Multisensory Integration in Memory:
Does Audiovisual Alignment Facilitate Memory Retrieval?

Deniz Sidi Sabanay
Erasmus University Rotterdam
Student number: 601271

April 3, 2026

Thesis advisor: Diane Pecher
Independent Reviewer: René Zeelenberg

word count: 6628

Abstract

Multisensory integration is often assumed to enhance memory performance by strengthening the encoding of information presented across multiple sensory modalities. The present study investigated whether audiovisual congruency during encoding improves recognition memory by directly comparing congruent audiovisual, incongruent audiovisual, and unimodal visual encoding conditions. Recognition memory and confidence ratings were subsequently assessed using an old/new recognition task in which participants were instructed to distinguish previously encountered stimuli from novel, distractor items. Importantly, participants were informed that some of the videos had been studied with sound, but that the test stimuli would be presented in a unimodal, visual-only format. Contrary to the hypothesis derived from multisensory integration theory, recognition memory and confidence ratings did not differ significantly across congruent audiovisual, incongruent audiovisual, and unimodal conditions. These findings indicate that integration of multiple sensory modalities alone does not necessarily enhance memory representations. Overall, the findings of the current study suggest that multisensory integration does not enhance recognition memory. Instead, the results suggest that future research should emphasize the importance of overlap between encoding and recognition conditions and use more complex, real-life, representative stimuli when investigating the effects of multisensory integration.

Keywords: multisensory integration, semantic congruency, audiovisual integration

Multisensory Integration in Memory: Does Audiovisual Alignment Facilitate Memory Retrieval?

Human perception is fundamentally multisensory. The brain constantly integrates information from diverse sensory modalities, including vision, hearing, and touch, to construct a coherent concept of the environment (Quak et al., 2015). Currently, the sensory information is presumed to be processed first in modality-specific regions and subsequently integrated in multisensory areas (Florio, 2025). This integration of various inputs from different sensory modalities to create a coherent mental representation is called multisensory integration, and it has been shown to enhance recognition and improve memory performance (Meyerhoff & Huff, 2016; Molholm et al., 2004; Thelen et al., 2015). The Multisensory Integration Theory (Stein & Meredith, 1993) offers a neurobiological framework for understanding this principle of how the brain integrates information across sensory modalities. According to this theory, the brain creates a coherent, unified representation by combining sensory information from temporally synchronized signals across multiple sensory modalities (Marucci et al., 2021). Behavioral studies have demonstrated the advantage of such integration, showing that multisensory stimuli often lead to faster and more accurate responses than the predicted cumulative effect of two unisensory inputs (Pannunzi et al., 2014).

From a neuroscientific perspective, the superficial layers of the superior colliculus correspond to visual space, and neurons in the deep layers respond to both auditory and somatosensory stimuli (Florio, 2025; May, 2006). Thus, integrating information from diverse sensory modalities, the superior colliculus and its associated cortical areas function as convergence zones, creating a unified, coherent representation (Florio, 2025; Stein et al., 2014; Woods & Newell, 2004). During this integration, spatially and temporally aligned signals originating from the same event converge at the same location on the motor map (Florio, 2025; Liu et al., 2022). The main principle that facilitates this integration is the unity assumption, the cognitive principle that two or more sensory cues will bind together if they are perceived as coming from a common source (Chen & Spence, 2017; Spence, 2007). The brain integrates different sensory inputs from different sensory modalities to create a single representation using the unity assumption (Chen & Spence, 2017; Spence, 2011). Therefore, it improves a memory trace of an event by functioning as a cognitive modulator of multisensory integration (Chen & Spence, 2017).

Semantic congruency is one of the higher-order cognitive factors that has been shown to further facilitate this integration (Laurienti et al., 2004; Molholm et al., 2004). The congruency occurs when auditory and visual inputs share matching semantic information, such as pairing a dog's bark with an image of a dog. Then the information from both sensory modalities is integrated, leading to faster and improved memory performance (Laurienti et al., 2004; Li & Deng, 2022). Congruent multisensory stimuli produce superadditive neural responses that improve perceptual processing and memory encoding. For example, Matusz et al. (2017) found that when auditory and visual information convey congruent information, this audiovisual congruency improves memory recall by activating long-term memory associations. Although their results did not specifically address whether the observed effect originated from processes occurring at encoding or retrieval, they suggested that semantically coherent multisensory presentations through encoding might strengthen memory representations. In contrast, when a semantically incongruent stimulus is presented, in which the auditory and visual cues conflict, such as when a dog picture is paired with a bird chirping sound, additional cognitive resources are required to resolve the conflict (Morey & Cowan, 2005). Therefore, in incongruent pairings, the efficacy of the memory encoding process and the accuracy of later retrieval are impaired (Roberts et al., 2024). Taken together, the evidence for semantic congruency effects in perception and memory lays a foundation for investigating the mechanisms by which multisensory alignment influences memory performance.

These empirical findings can further be explained by the Conceptual Short Term Memory (CSTM) model, which provides a theoretical framework to explain how encoding is improved for congruent stimuli, particularly when extended to the multisensory domain (Potter, 1993). According to this framework, CSTM functions as a memory buffer that rapidly combines recent perceptual information with previously stored information in the long-term memory to create new representations. Specifically, congruent visual information is rapidly recognized at the conceptual level and then activates the relevant information in long-term memory. The CSTM builds new connections among active concepts based on the perceptual characteristics of the stimuli, the current task goals, and existing long-term memory. After that, the structure can either be retained in working memory or integrated into long-term memory (Heikkilä et al., 2015; Potter, 1993).

