Acta Acustica 2024, **8**, 57

https://doi.org/10.1051/aacus/2024061

Available online at:
https://acta-acustica.edpsciences.org

SCIENTIFIC ARTICLE OPEN ACCESS

Non-uniform temporal weighting of intensity in audition and vision: The signature of an evidence integration process?

Daniel Oberfeld[1,*], Alexander Fischenich[1], and Emmanuel Ponsot[2]

[1] Experimental Psychology, Johannes Gutenberg-Universität Mainz, Wallstraße 3, 55122 Mainz, Germany
[2] Science & Technology of Music and Sound, IRCAM/CNRS/Sorbonne Université, F-75004 Paris, France

Received 26 January 2024, Accepted 6 September 2024

Abstract – Non-uniform temporal weights (TWs) are often reported regarding the perceptual evaluation of dynamic auditory and visual information, such as perceptual judgments of the overall intensity of time-varying stimuli. In particular, primacy effects, i.e., a stronger influence of early compared to later stimulus information on the perceptual decision, have been observed across a large number of studies. Yet, it is not clear whether these non-uniform patterns of TWs result from sensory or attentional processes that coincidentally produce similar time-courses, or whether they reflect the common signature of supra modal and subject-specific decision-making processes. The present study addresses the hypothesis that TWs in loudness (perceived auditory intensity) and brightness (perceived visual intensity) judgments result from a common supramodal evidence-integration process. In Experiment 1, we compared TWs for loudness and brightness judgments in the same participants, with task difficulty matched individually. The observed average temporal weighting profiles differed substantially between the two modalities. In Experiment 2, we assessed the additional contribution of modality-specific sensory and attentional processes to the observed differences between TWs by measuring intensity resolution at different temporal positions in the auditory and visual stimuli. We observed a significantly different dependence of sensitivity on temporal position in the two modalities, but these sensitivity differences only partially accounted for the temporal weighting differences observed in Experiment 1. The collective findings indicate that the TWs observed for loudness and brightness judgments cannot be attributed to a supramodal evidence-integration process alone. Instead, our results suggest that both sensory and decision-making processes shape patterns of TWs.

Keywords: Temporal weighting, Sensory decision making, Loudness, Brightness, Evidence integration

1 Introduction

When humans evaluate perceptual qualities of time-varying stimuli such as the overall magnitude of a given stimulus dimension, non-uniform temporal weighting patterns are commonly observed, reflecting the fact that the information conveyed through different temporal portions does not equally contribute to the perceptual decisions. Such non-uniform temporal weighting patterns have been reported within different sensory modalities and for a variety of perceptual qualities, such as in the auditory domain, for loudness or pitch judgments (e.g., [5, 58, 63]), or in the visual domain, for brightness or direction-of-motion judgments (e.g., [10, 40]). Furthermore, these non-uniform integration profiles are observed for stimuli covering a large range of durations, from a few milliseconds up to several seconds [55]. Yet, it is not clear which mechanism underlies these effects. Even more fundamentally, it remains unknown whether the same mechanisms are engaged across different modalities, sensory dimensions and durations, or whether we are encountering distinct mechanisms with comparable outcomes on observed temporal weighting.

Here, we were specifically interested in the temporal weighting underlying judgments of the *overall ("global") intensity* of a time-varying stimulus. Typically, the experimental design employed to address this question in the auditory domain consists in presenting broadband noises varying in sound pressure level over time to participants who are asked to judge the overall loudness (perceived intensity; [38]) of the sounds. In such loudness judgments, primacy effects, i.e., a stronger influence of early stimulus information (the initial 300–500 milliseconds of a sound) compared to the contribution later parts of the stimulus to the judgments has been observed very consistently across a large number of studies (e.g., [21, 56, 58, 63, 64]). Two alternative accounts for these primacy effects in loudness judgments have been discussed most widely in the literature

*Corresponding author: oberfeld@uni-mainz.de

(e.g., [21, 57, 58]): a) neural response patterns of the auditory nerve, which show an initial peak in the firing rate at the onset of a sound with a following decline to a lower steady-state response [39], and b) non-simultaneous masking effects on the intensity resolution [50, 59, 93, 94]. Each of these explanations can account for a significant part of the empirical data, but not for all aspects (for a detailed discussion of these possible explanations see [19, 21, 53, 58]). The first potential account clearly attributes the non-uniform temporal weighting patterns to *early, sensory mechanisms*. In the second potential account, the masking effects on intensity resolution are likely caused by more central effects [50, 59], although peripheral effects could play an additional role [94]. In any case, the underlying assumption of how masking effects on intensity resolution might affect temporal weights refers to an ideal-observer idea. The assumption is that observers place higher weights on stimulus components for which their intensity resolution is higher (e.g., [25, 57]), in order to maximize their performance in an intensity judgment task. Thus, the assignment of non-uniform temporal weights (TWs) is assumed to reflect an attentional process [5], rather than direct sensory effects as in in the first potential account described above.

A third alternative explanation for the primacy effect in loudness judgments, based on *decisional processes*, can also be brought forward. This account refers to *sequential evidence integration mechanisms,* proposed in decision models such as Stone [84], the accumulator model [89], the diffusion model [68, 69], or decision field theory [12]. In experiments measuring temporal loudness weights, participants typically decide whether a) the current stimulus is more or less intense compared to previous stimuli in the experiment (one-interval, two-alternative forced-choice; 1I, 2AFC), or b) which of two sounds presented in a two-interval task is louder (two-interval, two-alternative forced-choice; 2I, 2AFC). The basic assumption in evidence integration models of perceptual judgments is that during the presentation of a stimulus that varies in intensity over time and that has to be judged regarding two alternative responses, participants accumulate the evidence for each of the two alternatives in a random walk process.[1] In some variants of evidence integration models, it is assumed that participants make their decision as soon as one of the decision boundaries is reached, and ignore the remaining part of the ongoing stimulus [10]. If now a decision boundary is reached before the end of the stimulus on a substantial number of trials, this results in -on average- a higher influence of early stimulus parts on the perceptual decision compared to later stimulus parts, in other words, a primacy effect [19–21, 53]). It seems reasonable to assume that the decision stage for perceptual decisions is located in later, *supramodal* rather than earlier, sensory- specific structures. For instance, if the decision stage was not supramodal, it would be difficult to account for multisensory integration. In fact, evidence for supramodal physiological signals correlated with perceptual decisions has been reported (e.g., [27, 49, 81, 87]), predominantly in (pre-)frontal cortex. An interesting implication of the theoretical concept of a *supramodal* decision stage in which the evidence integration process operates is that if the decision stage receives equivalent input from two different sensory modalities (e.g., auditory versus visual), then similar temporal weighting patterns can be expected for different sensory qualities and in different sensory modalities, because exactly the same evidence integration process forms the decision based on the equivalent sensory representations of the stimuli arriving at the decision stage.

The assumed structure of the processing stages involved in the sensory decision is depicted in Figure 1. It is similar to the structure assumed by many previous works (e.g., [43]). We assume that the external stimuli are first transformed into sensory representations, and we assume the auditory and visual sensory processing to result in equivalent sensory representations when the stimuli are equivalent (i.e., show an identical intensity variation across an identical presentation duration). An important prerequisite for assuming equivalent sensory representations in the two different modalities is that the sensory sensitivity for the intensity variation is identical between the two modalities, i.e., that the strength of the "external noise" (the intensity variation of the stimulus) relative to the strength of the "internal noise" (representing the limited sensory sensitivity) is identical between modalities, resulting in identical "signal-to-noise ratios" (e.g., [43]). In the supramodal decision stage, the evidence integration process operates on the sensory representations it receives, and the outcome of the evidence integration process determines which response option is selected by the participant.

In relation to temporal weights, several empirical works provide support to this account. In the auditory domain, primacy effects were reported not only for loudness judgments, but also for judgments of the annoyance of sounds [17], to a weaker extent for pitch judgments (e.g., [5]), and for localization judgments (e.g., [82]) (although on a much shorter time scale). In the visual domain, primacy effects were observed for brightness judgments (e.g., [10]), motion direction discrimination (e.g., [34, 40, 86]) and judgments of spatial position [34]. In the domain of numerical cognition (number averaging tasks), some studies also found a primacy effect (e.g., [31]). Using a common set of broadband noises varying in level, Dittrich and Oberfeld [17] observed comparable primacy effects for loudness and annoyance judgments in the same participants (despite a small tendency towards an additional recency effect for annoyance judgments). However, it is important to point out that the direct comparison of the TWs derived from these two tasks should be interpreted with caution because loudness and annoyance are highly correlated perceptual qualities of sounds (e.g., [30]).

Beyond these studies reporting comparable primacy effects, there are also a number of studies that observed recency effects, i.e. higher weighting of late temporal

[1] We note that in a two-interval task it would be necessary to assume either a separate evidence integration per interval with a subsequent combination of their outcomes, or that the evidence integration operates on a memory representation of the sensory information. A detailed discussion of this aspect is beyond the scope of the paper.

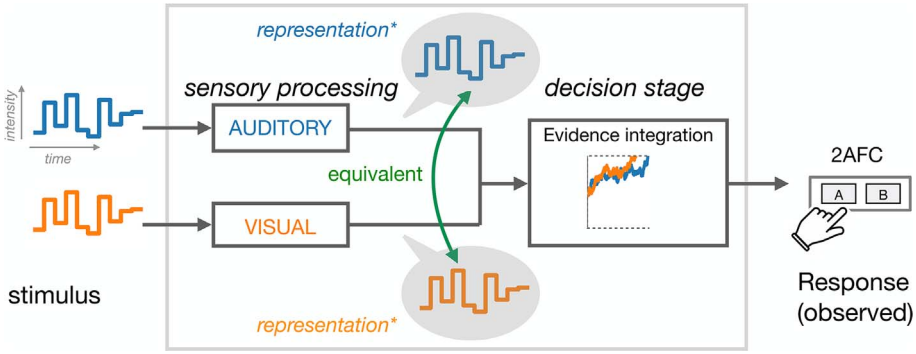

Figure 1. Schematic depiction of the assumed processing stages in perceptual judgments of the overall intensity of time-varying visual or auditory stimuli. Equivalent sensory input (i.e., stimuli with the same duration and the same intensity variation pattern across the stimulus duration) is received auditorily (blue) or visually (orange). At the sensory stage, the stimuli are transformed into (neural) representations, and we assume that under specific conditions, the auditory and visual sensory processing results in equivalent sensory representations, so that the decision stage (evidence integration process) receives equivalent auditory and visual input. The outcome of the common, *supramodal* evidence integration process, which is identical for both modalities, determines the participant's response in the 2AFC task.

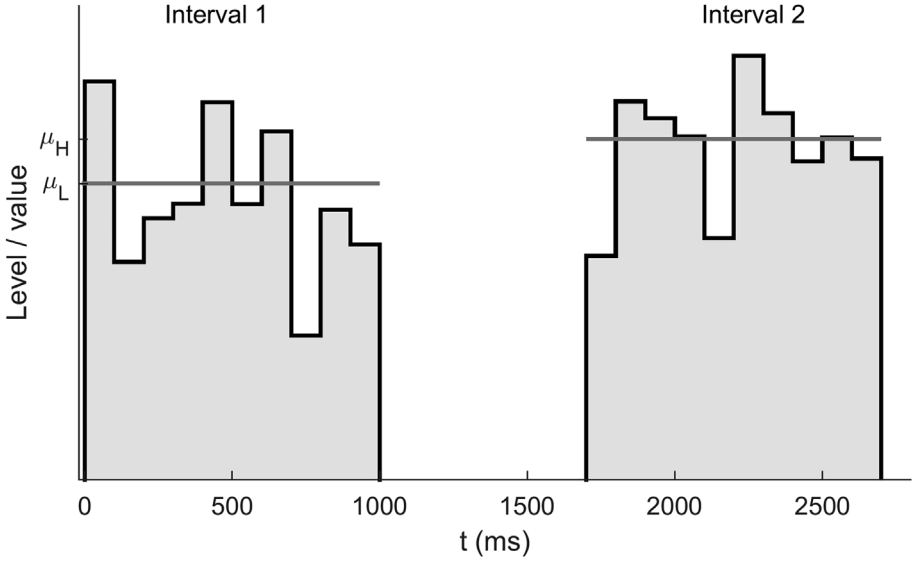

Figure 2. Experiment 1. Schematic depiction of the temporal fluctuations in sound level / RGB value in the two observation intervals. The gray horizontal lines display the mean level or RGB value of the distribution from which the stimulus components (ten contiguous 117.65-ms Gaussian wideband-noise segments per interval) were drawn in each interval. In this example, the distribution with higher mean (μ_H) is presented in the second observation interval and the distribution with lower mean (μ_L) in the first interval.

information, for instance when presenting stimuli of longer durations (e.g., [15, 34, 86]). In particular, a study by Bronfman et al. [10] observed a primacy effect in brightness judgments when stimulus duration was relatively short (1 s) but reported an additional recency effect at a longer duration (3 s). Do these results rule out an explanation of temporal weighting based on evidence integration mechanisms? The answer is no: Evidence integration models are able to produce a family of temporal weighting functions with different amounts and time-courses of primacy and recency effects depending on their computational formulation (e.g., the specific stochastic model of evidence integration, the type of decision boundaries) and parameters (e.g., [10, 65]), and it is plausible to assume that the dynamics of these processes vary with stimulus characteristics (e.g., duration). More critical to this evidence integration hypothesis, a significant number of studies have found different TWs for different to-be-judged stimulus dimensions using the exact same set of stimuli. For example, Hubert-Wallander and Boynton [34] measured TWs for judgments of the average size or spatial position of a series of dots in the same participants, and observed a primacy effect for size judgments but a recency effect for location judgments. In a study by Sato and Motoyoshi [77], observers were asked to judge the average of either the numerical value or the orientation of series of visually presented digits, in separate blocks. Their results showed a recency effect for judgments of the average orientation, but uniform TWs

for the numerical average. However, it might be the case that – within a given participant – the dynamics of the evidence integration process differ between stimulus dimensions. As such, these results cannot definitely rule out an explanation based on the evidence integration mechanism. Before that background, a more targeted empirical evaluation of this hypothesis would be to determine whether similar non- uniform temporal weighting patterns are observed for perceptual decisions involving the *same stimulus dimension* (e.g., intensity) in *different sensory modalities* (e.g., auditory versus visual), for which the dynamics of evidence integration mechanisms should remain unchanged. This assumption derives from the fact that decisional processes are generally conceived as being located at later, supramodal processing stages.

