference, which is a geometric moving average (Roberts, 1959) of the series of stimuli presented at the first temporal position within each trial. In other terms, it is a conglomerate of previous and current instances of the first stimulus presented in each trial, for which with a weighting factor regulates the relative contribution of prior and current information.

Following from this mechanism, the expected value of the internal reference is always equal to the standard s, as long as the values of c are distributed symmetrically around s. On $\langle cs \rangle$ trials, the internal representation of the first stimulus c will consequently be pulled towards s. In turn, this will decrease the subjective difference $\mathbf{D}_n$ and thus lower discrimination sensitivity compared to $\langle sc \rangle$ trials — as observed in the typical Type B effect. Moreover, this updating mechanism can also account for the above-mentioned assimilation effects such as the central tendency and sequential trial-by-trial modulations in perceptual judgments (Bausenhart, Dyjas, & Ulrich, 2014). Therefore, it seems promising to regard these different phenomena as caused by a common mechanism which mixes past and present stimulus representations.

As outlined above, this mechanism may reflect persisting memory traces from previous stimulation which are superposed by representations of current stimulation. In line with this assumption, the Type B effect diminishes with increasing intertrial interval (Ellinghaus, Gick, Ulrich, & Bausenhart, 2019). This can be attributed to decay of the internal reference from the previous trial with increasing delay between trials, which will result in a reduced influence of prior (i.e., a smaller weighting factor) relative to current stimulus information in the integration process. This interpretation is highly consistent with recent demonstrations that sequential assimilatory effects get more pronounced with increasing delay of responses to a current stimulus, and get less pronounced with increasing delay between previous and the current stimulus instances (Bliss et al., 2017; Fritsche, Mostert, & de Lange, 2017). Consequently, various experimental observations, such as the central tendency, sequential effects, and stimulus-order effects are likely to reflect the involvement of mnemonic processes which combine recent stimulus history with current stimulus input for perceptual decision making.

What exactly is “mixed” in memory?

Even though there is ample evidence in favor of the general concept of such memory-mixing, so far, not much is known about what exactly is encoded in the resulting “mixed” stimulus representation. Current mathematical accounts of memory mixing such as IRM operate on the internal representations of the specific stimulus feature relevant to the task at hand, as for example, the spatial orientations of the stimuli in an experiment probing perceived orientation of tilted Gabor patches (Fischer & Whitney, 2014), or the durations of the presented auditory intervals in a temporal discrimination task (Dyjas et al., 2012). Yet, any external stimulus is necessarily composed of a variety of features. For instance, a tilted Gabor patch has a certain frequency, location, contrast, color, size and envelope, or an auditory interval of a certain duration may be presented from a specific location and with a certain frequency, amplitude, and phase.

Such stimulus features are often cognitively related. For example, it is well-known that variation in task-irrelevant features may impair task performance (Garner, 1974). Moreover, task-irrelevant features can also directly affect perceptual judgements in the task-relevant domain. For example, perceived duration typically depends on physical interval duration, but also on other features as stimulus size (Mo & Michalski, 1972; Rammsayer & Verner, 2014), modality (Wearden, Edwards, Fakhri, & Percival, 1998), (relative) intensity (Goldstone, Lhamon, & Sechzer, 1978; Matthews, Stewart, & Wearden, 2011), spatial position (Kliegl & Huckauf, 2014), and novelty (Matthews, 2015). Therefore, undoubtedly, task-irrelevant features are encoded and processed along with task-relevant stimulus features.

Yet, it is unclear whether such task-irrelevant features also enter the memory-mixing mechanism. Theoretically, it is conceivable that participants can isolate and extract the task-relevant information from a stimulus and encode this information into an internal reference or perceptual prior [2]. One may in this case speak of *feature-based* coding. In this case, a single internal reference would be formed irrespective of variations in task-irrelevant features. Another viable assumption is that an internal reference may consist of a comprehensive stimulus representation including various stimulus features, that is, *stimulus-based* coding. In this latter case, it is conceivable that separate internal references are built for stimuli that differ along various dimensions, or otherwise, that the build-up of any internal reference is hampered when multiple references cannot be easily maintained due to decay of the respective memory traces (cf. Bliss et al., 2017; Ellinghaus et al., 2019; Fritsche et al., 2017).

So far, empirical evidence on this question is rather scarce and unequivocal (e.g., Baugh, Yak, Johansson, & Flanagan, 2016; Roach, Mcgraw, Whitaker, & Heron, 2017; Rhodes, Seth, & Roseboom, 2018). For example, in the domain of temporal processing, Rhodes et al. (2018) had their participants indicate whether a forth stimulus appeared early or late relative to a rhythm induced by a series of three preceding stimuli. These inducing stimuli could be either auditory or visual, and crucially, on each trial their interstimulus interval was drawn from a different duration distribution for each modality. For example, visual interstimulus intervals could be on average shorter (ranging around a mean of 650 msec) than auditory interstimulus intervals (ranging around a mean of 1550 msec), or vice versa. Importantly, auditory perceptual judgements were biased towards the mean of the auditory interstimulus interval distributions, whereas visual judgements were biased towards the mean of the visual distributions. This indicates that separate temporal priors for

---

[2] In the following, we will not distinguish further between these two terms and use them interchangeably.

the interstimulus intervals were formed based on the modality of the inducer stimuli. A similar result was obtained when the inducer stimuli varied between white noise and pure tones, but not for inducer stimuli varying between low and high pitch. In sum, the authors concluded that participants generalize across different duration distributions when the inducing stimuli just vary in pitch, but form separate perceptual priors based on the task-irrelevant stimulus features when these are perceptually more distinct.

This finding contrasts with a study by Roach et al. (2017), wherein participants had to manually reproduce intervals presented through varying stimuli (e.g., visually and auditorily marked intervals, or intervals presented to the left and right of fixation). Different stimuli were associated with different duration distributions (e.g., visual stimulus durations could be centered around a mean of 320 msec, and auditory stimulus durations could be centered around a mean duration of 1280 msec). Interestingly, when trials from both modalities were presented randomly interleaved, reproductions were shifted towards the mean of all presented durations. Thus, in contrast to the results of Rhodes et al. (2018), participants formed a general perceptual prior across the temporal distributions associated with the different task-irrelevant stimulus attributes as stimulus location and modality. Evidence for separate perceptual priors for the different temporal distributions was, however, observed when the intervals did not only differ in location, but also in the specific task assigned to each stimulus location (e.g., manual reproduction for stimuli presented on the left and binary discrimination for stimuli presented on the right). Therefore, priors might not be based on initial sensory experience but rather on information represented in decisional or motor processing stages.

Consistent with this notion, it has been suggested that assimilatory bíases towards previous trials do not emerge on a perceptual but on a decisional level (Fritsche et al., 2017; Pascucci, Mancuso, Santandrea, Libera, & Plomp, 2017). In detail, these authors argue that perceptual representations of current stimuli are contrasted away from preceding stimuli (i.e., repulsive after-effects), while assimilatory history effects are caused by a response bias towards responses given on earlier trials (but see Fornaciai & Park, 2018, for a demonstration of sequential assimiliation towards passively viewed stimuli). In this regard, investigation of the Type B effect as a proxy of memory-mixing seems advantageous, because the history of given responses should not differ systematically between stimulus orders in a 2AFC task, and therefore, the Type B effect cannot be explained as a response-driven phenomenon.

In sum, humans apparently can both generalize across and differentiate between stimulus distributions based on task-irrelevant features when making use of prior knowledge to guide perception, but it is unclear under which conditions one or the other of these two cases arises. The present series of experiments tackles this issue from the perspective of IRM. We investigated how variation in task-irrelevant (Experiments 1-3) and task-relevant stimulus features (Experiment 4) affects the build-up of an Internal Reference in a 2AFC task, as indicated by the magnitude of the Type B effect. More specifically, in Experiment 1, participants completed an auditory 2AFC duration discrimination task in which the frequency of the pure tones marking the interval duration either remained constant across trials or alternated from trial to trial between high and low pitch. The variation in task-irrelevant features was intensified by mixing 15 different tone frequencies in Experiment 2 and by mixing auditory and visual intervals in Experiment 3. Finally, in Experiment 4, we varied both task-relevant and task-irrelevant features in a task requiring auditory and visual intensity discrimination.

In general, if participants isolate and encode only task-relevant information into their internal references (*feature-based* coding), the build-up of the internal reference should not be influenced by variation of task-irrelevant features. Accordingly, the magnitude of the Type B effect should not differ between the conditions in which task-irrelevant-features remain constant within a block of trials and in which they alternate on a trial-by-trial basis. Alternatively, participants might also encode task-irrelevant features into the internal reference and hence construct of separate internal references for different stimulus types (*stimulus-based* coding). In this case, due to the trial-by-trial alternation of the different stimulus types, each of the separate internal references would have to be maintained in memory for a longer time before it can be updated in the next trial presenting the same stimulus type. Therefore, each internal reference would be subject to stronger memory decay in alternating compared to constant conditions, resulting in a reduced Type B effect in the former case (cf. Ellinghaus et al., 2019). A third hypothesis ($multiple - feature$ coding), which may be regarded as intermediate between these two extremes, will be outlined in the Introduction of Experiment 3, where it becomes relevant.

Experiment 1

In this experiment, participants performed a 2AFC duration discrimination task with a standard duration s of 500 ms and variable durations of the comparison stimuli c. The order of s and c varied randomly between trials, and discrimination thresholds for each stimulus order were determined according to an adaptive procedure (Kaernbach, 1991) that was previously successfully employed to investigate the Type B effect across a variety of stimulus attributes and modalities (Ellinghaus et al., 2018). The presented intervals were filled by low- and high-frequency pure tones, which either remained constant across trials in one half of the experiment or alternated from trial to trial in the other half. We expected to observe a typical Type B effect, that is, higher DL in trials with stimulus order $\langle cs \rangle$ than in trials with order $\langle sc \rangle$ when the tone pitch remains constant across trials. This effect should be reduced if the variation of tone frequency leads to the formation of separate, stimulus-based internal references, because the effectively longer inter-trial interval between two repetitions of the same stimulus frequency in the alternating compared to the constant presentation mode implies a stronger decay of each of the internal references (Ellinghaus et al., 2019).

Methods

Participants. 31 female and 17 male volunteers (mean age 22.9 ± 2.4 years) participated in the study. All participants provided written informed consent prior to the experiment. The data of six participants were replaced because their estimated DLs or PSEs were outside the predefined three-sigma range in one or more of the experimental conditions.

Apparatus and Stimuli. The experiment was written in Matlab (The MathWorks, Inc.) using the Psychophysics Toolbox extensions (Kleiner, Brainard, & Pelli, 2007). Instructions and feedback were presented on a computer screen in white on a black background. A white dot in the center of the screen (diameter 1 mm) served as fixation point. Auditory stimuli were 1200 Hz and 400 Hz pure tones with ramped 5 msec on- and offsets, presented binaurally through headphones with an intensity of approximately 60 dB(A). The duration of the standard s was kept constant at 500 msec. The duration of the comparison c adapted to participants' responses as explained in detail hereafter. The 'y' and 'm' keys of a standard QWERTZ keyboard served as response keys.