In line with this model, when a congruent auditory stimulus is presented concurrently with a visual stimulus, the sound provides supplementary coherent information, facilitating the encoding of the visual representation into the CSTM (Chen & Spence, 2010). This multisensory integration results in stronger associations and more enduring memory traces, as supported by the findings of Heikkilä et al. (2015). Their research showed that encoding with semantically congruent multimodal stimuli improved recognition memory compared to unisensory visual encoding (Heikkilä et al., 2015). In contrast, when auditory and visual stimuli are incongruent, the lack of semantic congruency hinders memory formation, making it harder to form strong associations. This mismatch often creates confusion and reduced confidence, even when participants' recognition is accurate (Laurienti et al., 2004; Matusz et al., 2017). This conclusion implies that congruent bimodal stimulation enhances both subjective confidence in the perceived target and objective performance. According to the principle of Bayesian causal inference, the brain simultaneously assesses whether the cues originate from a single event or from two distinct occurrences, as well as the perceptual content itself (Meijer & Noppeney, 2023). Therefore, the brain employs an interactive mechanism. When it chooses separate causes, it evaluates the cues independently, but when it chooses a single-cause interpretation, thereby integrating multiple sensory modalities, it forms a unified percept.

This interpretation of multisensory facilitation is further supported by the research of Molholm et al. (2004), which demonstrated that the neural response to congruent audiovisual stimuli, particularly the amplitude of the visual N1 component, was significantly greater than for unimodal or incongruent conditions. Their study revealed that participants were faster and more accurate in object recognition tasks when auditory and visual cues conveyed semantically congruent information, compared with when such cues were presented in an unimodal format or when they were paired incongruently (Molholm et al., 2004). Importantly, the heightened neural response was localized to regions of the ventral visual stream associated with initial early object recognition, and the authors interpreted this increase as evidence of more effective multisensory integration rather than increased processing difficulty. Their control analyses also showed that the observed impact resulted from the summation of unisensory responses, suggesting that the integration of congruent audiovisual information can enhance perceptual processing at the early stages of neural analysis (Molholm et al., 2004). This evidence from prior research suggests that multisensory integration strengthens memory representations through a redintegration process, in which congruent auditory and visual information enhances recall.

An alternative view, proposed by Paivio's Dual-Coding Theory (1971), is that the brain uses two separate systems to represent information: a verbal (linguistic) code and a nonverbal (imaginal) code. These systems function through distinct representational units known as logogens, which encode linguistic elements such as words, and imagens, which encode perceptual and non-verbal entities such as objects or scenes (Paivio, 1971). These two systems are functionally independent yet interrelated, indicating that each can function independently while also engaging through two types of connections. Associative connections, which form relationships within the same system (e.g., word to word or image to image associations), and referential connections, which link the two systems (e.g., a word eliciting an image or an image evoking a verbal label). The greater the number of referential connections our brain creates, the more elaborate the memory representation becomes (Schnotz & Horz, 2010). Existing referential connections also make it easier to interpret new stimuli and create further associations that link this new information to pre-existing cognitive structures.

On the other hand, the selective interference effect occurs when a person attempts to perform two mental tasks that depend on the same modality, such as two verbal or two imagery tasks. In such situations, individuals perform more slowly and make more errors than when one of the tasks involves a different code (e.g., a verbal task paired with an imagery task) (Laurienti et al., 2004; Li & Deng, 2022). This shows that tasks using the same sensory modality interfere with each other as they both share the same representational and processing resources, providing support for the dual-coding theory (Brooks, 1967; De Beni & Moè, 2003). On the other hand, when an individual is presented with a visual stimulus associated with the verbal information or can visualize a mental image that corresponds to the verbal content, their memory for that information is improved (Anderson & Bower, 2014). In line with this finding, Thompson and Paivio (1994) found that multimodal congruent stimuli were recalled better than stimuli presented in a single modality (unimodal). Thus, dual-coding theory explains the memory benefits of congruent audiovisual stimuli through parallel verbal and nonverbal codes, differentiating it from multisensory integration theory's focus on integrating different sensory modalities into a unified representation (Heikkilä et al., 2015).

However, empirical evidence on the benefits of multisensory memory remains mixed, as some studies fail to show substantial memory enhancements for multimodal presentations (Cohen et al., 2009). Heikkilä et al. (2015) reported mixed results when adult participants intentionally memorized pictures paired with an auditory stimulus. In one experiment, objects paired with congruent sounds were not remembered significantly better than those paired with incongruent sounds or no sound, suggesting null effects of semantic congruency for multisensory stimuli. Similarly, Pecher and Zeelenberg (2022) obtained little evidence of a memory benefit when comparing a bimodal (audio+visual) study to a unimodal study. Recognition performance for items learned with both sound and image was comparable to that for items learned with a single modality.

On the other hand, Meyerhoff and Huff (2016) reported benefits of semantic congruency; however, their findings also indicated that the memory advantage for audiovisual stimuli was not consistently robust across all conditions and that effects could be diminished in the absence of temporal synchrony or when auditory and visual inputs were presented separately. However, it is important to note that, across all their experiments, the presentation modality remained consistent between the study and test phases. In their first experiment, participants were shown audiovisual, visual-only, and auditory-only clips during the study phase. Subsequently, during the test phase, participants encountered the stimuli in the same modality as in the study phase; for example, if they encountered a stimulus in an audiovisual format during the study phase, they were again presented in the audiovisual format during the testing phase (Meyerhoff & Huff, 2016). Similarly, in their second experiment, when they manipulated semantic congruency in the stimuli, congruent and incongruent audiovisual combinations were the same for each participant during the learning and testing phases (Meyerhoff & Huff, 2016). Therefore, the findings do not explain the performance differences when the sensory modality differs between the study and test phases.

The benefit of the study-test modality overlap can be explained by the principles of encoding specificity and transfer-appropriate processing. The encoding specificity principle states that memory performance will be enhanced if the context present at encoding and that present at subsequent retrieval are congruent (Tulving & Thomson, 1973). Therefore, memory is influenced not only by the encoding experience but also by the similarity between encoding and retrieval contexts (Park & Rugg, 2007). Building on this, the framework of transfer-appropriate processing states that memory performance is enhanced when the cognitive processing engaged during encoding is re-engaged during testing (Morris et al., 1977). Together, the significance of study-test overlap in understanding the interaction between encoding and retrieval is fundamental to both principles (Park & Rugg, 2007). These findings suggest that the advantage of multisensory memory may depend on additional components of multisensory integration, including temporal alignment and the encoding-retrieval context, rather than being limited to semantic congruency.

