Memory & Cognition (2023) 51:349–370
https://doi.org/10.3758/s13421-022-01355-6

Long-term memory representations for audio-visual scenes

Hauke S. Meyerhoff[1,2] · Oliver Jaggy[2] · Frank Papenmeier[3] · Markus Huff[2,3]

Accepted: 26 August 2022 / Published online: 13 September 2022

Abstract

In this study, we investigated the nature of long-term memory representations for naturalistic audio-visual scenes. Whereas previous research has shown that audio-visual scenes are recognized more accurately than their unimodal counterparts, it remains unclear whether this benefit stems from audio-visually integrated long-term memory representations or a summation of independent retrieval cues. We tested two predictions for audio-visually integrated memory representations. First, we used a modeling approach to test whether recognition performance for audio-visual scenes is more accurate than would be expected from independent retrieval cues. This analysis shows that audio-visual integration is not necessary to explain the benefit of audio-visual scenes relative to purely auditory or purely visual scenes. Second, we report a series of experiments investigating the occurrence of study-test congruency effects for unimodal and audio-visual scenes. Most importantly, visually encoded information was immune to additional auditory information presented during testing, whereas auditory encoded information was susceptible to additional visual information presented during testing. This renders a true integration of visual and auditory information in long-term memory representations unlikely. In sum, our results instead provide evidence for visual dominance in long-term memory. Whereas associative auditory information is capable of enhancing memory performance, the long-term memory representations appear to be primarily visual.

Keywords Long-term memory · Audio-visual integration · Study-test congruency · Audio-visual advantage · Naturalistic scenes

Introduction

At any waking moment, an endless stream of multimodal sensory information shapes our mental representation of the outside world. Over the past two decades, more and more research has appreciated the multimodal nature of human perception, as well as crossmodal interactions between different sensory streams. Whereas visual information processing was originally supposed to be impenetrable by non-visual information such as audition (e.g., Rock & Victor, 1964; Warren et al., 1981), more recent research has revealed strong and persistent interactions between auditory and visual information during perception (for reviews, see Ernst & Bülthoff, 2004; Koelewijn et al., 2010; Spence, 2011) as well as short memory durations (for reviews, see Matusz et al., 2017; Shams et al., 2011). However, research that has investigated how the interplay of multiple modalities forms subsequent long-term memory representations in humans is rare (see Gibson & Maunsell, 1997, for evidence from non-human primates). In the following, we report a modeling approach as well as behavioral experiments that investigate the nature of human long-term memory representations for brief naturalistic scenes. More specifically, we ask whether these memory representations consist of integrated auditory and visual information or whether auditory and visual information contribute to long-term memory independently of each other. As evidence in favor of audio-visual integration, we would consider long-term memory performance that cannot be explained by independent contributions of auditory and visual information.

✉ Hauke S. Meyerhoff
hauke.meyerhoff@uni-erfurt.de

[1] University of Erfurt, Nordhäuser Str. 63, 99089 Erfurt, Germany

[2] Leibniz-Institut für Wissensmedien, Tübingen, Germany

[3] Department of Psychology, University of Tübingen, Tübingen, Germany

Audio-visual integration during perception

There is a continuous interplay between perceptual processes and long-term memory. Whereas perception is obviously necessary to acquire new long-term memory representations, previously established long-term memory representations influence ongoing perceptual and attentional processes (e.g., Biederman et al., 1982; Võ & Wolfe, 2013).

Regarding perceptual processes, auditory information is capable of altering the quantity as well as the quality of a visual percept. For instance, when two brief sounds coincide with one visual flash, observers tend to perceive two visual flashes (Shams et al., 2000), and when a brief sound coincides with the moment of overlap between two moving discs, observers tend to misperceive the spatial relations between the moving discs (Meyerhoff & Scholl, 2018), resulting in the impression of two discs bouncing off rather than streaming past each other (Sekuler et al., 1997). Importantly, this crossmodal influence is not unidirectional, but both sensory streams are integrated into one joint percept (e.g., Alais & Burr, 2004; McGurk & MacDonald, 1976). In fact, there is neuroanatomical (Falchier et al., 2002) as well as electrophysiological evidence (e.g., Giard & Peronnet, 1999; van der Burg et al., 2011) that the integration of auditory and visual signals starts at the earliest stages within the cortex. Crucially, however, the effectiveness of this integration process strongly depends on the synchrony of both signals with a tolerance of only ± 200 ms (Lewald et al., 2001; Meyerhoff & Suzuki, 2018; Powers et al., 2009; Stevenson et al., 2012; van Wassenhove et al., 2007).

It seems obvious that perceptual illusions also affect subsequent memory representations; however, audio-visual interactions might also influence subsequent memory representations less obviously by increasing the efficiency of sensory processing in one of the modalities. Indeed, there is a substantial body of research demonstrating such early interactions between semantically meaningful auditory information and visual perception (see also Taylor et al., 2009, for neuropsychological evidence). With auditory information preceding the visual stimuli by a few hundred milliseconds, congruent naturalistic sounds (Chen & Spence, 2011a, b) as well as spoken words (Chen & Spence, 2018; Edmiston & Lupyan, 2015; Lupyan & Thompson-Schill, 2012; Lupyan & Ward, 2013) facilitate the detection of the corresponding visual objects. However, even when auditory and visual information is presented in temporal alignment, sounds still enhance visual processing relative to unimodal or semantically mismatching audio-visual presentations. For instance, semantically matching sounds accelerate the fixation (Iordanescu et al., 2010) and detection (Iordanescu et al., 2008) of visual objects, as well as facilitate their identification (Amedi et al., 2005; Chen & Spence, 2011b; Mädebach et al., 2017). Remarkably, this holds true even when conscious processing of the visual scene is prevented (Tan & Yeh, 2015).

Furthermore, audio-visually synchronous events attract attention (Meyerhoff et al., 2022; Santangelo & Spence, 2007; van der Burg et al., 2008), thus further improving their processing. This enhanced processing also impacts subsequent processes. For instance, perceptual learning (i.e., performance improvements with practice in basic perceptual tasks) is more pronounced following multimodal presentations than unimodal presentations alone (Kim et al., 2008; Seitz et al., 2006). Likewise, it therefore seems plausible that audio-visually integrated information also shapes short-lived memory representations as well as long-term memory representations.

Audio-visual impact on working memory

Similar to perceptual processes, the ability to store information in working memory appears to benefit from multimodal stimulus presentations. For instance, Frick (1984) demonstrated an increased capacity limitation for digits when the information was distributed between the visual and auditory modality. Importantly, observers also recalled more information from working memory when auditory and visual information had been presented simultaneously rather than unimodally in isolation (Delogu et al., 2009; for similar results, see also Goolkasian & Foos, 2002; Lewandowski & Kobus, 1993). However, the critical question, namely, whether the benefits of multimodal information in working memory stem from audio-visual integration (Saults & Cowan, 2007) or independent storages for auditory and visual information (Baddeley & Logie, 1999; Fougnie & Marois, 2011; for a review, see Quack et al., 2015), has not yet been resolved. Indeed, there is also experimental evidence that limits the general validity of the advantage of audio-visual stimuli in memory. Depending on the moment of retrieval, it has also been shown that either visual (Ngo et al., 2010) or auditory information alone (Ngo et al., 2010) might dominate the remaining modality.

In a related line of research, Murray and colleagues (for a review, see Matusz et al., 2017) have been studying the impact of audio-visual presentations on unimodal retrieval using the continuous recognition paradigm. In this paradigm, the participants attended to a stream of briefly presented stimuli, indicating whether the currently presented stimulus had been presented among the previously attended stimuli. Whereas the repeated items were unimodal, the initial presentation of the stimuli could have been either multimodal or unimodal as well. Semantically congruent audio-visual information during the initial presentation had a beneficial effect on the recognition of visual (e.g., Lehmann & Murray, 2005; Thelen et al., 2015) as well as auditory information (e.g., Matusz et al., 2015; Thelen et al., 2015). Using electrophysiology (Murray et al., 2004) as well as neuroimaging techniques (Murray et al., 2005), it was demonstrated that visual object identification was more active during unimodal recall for stimuli with rather than without a multimodal history (see also von Kriegstein & Giraud, 2006). As there were no general differences in performance for unimodal and audio-visual stimuli at the initial presentation (e.g., Matusz et al., 2015), the authors argued that the improved recognition performance does not stem from general enhancements of attentional or encoding-related processes. This is further supported by the observation that the presence of meaningless tones during the initial presentation could have detrimental effects on subsequent unimodal recognition (e.g., Lehmann & Murray, 2005; Thelen et al., 2012). Please note, however, that a recent re-investigation has revealed little evidence for the generality of a beneficial impact of audio-visual encoding on visual recognition performance (Pecher & Zeelenberg, 2022).

Auditory and visual information in long-term memory

With regard to long-term memory (i.e., long retention intervals with clearly separated encoding and recall/recognition), auditory and visual material has been studied in isolation rather than in combined presentation formats. This research has revealed a tremendous capacity for visual memory representations. In fact, no capacity limitation has been reported yet. For instance, Standing (1973) observed a linearly increasing number of successfully recognized images up to 10,000 pictures, which was the maximum of presented items in his study. This tremendous capacity is not restricted to pictures but also arises with real-world objects at a remarkable level of detail (Brady et al., 2008; Brady et al., 2013) as well as with the location of objects within entire scenes (Hollingworth, 2004, 2005; Konkle et al., 2010). Further, visual long-term memory representations not only consist of static object information, but also contain dynamic information such as changes over time in a dynamic scene. Using brief excerpts from movies, Matthews et al. (2007); see also Goldstein et al., 1982) demonstrated that observers are more accurate in discriminating previously studied dynamic clips from novel clips than their static counterparts (or a series of static snapshots). This dynamic superiority effect still appeared with retention intervals of four weeks. Importantly, however, recognition performance was best when the clips were tested in the same dynamic state as they had been presented during the study session (Buratto et al., 2009; Matthews et al., 2010). In other words, dynamic scenes were recognized more accurately when they were tested dynamically rather than statically, and their static counterparts were recognized more accurately when they were tested statically rather than dynamically (i.e., study-test congruency). This finding indicates that the dynamic information is in fact part of the memory representation as presenting this information additionally during testing hurts the recognition performance for statically studied scenes.

Compared to visual memory, memory for auditory stimuli appears to be inferior (Cohen et al., 2009; see also Kassim et al., 2018); however, recent work from our group has demonstrated that information from both modalities interacts during the formation of long-term memory representations (Meyerhoff & Huff, 2016). In this study, the participants studied brief auditory, visual, or audio-visual tracks from movies. Recognition performance following retention intervals of 1 day or 1 week was in fact more accurate for audio-visual tracks than for their unimodal counterparts. However, this does not necessarily indicate audio-visual integration as audio-visual tracks also contain more retrieval cues that could improve memory performance independently of each other. One finding supporting the hypothesis of audio-visually integrated long-term memory representations was that semantically matching stimuli were recognized more accurately than mismatching stimuli (i.e., visual and auditory information from different movie clips). Nevertheless, even these semantically mismatching audio-visual tracks were recognized more accurately than purely visual tracks (i.e., unimodal). This indicates that visual and auditory information are both also capable of improving memory performance independently of each other. Finally, the role of audio-visual synchrony deviated remarkably from what would be predicted from the studies on audio-visual integration during perception (see above). Whereas audio-visual synchrony is a key variable during perception, it was strikingly irrelevant in this study. Even when matching auditory and visual tracks were presented sequentially, recognition performance remained at the level of tracks presented in audio-visual synchrony. Only when multiple tracks were presented in between the matching auditory and visual tracks, did recognition performance finally decline during audio-visual testing relative to the tracks that had been studied in audio-visual synchrony. In sum, this study demonstrates that auditory information can improve memory for visual scenes (i.e., the effect of semantic congruency); however, it remains inconclusive whether this advantage actually stems from an integration of auditory and visual information in long-term memory representations.