Here, to test the hypothesis that temporal weighting patterns observed for perceptual judgments are predominantly driven by a supra-modal evidence integration mechanism, rather than by specific sensory-related processes, we focused on a *single stimulus dimension*, namely intensity, and compared the temporal weighting patterns for judgments of *auditory* intensity (loudness) and *visual* intensity (brightness), within the same participants. Importantly, we carefully selected the experimental conditions to avoid differences in the sensory representations of the stimuli arriving at the decision stage between modalities. The temporal structure of the temporal variation on the dimension of interest was identical for the two tasks (same number and duration of the temporal components), which was important because the total stimulus duration [10] or the duration of the temporal components [56] can affect the TWs. The intensity variation of all temporal portions of the stimuli was set so that all components were equally reliable in a given task. Thus, an ideal observer not affected by sensory limitations (i.e., there is no "internal noise"; [85]) would have assigned uniform TWs in both tasks [5]. The level of difficulty was matched individually between tasks, because in, e.g., an evidence integration process with fixed absorbing boundaries, the boundaries would be reached earlier when the task is easier, resulting in a more pronounced primacy effect. In addition, the two tasks (loudness and brightness judgments) were randomly interleaved within each experimental block, rather than being presented in a blockwise fashion. This was done to prevent that participants establish different decision strategies or decision criteria for each block of trials. With this high degree of control and matching, we reasoned that a similarity between the temporal weighting patterns observed in the two tasks / modalities would be compatible with the predominant role of a supramodal decision stage (i.e., an evidence integration process), while differences would rather suggest modality-specific effects of earlier sensory and thus modality-specific processes (e.g., [61]).

In Experiment 1, TWs in loudness judgments of broadband sounds varying in sound level across time were compared to TWs in brightness judgments of a rectangle varying in luminance over time. The TWs in the two tasks differed significantly, with a primacy effect in loudness but a recency effect in brightness judgments, lending no immediate support to the assumption that the TWs are primarily caused by a supramodal evidence integration process. To gain a better understanding of the results of Experiment 1, in Experiment 2 we measured the sensitivity for intensity changes (loudness and brightness) at several different temporal positions within the longer auditory and visual stimuli presented in Experiment 1. This enabled us to estimate to which extent differences in sensitivity between temporal components and potential resulting attentional processes might have caused the different patterns of TWs in Experiment 1.

2 Experiment 1

In Experiment 1, participants judged the overall intensity of sounds varying in level over time (*loudness judgment task*) and rectangles varying in luminance (*brightness judgment task*). As discussed above, a fundamental requirement to investigate the hypothesis that a supramodal decisional process is the main source of the non-uniform pattern of TWs was that (i) the temporal structure, (ii) the amount of information provided by the different temporal stimulus components, and (iii) the level of difficulty is comparable for both tasks. Therefore, in both tasks the stimuli consisted of the same number of temporal stimulus components with the same duration, each component providing an equal amount of information concerning the required decision. Furthermore, the difficulty of the tasks was matched individually.

2.1 Method

2.1.1 Participants

We tested eight participants with normal hearing and normal or corrected-to-normal visual acuity (7 female, 1 male, age 22 – 28 years). They reported no history of hearing problems. Hearing thresholds were measured by Békésy audiometry with pulsed 270-ms pure tones. All participants showed thresholds less than or equal to 15 dB HL on both ears in the frequency range between 125 Hz and 8 kHz. Visual acuity was measured using the Freiburg Visual Acuity and Contrast Test (FrACT; [3]). All participants had a visual acuity of 1.0 or better. They were students from Johannes Gutenberg-Universität Mainz and received partial course credit for their participation. The experiments were conducted according to the principles expressed in the Declaration of Helsinki. All participants participated voluntarily and provided informed written consent, after the topic of the study and potential risks had been explained to them. They were uninformed about the experimental hypotheses. The Ethics Committee of the Institute of Psychology of the Johannes Gutenberg-Universität Mainz approved the study (reference number 2016-JGU-psychEK-002).

We are committed to the psychophysical tradition that recognizes the importance of collecting a sufficient number of trials per participant and experimental condition (1060 trials per task in this experiment, see below) to obtain

reliable individual data (e.g., [11, 79]). In previous studies on temporal weights, the combination of sample size and number of trials per participant and experimental condition we used here provided sufficient power to reliably detect effects of the experimental parameters (e.g., [20, 58, 64, 72]).

2.1.2 Stimuli and apparatus

In the loudness-judgment task, the level-fluctuating sounds consisted of ten contiguous 117.65-ms Gaussian wideband-noise segments. Level fluctuations were created by assigning each segment a sound pressure level drawn independently and at random from a normal distribution on each trial (see Sect. 2.2). Thus, a 1176.5 ms broadband noise with abrupt level changes every 117.65 ms was presented. All sounds were generated digitally, D/A-converted by an RME ADI/S with a sampling frequency of 44.1 kHz and 24-bit resolution, attenuated by a TDT PA5 programmable attenuator, buffered by a TDT HB7 headphone buffer, and presented diotically via Sennheiser HDA 200 circumaural headphones. The audio system was calibrated according to IEC 60318-1:1998 [36].

The stimulus for the brightness judgment task was a rectangle varying in brightness across time. A gray rectangle (digital values R = G = B) was presented with a visual angle of 5° in height and 3° in width (height = 5 cm, viewing distance ≈ 50 cm), in the center of the screen. The grand mean luminance was 38.18 cd/m^2, and the gray background had a luminance of 0.45 cd/m^2. Every 10 video frames (i.e., each 117.65 ms), the luminance of the rectangle was changed abruptly, similar to the abrupt sound level changes in the loudness judgment task. The grayscale RGB values (R = G = B) of the 20 temporal components were drawn independently and at random from a normal distribution on each trial (see below). The visual stimuli were presented on a luminance-calibrated CRT display (Dell M783p, resolution 1024 × 768 pixels, refresh rate 85 Hz). All visual stimuli and instructions were presented and timed via Psych Toolbox 3.0.16 in Matlab 2017b. Participants were tested in a double-walled sound-insulated chamber (IAC Acoustics).

2.2 Procedure

To estimate TWs in the loudness judgment task, we used an experimental paradigm based on previous experiments on loudness judgments (e.g., [18, 58, 63, 72]). On each trial, two level-fluctuating noises were presented. The segment levels presented in each interval were set by drawing each segment's level independently and at random from a normal distribution. All segment levels in one of the intervals were sampled from a level distribution with a higher mean whereas the segment levels in the other interval were sampled from a distribution with lower mean. The grand mean of both distribution means was 58 dB SPL. With equal a-priori probability, segment levels sampled from the distribution with higher mean were presented in the first or in the second interval. The standard deviation of both distributions was $\sigma = 2.5$ dB. The initial difference between the higher and the lower mean ($\Delta\mu$) in the first block of the experimental task in the first session was 1.5 dB and was individually adjusted within each session by either increasing $\Delta\mu$ by 0.075 dB whenever a participant had produced less than 65% of correct responses within the last 50 trials in the corresponding task, or by reducing $\Delta\mu$ by 0.075 dB whenever a participant had produced more than 75% of correct responses within the last 50 trials in the corresponding task. This adaptive procedure was used to maintain the two tasks at the same level of difficulty throughout the experiment. Participants started the first block of the task in the second session and in each of the following sessions with the $\Delta\mu$ value of the last block of the task within the previous session. To avoid overly loud or soft segments, the range of possible segment levels was limited to $\mu \pm 3 \cdot \sigma$.

The same random sampling procedure was used in the brightness judgment task. The grand mean grayscale RGB value was 140.25 RGB. The initial difference between the higher and the lower distribution ($\Delta\mu$) was 15.3 RGB and was adjusted in steps of 1.275 RGB. The standard deviation of the distributions was $\sigma = 17.85$ RGB. As for the loudness task, the range of possible RGB values was limited to $\mu \pm 3 \cdot \sigma$.

Table 1 shows the average individual differences between the higher and lower means of the distributions ($\Delta\mu$). Because the quotient of the difference in mean of the two distributions ($\Delta\mu$) to the standard deviation (σ) – that is, the maximal d' – was identical for each component value, each temporal component in principle provided the same amount of information concerning the decision, so that an ideal observer not affected by sensory limitations would assign uniform TWs (e.g., [5]).

Figure 2 depicts an example trial, showing the random fluctuations in sound level or RGB value in the two observation intervals. On each trial, participants decided whether the stimulus in the first or the second interval was on average louder (in case of the sounds) or brighter (in case of the rectangles) than the stimulus presented in the other interval. Thus, a two-interval, two-alternative forced-choice (2I, 2AFC) task was used. One could also describe it as a *sample discrimination task* [6, 41, 44, 80] where the participants decided in which interval the stimulus components had been drawn from the distribution with higher mean.

At the beginning of each trial, a visual symbol (task cue) indicating the task of the current trial was shown for 300 ms on the computer screen. In the brightness judgment task, this was a lightbulb symbol. In the loudness task, a headphones symbol was presented. Following the cue, a black screen was shown for 500 ms, and then the first observation interval was presented, containing either a level-fluctuating sound or a luminance-fluctuating rectangle. The inter-stimulus interval within each trial was 700 ms, followed by the second observation interval. After the second interval, the participants pressed one of two response buttons, corresponding to the interval that they had perceived as containing the stimulus with the higher loudness or higher brightness. Trial-by-trial feedback was

Table 1. Experiment 1. Individual values of the difference between the higher and lower means of the intensity distributions ($\Delta\mu$), the proportion of "2nd interval louder/brighter"-responses (p_{Resp2}) as a measure of response bias, and sensitivity (d') in the loudness and brightness task. Displayed are means (M) and standard deviations (SD) across the five experimental sessions. Values of $\Delta\mu$ are reported in dB and RGB-value for the loudness and brightness task, respectively.

		$\Delta\mu$		p_{Resp2}		d'	
Participant	Task	*M*	*SD*	*M*	*SD*	*M*	*SD*
1	Brightness	20.05	2.27	0.63	0.08	0.70	0.24
1	Loudness	1.17	0.24	0.36	0.07	0.86	0.17
2	Brightness	9.33	0.48	0.57	0.04	0.89	0.14
2	Loudness	1.08	0.07	0.48	0.06	0.77	0.13
3	Brightness	12.82	1.22	0.66	0.05	0.92	0.19
3	Loudness	1.00	0.17	0.54	0.12	0.78	0.09
4	Brightness	8.87	2.00	0.65	0.02	0.96	0.11
4	Loudness	0.97	0.11	0.59	0.09	0.81	0.17
5	Brightness	10.73	0.78	0.53	0.07	0.88	0.08
5	Loudness	1.23	0.16	0.48	0.06	0.84	0.13
6	Brightness	8.01	1.76	0.49	0.04	0.74	0.20
6	Loudness	1.00	0.17	0.41	0.04	0.90	0.09
7	Brightness	9.13	2.69	0.55	0.03	0.81	0.16
7	Loudness	1.04	0.06	0.50	0.04	0.87	0.06
8	Brightness	15.37	0.55	0.67	0.04	0.90	0.13
8	Loudness	1.55	0.11	0.71	0.06	0.70	0.19

provided during the first seven trials of each block so that participants could easily adopt a decision criterion for the potentially changed difference between the two distribution means. Those trials were not considered in the data analysis. A summarizing feedback was provided each time 30 trials were completed as well as at the end of a block, which contained 60 trials. The feedback contained the percentage of correct responses. A response was classified as correct if the response ("interval 1"/"interval 2") matched the interval in which the stimulus levels were drawn from the distribution with the higher mean. Trials from the two tasks (loudness and brightness) were randomly interleaved within each experimental block, to prevent potential differences in the decisional strategies between the tasks when the tasks are blocked.

For each participant, we collected a total of 2925 trials, distributed across 6 sessions. In addition to five main experimental sessions, there was an initial session in which hearing levels and visual acuity were measured and practice blocks were presented for both tasks. After exclusion of practice blocks and feedback trials, 2120 trials per participant (1060 trials per task, thus 106 trials per stimulus component × task; distributed evenly across five experimental sessions) entered the data analysis. The duration of each session was approximately 60 minutes, including a mandatory pause after 30 minutes.

2.2.1 Data analysis

The perceptual weights representing the importance of the temporal components for the decision in the sample discrimination task were estimated from the trial-by-trial data via multiple logistic regression. The decision model assumed that the participant compares a weighted sum of the 10 temporal component values (sound level or RGB value) in interval 2 and the negative values of the 10 temporal component values in interval 1 to a fixed decision criterion, and responds that the more intense stimulus was presented in interval 2 if the weighted sum of the exceeds the criterion (a detailed description of the assumed decision model is provided by [51]). If the weighted sum is higher than the criterion, then the model predicts that the participant classifies the stimulus presented in the second interval as more intense. In the data analysis, the binary responses ("interval 1" or "interval 2") served as the dependent variable. The predictors (i.e., the 20 temporal component intensities) were entered simultaneously. The regression coefficients were taken as the decision weight estimates. For a given temporal component, a regression coefficient equal to zero means that the temporal component had no influence at all on the decision. For the same component, a regression coefficient greater than zero means that the probability of responding that the stimulus in interval 2 was more intense increased with the intensity of the component for components presented in the second interval or decreased with the intensity of the component for components presented in the first interval.

A separate logistic regression model was fitted for each combination of participant and task. Because the *relative* contributions of the different component values to the decision were of interest rather than the absolute magnitude of the regression coefficients, the 20 regression coefficients were normalized for each fitted model and separately for the 10 components in each interval, such that the mean of the absolute values of the weights within each interval was 1.0.

A summary measure of the predictive power of a logistic regression model is the area under the Receiver Operating Characteristic (ROC) curve (for details see [17]). Areas of 0.5 and 1.0 correspond to chance performance and perfect performance of the model, respectively. Across the 16 (participant × task) fitted logistic regression models, the

area under the ROC curve ranged from AUC = 0.78 to 0.92 ($M = 0.85$, $SD = 0.04$), which can be viewed as a reasonably good fit.

The individual normalized TWs were analyzed with repeated-measures analyses of variance (rmANOVAs) using a univariate approach with Huynh-Feldt correction for the degrees of freedom [35]. The correction factor $\tilde{\varepsilon}$ is reported, and partial η^2 is reported as measure of association strength. An α-level of .05 was used for all analyses.

2.3 Results and discussion

Table 1 shows the individual sensitivity in terms of d', together with the proportion of "second interval" responses as a measure of response bias. For the calculation of d', a trial was treated as "signal" in the signal-detection theory sense when the stimulus in interval 1 was drawn from the distribution with higher mean, whereas when the stimulus in interval 2 was drawn from the distribution with the higher mean, the trial was treated as "noise". Thus, a "hit" was scored when the stimulus drawn from the distribution with higher mean was presented in the first interval, and the participant responded that the more intense stimulus had been presented in interval 1. To correct for potential extreme proportions (0.0 or 1.0), 0.5 was added to both the number of hits and the number of false alarms, and 1.0 was added to both the number of signal and noise trials (log-normal correction; [29]). An rmANOVA with the within-subjects factors task (loudness, brightness) and session (2–6) showed no significant effect of task on d', $F(1, 7) = 0.43$, $p = .535$, $\eta_p^2 = .058$, confirming that the adaptive procedure resulted in a comparable difficulty of the two tasks, as intended. There was no significant effect of session, $F(4, 28) = 0.25$, $\tilde{\varepsilon} = .860$, $p = .880$, $\eta_p^2 = .035$, and no significant task × session interaction $F(4, 28) = 2.40$, $\tilde{\varepsilon} = .750$, $p = .096$, $\eta_p^2 = .256$, indicating that the adaptive procedure resulted in relatively stable task-difficulty between sessions for both tasks.