One potential explanation for these mixed results is the complexity of the stimuli and the ecological validity of prior research. Many earlier studies relied on simplistic, artificial stimuli such as basic line drawings of objects paired with isolated sounds or spoken words presented out of context (Matusz et al., 2017; Roberts et al., 2024; Thelen et al., 2015). Although these simplified stimuli facilitate experimental control, they may not accurately represent the conditions in which encoding usually occurs in the real world. This raises concerns regarding the ecological validity of such studies and the generalizability of their findings. Multisensory benefits may emerge reliably only when stimuli are sufficiently complex or meaningful to engage typical integrative processes. Therefore, a clear research gap exists: it is necessary to determine whether the benefits of semantic congruency for memory persist under ecologically valid conditions using naturalistic stimuli.

The present study aims to address these gaps by using dynamic audiovisual stimuli (short video clips) intended to more accurately depict real-world situations. In addition, unlike many prior experiments, this study explicitly includes a unimodal condition (visual-only presentations) as a baseline for comparison. This allows us to examine whether congruent audiovisual pairs enhance recognition memory beyond what visual modality alone can achieve and whether incongruent pairs impair memory relative to unimodal exposure. Specifically, we compare three conditions at encoding: (1) Congruent Audiovisual, where a video clip is paired with a semantically congruent sound; (2) Incongruent Audiovisual, where the video clip is paired with an incongruent sound; and (3) Unimodal (visual only), where the video clip is presented without sound.

The main question addressed in this study is whether semantic congruence between auditory and visual stimuli during encoding leads to higher recognition accuracy and confidence in later memory retrieval compared to incongruent and unimodal conditions. Based on the dual-coding theory and previous studies on multisensory integration (Meyerhoff & Huff, 2016; Molholm et al., 2004; Paivio, 1971; Thelen et al., 2015), we hypothesize that recognition accuracy and confidence ratings will be highest in the congruent audiovisual condition, followed by the unimodal condition, and lowest in the incongruent audiovisual condition. Congruent audiovisual stimuli are predicted to result in the most enhanced memory performance by engaging both sensory visual and auditory modalities, using dual memory codes, and stimulating deeper semantic processing, therefore increasing confidence ratings in memory performance. The unimodal (visual-only) condition serves as the baseline in this experiment and is predicted to yield higher recognition accuracy and confidence ratings than the incongruent condition, as it does not employ competing or conflicting sensory cues during encoding. On the other hand, based on the interference account, incongruent audiovisual pairs are expected to induce cognitive conflict and load (Morey & Cowan, 2005), therefore preventing integration of different sensory cues into a unified representation. Therefore, it is predicted that incongruent pairs will yield lower memory accuracy compared to congruent and unimodal pairs.

Method

Participants

A total of 60 participants were recruited from the university community through communication platforms. The sample consisted of 10 males (16.7%) and 50 females (83.3%). A power analysis carried out using G*Power (Faul et al., 2009) determined that a minimum sample size of 31 participants was required to achieve 95% power with an effect size of $f = 0.30$ in a repeated-measures ANOVA design. However, to ensure sufficient statistical power and account for potential dropouts, 60 participants were recruited for the study. Participants were between 17 and 25 years old ($M = 21.70$), and all reported having normal or corrected-to-normal vision and hearing. Participants either received course credits for their participation or volunteered without compensation. Informed consent was obtained prior to participation, and participants were informed of their right to withdraw at any time.

Apparatus

Stimuli were presented using E-Prime software (version 3.0; Psychology Software Tools, Pittsburgh, PA) on a 22-inch computer monitor with a resolution of 1920 x 1200 pixels. The monitor was positioned approximately 60 cm from the participants at eye level, while auditory stimuli were presented through headphones. Responses were recorded using a standard keyboard positioned approximately 30 cm in front of the monitor.

Materials

This study consisted of video clips featuring objects paired with corresponding sounds (congruent condition), non-corresponding sounds (incongruent condition), or omitted sounds (unimodal condition). A comprehensive list of potential stimuli was first compiled, focusing on objects and actions that (a) are visually distinctive, (b) produce recognizable sounds, and (c) could be feasibly depicted with motion (see Appendix A). These stimuli were systematically drawn from diverse categories, including animals (e.g., dog barking, cat meowing), environmental events (e.g., rain falling, ocean waves), household objects (e.g., phone ringing, scissors cutting), vehicles/mechanical (e.g., motorcycle, car), and other actions (e.g., frying, pouring water, camera click). Each category contributed approximately an equal number of items, ensuring a broad sampling of sensory contexts. These 60 items were then systematically divided into three lists of 20 items each, corresponding to the three experimental conditions, while maintaining a comparable distribution of categories across lists (see Appendix B). In the incongruent condition, mismatched pairs were intentionally created by pairing items from different categories. This ensured that the audiovisual combinations could not be interpreted as belonging to the same semantic group. This procedure reduced the risk that within-category similarity would otherwise create, therefore strengthening the manipulation of incongruency.

To maintain both experimental control and ecological validity, stimuli were obtained from publicly available online databases. Employing real-world materials from platforms Pixabay and Pexels, along with open-access sound libraries, enhanced ecological validity by reflecting how sensory information is typically encountered in everyday life. Videos and sounds were carefully selected to maintain consistency in visual and acoustic quality while reducing variability that often occurs in naturalistic recordings, such as inconsistent camera angles and motion artifacts. This allowed any observed effects to be attributed to the experimental manipulation, specifically semantic congruency, rather than perceptual noise. After the video clips and audio files were sourced, the corresponding auditory stimuli were manually synchronized with the videos using CapCut video editing software. This process ensured that each audio file is properly aligned with the video content, preventing any unintended temporal mismatches.

Video clips depicted various objects or animals in naturalistic settings, with auditory stimuli systematically paired according to three experimental conditions: congruent, incongruent, and unimodal. In the congruent condition, auditory and visual stimuli were semantically congruent; a video of a dog was paired with the sound of a dog barking. In contrast, in the incongruent condition, audio-visual stimuli were intentionally mismatched (e.g., a video of a dog with a bird chirping sound). In the unimodal condition, participants viewed the video stimuli with the omitted sound. Each video clip lasted 2500 ms, providing sufficient time for participants to process both visual and auditory information.