Design and Procedure. Participants completed an adaptive 2AFC duration discrimination task. In each trial, the fixation cross was presented at the screen center and remained on the screen during the whole trial until the response. The two auditory stimuli, s and c, were presented successively, separated by an ISI of 1000 msec. The order of these two stimuli varied randomly from trial to trial, that is, in half of the trials, s preceded c (order $\langle sc \rangle$), and the other half, c preceded s (order $\langle cs \rangle$). Participants then indicated whether the first or the second stimulus' duration was longer by pressing 'y' or 'm', respectively. Immediately after the response, feedback was given for 400 msec by a centrally presented '1' or '2', respectively indicating whether the first or the second duration was physically longer. Then, the fixation point reappeared and remained on the screen for an ITI of 1600 msec before the next trial started.

During the whole experiment, the duration of c adapted to participants' responses following the weighted up-and-down method (Kaernbach, 1991). In detail, c values of 600 msec and 400 msec were chosen as start values for upper and lower runs, respectively. For the upper (lower) runs, the magnitude of c was decreased (increased) by 20 msec when c was judged as being larger (smaller) than s and increased (decreased) by 60 msec when c was judged as being smaller (larger) than s. The upper and lower runs respectively targeted the 75th and 25th percentile of the psychometric function. The two start values combined with the two trial orders $\langle sc \rangle$ and $\langle cs \rangle$ yielded four independent trial runs of 40 trials each, with all trials presented in random order. Thus, 160 trials were administered to assess the Type B effect in each of the four (high vs low pitch × constant vs. alternating presentation mode) experimental conditions.

These four experimental conditions were created as follows: In one half of the experiment, the frequency of the auditory stimuli remained constant at either 1200 Hz (high pitch block) or 400 Hz (low pitch block). In the other half of the experiment, the frequency alternated between 1200 Hz and 400 Hz from trial to trial (i.e., in a given trial, both s and c were low-pitched, and in the subsequent trial, both were high-pitched, and so on). Accordingly, the constant half consisted of two blocks (high pitch block and low pitch block) that were divided into 8 subblocks of 20 trials each. The alternating half of the experiment consisted of one block of 320 trials that were divided into 16 subblocks of 20 trials each. At the beginning of each block, 20 additional practice trials were presented. Practice trials did not enter data analysis and start values were reset after practice. Between subblocks, a self-terminated break was provided to allow participants to rest. Half of the participants started with the alternating and half of the participants started with the constant presentation mode. The order of the two constant blocks (i.e., high pitch block and low pitch block) was also counterbalanced between participants.

Results

The main hypotheses of this study are based on the Type B effect, that is, on the variation of DL as a function of stimulus order. For the sake of completeness, however, we also report the point of subjective equality (PSE) as a dependent measure, since stimulus order typically affects this parameter, too, producing a time-order error or Type A effect (Fechner, 1860; Hellström, 1985; Ulrich & Vorberg, 2009). For each condition and participant, the reversal points for each run were determined and averaged, yielding estimates of $x_{25\%}$ and $x_{75\%}$. DL was defined as $DL = (x_{75\%} - x_{25\%})/2$ and PSE was defined as $PSE = (x_{75\%} + x_{25\%})/2$. For statistical analysis, a 2 (stimulus order) $\times$ 2 (pitch) $\times$ 2 (presentation mode) repeated-measures analysis of variance (ANOVA) was conducted for both DL and PSE.

Difference Limen. The left panel of Figure 1 depicts mean DL as a function of stimulus order, pitch and presentation mode. The main effect of stimulus order was statistically significant, F(1, 47) = 88.24, $p <$ .001, $\eta_p^2 =$ .65, reflecting larger average DL in $\langle cs \rangle$ trials (M = 79.5 msec, SD = 24.9 msec) compared to $\langle sc \rangle$ trials (M = 55.5 msec, SD = 18.4 msec). This finding reflects the typical negative Type B effect. Most important to the present investigation, this Type B effect was not further affected by whether the pitch stayed constant or alternated from trial to trial, as the interaction of stimulus order and presentation mode was not significant, F(1, 47) = 0.36, p = .553, η_p^2 = .01.[3]

Likewise, there were neither significant interactions between stimulus order and pitch, F(1, 47) = 0.00, p = .978, $\eta_p^2 <$ .01, nor between all three factors, F(1, 47) = 0.77, p = .384, η_p^2 = .02. Moreover, DL did not differ between the alternating (M = 67.4 msec, SD = 23.2 msec) and the constant condition (M = 67.6 msec, SD = 19.2 msec), F(1, 47) = 0.01, p = .915, $\eta_p^2 <$ .01. Unexpectedly, DL was slightly larger in the high pitch condition (M = 69.7 msec, SD = 20.3 msec) than in the low pitch condition (M = 65.4 msec, SD = 21.8 msec), F(1, 47) = 5.31, p = .026, η_p^2 = .10. This effect was further modulated by presentation mode, F(1, 47) = 5.26, p = .026, η_p^2 = .10, reflecting that the aforementioned main effect of pitch originated from the alternating (Δ = 8.3 msec) rather than the constant (Δ = 0.3 msec) presentation mode.

Point of Subjective Equality. The right panel of Figure 1 illustrates mean PSE as a function of stimulus order, pitch, and presentation mode. Numerically, PSE was slightly larger for $\langle sc \rangle$ than for $\langle cs \rangle$ trials, namely 506.8 msec (SD = 13.0 msec) vs. 501.3 msec (SD = 18.0 msec). The corresponding main effect of stimulus order, however, was not significant, F(1, 47) = 1.71, p = .197, η_p^2 = .04, that is, we did not observe a reliable time-order error (Type A effect) averaged across conditions. Also, there was no main effect of presentation mode, F(1, 47) = 0.14, p = .706, $\eta_p^2 <$ .01. Yet, the aforementioned PSE decrease from $\langle sc \rangle$ to $\langle cs \rangle$ trials was marginally larger in the alternating condition (Δ = -10.7 msec) than in the constant condition (Δ = -0.3 msec), F(1, 47) = 3.08, p = .086, η_p^2 = .06.

---

[3] In addition we conducted a Bayesian analysis testing the main hypothesis of this study with the 'BayesFactor' package for R. In detail, a model including the main effects of order and presentation mode yielded a higher Bayes Factor than a model which additionally included the interaction term of order and presentation mode, namely $BF10 = 5.93e^{16} \pm 1.78$ vs. $BF10 = 1.44e^{16} + -2.49$. Reframing this result, the present data are 4.11 more likely under the model including only the main effects. Hence, the Bayesian analysis mirrors the ANOVA reported in the main body in suggesting that the Type B effect does not differ as a function of presentation mode.

The main effect of pitch was also marginally significant, $F(1, 47) = 3.23$, $p = .079$, $\eta_p^2 = .06$, corresponding to a slightly larger PSE in the high pitch condition ($M =$ 505.9 msec, $SD = 8.3$ msec) compared to the low pitch condition ($M = 502.1$ msec, $SD = 10.4$ msec). Pitch did not interact with stimulus order, $F(1, 47) = 0.86$, $p = .359$, $\eta_p^2 = .02$, but with presentation mode, $F(1, 47) = 21.82$, $p < .001$, $\eta_p^2 = .32$. Specifically, in the alternating condition, PSE was substantially higher in high pitch trials (509.0 msec, $SD =$ 12.7 msec) than in low pitch trials (498.4 msec, $SD = 12.6$ msec), whereas the reverse was true for the constant condition, i.e., a higher PSE was observed for the low pitch trials (505.8 msec, $SD = 13.3$ msec) than for the high pitch trials (502.9 msec, $SD = 9.6$ msec). This interaction was further modulated by stimulus order, i.e., the threefold interaction of all factors was significant, $F(1, 47) = 5.78$, $p = .020$, $\eta_p^2 = .11$. In detail, PSE for the high pitch intervals was about 3 msec larger for $\langle sc \rangle$ than for $\langle cs \rangle$ trials in both presentation modes. For the low pitch intervals, however, PSE was about 18 msec larger for $\langle sc \rangle$ than for $\langle cs \rangle$ trials in the alternating condition, and about 3 msec lower for $\langle sc \rangle$ than for $\langle cs \rangle$ trials in the blocked condition.

Discussion

This 2AFC duration discrimination experiment investigated whether integration of task-relevant information (i.e., duration) into an internal reference is modulated by variation of a task-irrelevant stimulus attribute (i.e., frequency). The frequency of the two stimuli to be discriminated either stayed constant at a high or a low pitch (constant condition), or alternated between a high and a low pitch from trial to trial (alternating condition). We observed a typical negative Type B effect, that is, higher DL for $\langle cs \rangle$ trials compared to $\langle sc \rangle$ trials, averaged across conditions. This indicates that participants integrated previous and current stimulus information into an internal reference, on which their decisions about stimulus durations were based. Most importantly, the magnitude of this Type B effect did not differ meaningfully between the constant and the alternating condition. Likewise, it was neither modulated by stimulus pitch nor by a combination of pitch and presentation condition. Therefore, one may conclude that the task-irrelevant feature in this task did not alter the Type B effect and consequently, the relative influence of past and current stimulus instances on the decision process. This result is well in line with the results of Rhodes et al. (2018) who also concluded that participants can integrate across temporal information marked by different pitch stimuli. While these authors had their participants judge the rhythmicity of empty intervals marked by brief auditory pulses of different pitches, we employed filled intervals - that is, in our study the durations themselves were conveyed by auditory stimulation. Nonetheless, both results point to a common internal reference for temporal information across different-pitch stimuli and thus, to a feature-based coding of the internal reference.

Similar results would be expected if participants had constructed separate, stimulus-based internal references for each of the two frequencies, if these were neither subject to any decay over time, nor to interference from internal representations of the respectively other stimulus type. Yet, based on previous results, this is highly unlikely, given that simply increasing the intertrial interval from 1600 to 3200 msec in a nearly identical task led to a significant reduction in the Type B effect (Ellinghaus et al., 2019). In comparison, in the present case, the effective intertrial interval between two trials of the same stimulus type was also 1600 msec in the constant condition, and on average more than 4 times longer (6600 msec + response time) in the alternating condition. Therefore, in case separate internal representations were formed for the different stimulus types, each internal reference most likely would have decayed to a significant extent before the same trial type was repeated in the alternating condition, and consequently the Type B effect should have diminished in the alternating compared to the constant condition.

Furthermore, the results do also not show any general decrement of performance caused by the trial-by-trial variation in pitch. That is, not only the Type B effect but also overall discriminations sensitivity was not affected the task-irrelevant attribute's variation. This also seems consistent with the view that participants can isolate and extract the duration information from the different auditory stimuli into one internal reference representation, i.e. feature-based coding of duration information. In the next experiment, we aimed at a conceptual replication of this general finding but with a stronger variation of the task-irrelevant feature.

Experiment 2

In this experiment, again, discrimination performance and the Type B effect were assessed in a 2AFC duration discrimination task. In contrast to Experiment 1, however, the frequency of the to be discriminated auditory intervals in the variable condition did not alternate between two values but randomly took on one of 15 possible values in each trial. The theoretical motivation for this manipulation was twofold.