An rmANOVA with the within-subjects factors task (loudness, brightness) and session (2–6) showed a significant effect of task on the proportion of "second interval louder/brighter" responses (i.e., the response bias), $F(1, 7) = 7.26$, $p = .031$, $\eta_p^2 = .509$. On average, the proportion of "second interval" responses was larger for the brightness task ($M = 0.59$, $SD = 0.067$) than for the loudness task ($M = 0.51$, $SD = 0.11$). There was also a significant effect of session, $F(4, 28) = 4.85$, $\tilde{\varepsilon} = .898$, $p = .006$, $\eta_p^2 = .409$. On average, the proportion of "second interval" responses was largest for session 5 ($M = 0.59$, $SD = 0.11$) and lowest for session 3 ($M = 0.51$, $SD = 0.07$). The task × session interaction was significant, $F(4, 28) = 4.35$, $\tilde{\varepsilon} = 1.015$, $p = .007$, $\eta_p^2 = .383$, indicating that the response bias differed between sessions and tasks. On average, the proportion of "second interval" responses was lowest for session 2 of the loudness task and largest for session 5 of the brightness task.

Figure 3 shows the mean normalized TWs in the two tasks, as a function of component number and averaged across intervals. For the loudness task (blue loudspeaker symbols), the pattern of weights showed a clear primacy effect with the highest weights being assigned to the first two components. For the brightness task (orange sun symbols), the mean weights showed a recency effect, with the highest weight observed on the last component. The first component received the lowest weight, and the assigned weight increased considerably from the first to the second component, giving rise to a weak "delayed primacy effect" [58].

We conducted an rmANOVA on the normalized weights, with the within-subjects factors task (loudness, brightness), component number (1–10), and interval (1, 2). The effect of component number was not significant, $F(9, 63) = 2.50$, $\tilde{\varepsilon} = .436$, $p = .067$, $\eta_p^2 = .263$, indicating that averaged across the two tasks the weights assigned to the temporal components did not vary very substantially. This test result can be attributed to the partly opposing weighting patterns in the two tasks. Importantly, the task × component number interaction was significant with a substantial effect size, $F(9, 63) = 7.41$, $\tilde{\varepsilon} = .760$, $p < .001$, $\eta_p^2 = .514$. Thus, the pattern of weights differed between the two tasks, which conflicts with the hypothesis of TWs in intensity judgments being independent of the sensory modality. Separate post-hoc rmANOVAs for each task, with the factors component number and interval, showed a significant main effect of component number for the rmANOVA in the loudness task, $F(9, 63) = 8.82$, $\tilde{\varepsilon} = .396$, $p < .001$, $\eta_p^2 = .371$ as well as in the brightness task $F(9, 63) = 2.81$, $\tilde{\varepsilon} = .666$, $p = .022$, $\eta_p^2 = .210$. Thus, participants assigned significantly non-uniform TWs to the ten temporal components within each interval in both of the tasks, but the pattern of weights differed significantly between tasks.

In the rmANOVA that included the TWs of both tasks, the component number × interval interaction was not significant, $F(9, 63) = 2.03$, $\tilde{\varepsilon} = .956$, $p = .054$, $\eta_p^2 = .225$, indicating that the weighting patterns did not differ substantially between the two intervals. The component × task × interval interaction was also not significant, $F(9, 63) = 1.78$, $\tilde{\varepsilon} = .643$, $p = .130$, $\eta_p^2 = .203$.

Figure 4 shows the normalized individual weights, averaged across interval. In the loudness task (blue loudspeaker symbols), all participants showed a primacy effect with higher weights on the first one or two segments than on the following segments. Only for participant 4, the weights also showed an additional recency effect, with the weight on the last component slightly exceeding that of the first segment. For participant 7, the weighting curve was relatively flat. In the brightness task (orange sun symbols), the inter-individual variation of the weighting patterns was larger than in the loudness task. For example, several participants showed a recency effect (participants 1, 4, 6, 7 and 8), while participants 1, 2, 3 and 5 (additionally) showed a type of "delayed primacy effect", with a low weight on the first component and higher weights on the following component(s).

In the next step, we quantitatively assessed the degree of similarity of the individual temporal weighting patterns between the two tasks, which we term the *within-subject, between-tasks similarity* (WSBT). We quantified it by

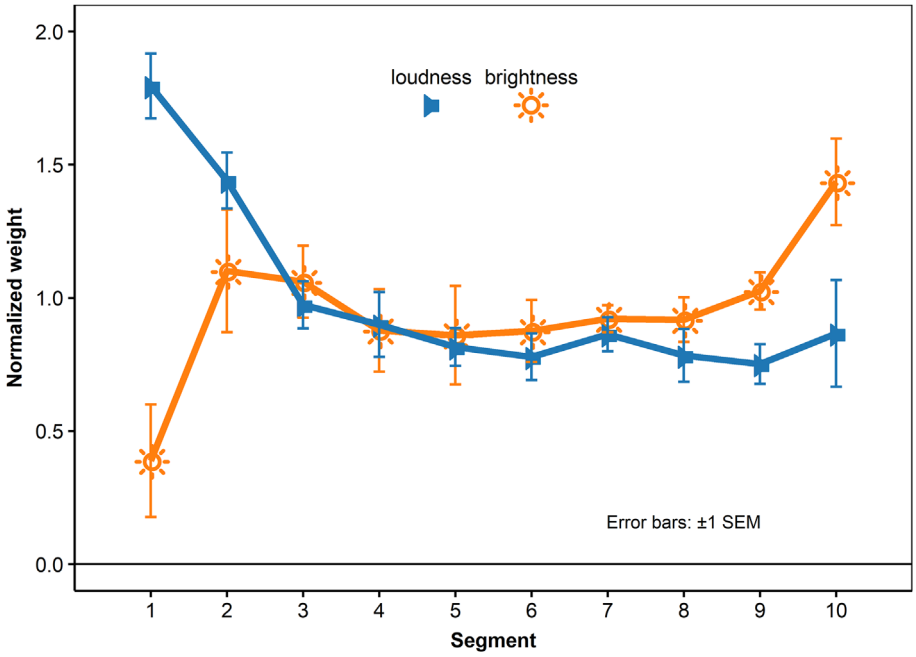

Figure 3. Experiment 1. Normalized temporal weights as a function of component number for the two tasks. Task is indicated by color and symbol (blue loudspeakers: loudness, orange suns: brightness). Error bars show ± 1 standard error of the mean (SEM) across the 8 participants.

computing the cosine similarity between the normalized individual weights averaged across interval, as plotted in Figure 4. The sequences of 10 TWs in the loudness and in the brightness tasks, respectively, are interpreted as two 10-dimensional vectors (**A** and **B**), and the cosine similarity is the cosine of the angle θ between the two vectors, conveniently computed as the inner product of two normalized vectors, $\cos(\theta) = \boldsymbol{A} \cdot \boldsymbol{B}(||\boldsymbol{A}||||\boldsymbol{B}||)$, where $||\boldsymbol{A}||$ denotes the l^2-norm of vector $\boldsymbol{A}$. A value of $\cos(\theta) = 1$ represents perfect similarity (i.e., the vectors are parallel), while a value of $\cos(\theta) = 0$ represents the maximum amount of dissimilarity that can be observed when all weights are positive (i.e., the vectors are orthogonal). We chose the cosine similarity over other possible similarity metrics because our goal here was to assess the similarity between the *shape* of the temporal weighting patterns, irrespective of their absolute magnitude. The cosine similarity is independent of the scaling of the vectors, because only the vectors' direction in space but not their magnitude is considered. Thus, the cosine similarity is not affected by the scaling of the perceptual weights.

To evaluate to which extent individual WSBT similarity values smaller than 1.0 reflect systematic differences between the TW patterns for loudness and brightness, we used a resampling approach and compared the WSBT similarity to the similarity of the TWs within each subject and task estimated for random splits of the individual data into two non-overlapping subsets ("folds"; similar to the data partitioning in a 2-fold cross-validation). The latter cosine similarity can be viewed as the split-half reliability of the TWs within subject and task. For each combination of subject and task, we randomly partitioned the data into two non-overlapping subsets (i.e., $k = 2$ folds), computed the temporal weights for each fold using the logistic regression approach described above, and repeated this random splitting $m = 1000$ times. For each participant, the blue and orange violin plots in Figure 5 show the distribution of the $m = 1000$ within-subject, within-task, between-folds (WSWTBF) cosine similarity values for the TWs in the loudness and brightness task, respectively. Values close to 1.0 and 0.0 represent high and low split-half reliability, respectively, of the estimated temporal weights. In addition, for each subject, we computed the similarity between the TWs obtained in the two tasks (within-subject, between-tasks similarity; WSBT). For each of the $k \times m$ random subsets of the individual loudness task trials, we computed the cosine similarity between a) the loudness TWs and b) the brightness TWs in one of the $k \times m$ random subsets of the individual brightness task trials. The gray violin plots in Figure 5 show the distribution of the resulting in $k \times m = 2000$ WSBT similarity samples per participant.

The mean individual WSBT cosine similarities are shown by the gray distributions in Figure 5. They support what could already be observed at a descriptive level in Figure 4, with the lowest WSBT similarities found for P1 (0.47) and P8 (0.52), and the highest values for P2 (0.89) and P7 (0.89). For each subject, the WSBT similarity (gray distributions in Fig. 5) was significantly smaller than the WSWTBF similarity in both the loudness and the brightness task (blue and orange distributions in Fig. 5, respectively), as indicated by Welch two-samples tests (all p-values $< .00001$). This analysis indicates that the only moderately high WSBT similarities can indeed be attributed to systematic differences between the TWs in the two

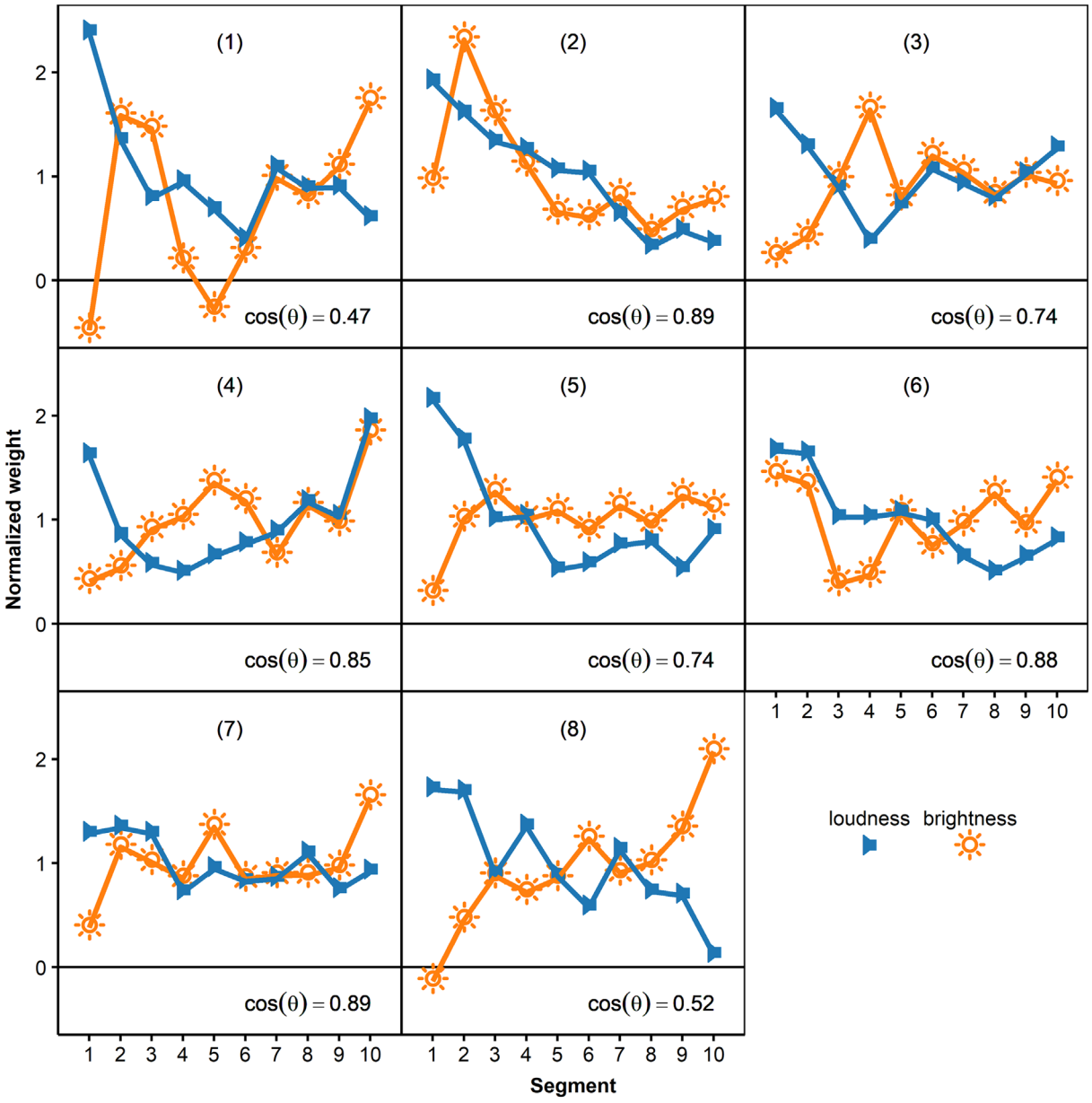

Figure 4. Experiment 1. Individual normalized temporal weights as a function of component number for the two tasks, averaged across intervals. The cos(θ) values are the cosine similarity between the two patterns of temporal weights within a given participant (WSBT similarity; see text). Task is indicated by color and symbol (blue loudspeakers: loudness, orange suns: brightness).

tasks, rather than to only the inherent noisiness of the data. This analysis also showed significantly lower WSWTBF similarity in the brightness task than in the loudness task for all subjects but participant 8 (all p-values $< .0052$).