Procedure

Upon arrival, participants received a brief introduction and detailed instructions about the experimental tasks. The experiment began with the study trials, each starting with a 500 ms fixation marker (+), followed by a 2500 ms stimulus. The order of stimulus presentation was fully randomized and counterbalanced across participants. After each stimulus presentation, a blank screen appeared for an inter-stimulus interval (ISI) of 1500 ms. During the study phase, participants watched 60 videos, of which 20 were congruent (video and audio matched), 20 were incongruent (video and audio mismatched), and 20 were unimodal (video-only).

Following the study phase, participants completed a filler task designed to ensure reliance on long-term memory processes during retrieval. This filler task consisted of a series of simple arithmetic problems (single-digit addition, subtraction, and multiplication) and was intended to occupy working memory and prevent rehearsal of the encoded items.

After the filler task, participants proceeded to the recognition test phase, in which they encountered 120 stimuli: 60 previously presented (old) and 60 novel (new). Participants were instructed that a list of videos they had to remember would be shown on the monitor and that they had to decide, for each video, whether it was 'old' or 'new'. Additionally, they were informed that some of the videos had been studied with sound, but that the old-new decision should be based only on the sound-omitted videos. Therefore, the test stimuli were presented in a unimodal format (visual only). As participants saw a stimulus, the question 'On a 6-point scale, how confident are you that this item is old or new' was presented. Participants were instructed to respond using a combined recognition-confidence judgment on a 6-point scale: 1 (sure new), 2 (probably new), 3 (guess new), 4 (guess old), 5 (probably old), and 6 (sure old). No feedback was provided during the experimental trials.

Data Processing and Statistical Analyses

All individual E-Prime output files were merged into a single dataset and filtered to include only the trials from the recognition phase. Each data point contained the subject number, age, gender, study condition (congruent, incongruent, unimodal, or distractor), the name of the stimulus studied, and a 6-point confidence rating indicating whether the item was judged as previously studied. No participants were excluded from the analysis.

Recognition responses (confidence ratings) were first recoded into a binary old/new decision. Confidence ratings of 1 through 3 were classified as "new," whereas ratings of 4 through 6 were classified as "old". The accuracy was then coded relative to each item's true status. For previously studied items (congruent, incongruent, and unimodal), an "old" response was coded as a hit, and a "new" response as a miss. For distractor items, a "new" response was coded as a correct rejection, and an "old" response was coded as a false alarm. Hit rates (H) were computed separately for each of the three study conditions, and a single false alarm rate (FA) was computed from responses to distractor items.

Because proportions of 0 or 1 yield infinite z-scores in signal detection analyses, all hit and false alarm rates were corrected using Snodgrass and Corwin (1988) log-linear adjustment. This correction adds 0.5 to the number of hits and false alarm, and 1 to the number of trials. Therefore, the corrected rates were computed as:

$$H_{\text{corrected}} = \frac{H + 0.5}{N_{\text{old}} + 1}$$

$$FA_{\text{corrected}} = \frac{FA + 0.5}{N_{\text{new}} + 1}$$

Sensitivity, measured as d-prime (d'), was then calculated for each participant and each study condition using the following standard signal detection formula:

$$d' = z(H_{\text{corrected}}) - z(FA_{\text{corrected}})$$

where z denotes the inverse of the standard normal distribution. This procedure yielded three d' values (one per study condition) per participant.

Inferential analyses were performed using SPSS (version 29). A one-way repeated-measures analysis of variance (ANOVA) was conducted with the study condition (congruent, incongruent, and unimodal) as the within-subject factor. The assumption of sphericity was evaluated using Mauchly's test. Planned pairwise comparisons were conducted using paired-samples t-tests to examine the differences in sensitivity between the different experimental conditions. All significance tests were two-tailed with $\alpha = .05$, and effect sizes were reported using Cohen's d, with 95% confidence intervals.

In addition to computing d', receiver operating characteristic (ROC) curves were generated for each study condition to obtain a threshold-independent assessment of the recognition performance. Whereas d' reflects sensitivity at only the specific decision threshold used to classify responses as "old" or "new", ROC analysis considers performance across all possible response criteria. For each confidence boundary, the corresponding hit and false-alarm rates were calculated, resulting in three curves per study condition that characterize recognition accuracy across the range of decision thresholds.

Results

Corrected hit rates were calculated using the Snodgrass and Corwin (1988) correction for each encoding condition to provide a descriptive summary of recognition performance. Corrected false-alarm rates were identical across conditions ($M = .27$, $SD = .18$), because new items were not associated with specific study conditions. The corrected hit rates presented in Table 1 revealed only minimal variation across congruent, incongruent, and unimodal items, indicating the pattern observed in the raw hit rates (see Appendix C, Table C1). Sensitivity scores (d') were then calculated for each study condition using the corrected hit and false alarm rates. The descriptive statistics indicated minimal variation across the three study conditions. As depicted in Table 1, the mean sensitivity was highest in the congruent condition, followed by the incongruent and unimodal conditions.

Table 1

Signal Detection Measures by Study Condition

Conditions	Corrected Hits		Sensitivity (*d'*)	
	Mean	SD	Mean	95% CI
Congruent	.83	.10	1.83	[1.58, 2.09]
Incongruent	.82	.10	1.73	[1.51, 1.96]
Unimodal	.80	.11	1.71	[1.48, 1.94]

