Title
The Effects of Multisensory Stimulus Presentation in Episodic Memory

Author
Murray, Carolyn

Publication Date
2023

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

The Effects of Multisensory Stimulus Presentation in Episodic Memory

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Psychology

by

Carolyn Murray

2023

ABSTRACT OF THE DISSERTATION

The Effects of Multisensory Stimulus Presentation in Episodic Memory

by

Carolyn Murray

Doctor of Philosophy in Psychology

University of California, Los Angeles, 2023

Professor Ladan Shams, Chair

A common desire in the modern world is to improve how much we remember about key daily events, and improving this requires understanding how information is processed in memory. One promising yet often overlooked method may be to utilize sensory integration. Previous work performed by multisensory research groups have shown that multisensory stimulus presentation can improve memory performance for facts and episodes. However, whether these findings are general and robust, what kind of tasks could benefit from multisensory encoding, and what the underlying mechanisms are questions still unanswered. Some of these limitations emerge from the limited number of studies investigating memory through a multisensory framework, and that these studies do not always replicate one another's results. Other limitations come from the interplay of these multisensory studies with existing memory theories, almost all of which do not acknowledge that sensory combination and sensory integration are distinct. Thus, I sought to answer remaining key questions in the field of multisensory memory encoding. First, with discrepant findings amongst multisensory research

groups regarding the presence of memory benefit, I investigated drift-diffusion modeling and other simultaneous measures of speed and accuracy as tools to quantify multisensory benefit (Study 1). This discovered that such measures were a sensitive and reliable measure of multisensory benefit, which was later applied to investigate if speed-accuracy tradeoffs were present in our empirical memory findings. Second, as multisensory memory benefit has only previously been explored in a limited variety of memory tasks for basic objects, we sought to expand the body of research to include more challenging associative memory tasks, specifically memory for face-name associations (Study 2) and Swahili-to-English vocabulary memorization (Study 3). Study 2 showed that multisensory stimulus presentation, specifically, is helpful for bolstering associative memory for faces and names. Study 3 provides an interesting case where multisensory presentation fails to produce better memory, providing insight to important border conditions that have not been previously discussed. Finally, as existing memory theories do not separate multisensory processes from the mere presence of information across senses, I investigated whether multisensory representations are stored in memory, and whether this is true for all individuals (Study 4). By testing participants' memory for an illusion, it was shown that, for the vast majority of participants, multisensory representations are coded, meaning most participants should specifically benefit in their memory performance from encoding information in a multisensory manner. These studies shed light on the mechanisms of encoding and retrieval in ecological multisensory experiences and may have translational implications for facilitation of memory in everyday tasks.

The dissertation of Carolyn Murray is approved.

Barbara Knowlton

Jesse A Rissman

Aaron Seitz

Ladan Shams, Committee Chair

University of California, Los Angeles

2023

Table of Contents

List of Figures & Tables

Figures

Tables

Acknowledgements

Chapter 1 has sections adapted from the following published manuscript:

C. A. Murray and L. Shams (2023). Crossmodal interactions in human learning and memory. *Frontiers in Human Neuroscience,* 17:1181760. doi: 10.3389/fnhum.2023.1181760

Chapter 2 is a version of the following published manuscript:

C. A. Murray, E. S. L. de Larrea-Mancera, A. Glicksohn, L. Shams, & A. R. Seitz. (2020). Revealing multisensory enhancement with diffusion modeling. *Journal of Mathematical Psychology, 99, 102449*. doi: 10.1016/j.jmp.2020.102449.

Chapter 3 is a version of the following published paper:

C. A. Murray**,** M. Tarlow, J. Rissman, & L. Shams (2022). Multisensory encoding of names via name tags facilitates remembering. *Applied Cognitive Psychology*, *36*(6), 1277-1291.E.

Chapter 4 is a version of an in-preparation manuscript:

C. A. Murray, G. Rahimi, S. Kannan, & L. Shams (*In prep*). Foreign vocabulary learning is assisted by cognitive but not sensory enrichment during encoding.