First, according to some researchers (e.g., Ashby & Maddox, 1994), Garner interference-like performance decrements are a result of the uncertainty caused by the larger number of stimuli in the variable condition (but see Burns, 2016). In our first experiment, however, stimulus uncertainty was virtually nonexistent, since tone frequency varied from one trial to the next between only two possible values in a fully predictable manner (i.e., a low pitch trial would always be followed by a high pitch trial and vice versa). Thus, one cannot exclude that task performance and the memory-mixing mechanism might still be affected by a task-irrelevant stimulus attribute that is more variable and crucially, less predictable than in Experiment 1.

Second, introducing a greater number of stimuli qualifies for a harder test of the hypothesis that multiple internal references are constructed. The memory load of maintaining 15 separate internal references would be extremely high and presumably exceed the limits of working memory by far, as severe decrements in the recall of interval durations have already been demonstrated for memory loads from two to five intervals, respectively (Manohar & Husain, 2016; Teki & Griffiths, 2014). Moreover, even though the effective intertrial interval between two repetitions of the same trial type in Experiment 1 should have been already sufficiently long to cause a significant decay in the respective internal reference, a random presentation of 15 different trial types should multiply the duration of this interval, on average, by a factor of 14.

Therefore, it seems reasonable to assume that each internal reference would most certainly have decayed when a the respective frequency is repeated after 15 trials on average, becoming evident in a reduced Type B effect. Thus, a conceptual replication of the first experiments's main result, i.e. a Type B effect not differing in magnitude between the constant and the variable condition, would provide further evidence for a single underlying internal reference which encodes the duration information independent of other stimulus attributes.

Methods

Participants. 36 female and 12 male volunteers (mean age 24.4 ± 4.8 years) participated in the study. Data of 3 participants were replaced because their estimated DLs were outside the predefined three-sigma range in at least one of the experimental conditions.

Apparatus and Stimuli. Auditory stimuli were 15 different pure tones (ramped 5 msec on- and offsets) ranging between 400 Hz and 1200 Hz, presented binaurally through headphones with an intensity of approximately 60 dB(A). Specifically, the frequencies (f) 400 Hz and 1200 Hz were converted into their respective MIDI pitches (p) 67.35 and 86.37 via the formula $p = 69 + 12\ log_2(f/440)$. Then, 15 MIDI pitches were selected linearly spaced between these extreme points, and converted back to their respective Hz frequencies, in order to create a continuum of 15 equidistant pitch stimuli. In all other aspects, apparatus and stimuli were identical to Experiment 1.

Design and Procedure. As in Experiment 1, participants fulfilled an adaptive 2AFC duration discrimination task. The temporal structure of a single trial was equal to Experiment 1. However, in this experiment, in one half of the experiment, the constant condition, the pitch of the auditory stimuli remained constant at 692 Hz (the midpoint pitch of the 15 pure tones). In the other half of the experiment, the mixed condition, the pitch randomly took one of the 15 different values. As in Experiment 1, the duration of c adapted to participants' responses following the weighted up and down method (Kaernbach, 1991). In detail, the two start values (600 msec and 400 msec) combined with the two trial orders $\langle sc \rangle$ and $\langle cs \rangle$ led to 4 trial runs of 45 trials each, which were presented randomly interleaved. Hence each tone was presented three times per run in the mixed pitch condition and 45 times in the constant pitch condition. Accordingly, each presentation mode (constant vs. mixed) consisted of 180 trials which were divided into 9 subblocks of 20 trials each, separated by self-terminated breaks.

At the beginning of each experimental half, 20 practice trials were presented. Practice trials did not enter data analysis and start values were reset after practice. The order of the presentation mode conditions was counterbalanced between participants.

Results

The same procedure as in Experiment 1 was used to estimate DL and PSE in each condition. For inference statistics, a 2 (stimulus order) × 2 (presentation mode) repeated-measures ANOVA was conducted for both DL and PSE.

Difference Limen. The left panel of Figure 2 depicts mean DL as a function of stimulus order and presentation mode. As in Experiment 1, we observed the typical negative Type B effect. That is, the main effect of stimulus order was significant, $F(1, 47) = 40.39$, $p < .001$, $\eta_p^2 = .46$, reflecting again larger average DL in $\langle cs \rangle$ trials (72.2 msec, $SD =$ 28.5 msec) compared to $\langle sc \rangle$ trials (53.1 msec, $SD = 17.8$ msec). Furthermore, on average, DL was almost identical for the mixed and the constant pitch condition, that is, 63.5 msec ($SD = 23.4$ msec) and 61.9 msec ($SD = 22.0$ msec), respectively, $F(1,47) = 0.48$, $p =$ .492, $\eta_p^2 = .01$. Most importantly, the interaction of stimulus order and presentation mode on DL was not significant, either, $F(1, 47) = 1.25$, $p = .270$, $\eta_p^2 = .03$. In detail, we observed Type B effects (increase from $\langle sc \rangle$ trials to $\langle cs \rangle$ trials) of similar magnitude for the constant ($\Delta = 21.0$ msec) and the mixed ($\Delta = 17.2$ msec) presentation mode.[4]

Point of Subjective Equality. The right panel of Figure 2 illustrates mean PSE as a function of stimulus order and presentation mode. The main effect of presentation mode was significant, $F(1, 47) = 6.04$, $p = .018$, $\eta_p^2 = .11$, corresponding to 505.0 msec ($SD = 7.9$ msec) in the constant condition vs. 500.7 msec ($SD = 9.7$ msec) in the mixed

---

[4] In addition we again conducted the same Bayesian analysis as for Experiment 1. The result of this analysis indicated that the present data are 3.66 more likely under the model including only the main effects of stimulus order and presentation mode as compared to a model which also included the interaction term of the two factors, thus again supporting the ANOVA results reported in the main body of the text.

condition. PSE was also larger for $\langle sc \rangle$ than for $\langle cs \rangle$ trials, namely 510.2 msec ($SD =$ 13.8 msec) vs. 495.5 msec ($SD =$ 17.5 msec), F(1, 47) $=$ 12.52, $p =$ .001, $\eta_p^2 =$ .21. This indicates a positive time-order error, that is, overestimation of the duration of the first compared to the second interval presented within a trial. The significant interaction between stimulus order and presentation mode qualified this time-order error, F(1, 47) $=$ 29.97, $p <$ .001, $\eta_p^2 =$ .39. Specifically, as can be seen in Figure 2, PSE decreased from $\langle sc \rangle$ to $\langle cs \rangle$ trials for the mixed presentation mode ($\Delta =$ -30.9 msec), but not for the constant presentation mode ($\Delta =$ 1.4 msec).

Discussion

In this auditory 2AFC duration discrimination experiment, the frequency of the to be discriminated tones either remained constant at an intermediate value (constant presentation mode) or varied randomly from trial to trial between 15 different frequencies (mixed presentation mode). Again, a reliable negative Type B effect was observed averaged across these conditions. Conceptually mirroring the main results of Experiment 1, the magnitude of this Type B effect did not differ between the constant and the mixed presentation mode. This result renders highly unlikely that participants constructed separate, stimulus-based internal references for the various frequencies — as outlined above, creating 15 separate internal references would overload working memory on the one hand, and each memory representation would additionally be subject to severe temporal decay. Therefore, again, the results point to a single internal reference computed across temporal intervals marked by different pitches. Neither this process nor discrimination sensitivity in general seem to be hampered by the unpredictable frequency variation in the present experiment.

Although not directly related to the present investigation, it is noteworthy that we observed a reliable positive time-order error (or Type A effect) for the mixed but not for the constant presentation mode (note that a similar trend was also evident in the alternating condition in Experiment 1). This indicates that when stimulus type changed between trials, the duration of the first interval was overestimated relative to the second one. This result of a positive time-order error is generally unusual in duration discrimination. Typically the duration of the first stimulus is underestimated relative to the second stimulus (i.e., a negative time-order error), potentially due to a fading of the memory representation of the first stimulus (Allan, 1979; Jamieson & Petrusic, 1975; Schab & Crowder, 1988; Wackermann & Ehm, 2006). The present pattern of results might, however, be explained in terms of stimulus expectancy. Specifically, a number of studies show that the duration of novel or unexpected stimuli is overestimated compared to the duration of repeated or expected ones (Birngruber, Schröter, & Ulrich, 2015; Matthews, 2011; Tse, Intriligator, Rivest, & Cavanagh, 2004). In in the alternating / mixed conditions of our experiments, the pitch of the first stimulus in each trial was novel (and in the randomly mixed conditions additionally unexpected) in relation to the previous stimulus' pitch, and the pitch of the second stimulus always was an (expected) repetition of the first stimulus' pitch. Accordingly, the first stimulus would appear longer than the second one, resulting in a positive time-order error. It is so far a matter of debate whether the subjective temporal increase results from additional attentional resources being allocated to the unexpected stimulus (Tse et al., 2004), from repetition suppression (Pariyadath & Eagleman, 2008, 2012), from predictive coding (Schindel, Rowlands, & Arnold, 2011), or from a combination of multiple mechanisms (for an overview, see Ulrich & Bausenhart, 2019). In any case, these results seem unrelated to the proposed mechanism for creating an internal reference from previous and current stimulus instances, as any stimulus order effects on PSE were not accompanied by corresponding changes in DL. This further underlines the independence of Type A and Type B effects, and thus, their underlying mechanisms, as already noted in previous studies (e.g., Dyjas, Bausenhart, & Ulrich, 2014; Yeshurun, Carrasco, & Maloney, 2008).

There is, however, a caveat to the interpretation of our results in terms of a single, feature-based internal reference which only incorporates the task-relevant duration information. Specifically, so far, we varied sound frequency as a task-relevant attribute. On the one hand, Rhodes et al. (2018) argued that intervals marked by stimuli of different frequencies may be too similar to stimulate the formation of different perceptual priors. Going beyond this potential lack of distinctiveness, pitch can, just as duration, be regarded as a continuous attribute (ranging from low to high pitch) and thus may be stored in the form of a quantitative code. Consequently, the representation of pitch can, in principle, be integrated across trials into an internal reference for pitch discrimination, just as subsequent interval durations can be encoded into an internal reference for duration. In fact, a pronounced Type B effect in a pitch discrimination task has been demonstrated empirically (Ellinghaus et al., 2018). Therefore, one may formulate a sort of intermediate hypothesis between purely feature-based and purely stimulus-based coding. Accordingly, neither a single reference containing just the task-relevant attribute nor multiple references representing different stimuli with different task-irrelevant features would be formed, but a single reference which integrates past and current stimulus information regarding multiple, task-relevant and task-irrelevant, features. For example, the internal reference formed in the present experiments might contain a conglomerate of all presented durations, stimulus frequencies and also other quantitative features, as for example, sound amplitude. In the following, we will refer to this possibility as *multiple-feature* coding, as opposed to the proposed single-feature coding outlined above. We will try to distinguish these two accounts in the following experiment.