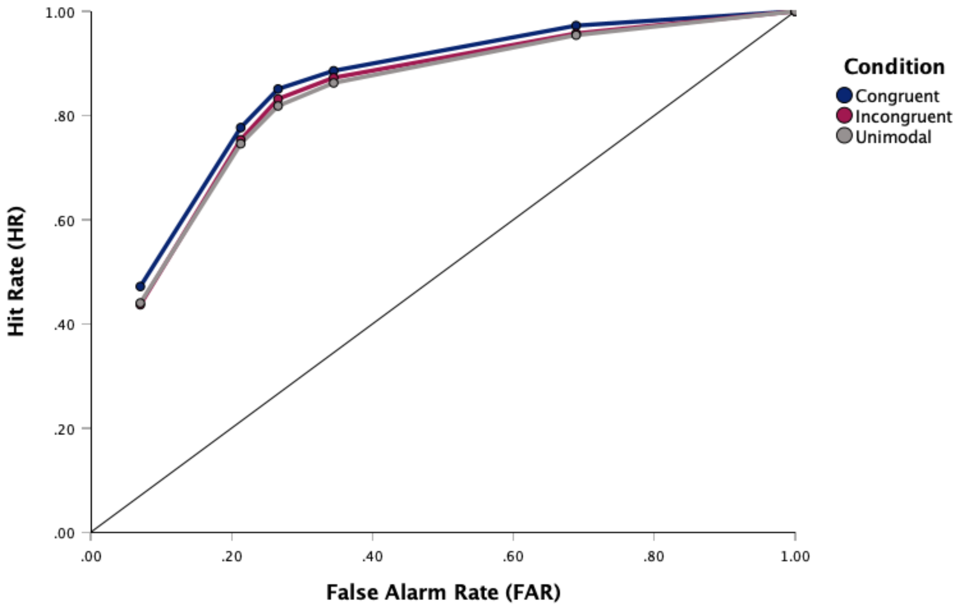
Although these means differed slightly, the descriptive pattern alone did not indicate any systematic effect of the study condition on recognition performance. Therefore, a repeated-measures analysis of variance (ANOVA) was conducted to determine whether recognition sensitivity differed across study conditions. The analysis revealed no significant effect of condition on sensitivity, $F(2, 118) = 1.57$, $p = .212$, partial $\eta^2 = .026$, suggesting that the study condition did not significantly affect recognition performance. The ANOVA results are presented in Appendix C (Table C2).

To further investigate the hypothesized differences among study conditions, paired-samples *t*-tests were conducted. None of the contrasts was statistically significant. Sensitivity did not differ between congruent and incongruent conditions, $t(59) = 1.41$, $p = .163$, Cohen's $d = 0.18$, between the congruent and unimodal conditions, $t(59) = 1.63$, $p = .109$, Cohen's $d = 0.21$ and between the incongruent and unimodal conditions, $t(59) = 0.35$, $p = .727$, Cohen's $d = 0.05$ (see Appendix C, Table C3). These results further support the conclusion that recognition sensitivity did not differ across different study conditions.

To further examine recognition performance independent of decision thresholds, ROC curves were computed for each study condition (Figure 1). All three ROC curves were above the diagonal, indicating that performance reliably exceeded chance. The curve for the congruent condition was slightly elevated relative to the incongruent and unimodal conditions, showing higher hit rates for comparable false alarm rates. Despite this, the overall disparity between the curves was small, consistent with the subtle differences seen in the sensitivity scores. These findings indicate limited variation in recognition discriminability across the different experimental conditions.

Figure 1

ROC Curves for Congruent, Incongruent, and Unimodal Study Conditions



Across all measures, including hit rates, corrected proportions, signal-detection indices, and ROC curves, the differences among the congruent, incongruent, and unimodal conditions were small and did not reach statistical significance. Although the results consistently showed a slight numerical advantage for congruent audiovisual pairings, this pattern was not supported by inferential analyses. Overall, the findings indicate that recognition performance remained largely comparable across study conditions and that the manipulation of semantic congruency did not produce systematic differences in recognition sensitivity.

Discussion

The aim of the present study was to investigate how multisensory information is represented in memory by examining whether congruency between auditory and visual information during encoding influences subsequent recognition memory. Specifically, we examined whether memory performance differed across stimuli encoded in multisensory (audiovisual) format, which were either semantically congruent or incongruent, or a unimodal (visual-only) format. By directly comparing these three different encoding conditions, the present research aimed to clarify whether multisensory memory benefits are explained by integrative accounts of multisensory processing or by alternative explanations emphasizing modality-specific mechanisms.

The results of the present study do not provide support for the initial hypothesis that recognition accuracy and confidence ratings will be highest in the congruent audiovisual condition, followed by the unimodal condition, and lowest in the incongruent audiovisual condition. No statistically significant differences in recognition performance were observed between the three encoding conditions. Therefore, the results also did not support the prediction that congruent audiovisual encoding would produce better recognition accuracy and increased confidence ratings in memory than in the incongruent and unimodal conditions. Comparably, the prediction that incongruent audiovisual stimuli would yield the lowest memory performance and confidence ratings is also not supported. Overall, the present findings do not support a significant benefit of audiovisual congruency for recognition memory.

The lack of benefit for congruent audiovisual stimuli suggests that, given the present experimental conditions, multisensory congruency did not enhance recognition performance. Integrative theories of multisensory memory suggest that a unified representation is created through the integration of synchronized signals from diverse sensory modalities (Marucci et al., 2021). Therefore, multimodal stimuli often lead to faster and more accurate responses than those predicted by the combined effects of two separate unisensory inputs (Pannunzi et al., 2014). Similarly, the unity assumption proposes that semantically and temporally aligned audiovisual inputs are integrated into unified representations during the encoding phase, which in turn, facilitates later recognition (Chen & Spence, 2017; Thelen et al., 2015). This assumption of unity suggests that audiovisual integration strengthens the memory trace of the event (Chen & Spence, 2017). If such integration resulted in stronger or more distinctive memory traces, a clear benefit for congruent audiovisual stimuli would be expected in sensitivity (d′). Although previous research has shown evidence of enhanced processing of congruent multisensory stimuli (Chen & Spence, 2017; Spence, 2011), the present findings indicate no evidence of multisensory integration.

The lack of advantage in semantic congruency also questions Paivio’s dual coding theory (Paivio, 1971). The framework states that through distinct but interconnected systems, verbal and nonverbal, information is encoded, making memory more robust and accessible (Paivio, 1971). When the verbal and non-verbal systems reinforce each other, a coherent mental representation can be created. In contrast, the current findings suggest that these systems for integrating multiple senses may be insufficient. A possible explanation for the absence of a significant congruency effect is that participants may have relied predominantly on nonverbal, visual information during both encoding and retrieval. Because the recognition task required discrimination of only visual presentations, the visual features may have provided sufficient information for recognition, consequently reducing the potential contributions of the auditory modality during the study phase. Furthermore, from prior research, it is also known that an individual's inherent preference for a specific modality can prioritize information encoding. Information from other senses may then be obscured by information from the dominant modality (Li & Deng, 2022). Under such circumstances, congruent auditory information may not significantly improve memory performance beyond what is already supported by visual information alone.