Taken together, the results from Experiment 1 indicate that the temporal weighting patterns in two-interval intensity-judgment tasks with matched difficulty and for stimuli with identical temporal characteristics, but presented either in the auditory or the visual modality, differ quite substantially within participants. The TWs in the loudness task consistently showed primacy effects (higher weights assigned to the beginning of the sound compared to later parts) and no or only weak recency effects (except for one participant). This is in line with previous findings on temporal loudness weights (e.g., [18, 21, 58, 64]). In contrast, the mean TWs in the brightness task showed a different pattern. The first component received a lower weight than the remaining components, and on average a clear recency effect was observed. The cosine similarity analysis demonstrated that this pattern of results reflects genuine between-task differences at the level of each subject. In addition, our results showed a larger variability of TW patterns across participants in the brightness task as compared to the loudness task. The only other study of temporal brightness weights we are aware of [10] found on average a primacy effect for stimulus durations of 1 to 2 seconds

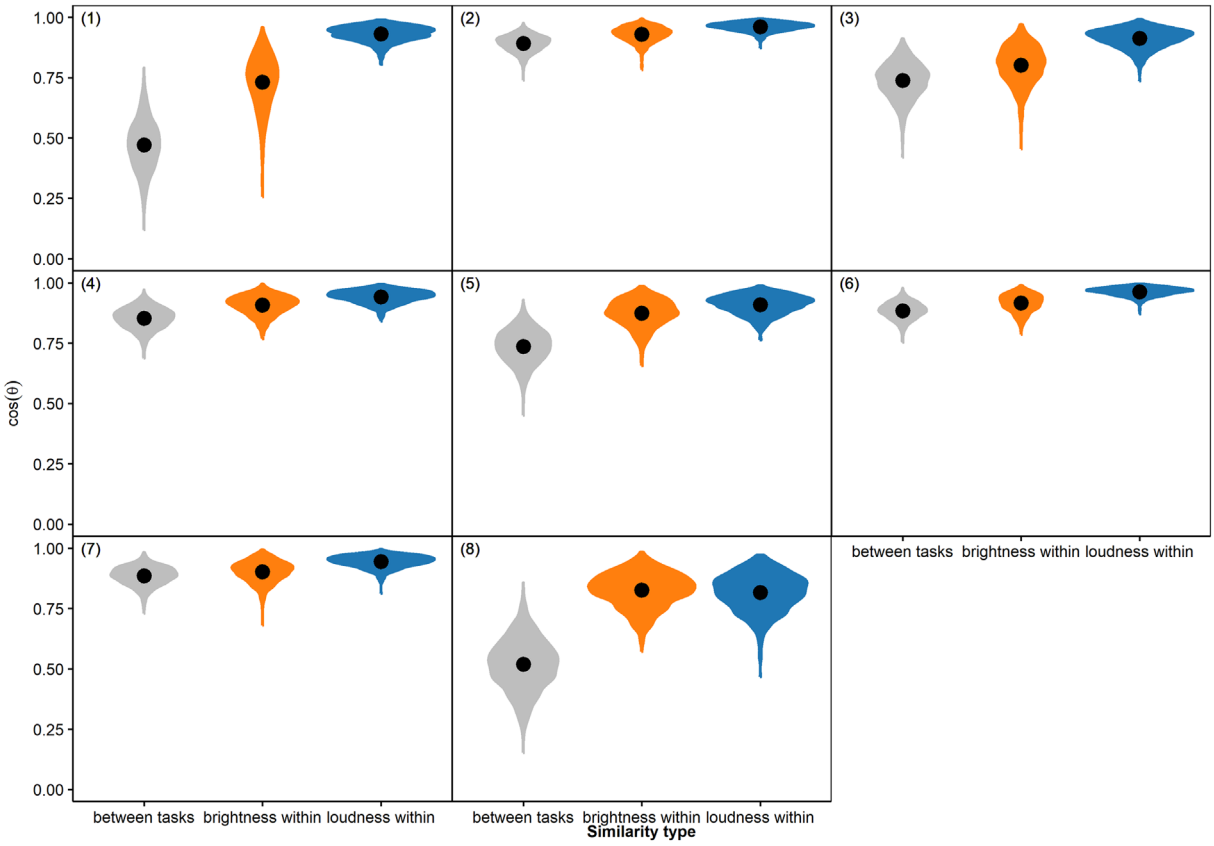

Figure 5. Exp. 1: Within-task (WSWTBF) and between-task similarities (WSBT) of individual temporal weighting patterns. Violin plots showing the distribution (kernel-smoothed density) of cosine similarity values computed for each subject (numbered in each panel) by comparing the temporal weighting patterns obtained for random splits of the trials into $k = 2$ non-overlapping folds within a given task (WSWTBF similarity; blue distribution = loudness, orange distribution = brightness), or between the two tasks (WSBT similarity; gray distribution); this process was repeated 1000 times to generate the distributions. The points show the means of the distributions.

and a primacy and a recency effect for longer durations of 3 seconds. Notwithstanding the fact that the literature suggests a higher variability in the TWs in a brightness judgment task than in a loudness judgment task, where a primacy effect is observed across a wide range of stimulus durations [55], the present observation that within subjects the TWs differed quite substantially between the two modalities is not what one would expect if the TW patterns in both tasks were caused exclusively by a common supramodal evidence integration process.

3 Experiment 2

The pronounced differences between the weighting patterns for the loudness and the brightness task observed within subjects in Experiment 1 are incompatible with the hypothesis that the TWs observed in judgments of overall perceived intensity are exclusively caused by a supramodal evidence integration process. However, the results cannot be taken as evidence against any involvement of an evidence integration process for the TWs *per se*, since they rely on the strong assumption of equivalent sensory representations of the auditory and visual stimuli arriving at the input of the decision stage (Fig. 1). Even if the same evidence integration mechanism was at play for the different tasks / modalities, differences in sensory processing would lead to different characteristics of the sensory representations arriving at the decision stage, which could result in differences between the temporal weighting patterns. For example, auditory nerve neurons show fast recovery time constants in the millisecond range [39], while the recovery of responsivity of retinal ganglion cells can take several seconds (e.g., [92]). Also, pronounced afterimages occur in the visual domain [2, 7, 9]. Comparable aftereffects are not observed in the auditory domain, where only under specific circumstances like a prolonged exposure to a pulse train a change in timbre of following sounds may be experienced [75], or where by a stimulation via a notched noise a special form of an "acoustic afterimage", the Zwicker tone, can be induced [95]. Furthermore, in the auditory domain, adaptation to intense stimuli, via the stapedius reflex, occurs within less than 100 milliseconds [1, 16] whereas in the visual domain, adaptation tends to require longer time constants for pupil and retina responses from several hundreds of milliseconds to several seconds [8, 13].

Due to these differences in the early visual and auditory sensory processing stages, the sensitivity for discriminating changes in auditory or visual intensity might depend in different ways on the temporal position of the relevant stimulus component within a longer stimulus. For instance, in the brightness judgment task, afterimage effects might have increased the intensity resolution for the final component, while adaptation due to the abrupt change in brightness at stimulus onset might have reduced the sensitivity for the first component. Because an ideal observer would place higher weights on stimulus components for which the sensitivity is higher (e.g., [25, 57]), such differences in sensitivity across temporal positions could imply different weighting patterns. Assuming that the evidence integration process produces a primacy effect when the sensitivity for each temporal component is identical, assigning a higher amount of attention to components which can be judged more precisely would modulate the pattern of TWs. For instance, while the final component should receive a low weight due to the primacy effect, when the sensitivity for intensity changes imposed on the final component is particularly high, participants might increase the weight assigned to the final component (i.e., direct attention to this particular component). For the brightness judgment task, a potential reduction in intensity resolution at stimulus onset (due to adaptation effects) and a potential increase in sensitivity at the offset (due to longer-lasting afterimages) would be compatible with the low weight on the first and the relatively high weights on the final components, respectively, in Experiment 1. In Experiment 2, we therefore measured the intensity resolution for temporal components at different temporal positions within the stimuli presented in the brightness and loudness task of Experiment 1, in order to identify differences in intensity resolution between temporal components in each task, and to determine the extent to which these task-specific differences might have contributed to the differences between the temporal weighting patterns observed in Experiment 1.

3.1 Method

3.1.1 Participants

We tested 8 participants with normal hearing and normal or correct-to-normal visual acuity (8 female, age 21 – 39 years), using the same inclusion criteria and study protocol as in Experiment 1. None of them had participated in Experiment 1.

3.1.2 Stimuli

Five different types of level-fluctuating sounds and luminance-varying visual stimuli (rectangles) were presented. The stimulus generation was identical to Experiment 1, except for the following differences. For four types of sounds and rectangles, the overall duration was $10 \cdot 117.65$ ms = 1176.5 ms, as in Experiment 1. In contrast to Experiment 1, only the sound level or luminance of a *single* temporal portion (with a duration of 117.65 ms) of the auditory and visual stimuli, respectively, varied between trials and observation intervals, while the remaining temporal portions of the stimuli were presented with constant sound level or luminance.

In the first condition (Fig. 6, panel A), the intensity-varying temporal portion, for which we use the term *target component* in the following, was presented with an onset at 0 ms and an offset at 117.65 ms. Thus, it corresponded to the first temporal component in the previous experiment (Fig. 2) and is therefore referred to as *target component 1* in the following. In the loudness task, the first component of the sound had received the highest average weight, while in the brightness task, the first component had received the lowest weight (see Fig. 3). In the second condition, the onset of the target component was at 117.65 ms, and thus corresponded to the second temporal component in Experiment 1. It is referred to as *target component 2* in the following. In the loudness task, the weight on the second component was slightly lower than the weight on the first component, while in the brightness task, the average weight assigned to the second component was substantially higher than the weight on the first component. In the third condition, the onset of the target component was at 470.59 ms, and thus corresponded to the fifth temporal component in Experiment 1. It is referred to as *target component 5* in the following. In the loudness task, the weight on the fifth component was substantially lower than the weight on the first and second components, while in the brightness task, the weight assigned to the fifth component was higher than the weight on the first component and lower than the weight on the second component. In the fourth condition, the onset of the target component was at 1058.82 ms, and thus corresponded to the final temporal component in Experiment 1. It is referred to as *target component 10* in the following. In the loudness task, the weight on the final component was substantially lower than the weight on the first and second components and similar to the weight on the fifth component, while in the brightness task, the final component had received the highest mean weight. In an additional, fifth condition, only a single, isolated 117.65-ms component was presented (Fig. 6, panel E), without any adjacent constant stimulus parts, to assess the "baseline" intensity resolution for loudness and brightness.

The intensity (sound level or luminance) of the target component varied randomly from trial to trial and between intervals, just as in Experiment 1. On each trial, the target intensity in one of the two intervals was drawn from a normal distribution with lower mean and the target intensity in the other interval was drawn from a distribution with higher mean. The remaining part(s) of the stimuli had the exact same and fixed sound pressure level of 58 dB SPL or luminance of 127.5 RGB in both intervals, respectively, corresponding to the grand mean of the higher and the lower distribution mean.

Figure 6 shows schematic depictions of the five conditions. This time, the differences between the means of the higher and lower distribution were set to a fixed, nonindividual value. For the sounds, in one of the two intervals (selected randomly), the target intensity was sampled from

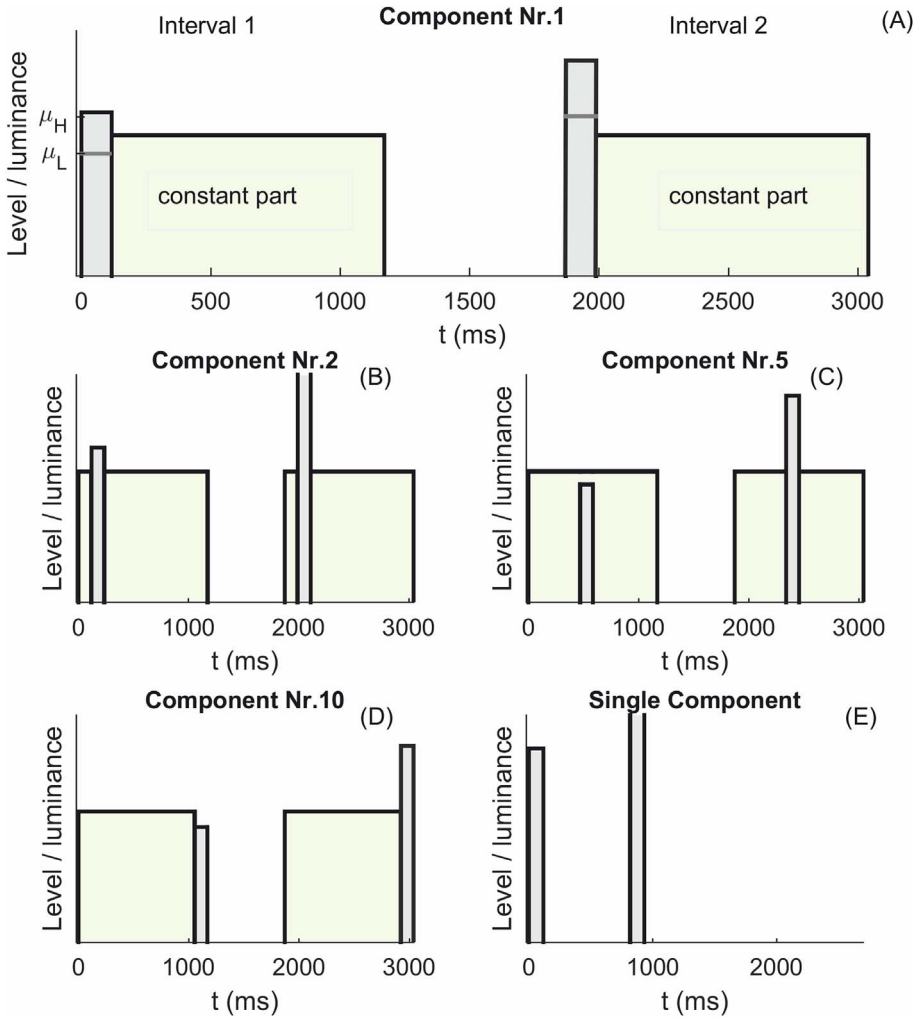

Figure 6. Experiment 2. Schematic depictions of the stimuli in the five different conditions. In each panel, the gray bars indicate the intensity of the *target component* (i.e., the temporal part of the stimulus that was varied in intensity), while the beige areas indicate the level of the constant part of the stimulus that had the same sound level/brightness in both intervals of a given trial. In this example, the target intensity is sampled from the distribution with higher mean (indicated by the gray horizontal line) in interval 2 and from the distribution with lower mean in interval 1. Panel A: target component onset at 0 ms, corresponding to the first temporal component in the time-varying stimuli in Experiment 1. Panel B: target component onset at 117.65 ms, corresponding to the second temporal component in Experiment 1. Panel C: target component onset at 470.59 ms, corresponding to the fifth temporal component in Experiment 1. Panel D: target component onset at 1058.82 ms, corresponding to the final (10th) temporal component in Experiment 1. Panel E: target component presented in isolation, without a surrounding constant-intensity stimulus part.

a lower normal level distribution with a mean of 56.46 dB SPL. The target intensity in the other interval was sampled from a higher normal level distribution with a mean of 59.6 dB SPL, with the restriction that the target intensity in the "higher" interval was not lower than the target intensity in the "lower" interval. Both distributions had a standard deviation of $\sigma = 2.15$ dB. For the rectangles, the mean of lower normal brightness distribution was 115.33 RGB and the mean of the higher normal brightness distribution was 139.95 RGB. Both distributions had a standard deviation of $\sigma = 15.5$ RGB. In both tasks, the range of possible intensity values was limited to $\mu \pm 3 \cdot \sigma$. Across all trials, the mean difference between the higher and the lower intensity was 3.03 dB (SD = 2.46 dB) in the loudness task and 24.62 RGB (SD = 18.1 dB) in the brightness task.