Experiment 3

While our results so far render unlikely the possibility of stimulus-based coding and thus multiple internal references, we cannot distinguish whether the internal reference formed in the previous experiments contained only task-relevant duration information (single-feature coding) or also a mixture of previous and current task-irrelevant stimulus features (multiple-feature coding). Therefore, in Experiment 3, we again assessed the effects of stimulus order in a duration-discrimination 2AFC task with task-irrelevant attributes either remaining constant across trials (blocked presentation mode) or alternating from trial to trial (alternating presentation mode), similar to Experiment 1. However, instead of sound frequency, in this experiment we varied the modality (audition vs. vision) of the presented intervals. Obviously, these two stimulus types are not only more distinct, which should, according to the reasoning and results of Rhodes et al. (2018), enable the formation of multiple duration priors, but also their specific task-irrelevant features cannot be easily arranged along a single perceptual quantitative continuum. In other words, one can easily imagine that two subsequent auditory stimuli (e.g., a low- and a high-pitched stimulus) might be mixed in memory to a conglomerate representation (e.g. a medium-pitch stimulus), but such an averaging process is not possible across the qualitative stimulus differences associated with different sensory modalities.

Methods

Participants. 32 female and 16 male volunteers (mean age 25.1 ± 6.2 years) participated in this experiment. Data of 6 participants were replaced because of too many incorrect responses, causing one or more of the adaptive staircases to run into negative values, or because their estimated DLs or PSEs were outside the predefined three-sigma range in at least one of the experimental conditions.

Apparatus and Stimuli. In this experiment, two different stimuli were employed. The standard stimulus of Experiment 2 served as auditory stimulus, that is, a 693 Hz pure tone (ramped 5 msec on- and offsets) presented binaurally through headphones with an intensity of approximately 60 dB (A). In addition, a grey circle (15 mm diameter) served as visual stimulus. For both modalities, the s intervals were presented for a duration of 500 msec, whereas c durations adapted to participants' responses (see below). In all other aspects, apparatus and stimuli were identical to the ones of Experiment 1.

Design and Procedure. As in Experiment 1, participants fulfilled an adaptive 2AFC duration discrimination task, however, with the following changes. First, the variation of stimulus pitch was replaced by a variation of stimulus modality. Consequently, participants performed one block of purely auditory duration discrimination trials (auditory blocked condition), one block of purely visual duration discrimination trials (visual blocked condition), and one block in which stimulus modality alternated between subsequent trials (auditory and visual alternating conditions). The temporal structure of the trials as well as the starting values (400 and 600 msec) and step-sizes (10 and 30 msec) of the auditory adaptive staircases were identical to the ones of Experiment 1 and 2. Due to the typically lower temporal resolution of the visual system (Ulrich, Nitschke, & Rammsayer, 2006), the lower and upper run of the visual staircases started at 300 and 700 msec, respectively, and step sizes of 20 and 60 msec were employed to adapt c duration after each response. All other aspects, including trial numbers, block length, feedback, randomization of runs and stimulus order, and counterbalancing of the different blocks were identical to Experiment 1, thus yielding a balanced 2 (stimulus order) × 2 (presentation mode) × 2 (modality) × 2 (runs) within-subjects design.

In addition, after completion of this main task, we asked participants to perform a brief control experiment in order to assess a potential overestimation of the auditory compared to the visual intervals (Ulrich et al., 2006; Wearden et al., 1998). Here, participants directly compared an auditory and a visual interval within each trial. All details of this experiment mirrored the visual blocked condition of the main experiment, except that an auditory stimulus (500 msec, 593 Hz) served as s, which was always presented at the first interval position and was followed by a visual c (i.e., stimulus order was always $\langle sc \rangle$). Following 10 practice trials, participants performed two randomly interleaved staircases (an upper and a lower run) of 40 trials each, divided in 4 sub-blocks of 20 trials each.

Results

As in the previous experiments, the reversal points for each run, and consequently, DL and PSE were determined for each condition and participant. For statistical analysis, a 2 (stimulus order) × 2 (modality) × 2 (presentation mode) repeated-measures ANOVA was computed for both DL and PSE.

Difference Limen. The left panel of Figure 3 depicts mean DL as a function of stimulus order, modality and presentation mode. Replicating Experiments 1 and 2, a significant negative Type B effect was evident in the main effect of stimulus order in the corresponding ANOVA, $F(1, 47) = 106.36$, $p < .001$, $\eta_p^2 = .69$, reflecting larger average DL in $\langle cs \rangle$ trials ($M =$ 105.7 msec, $SD =$ 35.3 msec) compared to $\langle sc \rangle$ trials ($M =$ 76.9 msec, $SD =$ 26.3 msec). Most important to the present investigation, the Type B effect was not further affected by whether the modality stayed constant or alternated from trial to trial, as the interaction of stimulus order and presentation mode was not significant,

F(1, 47) = 0.30, p = .587, η_p^2 = .01.[5]

Unexpectedly, DL was significantly lower in the alternating condition (M = 87.0 msec, SD = 28.5 msec) than in the constant condition (M = 95.6 msec, SD = 33.1 msec), F(1, 47) = 11.57, p = .001, η_p^2 = .20. The main effect of modality was significant, too, F(1, 47) = 218.77, $p <$ .001, η_p^2 = .82, reflecting a larger DL in the visual condition compared to the auditory condition, namely 119.6 msec (SD = 41.2 msec) vs. 62.9 msec (SD = 20.2 msec). This effect of modality was further modulated by stimulus order, F(1, 47) = 28.81, $p <$ .001, η_p^2 = .38, that is, the Type B effect was larger for the visual modality (Δ = 40.9 msec) than for the auditory modality (Δ = 16.6 msec). This replicates the one reported in Ellinghaus et al. (2019) and is rather unsurprising given the baseline difference in DL between the two modalities. Finally, the interaction of modality and presentation mode, F(1, 47) = 1.58, p = .215, η_p^2 = .03, and the threefold interaction of all factors, F(1, 47) = 1.31, p = .258, η_p^2 = .03, were not significant.

Point of Subjective Equality. The right panel of Figure 3 illustrates mean PSE as a function of stimulus order, modality, and presentation mode. Mean PSE did not differ significantly between the blocked (M = 511.3 msec, SD = 15.3 msec) and the alternating presentation mode (M = 507.9 msec, SD = 12.8 msec), F(1, 47) = 1.55, p = .219, η_p^2 = .03. The main effect of modality was significant, F(1, 47) = 19.85, p = < .001, η_p^2 = .30, corresponding to a larger PSE for the visual stimuli (M = 514.9 msec, SD = 17.1 msec) than for the auditory stimuli (M = 504.2 msec, SD = 8.1 msec). Furthermore, PSE was numerically larger for $\langle sc \rangle$ trials (M = 518.0 msec, SD = 22.1 msec) than for $\langle cs \rangle$ trials (M = 501.1 msec, SD = 25.3 msec), F(1, 47) = 7.57, p = .008, η_p^2 = .14. Thus, we again observed a reliable positive time-order error (Type A effect) averaged across conditions, and again, this effect was moderated by presentation mode, F(1, 47) = 8.16, p = .006, η_p^2 = .15, as before showing that this effect (decrease from $\langle sc \rangle$ to $\langle cs \rangle$ trials) was mainly due to

---

[5] In addition we again conducted the same Bayesian analysis as for Experiment 1. The result of this analysis indicated that the present data are 4.31 more likely under the model including only the main effects of stimulus order and presentation mode as compared to a model which also included the interaction term of the two factors, thus again supporting the ANOVA results reported in the main body of the text.

the alternating ($\Delta =$ -28.1 msec) rather than to the blocked condition ($\Delta =$ -5.8 msec). The twofold interaction between modality and presentation mode, as well as the twofold interaction between stimulus order and modality were nonsignificant, F(1, 47) $=$ 0.53, $p <$ .469, $\eta_p^2 =$ .01, and F(1, 47) $=$ 0.35, $p =$ .554, $\eta_p^2 =$ .01, respectively. The threefold interaction of all factors was non-significant, too, F(1, 47) $=$ 0.70, $p <$ .406, $\eta_p^2 =$ .01.

For the control experiment, PSE was determined analogously and a one-sample t-test against the standard duration of 500 msec was computed in order to assess differences in perceived stimulus duration between the modalities. This test reached significance, t(47) $=$ 4.80, $p <$ .001, reflecting the fact that the auditory comparison stimuli ($PSE =$ 565.0 msec) were overestimated relative to the visual 500 msecstandard stimuli.

Discussion

This third 2AFC duration discrimination experiment was similar to Experiment 1, except that the variation of pitch was replaced by a variation of modality. In detail, for the constant condition, participants performed both a block of visual duration discrimination and a block of auditory duration discrimination, and for the alternating condition, stimulus modality alternated between visual and auditory from trial to trial. The results conceptually replicated Experiment 1. That is, we observed the typical negative Type B effect averaged across conditions, and the magnitude of this effect did not meaningfully differ between the alternating and the constant condition. Therefore, it appears that humans can extract and isolate the task-relevant duration information not only from different stimuli within one modality (Experiments 1 and 2), but also from stimuli that stem from different modalities (Experiment 3). Following the same logic as in the previous experiments, this observation renders unlikely that participants constructed separate, stimulus-based internal references for the two modalities. On the contrary, the results are again consistent with the notion of a single, feature-based internal reference containing only the task-relevant duration information. Therefore, introducing a qualitative rather than quantitative variation in task-irrelevant stimulus features did not impact significantly on the mechanism underlying the Type B effect. Consequently, also the multiple-feature coding variant of the integration mechanism is not supported by the present data.

Regarding PSE, we again observed a positive time-order error (i.e., a relative overestimation of the first stimulus compared to the second stimulus), when stimulus type alternated from trial to trial rather than remained constant across the trials of an experimental block. As in Experiment 2, this may be attributed to the novelty of the modality of the first stimulus in each trial (relative to the modality of the stimuli presented in the preceding trial), as compared to the expected repetition of stimulus modality for the second stimulus within each trial (e.g., Birngruber et al., 2015; Matthews, 2011).

Another finding regarding stimulus modality was that discrimination sensitivity was higher in the visual compared to the auditory modality, which reflects a well documented finding in the time perception literature (e.g. Grondin, 1993; Grondin, Meilleur-Wells, Ouelette, & Macar, 1998). Theoretically, duration perception is often thought to be modulated via a pacemaker-accumulator type internal clock, where the perceived duration of a stimulus is reflected in a number of accumulated pulses (e.g. Bratzke & Ulrich, 2019; Gibbon, 1977; Gibbon, Church, & Meck, 1984; Treisman, 1963). Since the rate of the pulse-generating pacemaker is believed to be sped up by auditory stimulation, more pulses would be accumulated for auditory than for visual stimuli within the same time interval, causing auditory stimuli to be perceived as longer than visual ones (e.g. Bratzke & Ulrich, 2019; Ulrich et al., 2006; Wearden et al., 1998). The relative overestimation of the auditory compared to visual intervals in our control experiment is consistent with this assumption. Importantly, it has been argued that a faster rate of the pacemaker is also associated with higher temporal resolution (e.g., Rammsayer, 2008), and hence higher sensitivity in discrimination of auditory intervals compared to visual ones, as also observed in the present experiments, can also be attributed to this faster pacemaker rate (Jones, Poliakoff, & Wells, 2009, but see Williams, Yüksel, Stewart, & Jones, 2019).