This interpretation is also consistent with the principle of encoding specificity and transfer-appropriate processing. According to these frameworks, memory performance is largely dependent upon the similarity between encoding and retrieval tasks (Ritchey et al., 2012). Although the visual component in this study was identical across phases, maintaining significant study–test overlap, the overall alignment between encoding and retrieval was incomplete. The auditory modality present during the study phase was not reinstated during the test phase, which may have limited the extent to which multimodal representations could improve recognition performance. Therefore, the present findings are consistent with the prior research supporting modality overlap between study and test phases (Meyerhoff et al., 2022; Pecher et al., 2025; Pecher & Zeelenberg, 2022). For instance, Meyerhoff et al. (2022) investigated whether the modality available during the encoding phase matched the modality available at the test phase to assess the effect of study-test correspondence on congruency. Audiovisual clips used in the encoding phase were divided so that half of the stimuli were tested in the same audiovisual format and the other half were tested only visually. The same manipulation was also applied to clips encoded only visually; they were later tested in either the audiovisual or the visual-only condition. When the stimuli were studied audiovisually, recognition memory performance was more accurate when tested in audiovisual format than when tested only visually (Meyerhoff et al., 2022). This suggests that the presence of both modalities during testing can enhance memory if both were also present during the study.

Building on this perspective, Pecher et al. (2025) provided evidence that multisensory advantages are better explained by study–test modality overlap rather than by multisensory integration itself. Across several experiments, they showed that recognition performance improves when the test modality is the same as the study modality, regardless of whether the original encoding condition was multimodal or unimodal (Pecher et al., 2025). When the test format differed from the encoding format, performance reliably declined. Therefore, if retrieval primarily reactivates modality-specific regions, then additional congruent information during encoding may not provide a memory benefit if it is not reinstated at the test phase. Together, these studies indicate that multisensory congruency effects depend on whether the retrieval environment reinstates the sensory conditions present at encoding.

The results of this study further showed that recognition sensitivity to incongruent audiovisual stimuli did not differ significantly from that to congruent audiovisual or unimodal stimuli. The existing body of research suggests that multisensory input is inherently beneficial for memory (Meyerhoff & Huff, 2016; Molholm et al., 2004; Thelen et al., 2015). Conversely, the present findings contradict this view by demonstrating that multisensory information does not invariably enhance memory performance. The lack of such an advantage in the current study suggests that additional sensory information does not enhance memory representations. This finding is consistent with prior studies suggesting that incongruent audiovisual stimuli might not benefit from multisensory encoding (Chen & Spence, 2010; Laurienti et al., 2004). This is because the semantic incongruency prevents effective integration across different sensory modalities (Chen & Spence, 2010; Laurienti et al., 2004). However, the present data indicate that this lack of integration did not result in a measurable cost relative to unimodal encoding. This pattern suggests that incongruent multisensory input neither enhances nor substantially disrupts recognition sensitivity, in the context of the current recognition task.

On the other hand, the interference account, which predicts that incongruent stimuli would induce cognitive conflict between the components of a multisensory stimulus (Morey & Cowan, 2005) and therefore create cognitive load, is also not supported. If this account held strongly in this experimental design, we would expect incongruent stimuli to be remembered noticeably worse than unimodal items. However, the current results show that performance for incongruent and unimodal stimuli was relatively similar, and no clear disadvantage for incongruent trials was found. Prior research has shown that incongruent stimuli create conflicting mental representations, which reduce accuracy and require attentional resources to resolve the conflict (Li & Deng, 2022), though such interference effects were not found in the present recognition task.

Consequently, recognition sensitivity for unimodal stimuli was comparable to that observed for both congruent and incongruent multisensory stimuli. This finding contradicts the multisensory integration theory, which predicts a consistent advantage for multimodal encoding over unimodal encoding (Thelen et al., 2015). Therefore, the present findings align more consistently with previous research suggesting that multisensory presentation does not yield more robust memory traces (Meyerhoff et al., 2022). One possible explanation for these findings is that unimodal stimuli may allow for more focused encoding within a single modality, therefore avoiding potential attentional allocation or processing demands associated with integrating information across different sensory modalities. Therefore, the lack of multisensory benefit does not imply a weakness of unimodal encoding, but instead it highlights that increased sensory input does not necessarily enhance memory performance.

While the present study contributes to understanding how semantic congruency across visual and auditory sensory modalities influences memory and confidence judgments, certain limitations should be acknowledged. First, the experimental design used a controlled laboratory setting with a restricted set of audiovisual stimuli representing everyday objects and events. Although the 2.5-second clips used in this experiment were more ecologically representative of real-life events than the static images used in prior studies (Matusz et al., 2017; Roberts et al., 2024; Thelen et al., 2015), they were still simple in structure. Specifically, compared to the complex, detailed, narrative-driven movie scenes used in earlier studies (Meyerhoff & Huff, 2016; Meyerhoff et al., 2022), the present stimuli lacked social interactions, dialogue, and extended event structures. As real-world experiences are usually more complex and dependent on their context, the simplicity of the stimuli might have hindered the engagement of higher-order integrative processes during encoding and therefore affected the recognition performance. However, the ecological validity of the stimuli alone is unlikely to explain the absence of multisensory integration in this study, as several prior studies using similar stimuli have also failed to demonstrate a multisensory memory advantage (Meyerhoff & Huff, 2016; Meyerhoff et al., 2022). Thus, future studies could benefit from using stimuli with dynamic scenes or immersive environments, such as virtual reality, to more accurately represent real-world encoding conditions.

Secondly, the manipulation used in this study was limited to semantic congruency among auditory and visual stimuli. This experimental design allowed the isolation of semantic congruency, one key aspect of multisensory integration (Laurienti et al., 2004), while neglecting other factors, such as spatial correspondence, that may affect memory and confidence ratings. In this experimental design, temporal synchronization was kept constant to ensure coherent, real-life representative audiovisual stimuli. However, because temporal synchronization was not independently manipulated, this study could not examine how it might interact with semantic congruency in shaping memory performance. Therefore, it remains unclear whether the absence of a statistically reliable congruency effect reflects constraints of semantic integration or whether the combination of semantic and spatial coherence is beneficial for measurable differences in recognition performance. Future research incorporating this additional dimension may clarify whether the observed effects in this study are specific to semantic processing or if they are indicative of more extensive integrative mechanisms.