Interplay of auditory and visual information in long-term memory representations

Across different research traditions and paradigms addressing the interplay of distinct sources of information in

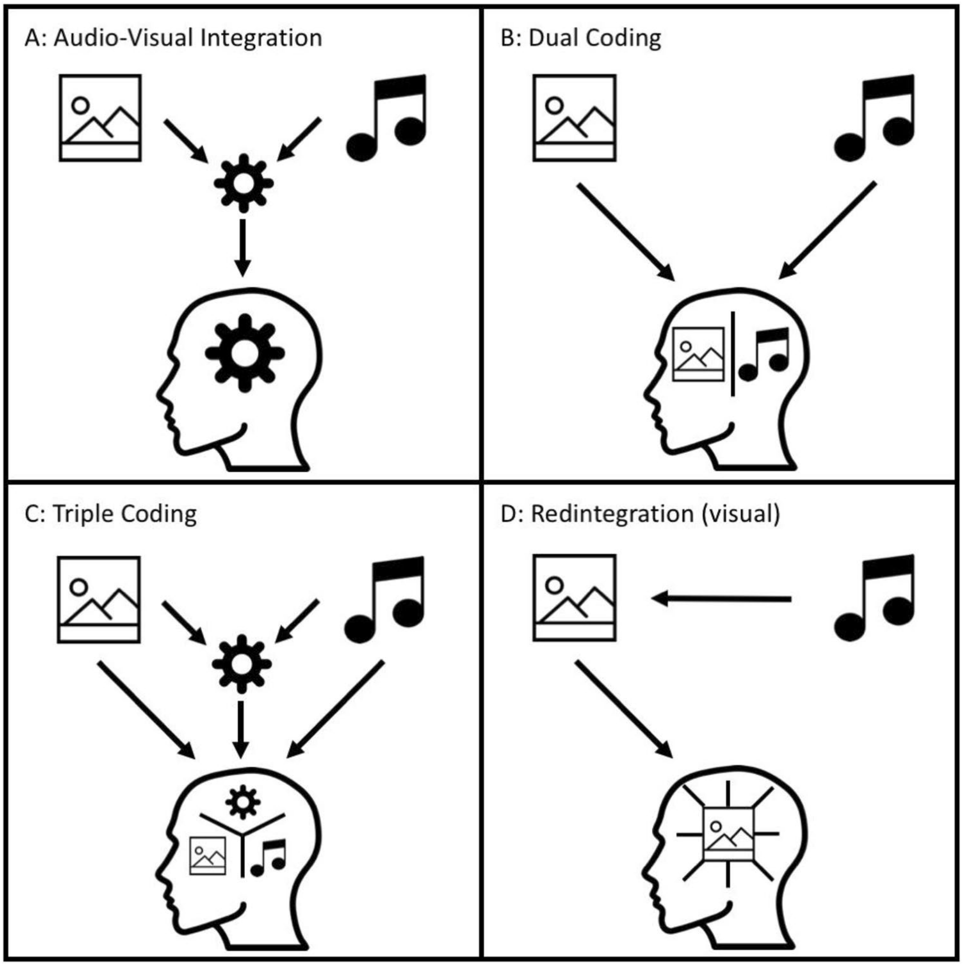

Fig. 1 Depiction of the central theoretical accounts addressing the interplay of auditory and visual information within memory representations. **A:** Audio-visual integration assumes that both modalities are merged into a unitary, amodal representation. **B:** Dual coding assumes that both modalities are encoded and stored separately. **C:** Triple coding assumes memory representations consist of auditory, visual, and integrated, amodal scene components. **D:** Redintegration assumes that audio-visual interactions enhance memory representations, which remain accessible during unimodal testing (depicted for visual memory). Icons originally designed by Smashicons from Flaticon

memory representations, several theories have been suggested in order to describe the structure of memory representations. Whereas many of the older studies have asked how verbal and pictorial information interact in memory, more recent work has more directly investigated the interplay of auditory and visual information (for a schematic overview, see Fig. 1).

Regarding the interplay of verbal and pictorial information, some accounts have proposed that memory representations are abstract and amodal in nature (i.e., modality-independent; Anderson, 1978, 1979; Kieras, 1978). However, Paivio and colleagues (Paivio & Csapo, 1973; Thompson & Paivio, 1994) observed that memory performance for audio-visual material in a recall task did not differ from what would be expected based on independent memory representations for the auditory and visual information. According to their interpretation and the corresponding *dual-coding theory* (i.e., modality-specific), the assumption of audio-visual integration in any form is therefore not necessary to explain memory performance. Since the empirical evidence remained conflicting with regard to modality-specific and modality-independent explanations across different tasks and material, there have also been suggestions for *triple code models* that include the modality-specific information as well as integrated information from the different representation formats (e.g., Glucksberg, 1984; Snodgrass, 1984; see also Dehaene, 1992; Dehaene et al., 1999).

Regarding the more direct interplay of auditory and visual information, the theory of redintegration (see Baddeley, 2007) has been proposed to explain the benefit of combined auditory and visual information relative to unimodal information. According to this theory, encoding congruent audio-visual stimuli results in more sophisticated memory representations than encoding auditory or visual stimuli in isolation. Importantly, this memory representation can be (fully) activated by unimodal retrieval cues (e.g., Heikkilä et al., 2015). Empirical support for this theory derives/comes from the studies by Murray and colleagues, who demonstrated more accurate recognition performance for visually tested stimuli that had been studied audio-visually rather than purely visually (e.g., Matusz et al., 2017; Thelen et al., 2015; but see Pecher & Zeelenberg, 2022). Interestingly, there is neuroimaging evidence suggesting that auditory areas of the brain are more active during visual retrieval of stimuli that were studied audio-visually than during visual retrieval of stimuli that were studied visually (Nyberg et al., 2000; Wheeler et al., 2000). However, these studies were not designed to detect behavioral differences in memory accuracy. Therefore, whether the increase in neural activation also elicits more accurate memory accuracy – to our knowledge – is still an open question.

Rationale of the current project

In this project, we aimed to investigate whether there is evidence for the integration of auditory and visual information during the formation of long-term memory representations. Approaching an answer to this question potentially could help to distinguish between the four theories presented in Fig. 1. All the theories with the exception of the dual-coding theory rely on the assumption that auditory and visual information are integrated (at least to some extend). As integration is a more complex mechanism than independent storages for auditory and visual information, empirical support for such integration is necessary to justify the assumption of integration processes. Surprisingly, however, this assumption has rarely been studied in the context of long-term memory (see Meyerhoff & Huff, 2016). We therefore address this lack of evidence with long-term memory experiments in which we tested memory performance for brief excerpts from movies in order to achieve a good balance between experimental control and ecological validity.

In particular, we were interested in two predictions that follow from the assumption of audio-visually integrated long-term memory representations. The first prediction addresses the recognition performance for audio-visual scenes relative to their unimodal counterparts. In particular, we ask whether the superior memory performance for audio-visual scenes relative to purely visual or purely auditory scenes (Meyerhoff & Huff, 2016) is large enough to rule out explanations based on the simple summation of retrieval cues. While multimodal integration, triple coding, and redintegration implicitly assume a non-additive summation, the presence of such a substantial increase potentially could rule out the dual-coding account. In order to address this prediction, we have reanalyzed the results of the first experiment by Meyerhoff and Huff (2016; retention intervals of 1 day or 1 week) using a Bayesian framework testing whether recognition performance for audio-visual scenes exceeds the performance expected from the recognition performance for auditory or visual tracks alone. To anticipate our results, the observed memory advantage of audio-visual scenes will not be large enough to rule out dual coding.

The second prediction of audio-visually integrated memory representations (e.g., amodal; Anderson, 1978, 1979; Kieras, 1978) addresses the effect of study-test congruency for unimodal versus audio-visual scenes. Of particular interest is the recognition performance for unimodal visual scenes. Whereas all theoretical accounts of long-term memory predict that scenes studied and tested audio-visually elicit the most accurate memory performance (due to more retrieval cues in this condition), different predictions arise for the recognition of visual tracks that had been studied visually or audio-visually. First, multimodal integration predicts that deviations between study and test format cause less accurate recognition performance. For instance, and audio-visual representation in long-term memory would mismatch with a purely visual representation during the test session. Second, dual coding would predict that the unimodal information remains available after encoding audio-visual scenes because both modalities are stored independently. With regard to recognizing a visual or an auditory track, it should therefore be irrelevant if this track was studied unimodally or audio-visually. Third, redintegration would predict more accurate recognition performance for unimodal scenes that have been studied audio-visually as the additional information during encoding should result in a more sophisticated memory representation. We tested these predictions with a series of five experiments in which we manipulated the congruency of the presented modalities between study and test (this method has been successfully used to demonstrate that dynamic information as well as the viewing conditions during encoding are preserved in memory; see Buratto et al., 2009; Reingold, 2002; for a review, see Kent & Lamberts, 2008). To anticipate our results, they will mostly be consistent with the predictions from the dual-coding approach.

Modelling of recognition performance for audio-visual tracks

In the first part of this project, we test whether the recognition performance for audio-visual clips exceeds the threshold at which it cannot be explained with the assumption of independent memory representations for both modalities (i.e., with the availability of multiple retrieval cues). Previous research from our own lab (Meyerhoff & Huff, 2016, Experiment 1; data depicted in the dark bars of Fig. 2) has shown that audio-visual tracks elicit more accurate recognition

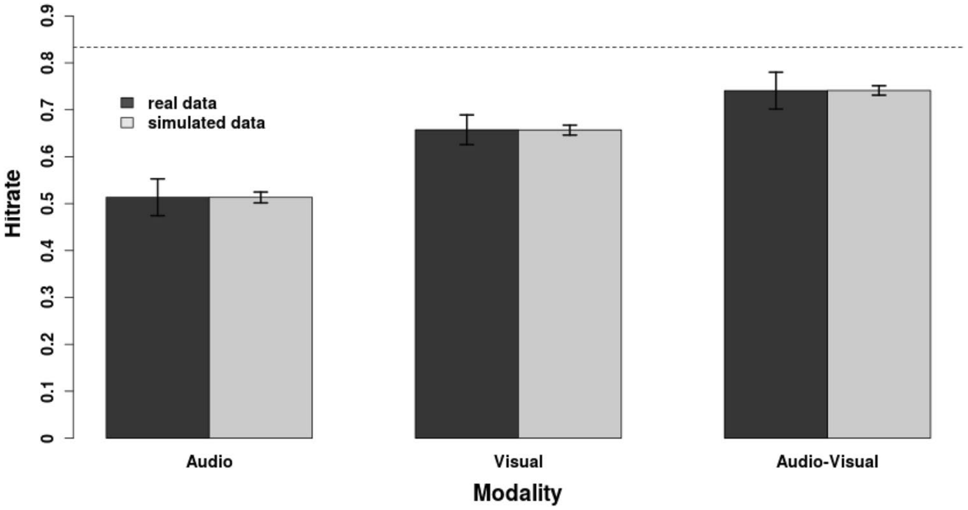

Fig. 2 Depiction of the recognition advantage of audio-visual scenes relative to their unimodal counterparts reported in Meyerhoff and Huff (2016). The dark gray bars display the data from the original experiment (the error bars indicate within-subject confidence intervals). The light gray bars display the simulation data of the present study (the error bars indicate highest density intervals). The dashed line represents the criterion of audio-visual integration. Audio-visual memory performance above this line cannot be explained by independent memory traces for auditory and visual information

performance than the unimodal counterparts of these clips. In this experiment, 48 participants performed an old/new recognition task with 900 brief excerpts from Hollywood movies (these excerpts are selected identically to those in the later experiments of this project; see the *Methods* of Experiment 1a for full details). One half of the clips were presented in a study session, whereas the full set was presented in the testing session after a retention interval of 1 day or 1 week. One-third of all stimuli were presented audio-visually (auditory and visual track were semantically congruent, i.e., from the same scene), one-third visually, and the remaining third auditorily during both study and test (i.e., maintaining study-test congruency). The stimuli were counterbalanced across participants so that each clip appeared in each modality equally often. The results showed the most accurate recognition performance for the audio-visual clips, followed by the visual tracks that were recognized more accurately than the auditory clips. Within this reanalysis, we tested whether the benefit of audio-visual clips relative to visual and auditory tracks is large enough to rule out explanations based on independent retrieval cues (Thompson & Paivio, 1994; see also Stevenson et al., 2014). To anticipate our results, the reanalysis will show that memory performance for audio-visual clips is substantially smaller than predicted by the criterion of audio-visual integration.

Methods

In order to test whether the benefit of audio-visual clips is large enough to rule out independent retrieval cues, we tested whether the hit rate for the audio-visual clips met the integration criterion proposed by Thompson and Paivio (Thompson & Paivio, 1994; see also Stevenson et al., 2014).[1] According to this criterion, performance gains stem from audio-visual integration if the probability to detect an audio-visual stimulus is larger than the complement probability of neither detecting the unimodal visual nor the unimodal auditory stimulus. Transferred to our current analysis of old/new recognition performance, this criterion states that dual coding could be ruled out if the probability of recognizing audio-visual clips is larger than recognizing the auditory and/or visual track of that clip independently of each other (this threshold is indicated by the dashed line in Fig. 2). In other words, a miss would occur only if a participant fails to recognize the visual as well as the auditory track of a clip. These considerations translate to Equation 1:

$$\hat{p}(av) = p(a) + p(v) - p(a)\,p(v) \tag{1}$$

Within this equation, *$\hat{p}(av)$* refers to the estimated recognition performance if both modalities contribute to the hit rate independently of each other. Please note that performance up to this criterion would not need any

[1] Please note that we simulated hit rates whereas we analyzed sensitivity in our previous work (Meyerhoff & Huff, 2016). We decided so simulate only hit rates because there is a clear theoretically motivated criterion for audio-visual integration for hit rates but not for false alarm rates. Nevertheless, when we repeated the simulation with sensitivity as dependent variable based on the empirically observed false alarm rates in the audio-visual condition, the results and conclusions were identical to the simulation of the hit rates (the full simulation script is available at https://osf.io/hywcz/).

additional assumption about audio-visual integration in order to explain the observed performance. Therefore, a clear demonstration of audio-visual integration would require recognition performance of audio-visual clips to exceed that criterion (Equation 2).

$$p(av) > \hat{p}(av) \tag{2}$$

We used a Bayesian modeling approach (Gelman et al., 2013; Kruschke, 2014; the modeled data is depicted in the light gray bars of Fig. 2) to estimate the criterion for audio-visual integration (1) as well as to test how the observed hit rates relate to that criterion (2). Following on from Equation 1, the criterion of audio-visual integration within this Bayesian framework could be expressed as in Equation 3.