3.1.3 Apparatus and procedure

Apparatus and procedure were largely the same as in Experiment 1 and thus only differences are reported. Per experimental block, only one combination of target component position (see Fig. 6) and task (brightness or loudness) was presented. Participants were informed that only one temporal part of the stimulus would differ in intensity between the two observation intervals, while the intensity of the remaining parts of the stimulus would remain constant and be identical in the two observation intervals. They were also informed about the temporal position of the target component before each block. The task was to decide if the first or the second stimulus presented on a given trial contained the target component with the higher

intensity (i.e., the louder or brighter target component). Three blocks with 40 trials each were presented per combination of participant, task, and target component position. Thus, we collected 120 trials per condition and participant, resulting in a total of 1200 trials per participant, distributed evenly across three experimental sessions. The blocks were presented in randomized order.

In the first session, hearing levels were measured and practice blocks were presented for all combinations of task and target position. The data were collected in the following three experimental sessions. The first block of each experimental session was a practice block. In practice blocks and on the first 7 trials of each experimental block, participants received visual trial-by-trial feedback indicating if the response was correct or incorrect. Data from the practice blocks as well as from the first 7 trials in each block (for which trial-by-trial feedback was provided) were not included in the analysis, leaving a total of 99 trials per participant and condition for the analysis.

3.2 Results

For each combination of participant, task, and target component, we fitted a logistic psychometric function (PMF) relating the difference between the target component intensity presented in interval 2 and the target component intensity presented in interval 1 ($\Delta I_{2\text{-}1}$) to the probability of the participant responding that the target component with the higher intensity was presented in the second interval (i.e., p("Interval 2") ~ $\Delta I_{2\text{-}1}$), using a maximum-likelihood approach (logistic regression). As a measure of sensitivity, we analyzed the slope of the psychometric function at the inflection point (where p("Interval 2") = ½). For the logistic function, this slope is proportional to the estimated regression coefficient for $\Delta I_{2\text{-}1}$ ($b_{\Delta 2\text{-}1}$). The intensity difference limen (DL) defined as half the difference between the 75%- and the 25%-point on the PMF is inversely proportional to the slope of the logistic PMF, $DL = \ln(3)/b_{\Delta 2\text{-}1}$. Large slopes (and thus small intensity DLs) indicate high intensity resolution.

Figure A1 shows the mean intensity DLs for the five different target components in both tasks. For the loudness task, the mean intensity DL for the isolated component (117.65 ms broadband noise burst) was about 1 dB, compatible with the literature (e.g., [47, 67]). The mean DL for the target component presented at the offset of the longer stimulus (component 10) was slightly lower than 2 dB. For the target component presented at the onset of the longer stimulus (component 1), and even more so for the target components 2 and 5 presented within the longer stimulus, the DL was increased considerably compared to the DL for the isolated component, showing DL elevations of up to 5 dB. These results are difficult to explain in terms of non-simultaneous masking because the mean sound level of the target component was identical to the sound level of the steady, constant-intensity parts of the stimulus, resulting in only relatively few trials on which the level of the constant-intensity parts exceeded the level of the target component by more than 10 dB. At such small masker-target level differences, only relatively weak DL elevations have been reported (e.g., [50]). A plausible but speculative explanation for the increased DLs for components 2 and 5 would be that participants had difficulty to direct their attention to the temporal components presented within a longer sound. For the brightness task, no comparison data concerning intensity resolution for an isolated stimulus were available. The DL for the isolated component was virtually identical to the DL for component 10. Afterimage effects at the offset of the longer stimulus might have contributed to this high sensitivity. Unlike in the loudness task, the DL for component 5 was also relatively small. In contrast, the DLs for component 1 and particularly component 2 were considerably higher than for the other components, showing a large inter-individual variability for component 2, mainly due to a near-zero PMF slope ($b_{\Delta 2\text{-}1}$) and thus a large DL for one participant (subject 7 in Fig. A2). At least for component 1, the increased DLs might be due to adaptation effects caused by the abrupt change in brightness at stimulus onset, as discussed above.

As in Experiment 1, what we were mainly interested in are the *patterns* of sensitivities (i.e., PMF slopes) across target positions in the two tasks, rather than the absolute magnitude of the PMF slopes. To investigate whether the patterns of sensitivities differed between the two tasks, we normalized the estimated slopes of the PMF in the same way as the TWs in Experiment 1. Per participant and task, we divided each estimated slope by the mean of the slopes across the four target components embedded in a longer sound (i.e., component 1, component 2, component 5, and component 10; excluding the isolated component), so that the mean of the resulting four normalized slopes was 1.0. Figure 7 shows the mean normalized PMF slopes from Experiment 2 together with the mean normalized weights from Experiment 1, replotted from Figure 3, to visualize whether the data support the notion that listeners assign high weights in judgments of global loudness (Exp. 1) to temporal components for which the intensity resolution is high (Exp. 2).

The patterns of normalized sensitivities observed in Experiment 2 partially, but not completely, reflect the temporal weighting patterns that were observed in Experiment 1. For the brightness judgments, the intensity resolution for component 1 (at stimulus onset) and component 2 was substantially lower than the sensitivity for component 5 and the final component 10. If one assumes that the evidence integration process causes a primacy effect and that this primacy-pattern is modulated by differences in sensitivity and the assignment of attention to components for which the intensity resolution is high, then the first component should have received a higher weight than the second component due to the primacy effect, and the slightly lower sensitivity for component 2 compared to component 1 should have amplified this effect. However, the TWs in the brightness task observed in Experiment 1 show a considerably higher weight on component 2 than on component 1. The pronounced increase in weight from component 1 to component 2 observed in Experiment 1 cannot be explained by the difference in sensitivity between the two

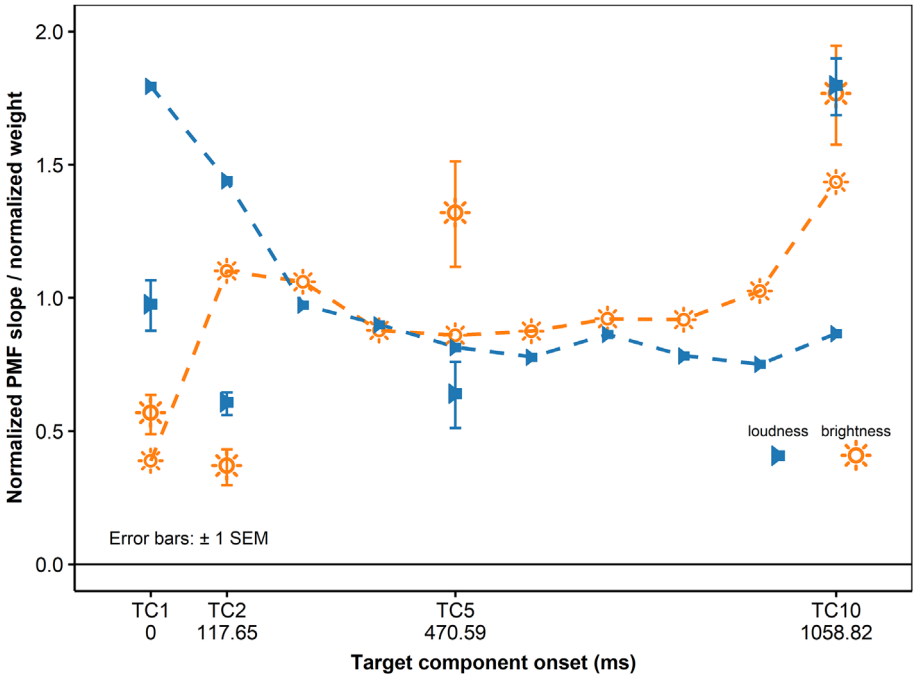

Figure 7. Mean normalized sensitivities (i.e., PMF slopes) from Experiment 2 (data points with error bars) and mean normalized temporal weights (dashed lines) from Experiment 1 (replotted from Fig. 3), as a function of target component onset. Blue speakers: loudness task. Orange suns: brightness task. Error bars show ± 1 SEM across the 8 participants of Exp. 2.

components (as measured in Experiment 2). For components 5 and 10, on the other hand, the primacy effect caused by the evidence integration process would result in relatively low weights. However, the intensity resolution for components 5 and 10 was substantially higher than for components 1 and 2, which might have compensated for the primacy effect and thus have contributed to the observed recency effect.

For the judgments of loudness, the intensity resolution for the first component exceeded the sensitivity for the second and fifth component. Thus, the decline in intensity resolution from component 1 over component 2 to component 5 might even have amplified the primacy-effect pattern in the TWs observed in Exp. 1. The highest sensitivity was observed for component 10, and this higher sensitivity might have partially compensated for the low weight assigned to this component due to the primacy effect, although an increase in weight towards the end of the stimulus was only barely visible in the mean TWs obtained in Experiment 1.

We conducted an rmANOVA on the normalized sensitivities (Fig. 7), with the within-subjects factors task (loudness, brightness) and target component position (component 1, component 2, component 5, component 10). Most important, the task × target position interaction was significant, $F(3, 21) = 8.00$, $\tilde{\varepsilon} = 0.75$, $p = .003$, $\eta_p^2 = .53$. Thus, the pattern of sensitivities across the four analyzed target component positions differed significantly between the two tasks. As can be seen in Figure 7, in the loudness task, the normalized sensitivity was higher for component 1 (onset at 0 ms) than for components 2 and 5, and again higher for component 10. In contrast, in the brightness task, the normalized sensitivity was relatively low for components 1 and 2 and higher for components 5 and 10. Across tasks, the effect of target component position on the normalized sensitivities was significant, $F(3, 21) = 23.60$, $\tilde{\varepsilon} = .50$, $p < .001$, $\eta_p^2 = .77$. Two post-hoc rmANOVAs computed separately for the two tasks showed a significant effect of target component both for loudness, $F(3, 21) = 24.49$, $\tilde{\varepsilon} = 0.56$, $p < .001$, $\eta_p^2 = .78$, and for brightness, $F(3, 21) = 15.22$, $\tilde{\varepsilon} = 0.78$, $p < .001$, $\eta_p^2 = .69$. Taken together, the differences in the sensitivity for intensity changes at different temporal positions within a longer stimulus observed in Experiment 2 are partially, but not completely, compatible with the differences in the TWs for loudness and brightness observed in Experiment 1.

In the next step, we conducted a similarity analysis identical to the one of Experiment 1. For each participant, the within-subject, within-task, between-folds (WSWTBF) similarity (i.e., split-half reliability) was computed as the cosine similarity between $m = 1000$ random splits of the individual data into two non-overlapping subsets (i.e., $k = 2$ folds). For each participant, the blue and orange violin plots in Figure 8 show the distribution of the $m = 1000$ WSWTBF cosine similarity values for the patterns of sensitivities in the loudness and brightness task, respectively, indicating high split-half reliabilities. In addition, for each subject, we computed the cosine similarity between the pattern of four sensitivities (target components 1, 2, 5, and 10; as plotted in Fig. A2) in the loudness task and the pattern of four sensitivities in the brightness task (within-subject, between-tasks similarity; WSBT). For each of the $k \times m$ random subsets of the individual loudness task trials, we computed the cosine similarity between a) the

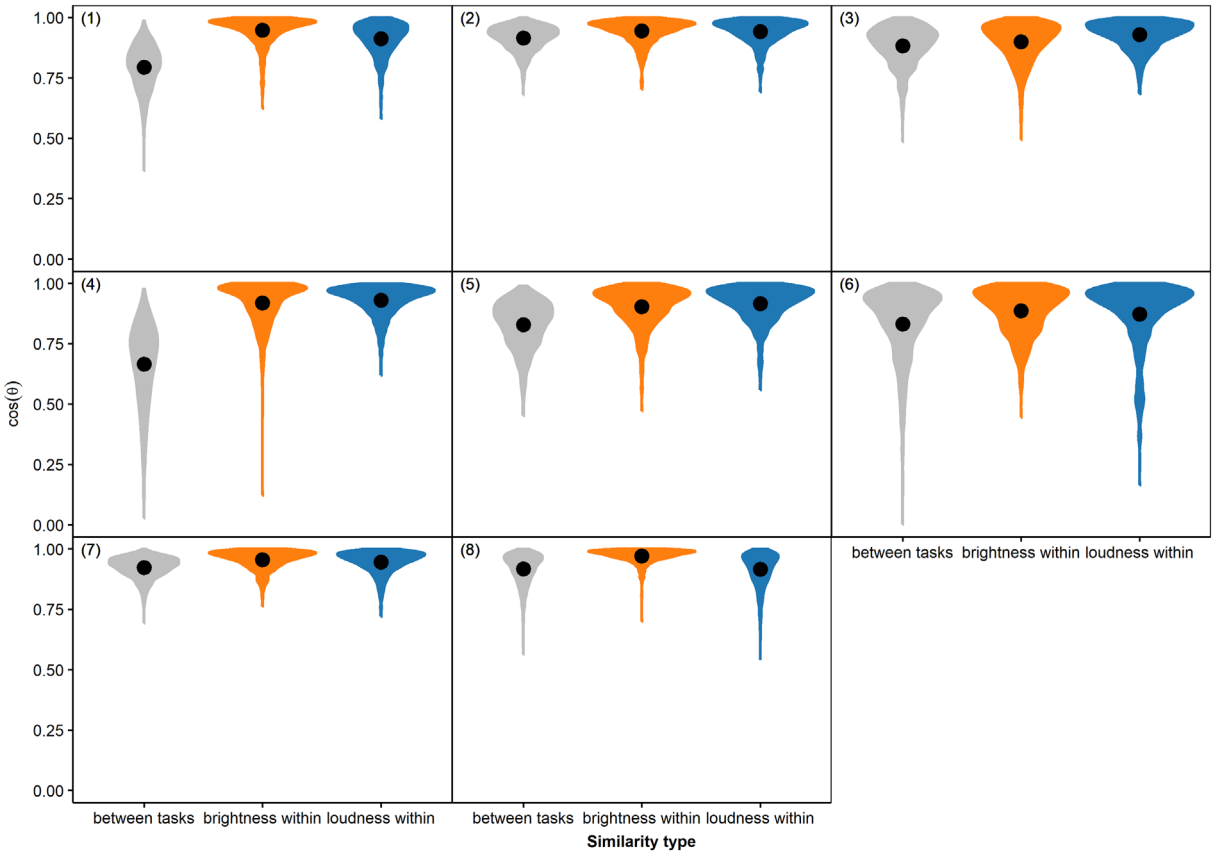

Figure 8. Exp. 2. Same format as Figure 5, but with cosine similarity values computed between individual patterns of sensitivities (Exp. 2; see Fig. A2).

loudness sensitivities and b) the brightness sensitivities in one of the $k \times m$ random subsets of the individual brightness task trials. The gray violin plots in Figure 8 show the distribution of the resulting in $k \times m = 2000$ WSBT similarity samples per participant.