Regarding the aim of the present experiments, it is interesting to note that the internal clock mechanism of such models is basically assumed to be central and amodal, with no dedicated sensory-specific sub-systems associated with the processing of visual and auditory durations, respectively. In other words, even though pulse rate may differ for different types of stimuli, temporal representations per se (i.e., the number of accumulated pulses) would be abstract and modality-unspecific. This conceptualization corresponds quite well with our experimental results so far — that is, abstract coding of the duration information into the internal reference. Of course, this raises the question of whether the present evidence for robust trial-by-trial integration of stimulus information irrespective of task-irrelevant stimulus features as pitch and modality is specific for the processing of temporal information, or whether such integration is also possible for other task domains. Moreover, our conclusion of feature-based coding across temporal stimuli of different pitches and different modalities is based on nonsignificant interactions of presentation mode and stimulus order. This conclusion would be strengthened by an empirical demonstration of such an interaction within the our experimental setup. Therefore, in Experiment 4, we chose to vary not only stimulus modality but, at the same time, also change the task-relevant feature, in order to enhance disctinctiveness between the alternating stimulus types.

Experiment 4

In this experiment, we assessed the magnitude of the Type B effect in an intensity discrimination task for visual and auditory stimuli. Therefore, not only task-irrelevant features but also the task-relevant stimulus feature was varied. Specifically, participants had to compare the loudness of two tones in one trial type and the brightness of two circles in the other trial type of a 2AFC task. These specific tasks were chosen for two reasons. First, robust Type B effects for both task types have been demonstrated empirically, indicating that loudness as well as brightness are both perceptual attributes that can be integrated across trials and enable the formation of an internal reference (Ellinghaus et al., 2018). Second, it would be particularly interesting to see whether an alternation of task type from trial to trial would hamper this process for these specific attributes. Actually, both attributes (loudness and brightness) are prothetic dimensions (e.g., S. S. Stevens, 1957), which can, somewhat similar to duration, be arranged along a quantitative continuum of magnitude (i.e., one can perceive a visual stimulus as more or less bright than another or an auditory stimulus as more or less loud than another). Moreover, a variation of these attributes, despite the different modalities associated with them, can be understood to affect the common dimension of stimulus intensity (Marks, Szczesiul, & Ohlott, 1986; S. S. Stevens, 1975). In fact, some researchers have argued for a common representational system for prothetic dimensions as space, time, number, and intensity (e.g., brightness and loudness in the visual and auditory domain, respectively), and this view is backed-up empirically by a number of studies on cross-sensory correspondences (cf., Marks, 1987, 1989; Spence, 2011; J. C. Stevens & Marks, 1965; Walsh, 2003). From this perspective, it is conceivable that, just as for duration, the intensity information represented in the internal reference might be stored as an common (abstract and magnitude-based) code across the different stimuli associated with the perceptual dimensions loudness and brightness. In this case, trial-by-trial integration, and thus the Type B effect could remain unimpaired even if stimulus/task type alternates from trial to trial. On the other hand, previous results as those of Rhodes et al. (2018) or Roach et al. (2017) described in the Introduction suggest that separate perceptual priors are formed when stimuli or the associated tasks are sufficiently distinct. Even if the variation of stimulus modality in the duration discrimination task of Experiment 3 apparently did not result in separate internal references, the variation of both task-irrelevant and task-relevant, attended stimulus attributes in the present intensity discrimination task now might yield perceptual codes distinct enough to hamper the integration across trials in the alternating condition.

Methods

Participants. A sample of 43 female and 5 male volunteers, M = 23.8 (SD = 3.5) years was tested for this experiment. Data of 3 participants were replaced because of too many incorrect responses, causing one or more of the adaptive staircases to run into negative values and/or resulting in estimated DL or PSE values outside of the predefined three-sigma range in at least one of the experimental conditions.

Apparatus and Stimuli. In this experiment, two different stimuli varying along different task-relevant attributes were employed. As auditory stimuli, 500 Hz, 500 msec pure tones (with ramped 5 msecon- and offsets) were presented binaurally through headphones.

As visual stimuli, grey circles (20 mm diameter) were presented for 500 msec on a black background at the centre of a gamma-calibrated CRT monitor. The intensity of the auditory standard stimulus was 50 dB(A), and the intensity of the visual standard stimulus was 34 cd/m^2. The intensity of the comparison stimuli in each modality varied according to an adaptive staircase procedure as in the previous experiments, with starting values and step sizes as outlined below). In all other aspects, apparatus and stimuli were identical to the one of Experiment 1.

Design and Procedure. In this experiment, participants fulfilled an adaptive 2AFC intensity discrimination task for visual and auditory stimuli. Specifically, participants had to indicate in each trial which of the two subsequently presented stimuli was more intense, that is, "brighter" in case of visual stimuli and "louder" in the case of auditory stimuli. The procedure was similar to Experiment 1 except for the following changes: First, the variation of stimulus pitch was replaced by a variation of stimulus modality /task. Consequently, participants performed one block of purely auditory intensity discrimination trials (auditory blocked condition), one block of purely visual intensity discrimination trials (visual blocked condition), and one block in which stimulus modality and task alternated between subsequent trials (auditory and visual alternating conditions).

The temporal structure of the trials was identical to the one of Experiment 1, except that the standard and comparison stimuli were each presented for 500 msec. The starting values for the lower and upper runs of the adaptive staircases were 27 and 41 cd/m^2 for the visual modality (step sizes of 0.88 and 2.63), and 40 and 60 dB(A) for the auditory modality (step sizes of 1 and 3). All other aspects, including trial numbers, block length, feedback, randomization of runs and stimulus order, and counterbalancing of the different blocks were identical to Experiment 1, thus yielding a balanced 2 (stimulus order) × 2 (presentation mode) × 2 (modality) × 2 (runs) within-subjects design.

Results

Again, the reversal points for each run, and consequently, DL and PSE, were determined for each condition and participant. Since these values are measured in different units for the two tasks, the estimates were z-transformed separately for each task before they were submitted to a 2 (stimulus order) × 2 (modality) × 2 (presentation mode) repeated-measures ANOVA.

Difference Limen. The left panel of Figure 4 depicts z-transformed values of DL as a function of stimulus order, modality and presentation mode. Expectedly, as a consequence of the z-transformation, there was no difference between the two modalities (F=0). Again, a negative Type B effect was indicated by a main effect of stimulus order, $F(1, 47) = 48.21$, $p < .001$, $\eta_p^2 = .51$, reflecting larger average DL in $\langle cs \rangle$ trials (M = 0.28, SD = 0.86) than in $\langle sc \rangle$ trials (M = -0.28, SD = 0.57). DL in the blocked presentation mode (M = 0.06, SD = 0.73) was marginally larger than in the alternating presentation mode (M = -0.06, SD = 0.68), $F(1, 47) = 3.87$, $p = .055$, $\eta_p^2 = .08$. Most importantly, however, the magnitude of the Type B effect was significantly reduced in the alternating ($\Delta = 0.38$) compared to the blocked condition ($\Delta = 0.73$), as indicated by an interaction of stimulus order and presentation mode, $F(1, 47) = 14.42$, $p < .001$, $\eta_p^2 = .23$.[6] This effect is clearly visible in Figure 4, as a steeper slope of the black (blocked conditions) compared to the grey lines (alternating conditions). Nonetheless, as indicated by post-hoc paired-samples t-tests conducted on separate data sets for the two presentation modes (aggregated across modality), the Type B effect was still significant for both the blocked, $t(47) = 8.26$, $p < .001$, as well as the alternating condition, $t(47) = 3.99$, $p < .001$.

An interaction between presentation mode and modality, $F(1, 47) = 8.33$, $p = .006$, $\eta_p^2 = .15$, indicated that DL for visual intensity was lower in the blocked (M = -0.05, SD = 0.82) than in the alternating (M = 0.05, SD = 0.90) condition, whereas DL for auditory intensity was lower in the alternating (M = -0.17, SD = 0.74) than in the blocked condition (M = 0.17, SD = 0.89). Finally, neither the interaction of stimulus order and modality, $F(1, 47) = 3.76$, $p = .058$, $\eta_p^2 = .07$, nor the threefold interaction of all factors,

---

[6] Again, we conducted the same Bayesian analysis as for Experiment 1. This time, this analysis indicated that the present data are 5.88 times more likely under the model including the interaction term of stimulus order and presentation mode in addition to the main effects, as compared to a model which only included the main effects. Again, this supports the ANOVA results reported in the main body of the text.

F(1, 47) = 1.81, p = .185, η_p^2 = .04, were significant.

Point of Subjective Equality. The right panel of Figure 4 depicts z-transformed values of PSE as a function of stimulus order, modality and presentation mode. Regarding PSE, there was of course also no main effect of modality ($F = 0$), and also no main effect of stimulus order, F(1, 47) = 1.00, $p <$ .322, η_p^2 = .02. However, a main effect of presentation mode indicated that PSE was lower in the blocked (M = -0.07, SD = 0.35) than in the alternating condition (M = 0.07, SD = 0.31), F(1, 47) = 4.40, $p <$ .041, η_p^2 = .09. In addition, there was an interaction of order and modality, F(1, 47) = 44.25, $p <$ .001, η_p^2 = .48, which indicated that PSE for visual intensity was lower in $\langle sc\rangle$ trials (M = -0.53, SD = 0.66) than in $\langle cs\rangle$ trials (M = 0.53, SD = 0.65), whereas PSE for auditory intensity was higher in $\langle sc\rangle$ trials (M = 0.44, SD = 0.58) than in $\langle cs\rangle$ trials (M = -0.44, SD = 0.81). In other words, participants underestimated the brightness of the first compared to the second stimulus in visual trials, but overestimated the loudness of the first compared to the second stimulus in auditory trials. None of the other interactions were significant (presentation mode × stimulus order: F(1, 47) = 3.30, p = .076, η_p^2 = .07, presentation mode × modality: F(1, 47) = 0.25, $p <$ .623, η_p^2 = .01, presentation mode × stimulus order × modality: F(1, 47) = 1.26, p = .268, η_p^2 = .03).

Discussion

In this final experiment, we assessed whether the integration of stimulus information into an internal reference would be affected when task-irrelevant stimulus features as well as the task-relevant one were varied. In contrast to the previous experiments, in which the task always was duration discrimination, now participants performed an intensity discrimination task for visual and auditory stimuli. Again, a typical negative Type B effect was observed averaged across conditions. However, in the present experiment, the Type B effect was meaningfully reduced when stimulus modality, and thus, participants' task varied from trial to trial rather than remained constant across the trials of an experimental block. This indicates that the formation of an internal reference was hampered by this variation. Consequently, one may conclude that the internal representations of stimulus intensity in the visual and the auditory domain, i.e., perceived brightness and perceived loudness, are not encoded in the same format - and therefore do not contribute to the same internal reference representation.