$$\theta_{crt} = \theta_a + \theta_v - \theta_a\,\theta_v \tag{3}$$

This approach allows us to simultaneously estimate the parameters for the auditory hit rate, the visual hit rate, the audio-visual hit rate, as well as the criterion for audio-visual integration given our data (y) from the preceding memory experiment (Equation 4)

$$\theta_a, \theta_v, \theta_{av}, \theta_{crt} \mid y \tag{4}$$

The relevant trials for this analysis were test trials presenting previously studied items (i.e., hits and misses). For these trials, the observed performance in the old/new recognition task is either a hit or a miss (i.e., a binomial variable). Therefore, the likelihood of our data could be expressed as in Equation 5, in which θ_m is the probability of success in the modality $m \in \{a, v, av\}$, n_m is the number of trials and y_m is the number of hits.

$$p(y_m \mid \theta_m) = \binom{n_m}{y_m} \theta_m^{y_m}(1-\theta_m)^{n_m-y_m} \tag{5}$$

In order to allow for an interpretation of the modeled θ_{crt}, we added an indicator variable signaling whether the hit rate for audio-visual clips meets the criterion for audio-visual integration. As can be seen in Equation 6, this indicator variable is $x_{ind} > 1$ if the hit rate for audio-visual clips exceeds the criterion for audio-visual integration and $x_{ind} \leq 1$ if the hit rate for audio-visual clips can be explained by independent memory traces for the unimodal tracks of the clips.

$$x_{ind} = \frac{\theta_{av}}{\theta_{crt}} \tag{6}$$

For the modeling itself, we used an informed prior on the indicator variable that prefers values around $x_{ind} = 1$ (i.e., neither preferring one of the different explanations; Equation 7) as well as mildly informed priors for the recognition probabilities of auditory and visual tracks (i.e., to reflect the knowledge that performance in these conditions falls between chance level and ceiling; see Equation 8). Please note that θ_{crt} inherited its priors from θ_a and θ_v, whereas θ_{av} inherited its prior from θ_a, θ_v, and x_{ind}.

$$x_{ind} \sim \mathrm{Normal}(\mu = 1, \ \ \sigma = .01) \tag{7}$$

$$\theta_{a/v} \sim \mathrm{Beta}(\alpha = 5.8, \ \ \beta = 4.2) \tag{8}$$

In order to calculate the joint probability given the data, we used the R package rjags (Plummer, 2016). We employed a Markov Chain Monte Carlo (MCMC) method (Gibbs Sampling; Gelfand & Smith, 1990) in order to derive samples from the posterior distribution. Following an adaption period of 1,000 samples as well as a burn-in period of 1,000 samples, four parallel chains performed 10,000 iterations with a thinning interval of 1. In order to check for convergence, autocorrelation, as well as effective sample size, we used the MCMC diagnostic tools included within the R package coda (Plummer et al., 2006).

Results

Our results were derived from the first MCMC chain. In order to determine whether differences in the observed means are meaningful, we calculated 95% highest density intervals (HDIs) for all reported parameters. As depicted in Fig. 3, our modeling approach indicated that the observed hit rates for audio-visual clips are clearly below the criterion for audio-visual integration proposed by Thompson and Paivio (1994). The parameter for the rate of correctly recognized audio-visual stimuli was $M = 0.74$, $SD = 0.005$, 95% HDI [0.73; 0.75], whereas the parameter for the estimation of the integration criterion was $M = 0.83$, $SD = 0.003$, 95% HDI [0.83; 0.84]. Consequently, the indicator variable x_{ind} revealed values smaller than 1, $M = 0.89$, $SD = 0.007$, 95% HDI [0.88; 0.90] (see Fig. 3).

Discussion

The reanalysis of our previous data showed that the recognition performance for audio-visual scenes was substantially below the criterion of audio-visual integration proposed by Thompson and Paivio (1994). This finding is remarkable because most theories on audio-visual integration would expect performance above this threshold (signaling audio-visual integration) or around this threshold (signaling dual coding). Performance below the threshold (i.e., sub-additivity), however, requires further elaboration because this observation implies that our participants performed less accurately than predicted by independent memory traces for auditory and visual tracks.

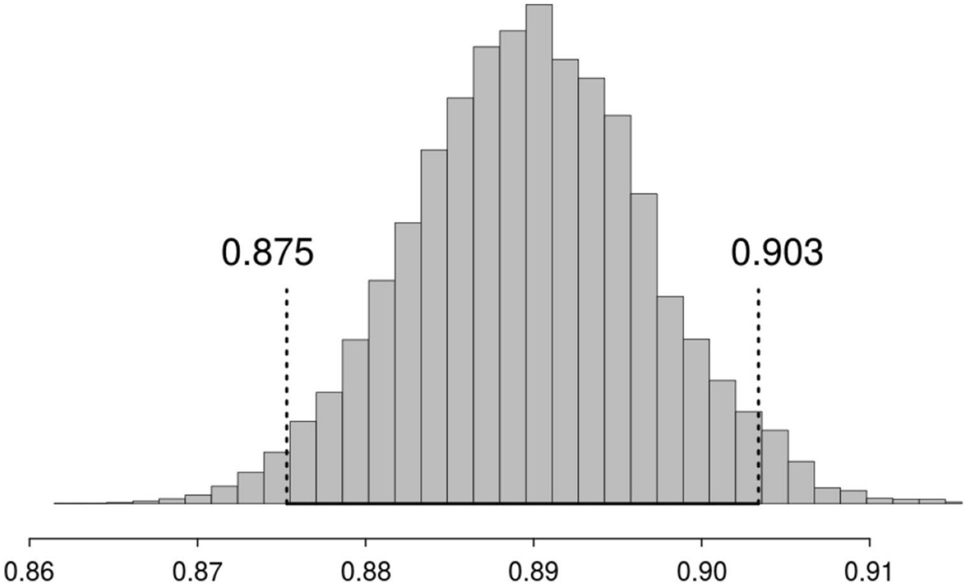

Fig. 3 Highest density interval of the indicator variable for the simulated integration criterion. Values smaller than 1 indicate hit rates below the expected performance for independent storages for auditory and visual information. Values larger than 1 indicate that the hit rates exceed the expected performance based on independent storages for auditory and visual information. Note that the entire highest density interval ranges from 0.875 to 0.903

So how can we explain this sub-additivity? In contrast to the predicted result patterns (performance at or above the threshold), performance below the threshold is less conclusive with regard to the question of whether auditory and visual information are integrated in long-term memory representations. On the one hand, research on the single-neuron level (Stein et al., 2009), as well as the level of neural activity in general (Sperdin et al., 2009) has suggested that sub-additivity might also reflect audio-visual integration as crossmodal depression (similar to crossmodal enhancement) indicates an interaction between the senses. On the other hand, however, this analogy does not necessarily hold true for memory processes as there are parsimonious alternative explanations that are independent of integration processes. For instance, with unimodal encoding, the total encoding time is twice as much as with audio-visual encoding, which might prevent audio-visual scenes from reaching the level of performance predicted by unimodal memory performance. Similarly, when auditory and visual track are stored separately, the necessity to do this at the same time for audio-visual clips would require allocating attentional resources between both modalities, which in return might result in less accurate memory performance than when both modalities are encoded in isolation.

In sum, this simulation does not provide strong evidence in favor of audio-visually integrated long-term memory representations. Due to the sub-additivity, however, a sub-optimal integration process remains possible, so that rejecting integration processes also might be premature. In the following experiments, we collect additional evidence, and we will return to the sub-additivity effect in the *General discussion*.

Experiment 1a

Following the simulation study, which could not rule out or prove the existence of audio-visual integration processes with regard to long-term memory representations, we strived to provide further experimental evidence that could help to resolve this ambiguity. In the following experiments, we therefore tested a second prediction from audio-visually integrated memory performance for dynamic scenes by investigating effects of study-test congruency. This approach has been used intensively to study the nature of memory representations. For instance, it has been used to demonstrate that memory representations also include contextual information (e.g., Godden & Baddeley, 1975; Grant et al., 1998) as well as dynamic information of scenes (Buratto et al., 2009).

If auditory and visual information are integrated in long-term memory representations, a successful recognition of the scene should be more likely when the modality during the study session matches the modality of the testing session. In other words, a clip that has been studied audio-visually should be recognized more accurately when tested audio-visually rather than visually, and a clip that has been studied visually should be recognized more accurately when tested visually rather than audio-visually. As the first of these predictions could also be explained by an increase in retrieval cues (see below), the second of these predictions provides a strong test for the nature of the memory representation. If a scene has been studied purely visually, combining auditory and visual information during testing should result in mismatching representations, thus resulting in less accurate recognition performance than when the scene is tested only visually.

In contrast, if both auditory and visual information are not integrated in long-term memory representations of naturalistic scenes, memory accuracy should follow a function of the retrieval cues that could be used in the testing session. In this case, the condition with audio-visual clips during study and test should elicit the most accurate memory performance because the visual and the auditory track serve as independent retrieval cues in this condition. All other conditions should perform equally well because only the visual information is present during study and test in these conditions.

Methods

Participants

Twenty-four students (all female; age 18–27 years) from the University of Tübingen participated in exchange for course credit or payment. The experimental procedure was approved by the institutional review board of the Leibniz-Institut für Wissensmedien, Tübingen, and all participants provided informed consent prior to their participation. This sample size was chosen to match previous experiments from our lab addressing audio-visual integration in long-term memory performance with similar materials (Meyerhoff & Huff, 2016). In these experiments, we observed large correlations between repeated measures that go along with rather large effect sizes (i.e., effects of the presentation modality ranged from $\eta_p^2 = .40$ to $\eta_p^2 = .72$). Transferred into power calculations, the sample size of 24 participants allows us to detect effects of $\eta_p^2 = .28$ reliably ($\alpha = .05$, power $(1- \beta) = .82$).

Apparatus, stimuli, and procedure

The experiment was coded in Python using the PsychoPy libraries (Peirce, 2007). The stimuli were presented on a 23-in. LCD monitor (60 Hz, 1,920 × 1,080 pixels) controlled by a MacMini at an unrestricted viewing distance of approximately 60 cm.

The stimuli consisted of 1,200 brief clips from 50 Hollywood movies (1935–2008). From each of the movies, we extracted 24 clips, which were equally distributed across lengths of 3, 3.5, or 4 s (i.e., eight clips per length and movie). The clips were initially selected randomly from the movies with the only restriction that there were no filmic cuts within the clips (cinemetrix database; http://www.cinemetrix.lv/). Very few clips were replaced due to not carrying auditory and/or visual information, or the auditory information apparently mismatched the visual information (see Meyerhoff & Huff, 2016). Therefore, the final set of stimuli reflected a representative sample from various films and genres depicting a large variety of different visual scenes and auditory tracks that could easily be identified (including human speech, naturalistic sounds, and background music). Due to this selection procedure, it is likely that the clips differ among physical (e.g., luminance, visual activity) and psychological dimensions (e.g., salience, memorability). In order to eliminate any potentially confounding influence from the clips, we counterbalanced the assignment of each clip to the four experimental conditions as well as to the set of targets or foils across the subgroups of eight participants. This ensured that each clip was presented equally often in each condition and equally often as the target and foil. Further, there was a foil from the same movie (of the same length) for each target within the same experimental condition (in order to prevent participants from recognizing the movie rather than a particular clip).

The experiment was divided into a study and a test session separated by 24 h. The participants were instructed that they were participating in a memory experiment and that they would need to recognize the studied items in the testing session. During the study session (approx. 1 h), the participants attended to 600 of the clips (four clips of each length from each movie). 300 of the clips were presented with the matching sound track, whereas the other half were presented in silence. During the testing session (approx. 2 h), the participants saw the full set of stimuli and indicated after each one whether it had been presented during the study session by pressing the corresponding button on a keyboard (i.e., old/new recognition). Importantly, half of the stimuli that had been presented with the auditory track during the study session were presented without the auditory track during the test session, and half of the tracks that had been presented without the auditory track during the study session were accompanied by the auditory track during testing. Before the study session, the participants were informed that they would need to recognize the presented clips in either the same or the other modality condition (i.e., they were instructed to mark tracks as "old" even when they were aware of additional or lacking auditory information).