The mean individual WSBT cosine similarities are shown in Figure 8. They ranged from 0.66 to 0.92. For each subject, the WSBT similarity (gray distributions in Fig. 8) was significantly smaller than the WSWTBF similarity in both the loudness and the brightness task (blue and orange distributions in Fig. 8, respectively), as indicated by Welch two-samples tests (all p-values $< .00001$), except for participant 8 in the loudness task. This analysis indicates that the only moderately high within-subjects, between-task similarities can indeed be attributed to systematic differences between the patterns of sensitivities in the two tasks, rather than to only the inherent noisiness of the data.

4 General discussion

The aim of the present study was to empirically assess the hypothesis that the TWs observed in perceptual judgments of time-varying stimuli are the result of a supramodal decision process, more specifically, an evidence integration process, as suggested in recent works (e.g., [10, 40, 86]). We specifically addressed this question in the case of intensity judgments of auditory and visual stimuli (i.e., loudness and brightness judgments).

Experiment 1 compared the TWs of loudness to the TWs of brightness in the same participants, and found that they differed substantially between the two modalities (see Fig. 3). For the loudness task, a clear primacy effect was observed (compatible with the literature; e.g., [55]), while for the brightness task, recency effects and relatively lower weights on the first few temporal portions of the stimuli were observed. A metric introduced to quantify the similarity between the patterns of TWs in the two modalities within each subject provided further evidence against the hypothesis of an evidence integration process as the unique cause of non-uniform patterns of TWs.

Incidentally, the TWs showed a higher inter-individual variability in the brightness judgment task compared to the loudness judgment task. Previous studies investigating TWs in brightness judgments also observed substantial inter-individual variability [10], but the present study allows for a direct comparison with TWs in loudness judgments. At present, the reason of the higher variability observed in the visual compared to the auditory modality remains unclear and should be addressed by future studies. Also, on average we observed a recency effect but no primacy effect in the brightness judgments (Exp. 1), while a study by Bronfman et al. [10] on temporal brightness weights reported a primacy effect reported at stimulus

durations of 1 and 2 seconds. However, in the latter study, abrupt onsets and afterimage effects were controlled by presenting video frames with grey and white stimuli before and after the actual to-be-judged stimuli as a sort of adapting field or mask, and this sole quite "minor" difference in stimulus characteristics might account for the large differences in the final shape of the patterns of TWs. In sum, the results of Experiment 1 did not support the hypothesis that patterns of TWs observed in brightness and loudness judgments can be accounted for exclusively by a single supramodal decision mechanism. Yet, they did not totally rule out that a supramodal decision process contributes to the observed non-uniform TWs, but rather suggest that TWs additionally reflect the effects of modality-specific mechanisms that differ between audition (loudness) and vision (brightness).

The goal of the Experiment 2 was thus to better appreciate such contributions of modality-specific mechanisms in the observed differences between patterns of TWs for brightness and loudness. We specifically investigated the possibility that the low weight at stimulus onset and the high weight at the offset observed in the brightness judgments in Experiment 1 could be due to characteristics of early visual processing. In other words, we empirically investigated to which extent our initial assumption of the decision process receiving equivalent sensory representations of the auditory and visual stimuli (Fig. 1) was incorrect. To this end, we measured the sensitivity for auditory or visual intensity changes imposed on the first, intermediate, and final temporal portions of auditory or visual stimuli with a duration identical to the stimuli presented in Experiment 1. The results showed a significantly different dependence of relative sensitivity on the temporal position of the target component within the stimulus between the two modalities (Fig. 7). Overall, the observed differences in relative sensitivity measured in Experiment 2 were partially, but not entirely, compatible with the differences in TWs observed in Experiment 1 (see Fig. 7). Taken together, results from these two experiments thus suggest that the temporal weighting patterns of loudness and brightness judgments cannot be accounted for completely by an evidence integration process, and that it is important to consider the additional contribution of sensory and attentional mechanisms that might substantially shape these temporal weighting patterns. Our results thus speak for a *combined* influence of sensory, attentional, and decisional processes on TWs in general, in line with earlier studies that also emphasized that the empirically observed TWs in sensory judgments could be due to early sensory processes, later attentional or decisional mechanisms, or a combination of each of the three types (e.g., [21, 57, 58, 61]).

From this point of a view, different temporal dynamics of auditory (loudness) and visual (brightness) intensity processing would imply that the traces of the sensory representations arriving at the decision stage (see Fig. 1) might differ between audition and vision, despite the carefully matched temporal dynamics of the presented stimuli. In a certain sense, one could say that the temporal resolution of the auditory system is higher than for the visual system (see below). First of all, auditory neurons show phase-locking to the acoustic temporal fine structure up to 4 kHz [74]. Anecdotally, audio enthusiasts typically agree that sampling rates of at least 44.1 kHz are required for faithful reproductions of acoustic signals such as music by digital audio systems, while most movie cinema visitors do not complain that the frame rate of classical analogue movies is only 24 or 25 Hz. However, it is important to note that there is not only a single, unequivocal value characterizing the temporal resolution of a sensory system. Instead, depending on the task or measure considered, quite different values for "temporal resolution" are obtained (e.g., [88]). For the perception of brightness and loudness, which is most relevant here, data on temporal integration in vision and audition indicate only relatively moderate differences between the temporal dynamics of the two senses. In vision, the brightness increases with increases in the duration of the stimulus up to between 5 and 100 ms and is largely independent of stimulus duration above that critical duration, for typical luminance levels, and even shorter critical durations are observed at higher luminance [45], consistent with measurements of cone responses [32]. Thus, there is temporal integration of brightness, up to durations of 5–100 ms. In audition, the critical duration for the temporal integration of loudness is about 150 to 200 ms [83, 97] and the amount of temporal integration depends on the sound intensity in a non-monotonic fashion [22]. Hots et al. [33] suggested that the dependence of loudness on stimulus duration can be well predicted by a temporal integration stage consisting of two parallel low-pass filters with different time constants. At the threshold of audibility/visibility, the time constants for temporal integration are somewhat longer than at suprathreshold levels, both in vision [23] and audition [88]. In vision, longer critical durations at the detection threshold were observed in dark-adapted compared to light-adapted states (e.g., [76]). In audition, the minimum detectable temporal gap between two sounds was reported to be in the range between 2 and 100 ms, depending on the signal frequency and bandwidth (e.g., [26, 78]). In vision, gap detection thresholds are in a similar range (10 to 100 ms) and depend on, e.g., the pulse duration and the adaptation state (e.g., [42, 66]). Studies measuring temporal modulation functions (TMTFs; i.e., the modulation depth necessary for detecting flicker in stimuli with sinusoidal intensity modulation) in vision found cut-off frequencies of between 7 Hz and 25 Hz, again showing higher temporal resolution at higher background luminance (i.e., in a light-adapted state), with a peak of sensitivity at 8–10 Hz emerging at high background luminance [23], the latter suggesting a filter with both low- and high-frequency cutoff properties. In audition, the TMTFs in normal-hearing listeners show a lowpass-filter characteristic with cutoff frequencies 40–65 Hz, which indicates a lowpass-filter time constant of about 2.5 ms [4, 90]. Thus, the temporal resolution of the auditory system for detecting amplitude modulations is somewhat higher than for the visual system. However, according to the TMTFs reported in the literature, the 8.5-Hz rectangular amplitude modulation in the stimuli of Exp. 1 due to the random changes in intensity

every 117.65 ms was well detectable in both modalities. Taken together, this literature indicates that our assumption of presenting auditory and visual stimuli with the same physical temporal structure would result in brightness/loudness representations with equal temporal traces at the input of the decision stage (Fig. 1) was probably too strong, but at least not violated too severely. It is also important to note that the observation of temporal integration of brightness and loudness does not imply that the relevant mechanisms are located in early, sensory processing stages. At least in hearing, it appears unlikely that the temporal integration of loudness or the temporal integration at the threshold of audibility is due to mechanisms located in the auditory nerve or the brainstem (i.e., during the sensory processing stages in Fig. 1). Instead, these phenomena could be caused by more central mechanisms. For instance, the "multiple-looks hypothesis" [91] for the duration-dependence of detection and discrimination thresholds localizes the relevant mechanisms in a decision stage rather than in early sensory processing stages, and is thus similar in spirit to sequential evidence integration.

The present work illustrates the importance of a careful consideration of the different stages involved in the processing of a sensory quantity (see Fig. 1), as they might all contribute to the final set of estimated TWs. In some conditions, the perceptual weights measured by psychophysical reverse correlation as in the present study show striking similarity with physiological reverse correlation data obtained in single neurons for example in primary visual cortex, suggesting that the psychophysical weights can be linked quite directly to early sensory processes (for a review, see [48]). In other parts of the literature, sensory processes are not even considered to contribute to perceptual weights measured by psychophysical reverse correlation, and the patterns of TWs are quite directly attributed to evidence integration processes, and thus higher-level, decisional mechanisms (e.g., [34, 77]). In a sense, this simple partitioning is not unreasonable considering that in the former case, the tasks were designed specifically to tap into early sensory processes such as judging the orientation of single Gabor patch, while in the latter case, the task was to, e.g., judge the temporal average ("overall") of the orientation of a temporal sequence of spatial arrays of Gabor stimuli.

As discussed above, in the literature on loudness, the TWs were never exclusively attributed to early sensory processes, but were thought to likely involve attentional or decisional processes (e.g.,[58, 63]). For instance, as mentioned in the Introduction, a detailed consideration of the characteristics of auditory nerve (AN) responses shows that these might contribute to but are unlikely to be the sole cause of primacy effects observed in previous psychophysical studies on TWs for loudness. The firing rate of fibers in the auditory nerve (AN) shows a clear onset peak [39]. If loudness is assumed to be related to the spike count elicited by the sound [71], and the onset causes more neural activity than later temporal parts, this could explain a higher loudness weight on the sound onset. Because the inner hair cells that innervate the AN fibers are frequency specific, the recovery of the firing rate is also frequency specific (e.g., [28]). This is compatible with results demonstrating that temporal loudness weights are applied in a frequency-specific manner [20], and that when the spectrum changes abruptly within a contiguous sound, a second primacy effect is observed on the second sound part [63]. However, because neurons with high spontaneous rates (SR) show a fast recovery, so that the onset peak occurs after silent inter-stimulus intervals of only a few milliseconds [28], the observation that the primacy effect in loudness judgments shows full recovery only after silent gaps of about 350 ms or more [21] seems, at least at first sight, to be at odds with this explanation. Yet, low-SR neurons exhibit a considerably slower recovery of the onset peak [70, 73], thus one could assume that the primacy effect is primarily driven by these neurons. Another result that is not easily accounted for by the onset peak in AN fibers is that varying the mean sound level from just above detection threshold to higher levels, or presenting to-be-judged sound in a continuous background noise has almost no effect on the TWs [19], while the AN responses are strongly influenced by sound level and simultaneous masking. On a more general level, the neuronal auditory pathway is quite complex and involves different types of neurons as well as efferent and afferent loops [37]. Additional research based on predictions from computational auditory models is thus needed to evaluate more exactly the extent to which processes in the auditory periphery might contribute to the primacy effect in loudness weights. Also, computational models for the loudness of dynamic, time-varying sounds expressed at a 'functional' level (i.e., not attempting a physiologically plausible description of the auditory periphery processes) [14, 24, 46, 96] including multiple temporal integration stages do not predict the observed primacy effects in loudness judgments [19, 62].

The present study suggests that for overall judgments of both loudness and brightness of stimuli extending over a time range of about 1 second, sensory as well as attentional or decisional processes contribute to perceptual weights. The proposed involvement of all three types of processes is in line with recent work that illustrated this aspect by combining visual psychophysical experiments and computational modeling (e.g., [61]). For example, in a face categorization task, considering the discriminability of visual features in addition to an evidence integration process was necessary to account for the TWs measured by psychophysical reverse correlation [60].

One consequence of this insight is that when comparing TWs (or more generally, perceptual weights) *across* different modalities or tasks– as it is the case of the present work –, the influence of sensory-specific processes on TWs cannot be ignored, or more precisely, *should* always be considered. In the same line of reasoning, if one would fit an evidence integration model to the psychophysical data, the estimated model parameters (e.g., the drift rate in a diffusion model) will reflect the *combined* effect of sensory processes and decisional processes. We would thus like to comment on several methodological points that need to be carefully considered if one aims to compare perceptual processes across modalities. The typical approach in the

previous literature was to simply compare the average temporal weighting patterns measured across conditions (e.g., [10]). A first prerequisite that has not always been considered in prior studies but is nevertheless critical for a meaningful comparison of TWs between tasks at an individual level is to ensure that the *task difficulty* is matched individually. Second, the finding of similar *average* temporal weighting profiles across tasks is a necessary but not a sufficient condition to demonstrate that they are mediated by a single evidence integration process. Indeed, similarity *at the level of each participant* is critical to provide a direct support to this evidence integration account of temporal weighting; and even when similar weighting patterns are observed at the level of the group, this does not guarantee that this similarity is also observed at the level of each participant. Here, we went a step further by quantifying the similarity of patterns of TWs between two tasks within each participant, and comparing it to the individual within-task similarity between random subsets of the trials (i.e., split-half reliability).

A clear limitation of the present study is that the TWs (Experiment 1) and the sensitivity for intensity changes on different temporal components (Experiment 2) were not measured in the same participants. Within-subjects comparisons of the sensitivity patterns and temporal weighting patterns would have allowed an even stronger test of the evidence integration hypothesis, or more precisely, an estimation of how strongly the different patterns of TWs can be attributed to relatively early factors such as differences in sensitivity. On a more general level, we believe that when investigating the origin of TWs (or other perceptual weights), it is essential to include additional experiments measuring differences in sensitivity for the different stimulus components (e.g., temporal segments, frequency components etc.), and these additional data should preferably be collected within the same subjects. An ideal approach to investigate the combined effect of sensory (e.g., differences in sensitivity) and decisional processes (e.g., evidence integration) on TWs would be to combine computational models of these two different processing stages. For instance, in the case of loudness, one interesting direction would be to combine an auditory nerve model with an evidence integration model.