Furthermore, the sample's composition, primarily consisting of university students within a narrow age range, could limit the generalizability of the results. Prior research has shown that experience-driven modifications in synaptic pathways targeting integrative structures, such as the superior colliculus, may affect the efficacy of sensory input integration (Stein et al., 2014). Due to limited understanding of how these circuits change with age and the lack of age diversity in our sample, we cannot determine whether the null congruency effect reflects young-adult processing or whether multisensory benefits might be more pronounced in other age groups. Therefore, extending this line of research to children, older adults, or clinical populations could offer insights about developmental or pathological differences in multisensory memory integration.

Summary and Conclusion

Multisensory integration is thought to enhance memory performance through integrating information from diverse sensory modalities, thereby providing more comprehensive representations than unisensory encoding. Semantically congruent audiovisual information is considered to enhance later recognition memory through facilitating the integration of information across sensory modalities. This study examined the impact of semantic congruency during encoding on subsequent recognition memory by comparing congruent audiovisual, incongruent audiovisual, and unimodal encoding conditions. It used short video clips that provide higher ecological validity than the static line drawings typically used in previous studies. The findings of this study demonstrate that recognition performance did not differ across three different encoding conditions, suggesting that multimodal presentation alone is insufficient to enhance recognition memory and increase confidence ratings. Therefore, potential benefits of multisensory encoding might depend on contextual factors such as audiovisual stimulus complexity and study-test correspondence. Overall, the present results contribute toward the body of research on multisensory memory by showing that semantic congruency alone is insufficient to enhance recognition, clarifying the limits of multisensory contributions under the conditions tested.

References

Anderson, J. R., & Bower, G. H. (2014). *Human associative memory*. Psychology Press.

Brooks, L. R. (1967). The suppression of visualization by reading. *Quarterly Journal of Experimental Psychology*, *19*(4), 289–299. https://doi.org/10.1080/14640746708400105

Chen, Y., & Spence, C. (2010). When hearing the bark helps to identify the dog: Semantically-congruent sounds modulate the identification of masked pictures. *Cognition*, *114*(3), 389–404. https://doi.org/10.1016/j.cognition.2009.10.012

Chen, Y., & Spence, C. (2017). Assessing the role of the 'Unity Assumption' on multisensory Integration: a review. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.00445

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*(14), 6008–6010. https://doi.org/10.1073/pnas.0811884106

De Beni, R., & Moè, A. (2003). Imagery and rehearsal as study strategies for written or orally presented passages. *Psychonomic Bulletin & Review*, *10*(4), 975–980. https://doi.org/10.3758/bf03196561

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/brm.41.4.1149

Florio, T. M. (2025). Emergent aspects of the integration of sensory and motor functions. *Brain Sciences*, *15*(2), 162. https://doi.org/10.3390/brainsci15020162

Heikkilä, J., Alho, K., Hyvönen, H., & Tiippana, K. (2015). Audiovisual semantic congruency during encoding enhances memory performance. *Experimental Psychology*, *62*(2), 123–130. https://doi.org/10.1027/1618-3169/a000279

Laurienti, P., Kraft, R., Maldjian, J., Burdette, J., & Wallace, M. (2004). Semantic congruence is a critical factor in multisensory behavioral performance. *Experimental Brain Research*, *158*(4). https://doi.org/10.1007/s00221-004-1913-2

Li, J., & Deng, S. W. (2022). Facilitation and interference effects of the multisensory context on learning: a systematic review and meta-analysis. *Psychological Research*, *87*(5), 1334–1352. https://doi.org/10.1007/s00426-022-01733-4

Liu, X., Huang, H., Snutch, T. P., Cao, P., Wang, L., & Wang, F. (2022). The Superior Colliculus: cell types, connectivity, and behavior. *Neuroscience Bulletin*, *38*(12), 1519–1540. https://doi.org/10.1007/s12264-022-00858-1

Marucci, M., Di Flumeri, G., Borghini, G., Sciaraffa, N., Scandola, M., Pavone, E. F., Babiloni, F., Betti, V., & Aricò, P. (2021). The impact of multisensory integration and perceptual load in virtual reality settings on performance, workload and presence. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-84196-8

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

May, P. J. (2006). The mammalian superior colliculus: laminar structure and connections. *Progress in Brain Research*, 151, 321–378. https://doi.org/10.1016/s0079-6123(05)51011-2

Meijer, D., & Noppeney, U. (2023). Metacognition in the audiovisual McGurk illusion: perceptual and causal confidence. Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 378(1886), 20220348. https://doi.org/10.1098/rstb.2022.0348

Meyerhoff, H. S., & Huff, M. (2016). Semantic congruency but not temporal synchrony enhances long-term memory performance for audio-visual scenes. *Memory & Cognition*, *44*(3), 390–402. https://doi.org/10.3758/s13421-015-0575-6

Meyerhoff, H. S., Jaggy, O., Papenmeier, F., & Huff, M. (2022). Long-term memory representations for audio-visual scenes. *Memory & Cognition*, *51*(2), 349–370. https://doi.org/10.3758/s13421-022-01355-6

Molholm, S., Ritter, W., Javitt, D. C., & Foxe, J. J. (2004). Multisensory Visual–Auditory Object Recognition in Humans: A High-density Electrical Mapping Study. Cerebral Cortex, 14(4), 452-465. https://doi.org/10.1093/cercor/bhh007

Morey, C. C., & Cowan, N. (2005). When do visual and verbal memories conflict? The importance of Working-Memory Load and Retrieval. *Journal of Experimental Psychology Learning Memory and Cognition*, *31*(4), 703–713. https://doi.org/10.1037/0278-7393.31.4.703

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*(5), 519–533. https://doi.org/10.1016/s0022-5371(77)80016-9

Paivio, A. (1971). Imagery and verbal processes. Holt, Rinehart & Winston.