Taken together, our experiment therefore follows a 2 × 2 × 3 within-subject design, with modality being manipulated orthogonally between study (visual vs. audio-visual) and test (visual vs. audio-visual; i.e., a study-test congruency experiment). Further, we manipulated the clip length. The manipulation of the clip length was motivated by practical

Table 1 Results of all experiments (study/test) for different clip lengths

	Clip length					
	3 s		3.5 s		4 s	
	d'	c	d'	c	d'	c
	M (SD)	*M (SD)*	*M (SD)*	*M (SD)*	*M (SD)*	*M (SD)*
Experiment 1a						
av / av	0.88 (0.62)	0.31 (0.39)	1.05 (0.67)	0.27 (0.37)	1.17 (0.76)	0.25 (0.38)
av / v	0.83 (0.65)	0.49 (0.37)	0.82 (0.67)	0.40 (0.33)	0.80 (0.61)	0.37 (0.35)
v / av	0.70 (0.68)	0.49 (0.39)	0.70 (0.75)	0.36 (0.33)	0.79 (0.59)	0.38 (0.35)
v / v	0.78 (0.55)	0.46 (0.31)	0.74 (0.71)	0.41 (0.33)	0.86 (0.62)	0.43 (0.33)
Experiment 1b						
v / av	1.00 (0.38)	0.52 (0.50)	0.90 (0.36)	0.47 (0.45)	1.04 (0.40)	0.49 (0.47)
v / v	0.89 (0.34)	0.55 (0.45)	0.97 (0.41)	0.50 (0.46)	1.02 (0.43)	0.52 (0.44)
Experiment 2						
av / av	0.98 (0.82)	0.01 (0.30)	1.00 (0.95)	-0.07 (0.27)	1.05 (0.92)	-0.15 (0.36)
av / a	0.60 (0.54)	0.60 (0.24)	0.60 (0.60)	0.46 (0.31)	0.61 (0.73)	0.38 (0.34)
a / av	0.29 (0.41)	0.35 (0.27)	0.34 (0.55)	0.30 (0.40)	0.35 (0.54)	0.30 (0.36)
a / a	0.50 (0.51)	0.63 (0.35)	0.44 (0.61)	0.51 (0.25)	0.57 (0.70)	0.41 (0.24)
Experiment 3						
a / a	0.62 (0.49)	0.79 (0.43)	0.74 (0.48)	0.67 (0.39)	0.86 (0.51)	0.62 (0.45)
a / av	0.30 (0.35)	0.54 (0.36)	0.41 (0.49)	0.50 (0.40)	0.43 (0.42)	0.40 (0.38)
v / v	0.90 (0.47)	0.41 (0.41)	0.91 (0.55)	0.38 (0.47)	0.96 (0.56)	0.41 (0.45)
v / av	0.72 (0.44)	0.33 (0.42)	0.78 (0.51)	0.31 (0.40)	0.73 (0.48)	0.29 (0.47)
Experiment 4 (normal)						
av / av	1.32 (0.78)	0.12 (0.46)	1.33 (0.74)	-0.02 (0.44)	1.34 (0.71)	-0.03 (0.41)
av / v	0.97 (0.58)	0.08 (0.40)	1.12 (0.62)	0.07 (0.52)	1.11 (0.63)	0.15 (0.41)
v / av	0.83 (0.58)	0.21 (0.40)	0.97 (0.62)	0.12 (0.44)	1.01 (0.58)	0.15 (0.36)
v / v	0.97 (0.60)	0.18 (0.41)	1.18 (0.69)	0.11 (0.43)	1.18 (0.60)	0.17 (0.46)
Experiment 4 (degraded)						
av / av	0.87 (0.63)	0.35 (0.52)	0.90 (0.40)	0.30 (0.49)	1.00 (0.69)	0.25 (0.40)
av / v	0.42 (0.48)	0.54 (0.54)	0.47 (0.49)	0.54 (0.53)	0.42 (0.34)	0.42 (0.44)
v / av	0.35 (0.39)	0.60 (0.52)	0.40 (0.45)	0.52 (0.52)	0.37 (0.41)	0.43 (0.51)
v / v	0.50 (0.33)	0.56 (0.56)	0.35 (0.51)	0.49 (0.56)	0.49 (0.54)	0.55 (0.54)

M mean, *SD* standard deviation, *d'* sensitivity, *c* response criterion, *av* audio-visual, *v* visual, *a* auditory

reasons. Whereas we aimed at testing clips without filmic cuts, it was difficult to find a sufficient number of clips lasting 4 s (the analysis will show that this had no effect on the outcome of the study). The clips were presented in their original resolution (768 × 576 pixels or 1,024 × 576 pixels) in the center of the screen.

Following the testing session, the participants received a list of the 50 movies and were asked to mark those they had seen within the last 5 years. Across all reported experiments, this number varied from none to 34 movies.[2] Excluding familiar movies (for each participant individually) from the analysis did not affect any of the effects of the study or testing modality (nor any interactions). Therefore, we will not discuss this issue any further.

Results

The results show that clips studied and tested audio-visually elicited the most accurate memory performance. However, there was no full study-test congruency effect as clips studied visually did not differ between audio-visual and visual testing.

We conducted a repeated-measures ANOVA with the study modality, the test modality, and clip lengths as the independent variables and the sensitivity measure d' as the dependent variable (see Table 1 for means). Regarding the main effects, we observed more accurate memory

[2] Familiarity data are missing for one participant.

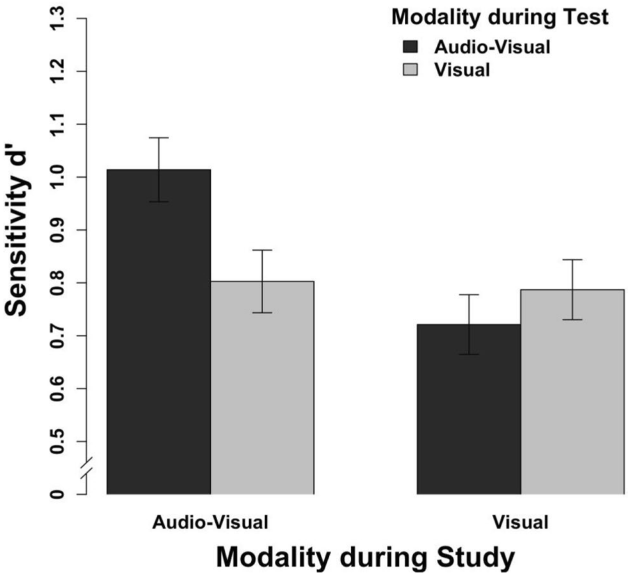

Fig. 4 Results of Experiment 1a aggregated across the different clip lengths. The error bars indicate within-subject confidence intervals

performance with audio-visual than purely visual clips during the study session, $F(1, 23) = 16.39$, $p < .001$, $\eta_p^2 = .42$, 95% confidence interval (CI) [.11; .61], as well as increasing memory accuracy with an increasing length of the clips, $F(2, 46) = 3.47$, $p = .040$, $\eta_p^2 = .13$, 95% CI [0; .29]. However, as the CI includes 0, the increase from 3 to 4 s in clip duration might be too short to induce reliable effects. Further, the main effect modality during the testing session approached, but did not reach, significance, $F(1, 23) = 3.77$, $p = .065$. Most importantly, we observed an interaction between the study and the test modality (see Fig. 4), $F(1, 23) = 10.18$, $p = .004$, $\eta_p^2 = .31$, 95% CI [.04; .53]. However, this interaction did not indicate a full study-test congruency effect. Whereas clips that were studied audio-visually were recognized more accurately when the test was audio-visual rather than visual, $t(23) = 3.57$, $p = .002$, clips that were studied purely visually did not differ when they were tested audio-visually versus purely visually, $t(23) = 1.21$, $p = .238$. This pattern of results indicates that more retrieval cues present during study and test elicit a more accurate memory performance. Importantly, this does not require the assumption of an audio-visually integrated memory representation. None of the two- and three-way interactions including the clip length reached significance, all $Fs(1, 23) < 2.49$, all $ps > .094$.

The analyses of the response criterion c did not reveal anything of interest. In general, reduced sensitivity came along with a larger response bias, indicating that our participants tended to indicate a clip as new when they were not sure whether it had been presented in the study session. Therefore, we do not report these analyses in full detail; however, the corresponding values are summarized in Table 1.

Experiment 1b

Experiment 1a did not reveal a study-test congruency effect because the testing modality was irrelevant for clips that had been studied purely visually. Although far from statistical significance, there was a numerical trend towards a full study-test congruency effect. In order to exclude the possibility that we did not observe the effect of the testing modality on clips that had been studied visually due to a lack of power, we replicated these two conditions with twice as many clips.

Methods

Participants

Twenty-four new students (16 female; age 18–45 years) participated in Experiment 1b.

Apparatus, stimuli, and procedure

Apparatus, stimuli, and procedure were identical to Experiment 1a with the following exceptions. During the study session, 600 of the clips were presented visually. During the testing session, the participants indicated for all 1,200 clips whether they had been presented in the study session. One half of all clips in total as well as one half of the clips from the study session were presented audio-visually, whereas the remaining half of the clips were presented visually.

Results

In agreement with the results of Experiment 1a, we did not observe an impact of the testing modality for clips that had been studied visually (see Fig. 5). We conducted a repeated-measures ANOVA with the modality during the testing session as well as clip lengths as the independent variables and the sensitivity measure d’ as the dependent variable. Most importantly, there was no evidence that the modality during the testing session had any influence on memory performance when the clips were studied purely visually, $F(1, 23) < 1$, thus replicating the results from Experiment 1a. The main effect of the length of the clip approached but did not reach significance, $F(2, 46) = 2.90$, $p = .065$. Unexpectedly, the interaction between clip length and the modality during the testing session reached significance, $F(2, 46) = 3.42$, $p = .041$, $\eta_p^2 = .13$, 95% CI [0; .29]. A closer inspection of the mean values (see Table 1) showed that this appears to be

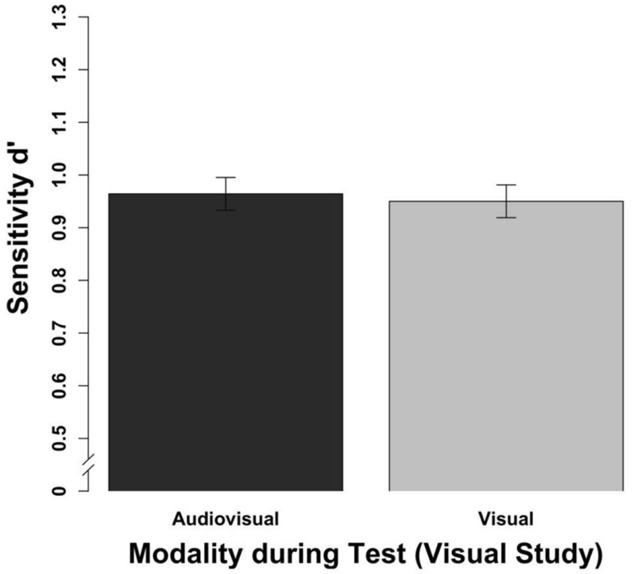

Fig. 5 Results of Experiment 1b aggregated across the different clip lengths. The error bars indicate within-subject confidence intervals

unsystematic. Further, as the CI includes 0, this effect probably reflects noise, so we will not discuss it any further.

Experiment 2

Experiments 1a and 1b revealed that visual and audio-visual testing elicited the same memory performance for clips that had been studied visually. There are at least two possible explanations for the absence of this effect. First, it is possible that information from different modalities is not integrated in long-term memory representations for naturalistic scenes but contribute to memory performance rather independently of each other. Alternatively, the absence of a study-test congruency effect might be a peculiarity of visual memory. From an ecological point of view, this seems reasonable as the auditory information might differ strikingly between study and test due to multiple reasons, such as the presence of additional noise or a varying distance to the scene. In order to differentiate between these two possibilities, we repeated Experiment 1a with auditory instead of visual scenes (i.e., auditory vs. audio-visual scenes). If the absence of study-test congruency effects is a peculiarity of visual memory, we should observe a full study-test congruency effect in this experiment. In contrast, the absence of a study-test congruency effect would signal that distinct modalities contribute to long-term memory performance rather independently of each other.

Methods

Participants

Twenty-four new students (17 female; age 18–34 years) participated in Experiment 2.

Apparatus, stimuli, and procedure

Apparatus, stimuli, and procedure were identical to Experiment 1a with the following exceptions. Instead of visual scenes, we presented the auditory track of the same scenes in this experiment. During the study session, 300 of the clips were presented auditorily, whereas the remaining 300 clips were presented audio-visually. During the testing session, the assignment of the modality was reversed for one half of the clips. The 600 foils were divided equally between the two modality conditions.

Results

Our results show a full study-test congruency effect for the combination of auditory and audio-visual clips. Without exception, a match in the modality conditions between study and test elicited more accurate performance than a corresponding mismatch.

We conducted a repeated-measures ANOVA with the study modality, the test modality, and clip lengths as the independent variables and the sensitivity measure d' as the dependent variable (see Table 1). Regarding the main effects, we observed more accurate memory performance with audio-visual clips rather than auditory clips during the study session, $F(1, 23) = 44.38$, $p < .001$, $\eta_p^2 = .66$, 95% CI [.38; .78], as well as more accurate memory performance for audio-visual clips rather than auditory clips during the testing session, $F(1, 23) = 11.67$, $p = .002$, $\eta_p^2 = .34$, 95% CI [.05; .55]. Neither the main effect of clip length nor any two-way nor three-way interaction including length reached significance, all Fs(2, 46) < 1. Most importantly, however, we observed an interaction between the study and the test modality (see Fig. 6), $F(1, 23) = 32.87$, $p < .001$, $\eta_p^2 = .59$, 95% CI [.28; .73]. In contrast to Experiment 1a, this interaction indicates a full study-test congruency effect because clips that were studied audio-visually elicited more accurate memory performance when they were tested audio-visually than auditorily, $t(23) = 5.61$, $p < .001$, and clips that were studied auditorily elicited more accurate memory performance when they were tested auditorily than audio-visually, $t(23) = 3.95$, $p < .001$. This finding indicates that study-test congruency effects for matching and mismatching modalities could be observed in general.