An alternative approach would be to modify the model of the evidence integration process to account for differences in sensitivity for the different stimulus components. While there is previous work on accounting for sensitivity differences between qualitatively distinct stimulus features in relation to perceptual weights, as for example different frequency channels in the case of loudness [54] or different facial features in the case of the categorization of emotional facial expressions [60], it remains an open question how to account for differences in sensitivity *across time* in the framework of evidence integration models. Finally, it would be interesting to investigate the temporal weighting in intensity judgments of combined auditory and visual stimuli.

To conclude, the present study showed that even for a task requiring an overall judgment of only a single sensory dimension varying across time (intensity), the patterns of TWs differ between sensory modalities (auditory versus visual), within participants. This observation is incompatible with the idea that the TWs in loudness and brightness are driven exclusively by a common, supramodal and subject-specific decision process. Instead, it is compatible with the view that the observed TWs are the result of a combination of sensory, attentional, and decisional processes. Analyzing individual differences both within tasks/modalities and between tasks/modalities and combining psychophysical reverse correlation measurements with data on sensory sensitivity appears to be a promising approach for gaining a better insight into the origin of TWs, or perceptual weights in general.

Acknowledgments

This work was supported by a grant from Deutsche Forschungsgemeinschaft (DFG; www.dfg.de) to Daniel Oberfeld (OB 346/6-1). E.P was supported by a grant from the Agence Nationale de la Recherche (ANR-22-CE28-0010). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We are grateful to Elena Gorbunov for her help with the data collection for Exp. 2.

Conflicts of interest

The authors declare that they have no conflicts of interest.

Data availability statement

The data from Exp. 1 and Exp. 2 are available in OSF, under the reference [52].

References

1. S.J. Aiken, J.N. Andrus, M. Bance, D.P. Phillips: Acoustic stapedius reflex function in man revisited, Ear and Hearing 34, 4 (2013) E38–E51. https://doi.org/10.1097/AUD.0b013e31827ad9d3.
2. M. Alpern, J.J. Faris: Luminance-duration relationship in the electric response of the human retina, Journal of the Optical Society of America 46, 10 (1956) 845–850. https://doi.org/10.1364/josa.46.000845.
3. M. Bach: The Freiburg Visual Acuity test: automatic measurement of visual acuity, Optometry and Vision Science 73, 1 (1996) 49–53. https://doi.org/10.1097/00006324-199601000-00008.
4. S.P. Bacon, N.F. Viemeister: Temporal modulation transfer-functions in normal-hearing and hearing-impaired listeners, Audiology 24, 2 (1985) 117–134.
5. B.G. Berg: Analysis of weights in multiple observation tasks, Journal of the Acoustical Society of America 86, 5 (1989) 1743–1746.
6. B.G. Berg, D.E. Robinson: Multiple observations and internal noise, Journal of the Acoustical Society of America 81 (1987) S33.
7. W.R. Biersdorf: Critical duration in visual brightness discrimination for retinal areas of various sizes, Journal of the Optical Society of America 45, 11 (1955) 920–925. https://doi.org/10.1364/josa.45.000920.
8. P. Binda, M. Pereverzeva, S.O. Murray: Attention to bright surfaces enhances the pupillary light reflex, Journal of Neuroscience 33, 5 (2013) 2199–2204. https://doi.org/10.1523/jneurosci.3440-12.2013.

9. G.S. Brindley: The discrimination of after-images, Journal of Physiology-London 147, 1 (1959) 194–203. https://doi.org/10.1113/jphysiol.1959.sp006234.
10. Z.Z. Bronfman, N. Brezis, M. Usher: Non-monotonic temporal-weighting indicates a dynamically modulated evidence-integration mechanism, PLoS Computational Biology 12, 2 (2016) e1004667. https://doi.org/10.1371/journal.pcbi. 1004667.
11. M. Brysbaert, M. Stevens: Power analysis and effect size in mixed effects models: a tutorial, Journal of Cognition 1, 1 (2018) 9. https://doi.org/10.5334/joc.10.
12. J.R. Busemeyer, J.T. Townsend: Decision field theory: a dynamic cognitive approach to decision-making in an uncertain environment, Psychological Review 100, 3 (1993) 432–459. https://doi.org/10.1037//0033-295x.100.3.432.
13. F.W. Campbell, A.H. Gregory: Effect of size of pupil on visual acuity, Nature 187, 4743 (1960) 1121–1123. https://doi.org/10.1038/1871121c0.
14. J. Chalupper, H. Fastl: Dynamic loudness model (DLM) for normal and hearing-impaired listeners, Acta Acustica United with Acustica 88, 3 (2002) 378–386.
15. S. Cheadle, V. Wyart, K. Tsetsos, N. Myers, V. de Gardelle, S.H. Castanon, C. Summerfield: Adaptive gain control during human perceptual choice, Neuron 81, 6 (2014) 1429–1441. https://doi.org/10.1016/j.neuron.2014.01.020.
16. G.T. Church, E.A. Cudahy: The time course of the acoustic reflex, Ear and Hearing 5, 4 (1984) 235–242. https://doi.org/10.1097/00003446-198407000-00008.
17. K. Dittrich, D. Oberfeld: A comparison of the temporal weighting of annoyance and loudness, Journal of the Acoustical Society of America 126, 6 (2009) 3168–3178. https://doi.org/10.1121/1.3238233.
18. W. Ellermeier, S. Schrödl: Temporal weights in loudness summation. In: C. Bonnet (Ed.), Fechner Day 2000. Proceedings of the 16th annual meeting of the international society for psychophysics. Strasbourg: Université Louis Pasteur, 2000, pp. 169–173.
19. A. Fischenich, J. Hots, J. Verhey, D. Oberfeld: Temporal weights in loudness: investigation of the effects of background noise and sound level, PLoS One 14, 11 (2019) e0223075. https://doi.org/10.1371/journal.pone.0223075.
20. A. Fischenich, J. Hots, J. Verhey, D. Oberfeld: Temporal loudness weights are frequency specific, Frontiers in Psychology 12 (2021) 588571. https://doi.org/10.3389/fpsyg.2021.588571.
21. A. Fischenich, J. Hots, J.L. Verhey, D. Oberfeld: The effect of silent gaps on temporal weights in loudness judgments, Hearing Research 395 (2020) 108028. https://doi.org/10.1016/j.heares.2020.108028.
22. M. Florentine, S. Buus, T. Poulsen: Temporal integration of loudness as a function of level, Journal of the Acoustical Society of America 99, 3 (1996) 1633–1644.
23. L. Ganz: Temporal factors in visual perception. In: E.C. Carterette, M.P. Friedman (Eds.), Handbook of perception (Vol. 5: Seeing). New York, San Francisco, London: Academic Press, 1975, pp. 169–231.
24. B.R. Glasberg, B.C.J. Moore: A model of loudness applicable to time-varying sounds, Journal of the Audio Engineering Society 50, 5 (2002) 331–342.
25. D.M. Green: Detection of multiple component signals in noise, Journal of the Acoustical Society of America 30, 10 (1958) 904–911.
26. J.H. Grose, D.A. Eddins, J.W. Hall: Gap detection as a function of stimulus bandwidth with fixed high-frequency cutoff in normal-hearing and hearing-impaired listeners, Journal of the Acoustical Society of America 86, 5 (1989) 1747–1755. https://doi.org/10.1121/1.398606.
27. S. Haegens, J. Vergara, R. Rossi-Pool, L. Lemus, R. Romo: Beta oscillations reflect supramodal information during perceptual judgment. Proceedings of the National Academy of Sciences of the United States of America 114, 52 (2017) 13810–13815. https://doi.org/10.1073/pnas.1714633115.
28. D.M. Harris, P. Dallos: Forward masking of auditory-nerve fiber responses, Journal of Neurophysiology 42, 4 (1979) 1083–1107.
29. M.J. Hautus: Corrections for extreme proportions and their biasing effects on estimated values of d', Behavior Research Methods Instruments & Computers 27, 1 (1995) 46–51.
30. R.P. Hellman: Loudness, annoyance, and noisiness produced by single-tone-noise complexes, Journal of the Acoustical Society of America 72, 1 (1982) 62–73.
31. C. Hendrick, A.F. Costantini: Number averaging behavior: a primacy effect, Psychonomic Science 19, 2 (1970) 121–122. https://doi.org/10.3758/bf03337452.
32. D.C. Hood, B.G. Grover: Temporal summation of light by a vertebrate visual receptor, Science 184, 4140 (1974) 1003–1005. https://doi.org/10.1126/science.184.4140.1003.
33. J. Hots, J. Rennies, J.L. Verhey: Modeling temporal integration of loudness, Acta Acustica United with Acustica 100, 1 (2014) 184–187. https://doi.org/10.3813/aaa.918697.
34. B. Hubert-Wallander, G.M. Boynton: Not all summary statistics are made equal: evidence from extracting summaries across time, Journal of Vision 15, 4 (2015) 5. https://doi.org/10.1167/15.4.5.
35. H. Huynh, L.S. Feldt: Estimation of the box correction for degrees of freedom from sample data in randomized block and split-plot designs, Journal of Educational Statistics 1, 1 (1976) 69–82. https://doi.org/10.2307/1164736.
36. IEC 60318–1:1998: Electroacoustics – simulators of human head and ear. Part 1: Ear simulator for the measurement of supra-aural and circumaural earphones. Geneva: International Electrotechnical Commission, 1998.
37. S.G. Jennings, M.G. Heinz, E.A. Strickland: Evaluating adaptation and olivocochlear efferent feedback as potential explanations of psychophysical overshoot, JARO-Journal of the Association for Research in Otolaryngology 12, 3 (2011) 345–360. https://doi.org/10.1007/s10162-011-0256-5.
38. W. Jesteadt, L. Leibold: Loudness in the laboratory, Part I: Steady-state sounds. In: M. Florentine, A.N. Popper, R.R. Fay (Eds.), Loudness, Springer, New York, NY, 2011, pp. 109–144. https://doi.org/10.1007/978-1-4419-6712-1_1.
39. N.Y.S. Kiang, T. Watanabe, E.C. Thomas, L.F. Clark: Discharge patterns of single fibers in the cat's auditory nerve, M.I.T. Press, Cambridge, MA, 1965.
40. R. Kiani, T.D. Hanks, M.N. Shadlen: Bounded integration in parietal cortex underlies decisions even when viewing duration is dictated by the environment, Journal of Neuroscience 28, 12 (2008) 3017–3029. https://doi.org/10.1523/jneurosci.4761-07.2008.
41. M. Kubovy, A.F. Healy: The decision rule in probabilistic categorization: what it is and how it is learned? Journal of Experimental Psychology: General 106, 4 (1977) 427–446. https://doi.org/10.1037//0096-3445.106.4.427.
42. M.F. Lewis: Two-flash thresholds as a function of luminance in the dark-adapted eye, Journal of the Optical Society of America 57, 6 (1967) 814–815. https://doi.org/10.1364/josa.57.000814.

43. Z.L. Lu, B.A. Dosher: Characterizing observers using external noise and observer models: assessing internal representations with external noise, Psychological Review 115, 1 (2008) 44–82.
44. R.A. Lutfi: Informational processing of complex sound. I: Intensity discrimination, Journal of the Acoustical Society of America 86, 3 (1989) 934–944. https://doi.org/10.1121/1.398728.
45. R.J. Mansfield: Brightness function: effect of area and duration, Journal of the Optical Society of America 63, 8 (1973) 913–920. https://doi.org/10.1364/josa.63.000913.
46. B.C.J. Moore, M. Jervis, L. Harries, J. Schlittenlacher: Testing and refining a loudness model for time-varying sounds incorporating binaural inhibition, Journal of the Acoustical Society of America 143, 3 (2018) 1504–1513. https://doi.org/10.1121/1.5027246.
47. B.C.J. Moore, D.H. Raab: Pure-tone intensity discrimination: some experiments relating to the "near-miss" to Weber's law, Journal of the Acoustical Society of America 55, 5 (1974) 1049–1054.
48. P. Neri, D.M. Levi: Receptive versus perceptive fields from the reverse-correlation viewpoint, Vision Research 46, 16 (2006) 2465–2474. https://doi.org/10.1016/j.visres.2006.02.002.
49. R.G. O'Connell, P.M. Dockree, S.P. Kelly: A supramodal accumulation-to-bound signal that determines perceptual decisions in humans, Nature Neuroscience 15, 12 (2012) 1729–1735. https://doi.org/10.1038/nn.3248.
50. D. Oberfeld: The mid-difference hump in forward-masked intensity discrimination, Journal of the Acoustical Society of America 123, 3 (2008) 1571–1581. https://doi.org/10.1121/1.2837284.
51. D. Oberfeld: Are temporal loudness weights under top-down control? Effects of trial-by-trial feedback, Acta Acustica United with Acustica 101, 6 (2015) 1105–1115. https://doi.org/10.3813/aaa.918904.
52. D. Oberfeld, A. Fischenich, E. Ponsot: Dataset: trial-by-trial data from two psychophysical experiments on temporal weights in loudness and brightness judgments, 2023. https://doi.org/10.17605/OSF.IO/X5KD9.
53. D. Oberfeld, A. Fischenich, E. Ponsot, J. Verhey, J. Hots: What causes the primacy effect in temporal loudness weights? In: E. Parizet (Ed.), Proceedings of the eForum acusticum. Lyon, 2020, pp. 3411–3415. https://doi.org/10.48465/fa.2020.0835.
54. D. Oberfeld, W. Heeren, J. Rennies, J. Verhey: Spectro-temporal weighting of loudness, PLoS One 7, 11 (2012) e50184. https://doi.org/10.1371/journal.pone.0050184.
55. D. Oberfeld, J. Hots, J.L. Verhey: Temporal weights in the perception of sound intensity: effects of sound duration and number of temporal segments, Journal of the Acoustical Society of America 143, 2 (2018) 943–953. https://doi.org/10.1121/1.5023686.
56. D. Oberfeld, L. Jung, J.L. Verhey, J. Hots: Evaluation of a model of temporal weights in loudness judgments, Journal of the Acoustical Society of America 144, 2 (2018) EL119–EL124. https://doi.org/10.1121/1.5049895.
57. D. Oberfeld, M. Kuta, W. Jesteadt: Factors limiting performance in a multitone intensity-discrimination task: disentangling non-optimal decision weights and increased internal noise, PLoS One 8, 11 (2013) e79830. https://doi.org/10.1371/journal.pone.0097209.
58. D. Oberfeld, T. Plank: The temporal weighting of loudness: effects of the level profile, Attention, Perception, & Psychophysics 73, 1 (2011) 189–208. https://doi.org/10.3758/s13414-010-0011-8.
59. D. Oberfeld, P. Stahn, M. Kuta: Why do forward maskers affect auditory intensity discrimination? Evidence from "molecular psychophysics", PLoS One 9, 6 (2014) e99745. https://doi.org/10.1371/journal.pone.0099745.
60. G. Okazawa, L. Sha, R. Kiani: Linear integration of sensory evidence over space and time underlies face categorization, Journal of Neuroscience 41, 37 (2021) 7876–7893. https://doi.org/10.1523/jneurosci.3055-20.2021.
61. G. Okazawa, L. Sha, B.A. Purcell, R. Kiani: Psychophysical reverse correlation reflects both sensory and decision-making processes, Nature Communications 9 (2018) 3479. https://doi.org/10.1038/s41467-018-05797-y.
62. B. Pedersen: *Auditory temporal resolution and integration. Stages of analyzing time-varying sounds.* Ph.D. thesis, Aalborg University, Aalborg, DK, 2006.
63. B. Pedersen, W. Ellermeier: Temporal weights in the level discrimination of time-varying sounds, Journal of the Acoustical Society of America 123, 2 (2008) 963–972. https://doi.org/10.1121/1.2822883.
64. E. Ponsot, P. Susini, G. Saint Pierre, S. Meunier: Temporal loudness weights for sounds with increasing and decreasing intensity profiles, Journal of the Acoustical Society of America 134, 4 (2013) EL321–EL326. https://doi.org/10.1121/1.4819184.
65. G. Prat-Ortega, K. Wimmer, A. Roxin, J. de la Rocha: Flexible categorization in perceptual decision making, Nature Communications 12, 1 (2021) 1283. https://doi.org/10.1038/s41467-021-21501-z.
66. D.G. Purcell, A.L. Stewart: The two-flash threshold: an evaluation of critical-duration and visual-persistence hypotheses, Perception & Psychophysics 9, 1A (1971) 61–64. https://doi.org/10.3758/bf03213029.
67. D.H. Raab, I.A. Goldberg: Auditory intensity discrimination with bursts of reproducible noise, Journal of the Acoustical Society of America 57, 2 (1975) 437–447. https://doi.org/10.1121/1.380467.
68. R. Ratcliff: A theory of memory retrieval, Psychological Review 85, 2 (1978) 59–108. https://doi.org/10.1037/0033-295x.85.2.59.
69. R. Ratcliff, P.L. Smith, S.D. Brown, G. McKoon: Diffusion decision model: current issues and history, Trends in Cognitive Sciences 20, 4 (2016) 260–281. https://doi.org/10.1016/j.tics.2016.01.007.
70. E.M. Relkin, J.R. Doucet: Recovery from prior stimulation. I: Relationship to spontaneous firing rates of primary auditory neurons, Hearing Research 55, 2 (1991) 215–222.
71. E.M. Relkin, J.R. Doucet: Is loudness simply proportional to the auditory nerve spike count? Journal of the Acoustical Society of America 101, 5 Pt 1 (1997) 2735–2740.
72. J. Rennies, J.L. Verhey: Temporal weighting in loudness of broadband and narrowband signals, Journal of the Acoustical Society of America 126, 3 (2009) 951–954. https://doi.org/10.1121/1.3192348.
73. W.S. Rhode, P.H. Smith: Characteristics of tone-pip response patterns in relationship to spontaneous rate in cat auditory nerve fibers, Hearing Research 18, 2 (1985) 159–168. https://doi.org/10.1016/0378-5955(85)90008-5.
74. J.E. Rose, J.F. Brugge, D.J. Anderson, J.E. Hind: Phase-locked response to low-frequency tones in single auditory nerve fibers of the squirrel monkey, Journal of Neurophysiology 30, 4 (1967) 769–793. https://doi.org/10.1152/jn.1967.30.4.769.
75. W.A. Rosenblith, G.A. Miller, J.P. Egan, I.J. Hirsh, G.J. Thomas: An auditory afterimage, Science 106, 2754 (1947) 333–335. https://doi.org/10.1126/science.106.2754.333.