Pannunzi, M., Pérez-Bellido, A., Pereda-Baños, A., López-Moliner, J., Deco, G., & Soto-Faraco, S. (2014). Deconstructing multisensory enhancement in detection. *Journal of Neurophysiology*, *113*(6), 1800–1818. https://doi.org/10.1152/jn.00341.2014

Park, H., & Rugg, M. D. (2007). The Relationship between Study Processing and the Effects of Cue Congruency at Retrieval: fMRI Support for Transfer Appropriate Processing. *Cerebral Cortex*, *18*(4), 868–875. https://doi.org/10.1093/cercor/bhm130

Pecher, D., & Zeelenberg, R. (2022). Does multisensory study benefit memory for pictures and sounds? *Cognition*, *226*, 105181. https://doi.org/10.1016/j.cognition.2022.105181

Pecher, D., Keytel, B., & Zeelenberg, R. (2025). Study-test overlap rather than multisensory integration benefits memory. *Memory & Cognition*. https://doi.org/10.3758/s13421-025-01759-0

Potter, M. C. (1993). Very short-term conceptual memory. *Memory & Cognition*, *21*(2), 156–161. https://doi.org/10.3758/bf03202727

Quak, 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

Ritchey, M., Wing, E. A., LaBar, K. S., & Cabeza, R. (2012). Neural Similarity Between Encoding and Retrieval is Related to Memory Via Hippocampal Interactions. *Cerebral Cortex*, *23*(12), 2818–2828. https://doi.org/10.1093/cercor/bhs258

Roberts, K., Jentzsch, I., & Otto, T. U. (2024). Semantic congruency modulates the speed-up of multisensory responses. *Scientific Reports*, *14*(1), 567. https://doi.org/10.1038/s41598-023-50674-4

Schnotz, W., & Horz, H. (2010). New Media, Learning from. In *Elsevier eBooks* (pp. 140–149). https://doi.org/10.1016/b978-0-08-044894-7.00740-5

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology General*, *117*(1), 34–50. https://doi.org/10.1037/0096-3445.117.1.34

Spence, C. (2007). Audiovisual multisensory integration. *Nippon Onkyo Gakkaishi/Acoustical Science and Technology/Nihon Onkyo Gakkaishi*, *28*(2), 61–70. https://doi.org/10.1250/ast.28.61

Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention Perception & Psychophysics*, *73*(4), 971–995. https://doi.org/10.3758/s13414-010-0073-7

Stein, B. E., & Meredith, M. A. (1993). *The merging of the senses.* The MIT Press.

Stein, B. E., Stanford, T. R., & Rowland, B. A. (2014). Development of multisensory integration from the perspective of the individual neuron. *Nature Reviews. Neuroscience*, *15*(8), 520–535. https://doi.org/10.1038/nrn3742

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*(3), 380–398. https://doi.org/10.1037/1196-1961.48.3.380

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, *80*(5), 352–373. https://doi.org/10.1037/h0020071

Woods, A. T., & Newell, F. N. (2004). Visual, haptic and cross-modal recognition of objects and scenes. *Journal of Physiology-Paris*, *98*(1–3), 147–159. https://doi.org/10.1016/j.jphysparis.2004.03.006

Appendix A
Stimulus Lists (Test and Distraction Items)

Test Items

dog, phone, baby, piano, knocking, motorcycle, cat, rain, guitar, blowdry, keyboard, bird, water, scissor, skating, microwave, toothbrush, clock, ocean, train, basketball, squirrel, camera, car, parrot, doorbell, music, writing, filling glass, hammer, snake, computer mouse, swimming, pan, duck, fire, tiger, book, coffee, boat, money, typing, helicopter, rabbit, stapler, bee, running, gun, drum, coughing, bottle opener, bubble wrap, lion, cutting, bell, volleyball, cycling, lighter, egg, zipper

Distraction Items

applause, archery, baboons, boiling, bouncing, bus, carpenter, cheers, chess, cleaning, confetti, cow, dj, dragonfly, drone, elephant, erasing, flamingo, frog, gaming, graffiti, grate, grind, horse, hourglass, insect, lawn mower, machine, metro, metronome, mixer, monkey, mouse, nail trimmer, pasta, photos, pill, present, rooster, sand, screwdriver, sealion, service bell, singingbowls, ski, socket, soil, sprey, stirring, stove, surf, tea, tennis, toaster, tractor, typewriter, vacuum, washing machine, waterfall, windmill

Appendix B

Distribution of Test Stimuli into Three Balanced Lists

List 1: dog, phone, baby, piano, knocking, motorcycle, cat, rain, guitar, blowdry, keyboard, bird, water, scissor, skating, microwave, toothbrush, clock, ocean, train

List 2: basketball, squirrel, camera, car, parrot, doorbell, music, writing, filling glass, hammer, snake, computer mouse, swimming, pan, duck, fire, tiger, book, coffee, boat

List 3: money, typing, helicopter, rabbit, stapler, bee, running, gun, drum, coughing, bottle opener, bubble wrap pop, lion, cutting, bell, volleyball, cycling, lighter, egg, zipper

Appendix C

Table C1

Mean Hit Rates by Study Condition

	M	*SD*
Congruent	.85	.11
Incongruent	.83	.11
Unimodal	.82	.12

Table C2

Repeated-Measures ANOVA for Sensitivity (d')

Source	Sum of Squares	*df*	Mean Square	*F*	*p*	n^2_p
Condition	.511	2	.256	1.570	.212	.026
Error	19.211	118	.163			

Note. Mauchly's test indicated that the assumption of sphericity was met; sphericity-assumed values are reported.

Table C3

Paired Samples t-test Between Study Conditions

		Paired Differences					Significance
				95% Confidence Interval of the Difference			
		Mean	Std. Error Mean	Lower	Upper	*t*	Two-Sided *p*
Pair 1	Congruent - Incongruent	.10	.07	-.04	.24	1.41	.163
Pair 2	Congruent - Unimodal	.12	.08	-.03	.28	1.63	.109
Pair 3	Incongruent - Unimodal	.03	.08	-.12	.18	0.35	.727