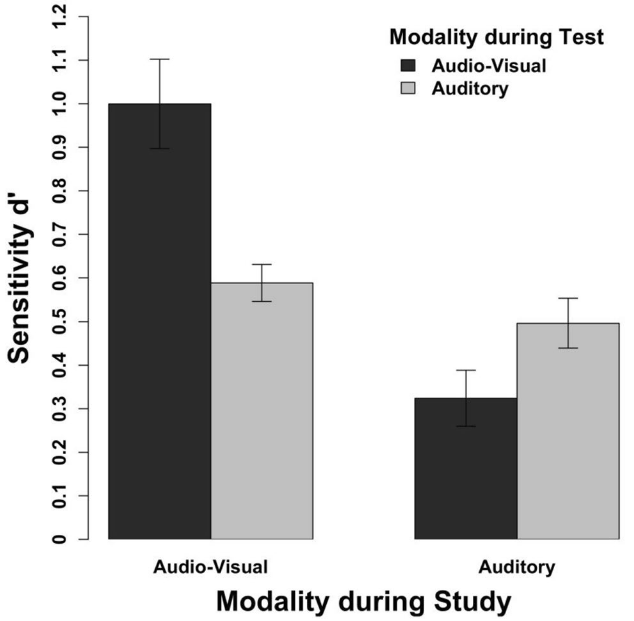

Fig. 6 Results of Experiment 2 aggregated across the different clip lengths. The error bars indicate within-subject confidence intervals

Cross-experimental analysis

The presence of a full study-test congruency effect for auditory versus audio-visual scenes suggests that the consistent absence of the study-test congruency effect in Experiments 1a and 1b is a peculiarity of memory representations including visual information. However, because the presence versus absence of an effect between two experiments cannot be interpreted as an interaction (Nieuwenhuis, Forstmann, & Wagenmakers, 2011), we ran a cross-experimental analysis to prove whether the observed study-test congruency effects differ between Experiment 1a and Experiment 2a. Such an analysis between these experiments is possible as they are structurally equivalent. In both experiments, we presented half of the clips unimodally and the remaining half of the clips audio-visually. The only difference to a full between-subject design is the lack of a random assignment of the participants to the two experiments (however, both groups of participants stem from the same pool of students).

For the cross-experimental analysis, we ran a mixed ANOVA with experiment (between-subject; Exp. 1a vs. 2a), study modality (within-subject; unimodal vs. audio-visual) and the test modality (within-subject; unimodal vs. audio-visual) as independent variables and the sensitivity measure d' as the dependent variable. Most importantly, we observed a significant three-way-interaction, $F(1, 46) = 9.82$, $p = .003$, $\eta_p^2 = .18$, 95% CI [.02; .36], indicating that the study-test congruency effects were differently pronounced between the two experiments. Furthermore, all two-way interactions as well as the main effects of study modality and test modality reached significance, all $Fs(1, 46) > 18.23$, all $ps < .001$, whereas the main effect of the experiment did not, $F(1, 46) = 1.88$, $p = .177$. Due to the three-way interaction, the results of this analysis are consistent with the interpretation that the recall of visual information is not susceptible to irrelevant additional auditory information during testing, but that auditory information is susceptible to irrelevant additional visual information during testing.

Experiment 3

The results of Experiment 2 revealed that irrelevant additional visual information during testing is detrimental for an accurate recognition of auditory scenes. This pattern of results is relevant as it implies that recognition performance for auditory tracks might not operate independently from visual information as suggested by the dual-coding theory (i.e., even information that was not even present during encoding disrupts auditory recognition). In contrast, however, visual information could be recognized independently of irrelevant auditory information. Indeed, the cross-experimental analysis confirmed that this contrasts with the findings of Experiments 1a and 1b, which showed that irrelevant additional auditory information has little to no effect on the recognition performance for visual scenes.

Nevertheless, because cross-experimental analyses per definition do not include randomized sampling, we aimed to replicate the stronger susceptible influence of additional visual information on auditory recognition than vice versa while further probing the boundary conditions of this effect as well as the dual-coding framework. In the previous experiments, one modality was irrelevant for solving the task (i.e., in Experiments 1a and 1b visual information was sufficient whereas auditory information was irrelevant; in Experiment 2a, auditory information was sufficient whereas visual information was irrelevant). In contrast to this, we designed this experiment so that both modalities are relevant during the testing session. The participants studied auditory or visual clips in this experiment. During testing, half of the clips were presented in the same modality whereas the other half were presented audio-visually. For the unimodal clips, the participants simply reported whether the clip had been presented during the study session. For the audio-visual clips, the participants reported whether either the auditory or the visual track had been presented during the study session. Because the participants do not know whether the auditory or the visual track (if any) of an audio-visual clip had been presented during the study session, this implies that there are no irrelevant modalities in this experiment. Therefore, participants cannot simply focus on one modality during the test of audio-visual clips but need to process both of them simultaneously. Whereas a strong version of the dual-coding framework would predict that both tracks can be processed independently, the sub-additivity between the two modalities observed in the simulation study suggests a general decrease in recognition performance for auditory and visual tracks that are tested audio-visually. Most importantly, however, the detrimental effect of additional visual information on auditory recognition performance should be more pronounced than the detrimental effect (if any) of additional auditory information on visual recognition performance (i.e., a statistical interaction).

Methods

Participants

Twenty-four new students (18 female; age 20–31 years) participated in Experiment 3.

Apparatus, stimuli, and procedure

Apparatus, stimuli, and procedure were identical to Experiment 1a with the following exceptions. During the study session, 300 of the clips were presented auditorily, and 300 clips were presented visually. During the testing session, half of these clips were presented in the same modality as during the study session, whereas the remaining half were presented together with the modality that was absent during the study session (i.e., audio-visually). The additional 600 foils were proportionally distributed across these modality conditions (i.e., 150 auditory foils, 150 visual foils, and 300 audio-visual foils).

Results

Our results show that in general the recognition performance of unimodally studied tracks was lower when the additional irrelevant modality was present during the test. Most importantly, however, this susceptibility to the irrelevant additional modality is clearly more pronounced for auditory than for visual information.

We conducted a repeated-measures ANOVA with the study modality (auditory, visual), the presence of the irrelevant additional modality during the test (present, absent) and clip lengths as the independent variables, and the sensitivity measure d’ as the dependent variable (see Table 1). Regarding the main effects, we observed more accurate recognition performance for visually than auditorily studied clips, $F(1, 23) = 36.95$, $p < .001$, $\eta_p^2 = .62$, 95% CI [.32; .75], as well as more accurate recognition performance for unimodal testing (i.e., same modality as during study) than for audio-visual testing (i.e., with the irrelevant additional modality), $F(1, 23) = 49.90$, $p < .001$, $\eta_p^2 = .68$, 95% CI [.41; .80]. Neither the main effect of clip length nor any two-way nor three-way interaction including length reached significance, all Fs(2,46) < 2.35, all ps > .107. Importantly, however, we observed an interaction between the modality during study and the presence of irrelevant additional modality during testing (see Fig. 7), $F(1, 23) = 5.99$, $p = .022$, $\eta_p^2 = .21$, 95% CI [.002; .45]. Although the effect size is rather small, this is consistent with the preceding cross-experimental analysis. This interaction shows that recognition performance for auditory clips is more susceptible for irrelevant additional visual information during testing than recognition performance for visual clips for irrelevant additional auditory information during testing. Nevertheless, the detrimental effect of the irrelevant additional modality during testing was present for auditorily, $t(23) = 6.69$, $p < .001$, as well as visually studied clips, $t(23) = 3.55$, $p = .002$. While this contradicts a completely independent processing of both modalities, it matches well with the results of the simulation study, which showed that recognition of audio-visual information is sub-additive relative to the isolated modalities.

Experiment 4

The preceding experiments show that memory for visual scenes is less susceptible to additional auditory information during testing than memory for auditory scenes to additional visual information during testing. On the one hand, this suggests a dominant role of visual information for the formation of long-term memory representations of naturalistic scenes. On the other hand, however, it remains possible that the dominant role of visual information in our experiments arises from a higher discriminability or memorability of the visual tracks than the auditory tracks. In order to test this possibility, we repeated Experiment 1a with an additional manipulation of the quality of the visual tracks (normal vs. degraded). If the dominant role of visual information in our experiments arose from a superior discriminability and/or memorability of the visual track, visually degraded tracks should be more susceptible for additional auditory information during testing than regular visual tracks. In contrast, if visual information was generally dominant in memory representations, both degraded and normal tracks should be equally unsusceptible to additional auditory information during testing.

Methods

Participants

Twenty-four new students (19 female; age 20–28 years) participated in Experiment 4.

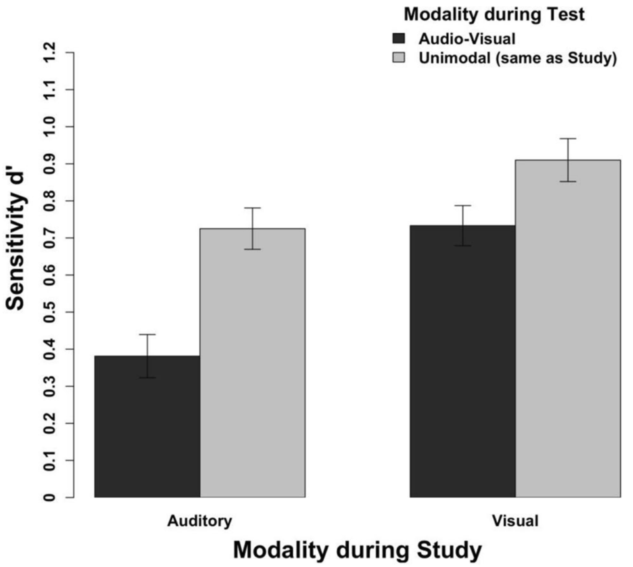

Fig. 7 Results of Experiment 3 aggregated across the different clip lengths. The error bars indicate within-subject confidence intervals

Apparatus, stimuli, and procedure

Apparatus, stimuli, and procedure were identical to Experiment 1a with the following exceptions. Using the framework Frei0r (2018) with the Pixeliz0r filter set to a pixelization degree of .03 × .03, we generated visually degraded versions of our clips. As in Experiment 1a, 300 of the clips were presented visually whereas the remaining 300 clips were presented audio-visually during the study session (equally split between normal and visually degraded stimuli).[3] During testing, half of the visually studied clips were presented audio-visually, and vice versa (equally split between normal and visually degraded stimuli). The additional 600 foils were proportionally distributed across all conditions.

Results

Our results confirm the conclusions from Experiments 1a, 1b, and 2. Most importantly, reducing the quality of the visual tracks did not increase the susceptibility to additional auditory information during testing (see Fig. 8).

We conducted a repeated-measures ANOVA with the visual quality (normal, degraded), the study modality (visual, audio-visual), the test modality (visual, audio-visual), and clip lengths as the independent variables and the sensitivity measure d' as the dependent variable (see Table 1). Neither the four-way interaction nor any of the three-way interactions reached significance, all Fs < 1.30, all ps > .283. Most importantly, we observed two two-way interactions (see Fig. 6). First, we observed an interaction between study modality and test modality, $F(1, 23) = 60.98$, $p < .001$, $\eta_p^2 = .73$, 95% CI [.48; .82]. This interaction indicates that a study-test congruency effect could be present in our data. Second, we observed an interaction between the test modality and the visual quality of the clips, $F(1, 23) = 5.61$, $p = .026$, $\eta_p^2 = .20$, 95% CI [0; .44]. This interaction indicates that the effect of the test modality differs between the normal and the degraded visual quality of the clips.

In a series of post hoc t-tests, we further pursued these interactions. Clips that were studied audio-visually elicited more accurate recognition performance when they were also tested audio-visually rather than visually in isolation for both the normal, $t(23) = 5.97$, $p < .001$, and the degraded visual quality, $t(23) = 5.82$, $p < .001$. In contrast, visually studied clips were not recognized more accurately when tested visually than audio-visually for both the normal, $t(23) = 2.00$, $p = .058$, and the degraded visual quality, $t(23) = 1.80$, $p = .086$. The overall pattern of these t-tests therefore suggests that the interaction between the study modality and the test modality arises from a partial but not a full study-test congruency effect (i.e., there is no detrimental effect of the addition of auditory information during testing). This replicates the findings of Experiments 1a and 1b. The interpretation of the interaction of the test modality and the visual quality is more complicated because the effect is so small that the lower bound of the CI of the effect size is 0. Further, there were no effects of the test modality in the post hoc tests for both visual qualities. If anything, however, the numerical effect is contrary to the prediction that an increased memorability induces the dominance of the visual information in memory representations (i.e., more pronounced at the normal rather than the degraded visual quality). Therefore, there is no evidence in the data that a reduction in the memorability of visual information alters the impact of additional auditory information on recognition accuracy.

None of the remaining two-way interactions reached significance, all Fs < 2.11, all ps > .160. Regarding the main effects, we observed more accurate recognition performance for audio-visually than visually studied clips, $F(1, 23) = 36.67$, $p < .001$, $\eta_p^2 = .61$, 95% CI [.32; .75], as well as more accurate recognition performance for audio-visually than visually tested clips, $F(1, 23) = 14.00$, $p = .001$, $\eta_p^2 = .38$, 95% CI [.08; .59]. Finally, we observed a main effect of the visual quality, $F(1, 23) = 129.85$, $p < .001$, $\eta_p^2 = .85$, 95% CI [.69; .90], indicating that the pixelating procedure indeed reduced the memorability of the visual tracks. The main effect of the clip length did not reach significance, $F(1, 23) = 2.36$, $p = .106$.