The present results are thus well in line with an interpretation in terms of separate internal references being constructed for the two task-relevant stimulus attributes loudness and brightness. Due to the relatively long effective intertrial interval between two repetitions of the same trial type in the alternating condition, each of these internal references would have already substantially decayed before it could be updated with the novel incoming stimulus information, which in turn would have reduced the magnitude of the Type B effect (cf. Ellinghaus et al., 2019).

It should be noted that our results are not directly in conflict with the assumption that the intensity of different-modality stimuli can in principle be represented as a common, magnitude-based code (Marks, 1989; Walsh, 2003); however, they render unlikely that such a supramodally coded intensity representation formed the basis for the internal reference in the present case. This view could nonetheless be reconciled with the present results, if one assumes that transforming the stimulus information into a common representational code would (a) only take place (or become relevant) in alternating blocks, (b) be costly and thus lead to a noisy or incomplete representation, and (c) only operate on the information stored in the internal reference but not on the direct internal representation of each stimulus. The resulting noisy internal reference would in turn receive less weight in the trial-by-trial integration process, and thus lead to a reduced Type B effect, without an accompanying general sensitivity impairment in alternating compared to blocked trials, as observed in the present study. Especially in the light of repeated demonstrations that different internal references or perceptual priors can be created and maintained even across more similar stimuli (e.g., Rhodes et al., 2018; Roach et al., 2017), however, an interpretation in terms of separate internal references for visual and auditory intensity discrimination appears more parsimonious and plausible to account for the present results.

In any case, the present experiment demonstrates a decrease of the Type B effect for the alternating compared to the blocked presentation mode, and therefore suggests that the formation of the internal reference is disturbed by the present variation of task-relevant and task-irrelevant attributes. Importantly, this empirical demonstration of a reduced Type B also lends further support to the conclusions based on the unaltered Type B effects observed Experiments 1-3, by demonstrating that the present experimental design is basically well-suited to unveil such interactive effects.

The result pattern regarding PSE is also different from the ones observed in the previous experiments, in that the effects of stimulus order (i.e., the time-order effect) did not differ between the blocked and the alternating presentation mode. The lack of this specific interaction fits well within our previous interpretation in terms of stimulus repetition vs. novelty, which is a well-established finding specifically regarding perceived duration (e.g., Birngruber et al., 2015; Matthews, 2011). Rather, in the present experiment there was a negative time-order error for visual intensity discrimination (i.e., underestimation of the brightness of the first compared to the second stimulus) but a positive time order error for auditory intensity discrimination (i.e., overestimation of the loudness of the first compared to the second stimulus). The maybe most simple explanation would be a differential response bias (i.e, participants tended to press the left response key for auditory stimuli and the right response key for visual stimuli, when in doubt about the correct response), however, it is of course also conceivable that genuine perceptual differences underlie these effects (Hellström, 1978; Patching, Englund, & Hellström, 2012). Most importantly, however, as in our previous experiments, the observed pattern of time-order errors and its variation with the experimental conditions seems completely unrelated to the observed changes of discrimination sensitivity in terms of DL. This again suggests that these two phenomena, namely Type A and Type B effects, are not mediated by the same mechanism.

General Discussion

Psychophysical research has provided several lines of evidence indicating that perceptual decisions are based on both present and prior stimulus information. The general principle assumed to underlie these effects may be summarized as memory-mixing. Yet, it is still unclear what exactly is mixed in memory, or framed differently, whether (and which) different types of stimuli can be mixed into a common memory representation or not.

In the present work, we therefore investigated such memory-mixing in the classic 2AFC task for the discrimination of duration (Experiments 1-3) and of intensity (Experiment 4). According to the Internal Reference Model (Dyjas et al., 2012), the Type B effect typically observed in this task indicates the psychological reality of an internal reference as a continuously updated conglomerate of past and present stimulus information. The core question of the present experiments was whether the information represented in this internal reference would be rather stimulus-based (i.e., an object-like representation containing various stimulus features, and therefore, separate internal references would be formed for different stimulus types) or rather feature-based (i.e., only the task-relevant feature is encoded, resulting in a common internal reference for different stimulus types). The results show that the Type B effect for duration discrimination was neither reduced when the stimuli in successive trials varied within a modality (i.e., pitch of auditory intervals, Experiments 1 and 2) nor when the stimuli varied between modalities (i.e., auditory and visual intervals, Experiment 3). This finding is difficult to reconcile with the idea that participants formed separate, stimulus-based references for the different stimulus types, since the reference information would have decayed substantially when a stimulus type is repeated after two (Experiments 1 and 3) or 15 (Experiment 2) trials (Ellinghaus et al., 2019), resulting in a reduced Type B effect. Also, the intermediate hypothesis, that is, a stimulus-based reference incorporating and mixing multiple stimulus features in memory, was rendered unlikely by the results of Experiment 3, since the qualitative differences associated with different-modality intervals would preclude trial-by-trial integration of aspects other than duration. In sum, these data are well in line with the idea that participants isolated the duration information and thus generalized across different temporal stimuli when integrating the task-relevant information into an internal reference, that is, feature-based coding.

This interpretation corresponds well with the mechanism described in terms of pacemaker-accumulator models of time perception, which posit that duration information is represented as an abstract amodal code (e.g., a number of pulses, Bratzke & Ulrich, 2019; Gibbon et al., 1984; Treisman, 1963; Wearden, Todd, & Jones, 2006). In fact, Shi et al.

(2013) have outlined a general framework of how such models may be extended to account for memory-mixing and dynamical memory updating processes as specified by Bayesian accounts or IRM. Others have recently suggested that temporal information, irrespective of the modality from which it is acquired, may also be represented as a supramodal auditory code (Guttman, Gilroy, & Blake, 2005; Kanai, Lloyd, Bueti, & Walsh, 2011; Bratzke, Seifried, & Ulrich, 2012; Bratzke, Quinn, Ulrich, & Bausenhart, 2016). This view holds that temporal information from different modalities is directly translated to an audition-based representational format. Even though less abstract, this may be effectively regarded as a variant of feature-based, as opposed to stimulus-based, coding. Consequently, even such a supramodal conceptualization of temporal representation is compatible with the idea that memory mixing may occur based on a common temporal code, and thus across intervals conveyed by different-modality stimuli as the visual and auditory intervals employed in the present Experiment 3.

In either case, the present results are consistent with previous results showing that humans may generalize across different temporal stimuli from separate duration distributions (as, e.g., visual and auditory intervals) in temporal reproduction tasks (Roach et al., 2017). However, the opposite has also been documented, that is, participants under some circumstances formed separate temporal priors based on (task-irrelevant) characteristics of stimuli from different duration distributions. Specifically, as outlined in the Introduction, Rhodes et al. (2018) obtained evidence for separate duration priors both when stimuli differed either within a modality (e.g., pure tones vs. white noise) or across modalities (i.e., visual vs. auditory stimuli) in a perceptual judgment task. Apparently, different temporal stimuli from different time ranges may either be differentiated to form separate perceptual priors, or they may be mixed in memory to form a single, common perceptual prior, and it is unclear what exactly separates the former from the latter cases.

Several factors may play a role in stimulating the formation of separate priors for stimuli from different stimulus distributions. Among these are, for example, differences in the specific tasks or motor outputs associated with different stimuli (Nagai, Suzuki, Miyazaki, & Kitazawa, 2012; Roach et al., 2017), greater perceptual distinctiveness of the stimuli (Cicchini, Mikellidou, & Burr, 2018; Rhodes et al., 2018), extensive practice (Roach et al., 2017; Baugh et al., 2016), or the distinctiveness of the stimulus distributions themselves (Gekas, Chalk, Seitz, & Seriès, 2013). Given the methodological variation between the respective studies, it seems difficult to pinpoint the exact contribution of each of these factors. As Rhodes et al. (2018, p. 21) point out, the "strength of available evidence" on whether stimuli belong to common or to separate sources or categories may determine whether separate or common priors are formed. Accordingly, also higher-level cognitive factors may have contributed to previous demonstrations of multiple perceptual priors, especially since the measures of perceived duration (i.e., reproductions and simultaneity judgments) employed in these studies may be prone to such influences (e.g., the knowledge that visual intervals are usually rather short may lead to deliberately shorter reproductions). In this regard, the present approach provides an independent methodological perspective by assessing memory mixing through variations in discrimination sensitivity, which likely reflects genuine perceptual differences, and by allowing to investigate memory mixing for different stimulus types while employing identical or strongly overlapping temporal distributions for the different stimulus types. In future research, it would therefore be interesting to see whether the Type B effect for duration discrimination would be reduced or eliminated if stimulus categories were made even more distinctive, for example by also using different duration distributions, different response modes, or even explicit information about the different stimulus types.

Unlike for temporal representations, the reduced Type B effect in the alternating compared to the constant presentation mode of Experiment 4 suggests that trial-by-trial integration across the visual and auditory intensity representations was prevented or at least strongly impaired. A straightforward interpretation of this effect in terms of IRM is that separate internal references for the two stimulus types were formed whereupon the relative contribution of previous compared to current stimulus information was reduced due to temporal decay (Ellinghaus et al., 2019, see also Bliss et al., 2017; Fritsche et al., 2017). Several potential reasons for this apparent segregation should be considered. On the one hand, it seems plausible that visual and auditory intensity in the internal reference are not represented as a common, magnitude-based code and thus, trial-by-trial integration across the varying codes was impossible. On the other hand, and as already outlined in the Discussion of Experiment 4, the present results could also be reconciled with the notion a common code for intensity representation under certain auxiliary assumptions. In this case, other factors might have contributed to formation of separate internal representations, as for example, the trial-by-trial alternation directly concerned the task-relevant attribute, which may have caused attention allocation directly to the varied attribute and thus may have promoted experienced distinctiveness of the different stimulus types.

On a more general level, the present study contributes to a growing body of knowledge on the interrelation of memory and perception (for a review, see Kiyonaga et al., 2017). Accordingly, assimilatory dependencies in perceptual judgments, interference effects in short term or working memory, and stimulus order effects as the Type B effect, although studied within separate lines of research, seem to be closely related. Classical psychophysical tools as the 2AFC task (Hegelmaier, 1852) may prove useful to shed further light on the mechanisms and principles underlying these effects, and memory mixing models such as IRM might offer a general theoretical framework under which these phenomena can be reconciled.

References

Alais, D., Leung, J., & Burg, E. V. D. (2017). Linear summation of repulsive and attractive serial dependencies : Orientation and motion dependencies sum in motion perception. *The Journal of Neuroscience*, *37*, 4381–4390.

Allan, L. G. (1979). The perception of time. *Perception & Psychophysics*, *26*, 340–354.

Arzounian, D., Kerangal, M. D., & Cheveign, A. D. (2017). Sequential dependencies in pitch judgments. *The Journal of the Acoustical Society of America*, *142*, 3047–3057.

Ashby, F. G., & Maddox, W. T. (1994). A response time theory of seperability and integrality in speeded classification. *Journal of Mathematical Psychology*, *38*, 423–466.

Baugh, L. A., Yak, A., Johansson, R. S., & Flanagan, J. R. (2016). Representing multiple object weights: Competing priors and sensorimotor memories. *Journal of Neurophysiology*, *116*, 1615-1625.

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.