Chapter 5 is a version of an in-preparation manuscript:

C. A. Murray, X. Guo, & L. Shams (*In prep*). Sensory inputs are encoded in memory after integration with other senses for most individuals.

I would like to express my deepest gratitude and thanks to my advisor, Dr. Ladan Shams, for all her guidance and endless patience through my time as a graduate student. It has truly been an honor to be mentored by someone who is so dedicated to producing excellent science, and who also allowed me to focus on research projects that excited me. I would also like to express my gratitude to my committee members Dr. Barbara Knowlton, Dr. Jesse Rissman, and Dr. Aaron Seitz. Your guidance and expertise have been critical to conducting the

best research possible, and much of the work done during my Ph.D. journey has benefitted immensely from your input.

I would additionally like to thank my co-authors from several of these manuscripts, both published and in prep: Arit, Golbarg, Hannah, Maisy, Sebastian, and Shreya. It has been an honor to work with you on these projects and receive your input and time.

I am also deeply grateful to my undergraduate research assistants—their patience, reliability, and enthusiasm was invaluable to keeping these projects moving. Not only have I benefitted from having extra hands to collect and score data, but I have benefitted from working with such enthusiastic and driven students, whose dedication to research has often reinvigorated my own love of what I do. Adele, Ashkan, Emma, Fatima, Isha, Janvi, Jasmine, Jessica, Jose, Kasra, Laleh, Matisse, Sanam, Sarah, and Varsha, this work would not have been the same without your presence on the team.

Finally, I would like to thank my lab mates, department colleagues, and other friends and family who have supported me both in science and in life's ups and downs throughout my time at UCLA—Akila, Alessandra, Alex H, Alex K, Alex L, Brandon, Charlotte, Chris, Danny, Egamaria, Elizabeth, Geneviève, Jody, Kimia, Lara, Maggie, Mary, Megan, Sashel, Saskia, Saul, Stefany, Stephen, and Veronica, thank you so much for your patience, support, and time. I wouldn't be where I am in life without you.

Curriculum Vita

Carolyn A. Murray

Education

M. A. in Psychology, University of California, Los Angeles Dec. 2017

B. S. in Neurobiology, Physiology & Behavior, University of California, Davis June 2016

Research & Teaching Experience

Graduate Student Researcher, UCLA, Dept. of Psychology, Sept 2016- Present

Instructor, UCLA Dept. of Psychology, Aug-Sept 2021

Lead Teaching Assistant, UCLA Dept.of Psychology, April 2019-March 2020

Teaching Assistant, UCLA Dept.of Psychology, Sept 2017-Sept 2023

Peer-Reviewed Publications

- **C. A. Murray**, X. Guo, L. Shams (*In preparation)*. Sensory inputs are encoded in memory after integration with other senses.
- **C. A. Murray**, G. Rahimi, S. Kannan, & L. Shams (*In preparation)*. Multisensory influences on novel language learning.
- **C. A. Murray** and L. Shams (2023). Crossmodal interactions in human learning and memory. *Frontiers in Human Neuroscience,* 17:1181760. doi: 10.3389/fnhum.2023.1181760

- **C. A. Murray,** M. Tarlow, J. Rissman, & L. Shams (2022). Multisensory encoding of names via name tags facilitates remembering. *Applied Cognitive Psychology*, *36*(6), 1277-1291.E.
- E. Chau, **C. A. Murray**, L. Shams (2021). Hierarchical drift diffusion modeling uncovers multisensory benefit in numerosity discrimination tasks. *PeerJ, 9:e12273,* https://doi.org/10.7717/peerj.12273
- **C. A. Murray,** E. S. L. de Larrea-Mancera, A. Glicksohn, L. Shams, & A. R. Seitz. (2020). Revealing multisensory enhancement with diffusion modeling. *Journal of Mathematical Psychology, 99, 102449*. doi: 10.1016/j.jmp.2020.102449.

Talks & Presentations

- "Do Sensory Inputs Get Encoded in Memory Before or After Integration with Other Senses?" CogFog Lab Meeting (UCLA Bjork Lab), Los Angeles, CA, May 2022. Invited Talk.