[3] Due to a codec error, we lost data for one clip for eight participants. Due to the counterbalancing procedure this affects all modality conditions equally.

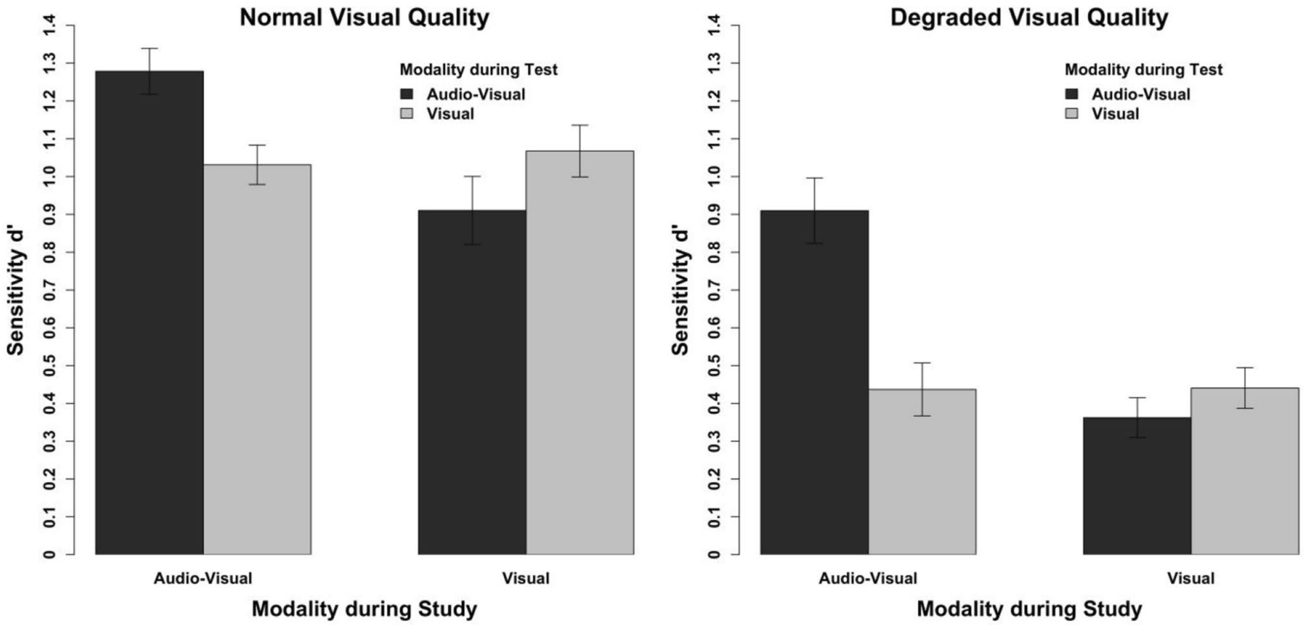

Fig. 8 Results of Experiment 4 aggregated across the different clip lengths for clips of normal visual quality (left) and degraded visual quality (right). The error bars indicate within-subject confidence intervals

General discussion

The present simulation and experiments were set out to probe whether auditory and visual information of dynamic scenes are integrated in long-tern memory representations. With providing initial evidence for this question, we were aiming at distinguishing between various theoretical accounts that potentially could explain long-term memory performance for audio-visual scenes (see Fig. 1). With regard to this question, there were four key findings in the presented results that promote the understanding of the interplay of auditory and visual information in long-term memory. First, the simulation showed that recognition performance for audio-visual scenes is actually less accurate than one would expect based on the recognition performance for their unimodal counterparts. This finding shows that superior memory performance for audio-visual relative to unimodal scenes does not necessarily require the assumption of integrative processes. Second, we did not observe a full study-test congruency effect for audio-visual versus purely visual scenes. In particular, for scenes that had been studied visually, the testing modality was irrelevant (Experiments 1a, 1b, and 4). In this case, audio-visual integration predicts less accurate recognition performance for visually studied scenes that are tested audio-visually. Third, there is a full study-test congruency effect for auditory versus audio-visual scenes. This finding demonstrates not only that study-test congruency effects in principle could be observed in the context of audio-visual scenes (Experiment 2), but also suggests that visual memory differs from auditory memory in its susceptibility to the other modality (Experiment 3). Fourth, reducing the visual quality of the clips in order to reduce the memorability of the visual track did not increase the susceptibility of visual recognition performance to additional auditory information during testing (Experiment 4). We elaborate further on these key findings in the following two paragraphs. Then we discuss the impact of our findings for theorizing on the nature of long-term memory representations.

The whole is smaller than the sum of its parts

Previous studies addressing the dual coding of auditory and visual information in the context of memory formation have observed that performance in the audio-visual condition reaches the accuracy predicted from independent memories for both modalities (Paivio & Csapo, 1973; Thompson & Paivio, 1994). However, in our analysis recognition performance was even lower than predicted by fully independent retrieval cues. There are two alternatives to how this effect could be reconciled within the framework of independent contributions of auditory and visual tracks to memory performance. First, presenting both modalities in isolation doubles the encoding duration. However, previous work from our lab (Meyerhoff & Huff, 2016) has shown that memory for audio-visual scenes is more accurate when the isolated visual and auditory tracks are presented immediately after each other rather than temporally separated. Because encoding durations are identical in both cases, this rules out an explanation solely based on encoding duration. Second, within the framework of independent retrieval cues, the simultaneous presentation of auditory and visual information might require splitting attention between the two modalities that could in return explain the lower recognition performance (e.g., Craik et al., 1996; Fernandes & Moscovitch, 2000; Kane & Engle, 2000).

The overall pattern of our results allows further speculation on this attentional splitting account. In Experiments 1a, 2, and 4, unimodal memory probes elicited roughly the same recognition performance whether they were studied audio-visually or unimodally. This indicates that there were no costs arising from attentional splitting during the encoding of audio-visual clips. In contrast, when both modalities are presented simultaneously during the testing session, processing both modalities appears to be associated with costs from attentional splitting. For audio-visual targets this becomes apparent from the sub-additive performance revealed by our simulation of audio-visual target recognition. For auditory studied clips, this becomes apparent in the lower recognition rates when the auditory target is embedded in additional visual information. The remarkable exception is visually studied targets embedded into additional auditory information during testing. For these targets, the additional auditory information does not impair recognition performance. This suggests that the costs from attentional splitting mostly draw upon auditory recognition (we discuss this in more detail within the next paragraph).

An asymmetric attentional splitting explanation also matches with one of our findings in Experiment 3. When the participants studied the scenes unimodally (i.e., auditory or visual), adding the remaining modality reduced recognition performance. Although this effect was larger for auditorily studied tracks, it was also present for visually studied tracks. Remarkably, such a detrimental effect of auditory information on visual recognition was present only in this experiment, not in Experiments 1a, 1b, and 4. A central difference between these experiments is that the additional auditory information was relevant in Experiment 3 (because it could also be an auditory target with additional visual information), but not in the other experiments. Thus, the necessity of attending to the visual and the auditory information (of the same clip) appears to lower recognition performance in general. This finding again matches remarkably well with the simulation study, which showed that recognition performance for audio-visual scenes is generally lower than predicted by the recognition probability of the individual modalities in isolation. Further, this finding is also in line with the assumption of a central attentional bottleneck (Tombu et al., 2011). Indeed, attentional involvement in one sensory stream such as detecting a target within a rapid series of images is capable of withdrawing attentional processing from the auditory stream and vice versa (Arnell & Jolicoeur, 1999; Ptito et al., 2008). Of course, our memory task did not require participants to detect targets, but it seems likely that salient events in one or the other sensory stream elicits similar effects.

It seems important to note that the overall recognition performance for audio-visual scenes does not rule out the possibility of an actual integration of both modalities. In fact, the detrimental effects of reduced encoding duration as well as effects of split attention might overshadow the potentially beneficial effects of audio-visual integration. Nevertheless, this analysis shows that the level of recognition performance for audio-visual scenes could be explained without the assumption of integrated auditory and visual information.

Study-test congruency effects for audition, but not for vision

A further observation that is hard to reconcile with audio-visually integrated memory representations is the asymmetric study-test congruency for visual and auditory scenes relative to audio-visual scenes. Whereas memory representations of visual scenes were immune to the additional presence of auditory information during testing (when the auditory information could not be the target itself), the additional presence of visual information during testing interfered with memory for auditory scenes. In other words, visual information affected auditory memory but not vice versa. Importantly, this cannot be explained with a better memorability of the visual rather than the auditory tracks as a reduction in the visual quality of the tracks did not decrease the dominance of the visual information (Experiment 4). A comparable dominance of visual information has been reported in simple reaction time tasks in which the presence of visual stimuli undermined the perception of simultaneously presented auditory stimuli (i.e., Colavita effect; Colavita, 1974; Colavita & Weisberg, 1979; Egeth & Sager, 1977; Hecht, Reiner, & Karni, 2009). Importantly, the preference for visual information indeed reflects differences in the perceptual sensitivity rather than just a shift in the criterion towards the visual modality (Koppen et al., 2009), which is not restricted to simple displays but also emerges during object identification (Ngo et al., 2010; Yuval-Greenberg & Deouell, 2009) as well as in semantically meaningful stimuli (Sinnett et al., 2007; Stubblefield et al., 2013; but see also Koppen et al., 2008).

Importantly, a similar dominance of visual information has been reported for short memory durations (Posner, 1967; see also Posner et al., 1976). Particularly, non-visual information can be ignored (after corresponding instructions) more easily than visual information (Klein & Posner, 1974). Interestingly, previous work from the continuous recognition paradigm showed that semantically matching crossmodal information had a beneficial effect on subsequent unimodal recognition (Lehmann & Murray, 2005; Thelen et al., 2015; but see Pecher & Zeelenberg, 2022). Despite the obvious methodological differences between this paradigm and our current set of experiments, the most interesting factor is the temporal delay between study and test. In the continuous recognition paradigm, recognition occurred immediately and intermixed with initial encounters whereas there is a delay of a full day in our studies. For future research, this raises the interesting question whether auditory contributions to visual information might decay faster than the visual representation itself.

Of course, it remains undoubted that auditory information in principle is capable of altering visual processing (Sekuler et al., 1997; Shams et al., 2000). In order to explain under which circumstances one modality might dominate the other, Welch and Warren (1980) formulated the modality appropriateness principle. According to this principle, the task-specific acuity of the involved modalities affects how they are integrated (Bertelson et al., 2000; Vroomen et al., 2001). With regard to our results for long-term memory representations, such an interpretation would suggest that long-term memory for auditory information is so unreliable (see Cohen et al., 2009) that it has no effect on visual information in a weighted integration. In any case, what seems clear from the asymmetric occurrence of study-test congruency effects is that memory representations are not just the product of an equally weighted integration of auditory and visual information.

Theoretical implications

As outlined in the *Introduction*, numerous attempts have been made to explain memory performance for audio-visual material. These explanations encompass an amodal integration of different sensory channels (Anderson, 1978; Kieras, 1978), redintegration (i.e., enhanced memory representations following audio-visual encoding that could be reactivated by unimodal retrieval cues; see Baddeley, 2007), dual coding (i.e., independent memory representations for auditory and visual components of the same stimuli; Thompson & Paivio, 1994), and triple coding (i.e., independent memories with additional associations between them; Glucksberg, 1984; Snodgrass, 1984). Interestingly, our results cannot be fully resolved by any of these theoretical accounts. First, the strong persistence of visual information is not in line with fully amodal memory representations. Second, audio-visually encoded scenes elicit the same memory performance in visual recognition as purely visually encoded scenes. In contrast, redintegration would have predicted more accurate memory performance for audio-visual scenes during visual recognition due to the supposedly more sophisticated memory representations. Finally, in contrast to the visual component, auditory information was not immune to the influence of the other modality, therefore questioning an independent memory trace for auditory information as proposed by the dual- and triple-coding models. As mentioned previously, we have studied dynamic scenes rather than static images mostly because we considered them to reflect a higher degree of ecological validity (while preserving experimental control). Given these striking differences in the material used, it might not be too surprising that previous accounts are not in full accord with our results.

Nevertheless, there are components of the dual-coding theory that fit in rather well with our results and should thus be incorporated into an explanation of memory for naturalistic scenes. First, as proposed by the dual-coding theory, auditory information enhanced memory performance, although there was no evidence of an actual integration of both of them. Second, visual information remained accessible rather independently even when auditory information was present during encoding (i.e., the absence of full study-test congruency effects for visual vs. audio-visual material). Considering the modality appropriateness principle (Welch & Warren, 1980) as well as the generally inferior auditory memory (Cohen et al., 2009), our study suggests that the core memory representation of dynamic scenes is visual in nature. What is more puzzling is the role of coinciding auditory information. Whereas auditory information is obviously stored (performance is well above chance level in all conditions), it appears not to be independent as soon as visual information is present. It thus seems more likely that auditory information – although not integrated with visual information – is associated with the visual information. We consider this to be a reformulation of the dual-coding theory with only one independent storage for visual information and an associated storage for additional auditory information.