76. J.A.J. Roufs: Dynamic properties of vision–I. Experimental relationships between flicker and flash thresholds, Vision Research 12, 2 (1972) 261–278. https://doi.org/10.1016/0042-6989(72)90117-4.
77. H. Sato, I. Motoyoshi: Distinct strategies for estimating the temporal average of numerical and perceptual information, Vision Research 174 (2020) 41–49. https://doi.org/10.1016/j.visres.2020.05.004.
78. M.J. Shailer, B.C.J. Moore: Gap detection as a function of frequency, bandwidth, and level, Journal of the Acoustical Society of America 74, 2 (1983) 467–473. https://doi.org/10.1121/1.389812.
79. P.L. Smith, D.R. Little: Small is beautiful: in defense of the small-N design, Psychonomic Bulletin & Review 25, 6 (2018) 2083–2101. https://doi.org/10.3758/s13423-018-1451-8.
80. R.D. Sorkin, D.E. Robinson, B.G. Berg: A detection theory method for the analysis of auditory and visual displays. In: Proceedings of the 31st annual meeting of the human factors society, 1987, pp. 1184–1188.
81. B. Spitzer, F. Blankenburg: Supramodal parametric working memory processing in humans, Journal of Neuroscience 32, 10 (2012) 3287–3295. https://doi.org/10.1523/jneurosci.5280-11.2012.
82. G.C. Stecker, E.R. Hafter: Temporal weighting in sound localization, Journal of the Acoustical Society of America 112, 3 (2002) 1046–1057. https://doi.org/10.1121/1.1497366.
83. J.C. Stevens, J.W. Hall: Brightness and loudness as functions of stimulus duration, Perception & Psychophysics 1, 9 (1966) 319–327. https://doi.org/10.3758/bf03215796.
84. M. Stone: Models for choice-reaction time, Psychometrika 25, 3 (1960) 251–260. https://doi.org/10.1007/bf02289729.
85. J.A. Swets, E.F. Shipley, M.J. McKey, D.M. Green: Multiple observations of signals in noise, Journal of the Acoustical Society of America 31, 4 (1959) 514–521.
86. K. Tsetsos, J. Gao, J.L. McClelland, M. Usher: Using time-varying evidence to test models of decision dynamics: bounded diffusion vs. the leaky competing accumulator model, Frontiers in Neuroscience 6 (2012) 79. https://doi.org/10.3389/fnins.2012.00079.
87. J. Vergara, N. Rivera, R. Rossi-Pool, R. Romo: A neural parametric code for storing information of more than one sensory modality in working memory, Neuron 89, 1 (2016) 54–62. https://doi.org/10.1016/j.neuron.2015.11.026.
88. J.L. Verhey: Temporal resolution and temporal integration. In: C.J. Plack (Ed.), The Oxford handbook of auditory science: hearing, Vol. 3, Oxford University Press, Oxford, 2010, pp. 105–122. https://doi.org/10.1093/oxfordhb/9780199233557.013.0005.
89. D. Vickers: Evidence for an accumulator model of psychophysical discrimination, Ergonomics 13, 1 (1970) 37–58. https://doi.org/10.1080/00140137008931117.
90. N.F. Viemeister: Temporal modulation transfer functions based upon modulation thresholds, Journal of the Acoustical Society of America 66, 5 (1979) 1364–1380.
91. N.F. Viemeister, G.H. Wakefield: Temporal integration and multiple looks, Journal of the Acoustical Society of America 90, 2 (1991) 858–865.
92. T.Y. Yeh, B.B. Lee, J. Kremers: The time course of adaptation in macaque retinal ganglion cells, Vision Research 36, 7 (1996) 913–931. https://doi.org/10.1016/0042-6989(95)00332-0.
93. F.G. Zeng, C.W. Turner: Intensity discrimination in forward masking, Journal of the Acoustical Society of America 92, 2 (1992) 782–787.
94. F.G. Zeng, C.W. Turner, E.M. Relkin: Recovery from prior stimulation II: effects upon intensity discrimination, Hearing Research 55, 2 (1991) 223–230.
95. E. Zwicker: Negative afterimage in hearing, Journal of the Acoustical Society of America 36, 12 (1964) 2413–2415. https://doi.org/10.1121/1.1919373.
96. E. Zwicker: Procedure for calculating loudness of temporally variable sounds, Journal of the Acoustical Society of America 62, 3 (1977) 675–682. https://doi.org/10.1121/1.381580.
97. J.J. Zwislocki: Temporal summation of loudness: an analysis, Journal of the Acoustical Society of America 46, 2 (1969) 431–441. https://doi.org/10.1121/1.1911708.

Appendix

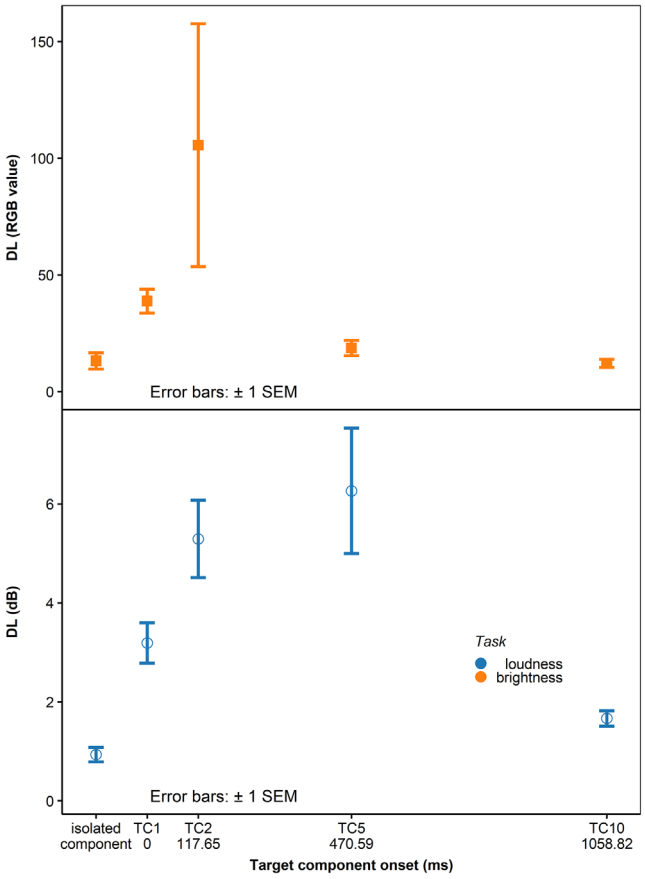

Figure A1. Exp. 2. Mean intensity difference limen (DL; defined as half the difference between the 75%- and the 25%-point on the psychometric function) as a function of target component for the two tasks. Top panel: brightness task, bottom panel: loudness task. Error bars show ± 1 SEM across the 8 participants.

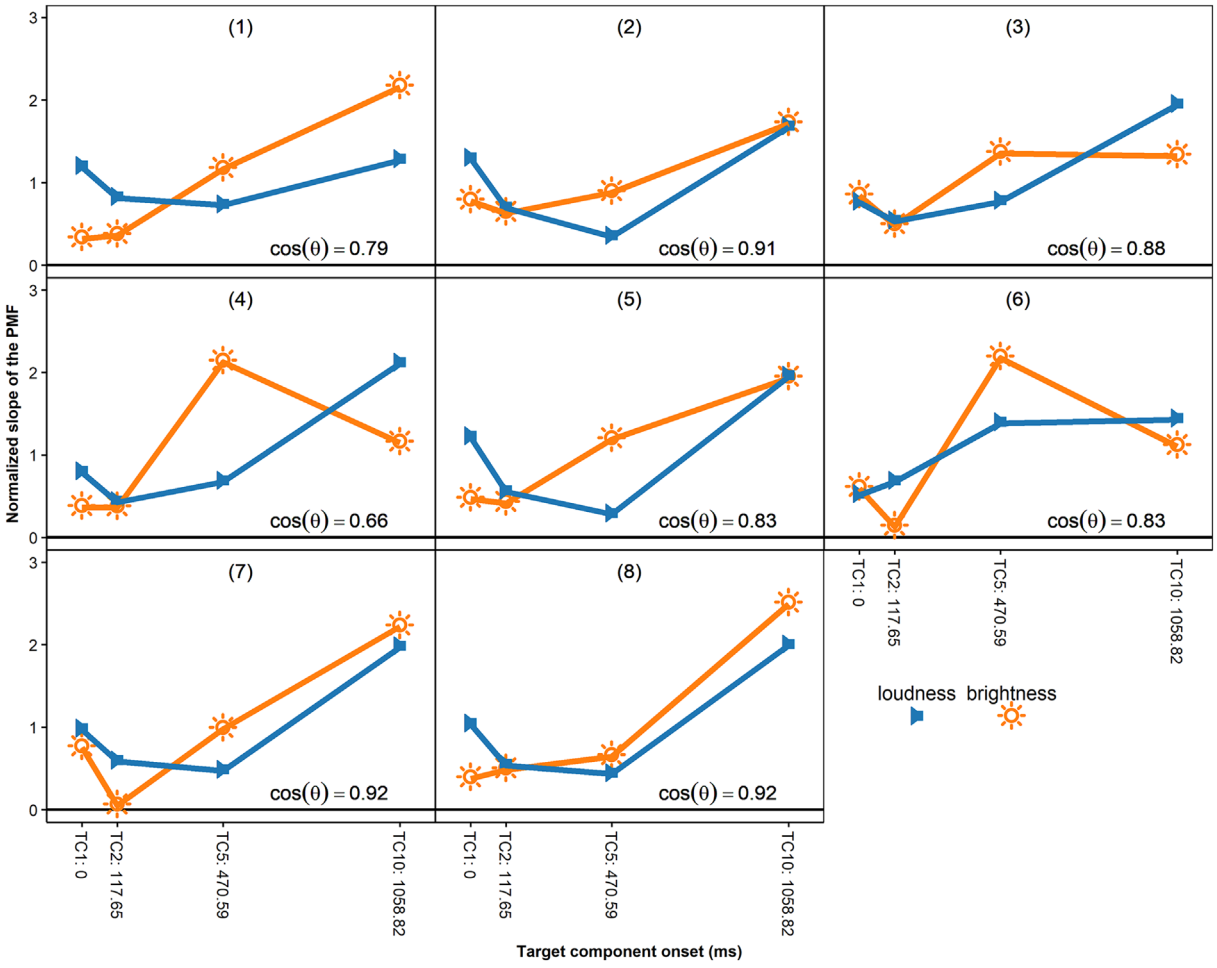

Figure A2. Exp. 2. Individual normalized sensitivity (PMF slope) as a function of target component for the two tasks (note, sensitivity for the isolated components not included). Task is indicated by color and symbol (blue loudspeakers: loudness, orange suns: brightness). Participant number is indicated in the panels. The cos(θ) values are the cosine similarity between the two patterns of sensitivities within a given participant (WSBT; see text).

Cite this article as: Oberfeld D. Fischenich A. & Ponsot E. 2024. Non-uniform temporal weighting of intensity in audition and vision: The signature of an evidence integration process?. Acta Acustica, **8**, 57. https://doi.org/10.1051/aacus/2024061.