Birngruber, T., Schröter, H., & Ulrich, R. (2015). The influence of stimulus repetition on duration judgments with simple stimuli. *Frontiers in Psychology*, *6*, 1213.

Bliss, D. P., Sun, J. J., & D'Esposito, M. (2017). Serial dependence is absent at the time of perception but increases in visual working memory. *Scientific Reports*, *7*, 14739.

Bratzke, D., Quinn, K. R., Ulrich, R., & Bausenhart, K. M. (2016). Representations of temporal information in short-term memory : Are they modality-specific? *Acta Psychologica*, *170*, 163–167.

Bratzke, D., Seifried, T., & Ulrich, R. (2012). Perceptual learning in temporal discrimination : asymmetric cross-modal transfer from audition to vision. *Experimental Brain Research*, *221*, 205–210.

Bratzke, D., & Ulrich, R. (2019). Temporal reproduction within and across the senses: Testing the supramodal property of the pacemaker-counter model. *Journal of Experimental Psychology: Human Perception and Performance*, *45*, 1218–1235.

Burns, D. M. (2016). Garner interference is not solely driven by stimulus uncertainty.

Psychonomic Bulletin & Review, *23*, 1846–1853.

Burr, D. C., & Cicchini, G. M. (2014). Vision: Efficient adaptive coding. *Current Biology*, *24*, R1096–R1098.

Cicchini, G. M., Anobile, G., & Burr, D. C. (2014). Compressive mapping of number to space reflects dynamic encoding mechanisms, not static logarithmic transform. *Proceedings of the National Academy of Sciences*, *111*, 7867–7872.

Cicchini, G. M., Mikellidou, K., & Burr, D. (2017). Serial dependencies act directly on perception. , *17*(4), 1–9.

Cicchini, G. M., Mikellidou, K., & Burr, D. C. (2018). The functional role of serial dependence. *Proceedings of the Royal Society B*, *285*, 20181722.

Dong, D. W., & Atick, J. J. (1995). Statistics of natural time-varying images. *Network: Computation in Neural Systems*, *6*, 345–358.

Durlach, N. I., & Braida, L. D. (1969). Intensity perception. I. Preliminary theory of intensity resolution. *The Journal of the Acoustical Society of America*, *46*, 372–383.

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.

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

Ellinghaus, R., Gick, M., Ulrich, R., & Bausenhart, K. M. (2019). Decay of internal reference information in duration discrimination : Intertrial interval modulates the type b effect. *Quarterly Journal of Experimental Psychology*, *72*, 1578–1586.

Ellinghaus, R., Ulrich, R., & Bausenhart, K. M. (2018). Effects of stimulus order on comparative judgments across stimulus attributes and sensory modalities. *Journal of Experimental Psychology: Human Perception and Performance*, *44*, 7–12.

Fechner, G. T. (1860). *Elemente der Psychophysik* [Elements of psychophysics]. Leipzig, Germany: Breitkopf und Härtel.

Fischer, J., & Whitney, D. (2014). Serial dependence in visual perception. *Nature Neuroscience*, *17*, 738–743.

Fornaciai, M., & Park, J. (2018). Attractive serial dependence in the absence of an explicit task. *Psychological Science*, *29*, 437–446.

Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society*, *364*, 1211–1221.

Fritsche, M., Mostert, P., & de Lange, F. P. (2017). Opposite effects of recent history on perception and decision. *Current Biology*, *27*, 590–595.

Garner, W. R. (1953). An informational analysis of absolute judgements of loudness. *Journal of Experimental Psychology*, *46*, 373–380.

Garner, W. R. (1974). *The processing of information and structure.* Oxford, England: Lawrence Erlbaum.

Gekas, N., Chalk, M., Seitz, A. R., & Seriès, P. (2013). Complexity and specificity of experimentally induced expectations in motion perception. *BMC Neuroscience*, *14*, P355.

Gibbon, J. (1977). Scalar expectancy theory and weber ' s law in animal timing. *Psychological Review*, *84*, 279–325.

Gibbon, J., Church, R. M., & Meck, W. H. (1984). Scalar timing in memory. *Annals of the New York Academy of Sciences*, *423*, 52–77.

Goldstone, S., Lhamon, W. T., & Sechzer, J. (1978). Light intensity and judged duration. *Bulletin of the Psychonomic Society*, *12*, 83-84.

Grondin, S. (1993). Duration discrimination of empty and filled intervals marked by auditory and visual signals. *Perception & Psychophysics*, *54*, 383–394.

Grondin, S., Meilleur-Wells, G., Ouelette, C., & Macar, F. (1998). Sensory effects on judgments of short time-intervals. *Psychological Research*, *61*, 261–268.

Gu, B.-M., & Meck, W. H. (2011). New perspectives on Vierordt's law: Memory-mixing in ordinal temporal comparison tasks. In A. Vatakis, A. Esposito, M. Giagkou,

F. Cummins, & G. Papadelis (Eds.), *Multidisciplinary aspects of time and time perception* (pp. 67–78). Berlin, Germany: Springer.

Guttman, S. E., Gilroy, L. A., & Blake, R. (2005). Hearing what the eyes see: Auditory encoding of visual temporal sequences. *Psychological Science*, *16*, 228–235.

Hegelmaier, F. (1852). Ueber das Gedächtniss für Linear-Anschauungen [On memory for visually perceived lines]. *Archiv für physiologische Heilkunde*, *11*, 844–853.

Hellström, Å. (1978). Factors producing and factors not producing time errors: An experiment with loudness comparisons. *Perception & Psychophysics*, *23*, 433–444.

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

Helson, H. (1964). *Adaptation-level theory*. New York: Harper & Row.

Hollingworth, H. (1910). The central tendency of judgment. *The Journal of Philosophy, Psychology and Scientific Methods*, *7*, 461–469.

Jamieson, D. G., & Petrusic, W. M. (1975). Presentation order effects in duration discrimination. *Perception & Psychophysics*, *17*, 197–202.

Jazayeri, M., & Shadlen, M. N. (2010). Temporal context calibrates interval timing. *Nature Neuroscience*, *13*, 1020–1026.

Jones, L. A., Poliakoff, E., & Wells, J. (2009). Good vibrations : Human interval timing in the vibrotactile modality. *The Quarterly Journal of Experimental Psychology*, *62*, 2171–2186.

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

Kanai, R., Lloyd, H., Bueti, D., & Walsh, V. (2011). Modality-independent role of the primary auditory cortex in time estimation. *Experimental Brain Research*, *209*, 465–471.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as bayesian inference. *Annual Review of Psychology*, *55*, 271-304.

Kersten, D., & Yuille, A. (2003). Bayesian models of object perception. *Current Opinion in Neurobiology*, *13*, 150-158.

Kiyonaga, A., Scimeca, J. M., Bliss, D. P., & Whitney, D. (2017). Serial dependence across perception, attention, and memory. *Trends in Cognitive Sciences*, *21*, 493–497.

Kleiner, M., Brainard, D., & Pelli, D. G. (2007). What's new in Psychtoolbox-3? *Perception*, *36*, 1.

Kliegl, K. M., & Huckauf, A. (2014). Perceived duration decreases with increasing eccentricity. *Acta Psychologica*, *150*, 136 - 145.

Knill, D. C., & Richards, W. (Eds.). (1996). *Perception as bayesian inference.* New York, NY, USA: Cambridge University Press.

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.

Maaß, S. C., Schlichting, N., & van Rijn, H. (2019). Eliciting contextual temporal calibration: The effect of bottom-up and top-down information in reproduction tasks. *Acta Psychologica*, *199*.

Maloney, L. T., Martello, M. F. D., Sahm, C., & Spillmann, L. (2005). Past trials influence perception of ambiguous motion quartets through pattern completion. *Proceedings of the National Academy of Sciences*, *102*, 3164–3169.

Manassi, M., Liberman, A., Chaney, W., & Whitney, D. (2017). The perceived stability of scenes : serial dependence in ensemble representations. *Scientific Reports*, 1–9.

Manohar, S. G., & Husain, M. (2016). Working memory for sequences of temporal durations reveals a volatile single-item store. *Frontiers in Psychology*, *7*, 1655.

Marks, L. E. (1987). On cross-modal similarity: Auditory-visual interactions in speeded discrimination. *Journal of Experimental Psychology: Human Perception and Performance*, *13*, 389–394.

Marks, L. E. (1989). On cross-modal similarity: The perceptual structure of pitch, loudness, and brightness. *Journal of Experimental Psychology: Human Perception and Performance*, *15*, 586–602.

Marks, L. E., Szczesiul, R., & Ohlott, P. (1986). On the cross-modal perception of intensity. *Journal of Experimental Psychology: Human Perception and Performance*, *12*, 517–534.

Matthews, W. J. (2011). Stimulus repetition and the perception of time : The effects of prior exposure on temporal discrimination, judgment, and production. *PLoS ONE*, *6*.

Matthews, W. J. (2015). Time perception : The surprising effects of surprising stimuli. *Journal of Experimental Psychology: General*, *144*, 172–197.

Matthews, W. J., Stewart, N., & Wearden, J. H. (2011). Stimulus intensity and the perception of duration. *Journal of Experimental Psychology: Human Perception and Performance*, *37*, 303-313.

Mo, S. S., & Michalski, V. A. (1972). Judgment of temporal duration of area as a function of stimulus configuration. *Psychonomic Science*, *27*, 97–98.

Morey, R. D. (2008). Confidence Intervals from Normalized Data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology*, *4*, 61–64.

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.

Nagai, Y., Suzuki, M., Miyazaki, M., & Kitazawa, S. (2012). Acquisition of multiple prior distributions in tactile temporal order judgment. *Frontiers in Neuroscience*, *3*, 1–7.

Pariyadath, V., & Eagleman, D. M. (2008). Brief subjective durations contract with repetition. *Journal of Vision*, *8*, 1–6.

Pariyadath, V., & Eagleman, D. M. (2012). Subjective Duration Distortions Mirror Neural Repetition Suppression. *PLoS ONE*, *7*, e49362.

Pascucci, D., Mancuso, G., Santandrea, E., Libera, C. D., & Plomp, G. (2017). Laws of concatenated perception : Vision goes for novelty, Decisions for perseverance. *bioRxiv*, 1–45.

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.

Penney, T. B., Gibbon, J., & Meck, W. H. (2000). Differential effects of auditory and visual signals on clock speed and temporal memory. *Journal of Experimental Psychology: Human Perception and Performance*, *26*, 1770–1787.

Rammsayer, T. H. (2008). Neuropharmacological approaches to human timing. In S. Grondin (Ed.), *Psychology of time* (pp. 295–320). Bingley: Emerald Group.

Rammsayer, T. H., & Verner, M. (2014). The effect of nontemporal stimulus size on perceived duration as assessed by the method of reproduction. *Journal of Vision*, *14*, 1–10.

Raviv, O., Lieder, I., Loewenstein, Y., & Ahissar, M. (2014). Contradictory behavioral biases result from the influence of past stimuli on perception. *PLoS Computational Biology*, *10*, e1003948.

Rhodes, D., Seth, A. K., & Roseboom, W. (2018). Multiple duration priors within and across the senses. *bioRxiv*, 467027.