Conclusion

Studying naturalistic dynamic scenes, the present study shows that long-term recognition performance for audio-visual scenes does not require assumptions about audio-visually integrated memory representations. Instead, our study suggests that visual information is dominant in long-term memory for such naturalistic scenes and that additionally presented auditory information is associated with that visual information rather than being stored independently. These findings deviate from previous research that has been using simpler study material, thus calling for further investigation of long-term memory using as ecologically valid stimuli as possible.

Open Practices Statement

All raw data and analysis scripts of the reported experiments are available at https://osf.io/hywcz/. None of the experiments was preregistered (data collection started in the year 2014).

Funding No funding was received for conducting this study.All analysis scripts of the reported experiments are available at https://osf.io/hywcz/.

Data availability Open Access funding enabled and organized by Projekt DEAL. All data of the reported experiments are available at https://osf.io/hywcz/.

Code availability All analysis scripts of the reported experiments are available at https://osf.io/hywcz/.

Declarations

Conflicts of interest The authors have no relevant financial or non-financial interests to disclose.

Ethics approval The experimental procedure was approved by the institutional review board of the Leibniz-Institut für Wissensmedien, Tübingen, Germany.

Consent to participate Informed consent was obtained from all individual participants included in the study.

Consent for publication Not applicable.

References

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology, 14*, 257–262. https://doi.org/10.1016/j.cub.2004.01.029

Amedi, A., von Kriegstein, K., van Atteveldt, N. M., Beauchamp, M. S., & Naumer, M. J. (2005). Functional imaging of human crossmodal identification and object recognition. *Experimental Brain Research, 166*, 559–571. https://doi.org/10.1007/s00221-005-2396-5

Anderson, J. R. (1978). Arguments concerning representations for mental imagery. *Psychological Review, 85*, 249–277. https://doi.org/10.1037/0033-295X.85.4.249

Anderson, J. R. (1979). Further arguments concerning representations for mental imagery: A response to Hayes-Roth and Pylyshyn. *Psychological Review, 86*, 395–406. https://doi.org/10.1037/0033-295X.86.4.395

Arnell, K. M., & Jolicoeur, P. (1999). The attentional blink across stimulus modalities: Evidence for central processing limitations. *Journal of Experimental Psychology: Human Perception and Performance, 25*, 630–648. https://doi.org/10.1037/0096-1523.25.3.630

Baddeley, A. (2007). *Working memory, thought, and action*. Oxford University Press.

Baddeley, A. D., & Logie, R. (1999). Working memory: The multiple component model. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 28–61). Cambridge University Press.

Bertelson, P., Vroomen, J., De Gelder, B., & Driver, J. (2000). The ventriloquist effect does not depend on the direction of deliberate visual attention. *Perception & Psychophysics, 62*, 321–332. https://doi.org/10.3758/BF03205552

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology, 14*, 143–177. https://doi.org/10.1016/0010-0285(82)90007-X

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences, 105,* 14325-14329. https://doi.org/10.1073/pnas.0803390105

Brady, T. F., Konkle, T., Gill, J., Oliva, A., & Alvarez, G. A. (2013). Visual long-term memory has the same limit on fidelity as visual working memory. *Psychological Science, 24*, 981–990. https://doi.org/10.1177/0956797612465439

Buratto, L. G., Matthews, W. J., & Lamberts, K. (2009). When are moving images remembered better? Study–test congruence and the dynamic superiority effect. *The Quarterly Journal of Experimental Psychology, 62*, 1896–1903. https://doi.org/10.1080/17470210902883263

Chen, Y. C., & Spence, C. (2011a). The crossmodal facilitation of visual object representations by sound: Evidence from the backward masking paradigm. *Journal of Experimental Psychology: Human Perception and Performance, 37*, 1784–1802. https://doi.org/10.1037/a0025638

Chen, Y. C., & Spence, C. (2011b). Crossmodal semantic priming by naturalistic sounds and spoken words enhances visual sensitivity. *Journal of Experimental Psychology: Human Perception and Performance, 37*, 1554–1568. https://doi.org/10.1037/a0024329

Chen, Y. C., & Spence, C. (2018). Audiovisual semantic interactions between linguistic and nonlinguistic stimuli: The time-courses and categorical specificity. *Journal of Experimental Psychology: Human Perception and Performance, 44*, 1488–1507. https://doi.org/10.1037/xhp0000545

Cohen, M. A., Horowitz, T. S., & Wolfe, J. M. (2009). Auditory recognition memory is inferior to visual recognition memory. *Proceedings of the National Academy of Sciences, 106*, 6008–6010. https://doi.org/10.1073/pnas.0811884106

Colavita, F. B. (1974). Human sensory dominance. *Perception & Psychophysics, 16*, 409–412.

Colavita, F. B., & Weisberg, D. (1979). A further investigation of visual dominance. *Perception & Psychophysics, 25*, 345–347. https://doi.org/10.3758/BF03198814

Craik, F. I., Govoni, R., Naveh-Benjamin, M., & Anderson, N. D. (1996). The effects of divided attention on encoding and retrieval processes in human memory. *Journal of Experimental Psychology: General, 125*, 159–180. https://doi.org/10.1037/0096-3445.125.2.159

Dehaene, S. (1992). Varieties of numerical abilities. *Cognition, 44*, 1–42. https://doi.org/10.1016/0010-0277(92)90049-N

Dehaene, S., Spelke, E., Pinel, P., Stanescu, R., & Tsivkin, S. (1999). Sources of mathematical thinking: Behavioral and brain-imaging evidence. *Science, 284*, 970–974. https://doi.org/10.1126/science.284.5416.970

Delogu, F., Raffone, A., & Belardinelli, M. O. (2009). Semantic encoding in working memory: Is there a (multi) modality effect? *Memory, 17*, 655–663. https://doi.org/10.1080/09658210902998054

Edmiston, P., & Lupyan, G. (2015). What makes words special? Words as unmotivated cues. *Cognition, 143*, 93–100. https://doi.org/10.1016/j.cognition.2015.06.008

Egeth, H. E., & Sager, L. C. (1977). On the locus of visual dominance. *Perception & Psychophysics, 22*, 77–86. https://doi.org/10.3758/BF03206083

Ernst, M. O., & Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Sciences, 8*, 162–169. https://doi.org/10.1016/j.tics.2004.02.002

Falchier, A., Clavagnier, S., Barone, P., & Kennedy, H. (2002). Anatomical evidence of multimodal integration in primate striate cortex. *The Journal of Neuroscience, 22*, 5749–5759. https://doi.org/10.3758/CABN.4.2.117

Fernandes, M. A., & Moscovitch, M. (2000). Divided attention and memory: Evidence of substantial interference effects at retrieval and encoding. *Journal of Experimental Psychology: General, 129*, 155–176. https://doi.org/10.1037/0096-3445.129.2.155

Fougnie, D., & Marois, R. (2011). What limits working memory capacity? Evidence for modality-specific sources to the simultaneous storage of visual and auditory arrays. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*, 1329–1341. https://doi.org/10.1037/a0024834

Frei0r (2018). [Computer Software] https://frei0r.dyne.org/

Frick, R. W. (1984). Using both an auditory and a visual short-term store to increase digit span. *Memory & Cognition, 12*, 507–514. https://doi.org/10.3758/BF03198313

Gelfand, A. E., & Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Aassociation, 85*, 398–409. https://doi.org/10.1080/01621459.1990.10476213

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Chapman and Hall/CRC.

Giard, M. H., & Peronnet, F. (1999). Auditory-visual integration during multimodal object recognition in humans: A behavioral and electrophysiological study. *Journal of Cognitive Neuroscience, 11*, 473–490. https://doi.org/10.1162/089892999563544

Gibson, J. R., & Maunsell, J. H. (1997). Sensory modality specificity of neural activity related to memory in visual cortex. *Journal of Neurophysiology, 78*(3), 1263–1275. https://doi.org/10.1152/jn.1997.78.3.1263

Glucksberg, S. (1984). Commentary: The functional equivalence of common and multiple codes. *Journal of Memory and Language, 23*, 100–104. https://doi.org/10.1016/S0022-5371(84)90529-2

Godden, D. R., & Baddeley, A. D. (1975). Context-dependent memory in two natural environments: On land and underwater. *British Journal of Psychology, 66*, 325–331. https://doi.org/10.1111/j.2044-8295.1975.tb01468.x

Goldstein, A. G., Chance, J. E., Hoisington, M., & Buescher, K. (1982). Recognition memory for pictures: Dynamic vs. static stimuli. *Bulletin of the Psychonomic Society, 20*, 37–40. https://doi.org/10.3758/BF03334796

Goolkasian, P., & Foos, P. W. (2002). Presentation format and its effect on working memory. *Memory & Cognition, 30*, 1096–1105. https://doi.org/10.3758/BF03194327

Grant, H. M., Bredahl, L. C., Clay, J., Ferrie, J., Groves, J. E., McDorman, T. A., & Dark, V. J. (1998). Context-dependent memory for meaningful material: Information for students. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, 12*, 617–623. https://doi.org/10.1002/(SICI)1099-0720(1998120)12:6<617::AID-ACP542>3.0.CO;2-5

Hecht, D., Reiner, M., & Karni, A. (2009). Repetition priming for multisensory stimuli: Task-irrelevant and task-relevant stimuli are associated if semantically related but with no advantage over uni-sensory stimuli. *Brain Research, 1251*, 236–244. https://doi.org/10.1016/j.brainres.2008.10.062

Heikkilä, J., Alho, K., Hyvönen, H., & Tiippana, K. (2015). Audiovisual semantic congruency during encoding enhances memory performance. *Experimental Psychology, 62*, 123–130. https://doi.org/10.1027/1618-3169/a000279

Hollingworth, A. (2004). Constructing visual representations of natural scenes: The roles of short-and long-term visual memory. *Journal of Experimental Psychology: Human Perception and Performance, 30*, 519–537. https://doi.org/10.1037/0096-1523.30.3.519

Hollingworth, A. (2005). The relationship between online visual representation of a scene and long-term scene memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 396–411. https://doi.org/10.1037/0278-7393.31.3.396

Iordanescu, L., Grabowecky, M., Franconeri, S., Theeuwes, J., & Suzuki, S. (2010). Characteristic sounds make you look at target objects more quickly. *Attention, Perception, & Psychophysics, 72*, 1736–1741. https://doi.org/10.3758/APP.72.7.1736

Iordanescu, L., Guzman-Martinez, E., Grabowecky, M., & Suzuki, S. (2008). Characteristic sounds facilitate visual search. *Psychonomic Bulletin & Review, 15*, 548–554. https://doi.org/10.3758/PBR.15.3.548

Kane, M. J., & Engle, R. W. (2000). Working-memory capacity, proactive interference, and divided attention: Limits on long-term memory retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 336–358. https://doi.org/10.1037/0278-7393.26.2.336

Kassim, A. A., Rehman, R., & Price, J. M. (2018). Effects of modality and repetition in a continuous recognition memory task: Repetition has no effect on auditory recognition memory. *Acta Psychologica, 185*, 72–80. https://doi.org/10.1016/j.actpsy.2018.01.012

Kent, C., & Lamberts, K. (2008). The encoding–retrieval relationship: Retrieval as mental simulation. *Trends in Cognitive Sciences, 12*, 92–98. https://doi.org/10.1016/j.tics.2007.12.004

Kieras, D. (1978). Beyond pictures and words: Alternative information-processing models for imagery effect in verbal memory. *Psychological Bulletin, 85*, 532–554. https://doi.org/10.1037/0033-2909.85.3.532

Kim, R. S., Seitz, A. R., & Shams, L. (2008). Benefits of stimulus congruency for multisensory facilitation of visual learning. *PLoS One, 3*, e1532. https://doi.org/10.1371/journal.pone.0001532

Klein, R. M., & Posner, M. I. (1974). Attention to visual and kinesthetic components of skills. *Brain Research, 71*, 401–411. https://doi.org/10.1016/0006-8993(74)90984-6

Koelewijn, T., Bronkhorst, A., & Theeuwes, J. (2010). Attention and the multiple stages of multisensory integration: A review of audiovisual studies. *Acta Psychologica, 134*, 372–384. https://doi.org/10.1016/j.actpsy.2010.03.010

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Scene memory is more detailed than you think the role of categories in visual long-term memory. *Psychological Science, 21*, 1551–1556. https://doi.org/10.1177/0956797610385359

Koppen, C., Alsius, A., & Spence, C. (2008). Semantic congruency and the Colavita visual dominance effect. *Experimental Brain Research, 184*, 533–546. https://doi.org/10.1007/s00221-007-1120-z

Koppen, C., Levitan, C. A., & Spence, C. (2009). A signal detection study of the Colavita visual dominance effect. *Experimental Brain Research, 196*, 353–360. https://doi.org/10.1007/s00221-009-1853-y

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.

Lehmann, S., & Murray, M. M. (2005). The role of multisensory memories in unisensory object discrimination. *Cognitive Brain Research, 24*, 326–334. https://doi.org/10.1016/j.cogbrainres.2005.02.005

Lewald, J., Ehrenstein, W. H., & Guski, R. (2001). Spatio-temporal constraints for auditory–visual integration. *Behavioural Brain Research, 121*, 69–79. https://doi.org/10.1016/S0166-4328(00)00386-7

Lewandowski, L. J., & Kobus, D. A. (1993). The effects of redundancy in bimodal word processing. *Human Performance, 6*, 229–239.

Lupyan, G., & Thompson-Schill, S. L. (2012). The evocative power of words: Activation of concepts by verbal and nonverbal means. *Journal of Experimental Psychology: General, 141*, 170–186. https://doi.org/10.1037/a0024904

Lupyan, G., & Ward, E. J. (2013). Language can boost otherwise unseen objects into visual awareness. *Proceedings of the National Academy of Sciences, 110*, 14196–14201. https://doi.org/10.1073/pnas.1303312110

Matthews, W. J., Benjamin, C., & Osborne, C. (2007). Memory for moving and static images. *Psychonomic Bulletin & Review, 14*, 989–993. https://doi.org/10.3758/BF03194133

Matthews, W. J., Buratto, L. G., & Lamberts, K. (2010). Exploring the memory advantage for moving scenes. *Visual Cognition, 18*, 1393–1419. https://doi.org/10.1080/13506285.2010.492706

Matusz, P. J., Thelen, A., Amrein, S., Geiser, E., Anken, J., & Murray, M. M. (2015). The role of auditory cortices in the retrieval of single-trial auditory–visual object memories. *European Journal of Neuroscience, 41*, 699–708. https://doi.org/10.1111/ejn.12804

Matusz, P. J., Wallace, M. T., & Murray, M. M. (2017). A multisensory perspective on object memory. *Neuropsychologia, 105*, 243–252. https://doi.org/10.1016/j.neuropsychologia.2017.04.008

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, 746. https://doi.org/10.1038/264746a

Mädebach, A., Wöhner, S., Kieseler, M. L., & Jescheniak, J. D. (2017). Neighing, barking, and drumming horses—Object related sounds help and hinder picture naming. *Journal of Experimental Psychology: Human Perception and Performance, 43*, 1629–1646. https://doi.org/10.1037/xhp0000415

Meyerhoff, H. S., & Huff, M. (2016). Semantic congruency but not temporal synchrony enhances long-term memory performance for audio-visual scenes. *Memory & Cognition, 44*, 390–402. https://doi.org/10.3758/s13421-015-0575-6

Meyerhoff, H. S., Gehrer, N. A., Merz, S., & Frings, C. (2022). The beep-speed illusion: Non-spatial tones increase perceived speed of visual objects in a forced-choice paradigm. *Cognition, 219*, 104978.

Meyerhoff, H. S., & Scholl, B. J. (2018). Auditory-induced bouncing is a visual (rather than a cognitive) phenomenon: Evidence from illusory crescents. *Cognition, 170*, 88–94. https://doi.org/10.1016/j.cognition.2017.08.007

Meyerhoff, H. S., & Suzuki, S. (2018). Beep, be-, or -ep: The impact of auditory transients on perceived bouncing/streaming. *Journal of Experimental Psychology: Human Perception and Performance, 44*, 1995–2004. https://doi.org/10.1037/xhp0000585

Murray, M. M., Foxe, J. J., & Wylie, G. R. (2005). The brain uses single-trial multisensory memories to discriminate without awareness. *Neuroimage, 27*, 473–478. https://doi.org/10.1016/j.neuroimage.2005.04.016

Murray, M. M., Michel, C. M., De Peralta, R. G., Ortigue, S., Brunet, D., Andino, S. G., & Schnider, A. (2004). Rapid discrimination of visual and multisensory memories revealed by electrical neuroimaging. *NeuroImage, 21*, 125–135. https://doi.org/10.1016/j.neuroimage.2003.09.035

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E. J. (2011). Erroneous analysis of interactions in neuroscience: A problem of significance. *Nature Neuroscience, 14*, 1105–1007. https://doi.org/10.1038/nn.2886

Ngo, M. K., Sinnett, S., Soto-Faraco, S., & Spence, C. (2010). Repetition blindness and the Colavita effect. *Neuroscience Letters, 480*, 186–190. https://doi.org/10.1016/j.neulet.2010.06.028

Nyberg, L., Habib, R., McIntosh, A. R., & Tulving, E. (2000). Reactivation of encoding-related brain activity during memory retrieval. *Proceedings of the National Academy of Sciences, 97*, 11120–11124. https://doi.org/10.1073/pnas.97.20.11120

Quack, M., London, R. E., & Talsma, D. (2015). A multisensory perspective of working memory. *Frontiers in Human Neuroscience, 9*, 197. https://doi.org/10.3389/fnhum.2015.00197

Paivio, A., & Csapo, K. (1973). Picture superiority in free recall: Imagery or dual coding? *Cognitive Psychology, 5*, 176–206. https://doi.org/10.1016/0010-0285(73)90032-7

Pecher, D., & Zeelenberg, R. (2022). Does multisensory study benefit memory for pictures and sounds? *Cognition, 226*. https://doi.org/10.1016/j.cognition.2022.105181

Peirce, J. W. (2007). PsychoPy—Psychophysics software in python. *Journal of Neuroscience Methods, 162*, 8–13. https://doi.org/10.1016/j.neumeth.2006.11.017

Plummer, M. (2016). RJAGS: Bayesian graphical models using MCMC. Retrieved from https://CRAN.R-project.org/package=rjags (R package, version 4-6)

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News, 6*, 7–11 Retrieved from https://journal.r-project.org/archive

Ptito, A., Arnell, K., Jolicœur, P., & Macleod, J. (2008). Intramodal and crossmodal processing delays in the attentional blink paradigm revealed by event-related potentials. *Psychophysiology, 45*, 794–803. https://doi.org/10.1111/j.1469-8986.2008.00677.x

Posner, M. I. (1967). Characteristics of visual and kinesthetic memory codes. *Journal of Experimental Psychology, 75*, 103–107. https://doi.org/10.1037/h0024911

Posner, M. I., Nissen, M. J., & Klein, R. M. (1976). Visual dominance: An information-processing account of its origins and significance. *Psychological Review, 83*, 157–171. https://doi.org/10.1037/0033-295X.83.2.157

Powers, A. R., Hillock, A. R., & Wallace, M. T. (2009). Perceptual training narrows the temporal window of multisensory binding. *Journal of Neuroscience, 29*, 12265–12274. https://doi.org/10.1523/JNEUROSCI.3501-09.2009

Reingold, E. M. (2002). On the perceptual specificity of memory representations. *Memory, 10*, 365–379. https://doi.org/10.1080/09658210244000199

Rock, I., & Victor, J. (1964). Vision and touch: An experimentally created conflict between the two senses. *Science, 143*(3606), 594–596. https://doi.org/10.1126/science.143.3606.594

Santangelo, V., & Spence, C. (2007). Multisensory cues capture spatial attention regardless of perceptual load. *Journal of Experimental Psychology: Human Perception and Performance, 33*, 1311–1321. https://doi.org/10.1037/0096-1523.33.6.1311

Saults, J. S., & Cowan, N. (2007). A central capacity limit to the simultaneous storage of visual and auditory arrays in working memory. *Journal of Experimental Psychology: General, 136*, 663–684. https://doi.org/10.1037/0096-3445.136.4.663

Seitz, A. R., Kim, R., & Shams, L. (2006). Sound facilitates visual learning. *Current Biology, 16*, 1422–1427. https://doi.org/10.1016/j.cub.2006.05.048

Sekuler, R., Sekuler, A. B., & Lau, R. (1997). Sound alters visual motion perception. *Nature, 385*, 308. https://doi.org/10.1038/385308a0

Shams, L., Kamitani, Y., & Shimojo, S. (2000). Illusions: What you see is what you hear. *Nature, 408*, 788. https://doi.org/10.1038/35048669

Shams, L., Wozny, D. R., Kim, R. S., & Seitz, A. (2011). Influences of multisensory experience on subsequent unisensory processing. *Frontiers in Psychology, 2*, 264. https://doi.org/10.3389/fpsyg.2011.00264

Sinnett, S., Spence, C., & Soto-Faraco, S. (2007). Visual dominance and attention: The Colavita effect revisited. *Perception & Psychophysics, 69*, 673–686. https://doi.org/10.3758/BF03193770

Stein, B. E., Stanford, T. R., Ramachandran, R., Perrault, T. J., & Rowland, B. A. (2009). Challenges in quantifying multisensory integration: Alternative criteria, models, and inverse effectiveness. *Experimental Brain Research, 198*, 113–126. https://doi.org/10.1007/s00221-009-1880-8

Stubblefield, A., Jacobs, L. A., Kim, Y., & Goolkasian, P. (2013). Colavita dominance effect revisited: The effect of semantic congruity. *Attention, Perception, & Psychophysics, 75*, 1827–1839. https://doi.org/10.3758/s13414-013-0530-1

Snodgrass, J. G. (1984). Concepts and their surface representations. *Journal of Memory and Language, 23*, 3–22. https://doi.org/10.1016/S0022-5371(84)90479-1

Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics, 73*, 971–995. https://doi.org/10.3758/s13414-010-0073-7

Sperdin, H. F., Cappe, C., Foxe, J. J., & Murray, M. M. (2009). Early, low-level auditory-somatosensory multisensory interactions impact reaction time speed. *Frontiers in Integrative Neuroscience, 3*, 2.

Standing, L. (1973). Learning 10000 pictures. *The Quarterly Journal of Experimental Psychology, 25*, 207–222. https://doi.org/10.1080/14640747308400340

Stevenson, R. A., Ghose, D., Fister, J. K., Sarko, D. K., Altieri, N. A., Nidiffer, A. R., et al. (2014). Identifying and quantifying multisensory integration: A tutorial review. *Brain Topography, 27*, 707–730. https://doi.org/10.1037/0033-2909.88.3.638

Stevenson, R. A., Zemtsov, R. K., & Wallace, M. T. (2012). Individual differences in the multisensory temporal binding window predict susceptibility to audiovisual illusions. *Journal of Experimental Psychology: Human Perception and Performance, 38*, 1517–1529. https://doi.org/10.1037/a0027339

Tan, J. S., & Yeh, S. L. (2015). Audiovisual integration facilitates unconscious visual scene processing. *Journal of Experimental Psychology: Human Perception and Performance, 41*, 1325–1335. https://doi.org/10.1037/xhp0000074

Taylor, K. I., Stamatakis, E. A., & Tyler, L. K. (2009). Crossmodal integration of object features: Voxel-based correlations in brain-damaged patients. *Brain, 132*, 671–683. https://doi.org/10.1093/brain/awn361

Thelen, A., Talsma, D., & Murray, M. M. (2015). Single-trial multisensory memories affect later auditory and visual object discrimination. *Cognition, 138*, 148–160. https://doi.org/10.1016/j.cognition.2015.02.003

Thompson, V. A., & Paivio, A. (1994). Memory for pictures and sounds: Independence of auditory and visual codes. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale, 48*, 380–398. https://doi.org/10.1037/1196-1961.48.3.380

Tombu, M. N., Asplund, C. L., Dux, P. E., Godwin, D., Martin, J. W., & Marois, R. (2011). A unified attentional bottleneck in the human brain. *Proceedings of the National Academy of Sciences, 108*, 13426–13431. https://doi.org/10.1073/pnas.1103583108

van der Burg, E., Olivers, C. N. L., Bronkhorst, A. W., & Theeuwes, J. (2008). Pip and pop: Non-spatial auditory signals improve spatial visual search. *Journal of Experimental Psychology: Human Perception and Performance, 34*, 1053–1065. https://doi.org/10.1037/0096-1523.34.5.1053

van der Burg, E., Talsma, D., Olivers, C. N., Hickey, C., & Theeuwes, J. (2011). Early multisensory interactions affect the competition among multiple visual objects. *NeuroImage, 55*, 1208–1218. https://doi.org/10.1016/j.neuroimage.2010.12.068

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia, 45*, 598–607. https://doi.org/10.1016/j.neuropsychologia.2006.01.001

Võ, M. L. H., & Wolfe, J. M. (2013). The interplay of episodic and semantic memory in guiding repeated search in scenes. *Cognition, 126*, 198–212. https://doi.org/10.1016/j.cognition.2012.09.017

von Kriegstein, K., & Giraud, A. L. (2006). Implicit multisensory associations influence voice recognition. *PLoS Biology, 4*, e326. https://doi.org/10.1371/journal.pbio.0040326

Vroomen, J., Bertelson, P., & De Gelder, B. (2001). The ventriloquist effect does not depend on the direction of automatic visual attention. *Perception & Psychophysics, 63*, 651–659. https://doi.org/10.3758/BF03194427

Warren, D. H., Welch, R. B., & McCarthy, T. J. (1981). The role of visual-auditory "compellingness" in the ventriloquism effect: Implications for transitivity among the spatial senses. *Attention, Perception, & Psychophysics, 30*, 557–564. https://doi.org/10.3758/BF03202010

Welch, R. B., & Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin, 88*, 638–667. https://doi.org/10.1037/0033-2909.88.3.638

Wheeler, M. E., Petersen, S. E., & Buckner, R. L. (2000). Memory's echo: Vivid remembering reactivates sensory-specific cortex. *Proceedings of the National Academy of Sciences, 97*, 11125–11129. https://doi.org/10.1073/pnas.97.20.11125

Yuval-Greenberg, S., & Deouell, L. Y. (2009). The dog's meow: Asymmetrical interaction in cross-modal object recognition. *Experimental Brain Research, 193*, 603–614. https://doi.org/10.1007/s00221-008-1664-6