Roach, N. W., Mcgraw, P. V., Whitaker, D. J., & Heron, J. (2017). Generalization of prior information for rapid Bayesian time estimation. *Proceedings of the National Academy of Sciences*, *114*, 412–417.

Roberts, S. W. (1959). Control chart tests based on geometric moving averages. *Technometrics*, *1*, 239–250.

Schab, F. R., & Crowder, R. G. (1988). The role of succession in temporal cognition: Is the time-order error a recency effect of memory? *Perception & Psychophysics*, *44*, 233–242.

Schindel, R., Rowlands, J., & Arnold, D. H. (2011). The oddball effect: Perceived duration and predictive coding. *Journal of Vision*, *11*, 17.

Shi, Z., Church, R. M., & Meck, W. H. (2013). Bayesian optimization of time perception. *Trends in Cognitive Sciences*, *17*, 556–564.

Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception & Psychophysics*, *73*, 971–995.

Stevens, J. C., & Marks, L. E. (1965). Cross-modality matching of brightness and loudness. *Proceedings of the National Academy of Sciences*, *54*, 407–411.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review*, *64*, 153–181.

Stevens, S. S. (1975). *Psychophysics: Introduction to its perceptual, neural and social prospects*. New York: Wiley.

Teki, S., & Griffiths, T. D. (2014). Working memory for time intervals in auditory rhythmic sequences. *Frontiers in Psychology*, *5*, 1329.

Thönes, S., Von Castell, C., Iflinger, J., & Oberfeld, D. (2018). Color and time perception: Evidence for temporal overestimation of blue stimuli. *Scientific Reports*, *8*, 1–8.

Treisman, M. (1963). Temporal discrimination and the indifference interval: Implications for a model of the "internal clock". *Psychological Monographs: General and Applied*, *77*, 1–31.

Tse, P. U., Intriligator, J., Rivest, J., & Cavanagh, P. (2004). Attention and the subjective expansion of time. *Perception & Psychophysics*, *66*, 1171–1189.

Ulrich, R., & Bausenhart, K. M. (2019). The illusions of time. philosophical and psychological essays on timing and time perception. In V. Arstila, A. Bardon, S. E. Power, & A. Vatakis (Eds.), (pp. 71–89). Cham: Palgrave Macmillan.

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

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

von Helmholtz, H. (1867). *Handbuch der Physiologischen Optik*.

Wackermann, J., & Ehm, W. (2006). The dual klepsydra model of internal time representation and time reproduction. *Journal of Theoretical Biology*, *239*, 482–493.

Walsh, V. (2003). A theory of magnitude: Common cortical metrics of time, space and quantity. *Trends in Cognitive Sciences*, *7*, 483-488.

Wearden, J. H., Edwards, H., Fakhri, M., & Percival, A. (1998). Why "sounds are judged longer than lights": Application of a model of the internal clock in humans. *Quarterly Journal of Experimental Psychology*, *51B*, 97–120.

Wearden, J. H., Todd, N. P. M., & Jones, L. A. (2006). When do auditory / visual differences in duration judgements occur ? *The Quarterly Journal of Experimental*

Psychology, *59*, 1709–1724.

Williams, E. A., Yüksel, E. M., Stewart, A. J., & Jones, L. A. (2019). Modality differences in timing and the filled-duration illusion : Testing the pacemaker rate explanation. *Attention, Perception, & Psychophysics*, *81*, 823–845.

Xia, Y., Leib, A. Y., & Whitney, D. (2016). Serial dependence in the perception of attractiveness. *Journal of Vision*, *16*, 28.

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.

Yuille, A., & Kersten, D. (2006). Vision as bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences*, *10*, 301 - 308.

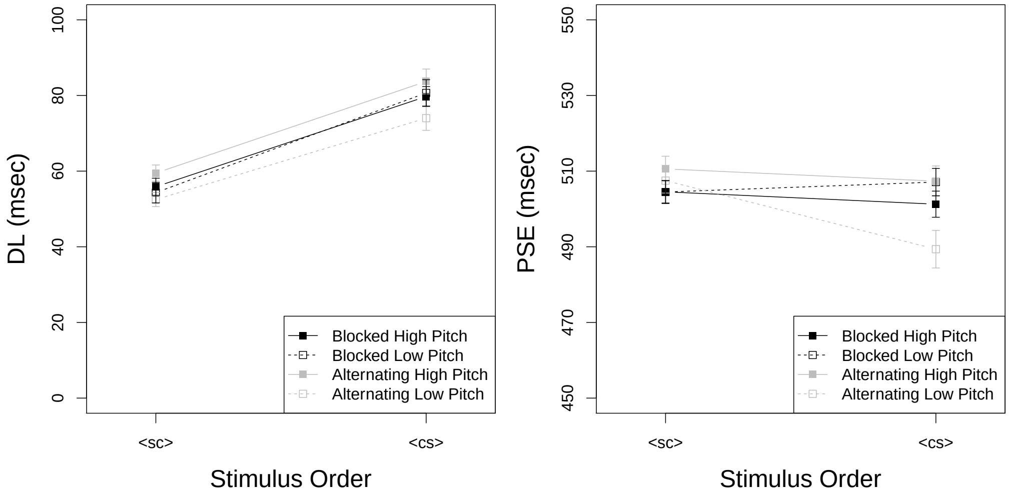

Figure 1. Mean *DL* (left panel) and *PSE* (right panel) observed in Experiment 1 as a function of stimulus order, presentation mode and pitch. Error bars reflect ± 1 within-subjects standard error of the mean according to a suggestion of Morey (2008) .

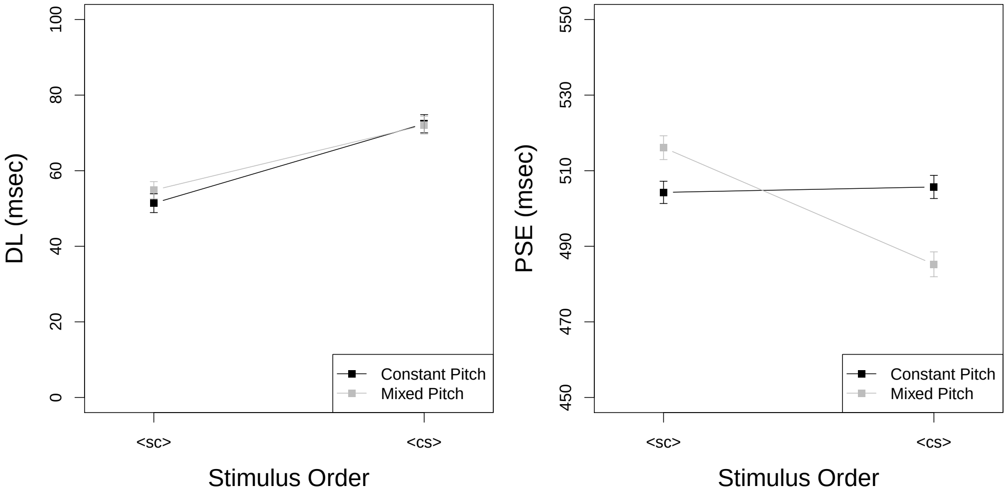

Figure 2. Mean *DL* (left panel) and *PSE* (right panel) observed in Experiment 2 as a function of stimulus order and presentation mode. Error bars reflect ± 1 within-subjects standard error of the mean according to a suggestion of Morey (2008).

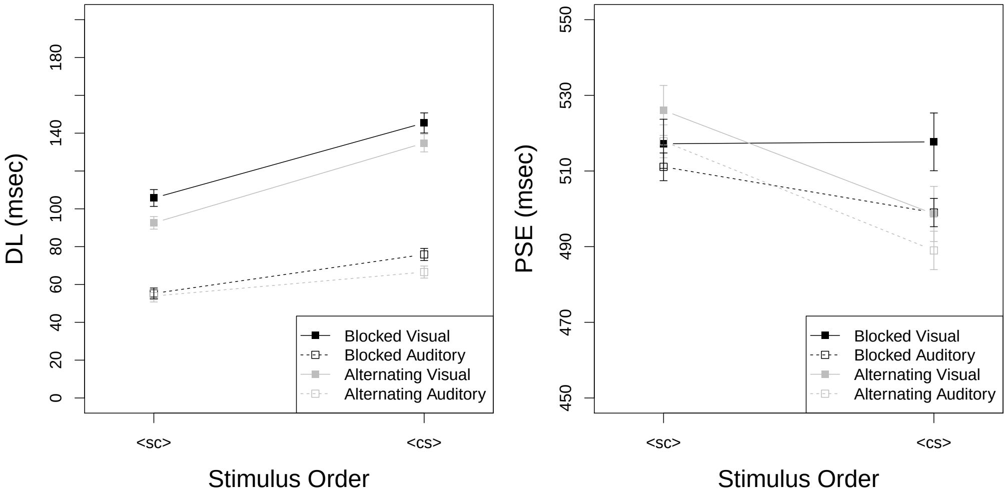

Figure 3. Mean *DL* (left panel) and *PSE* (right panel) observed in Experiment 3 as a function of stimulus order and presentation mode. Error bars reflect ± 1 within-subjects standard error of the mean according to a suggestion of Morey (2008).

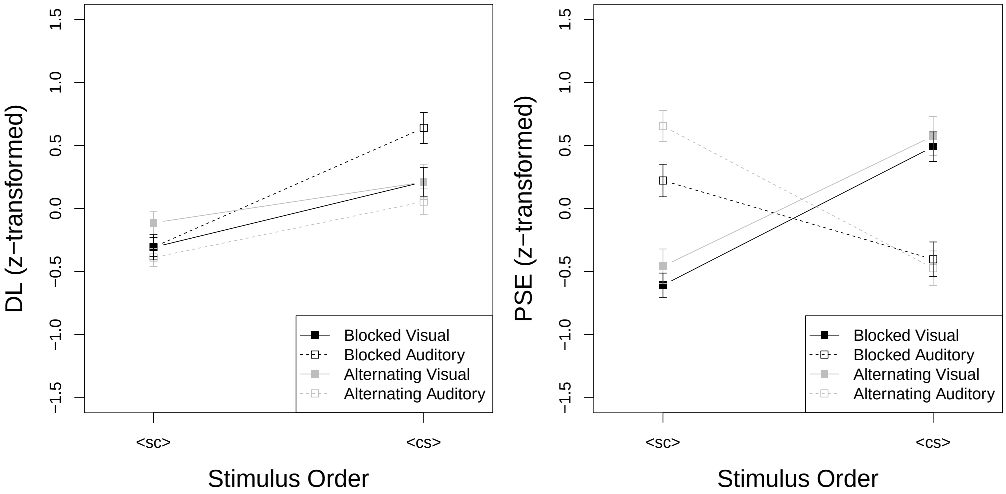

Figure 4. Mean *DL* (left panel) and *PSE* (right panel) observed in Experiment 4 as a function of modality, stimulus order and presentation mode. Values were z-transformed in order to enable a comparison across the different units associated with the loudness discrimination task in the auditory and the brightness discrimination task in the visual modality. Error bars reflect ± 1 within-subjects standard error of the mean according to a suggestion of Morey (2008).