- "Multisensory Contributions to Learning Face-Name Associations," Human Vision & Electronic Imaging, Burlingame, CA, January 2020. Invited Talk.

- "Multisensory Contributions to Learning Face-Name Associations," CogFog Lab Meeting (UCLA Bjork Lab), Los Angeles, CA, August 2019. Invited Talk.

- "Revealing Multisensory Enhancement with the Drift Diffusion Model," International Multisensory Research Forum, Toronto, Canada. June 2018. Poster

Chapter 1: General Introduction and Overview

The environment and set of tasks the human brain must complete throughout the course of our lives create an immense challenge for the nervous system. We live in dynamic environments, whose changes require a large variety of flexible behaviors to navigate. Moreover, the human body also changes through time, growing when we are young and deteriorating with age. The brain must recalibrate and adjust its functioning during all of these stages in life. The complexity of these systems is such that it is not possible for all behaviors to be hard-coded; the human genome only contains 20-25 thousand genes, which is far too few to code everything the brain must compute and perform. In addition, humans are social animals, which will require us to not just have a functional understanding of our physical environment, but of our social experiences and networks as well.

These complex environmental and developmental factors have thus necessitated the evolution of a brain that is capable of recalibration and learning. The human brain is, in fact, noted for being incredibly plastic (Calford, 2002; Kolb & Whishaw, 1998), and apt at both supervised and unsupervised learning (Knudsen, 1994). In addition, the human brain is accomplished in memory tasks that support learning about our environments and remembering our social interactions. As they are such fundamental functions of human behavior, both learning and memory have been studied extensively in humans over the decades in a variety of disciplines and using a variety of methods. However, the vast majority of these studies focus on studying one sense at a time (for overviews, see Fiser & Lengyel, 2022; Goldstone, 1998).

While situations that focus on the experiences of only one sense can be created in an experimental space, such work does not reflect the cues across many senses that would be available and working in concert in a natural environment. On a daily basis, we use information across multiple senses to learn about our environment and encode in our memories for later

use. The senses do not operate in a vacuum. If we drop a glass, we do not just see it fall, but we hear the impact and feel the lack of its weight in our hands. When talking to a friend, we do not just hear their voice, but see their facial expressions and smell their perfume. With such rich information available across senses about the same experience, it would make sense if the brain was capable of processing this information in a holistic way, without the boundaries of sensory modality and perhaps even exploiting the relationship between the sensory cues. Yet, the vast majority of studies of perceptual learning and memory have used unisensory stimuli and tasks.

Research over the last two decades, however, has greatly enhanced our understanding of how the brain is able to combine information across the senses, jump-starting the field of research in *multisensory perception*. Multisensory perception and multisensory integration, a specific case of multisensory perception wherein, beyond the combined use of cues from separate sources, these are combined into a separate, meaningful signal (Stein, 2012), are terms that can capture a large number of different sensory combinations and sensory relationships. Myriad studies have established that sensory pathways can influence one another, even at their earliest stages. For example, the presence of low-level multisensory illusions, such as the ventriloquist illusion (Bruns, 2019; Thurlow & Jack, 1973) and the sound-induced flash illusion (Hirst et al., 2020; Shams et al., 2000) indicate that the senses combine information early on and influence one another in ways that are observable at a behavioral level. Psychophysical studies have established that the interactions between the senses is ubiquitous. Correspondence between senses in the real world allow for the development of multisensory neurons and regions in the brain by the end of the first year of life (Neil et al., 2006), that are refined throughout middle childhood (Brandwein et al., 2011; Rohlf et al., 2020) before reaching adult levels. Adaptation of and use of crossmodal processing continue throughout the lifespan (e.g., Burr & Gori, 2012; McGovern et al., 2022; M. M. Murray, Lewkowicz, et al., 2016; Nardini et al., 2012; Setti et al., 2011), across many sensory modalities and tasks (e.g., Botvinick &

Cohen, 1998; Bruns, 2019; Peters et al., 2015; Shams et al., 2000; Wozny et al., 2008), Accordingly, brain studies have revealed interactions between the senses at a variety of processing stages, in all processing domains (Ferraro et al., 2020; Gau et al., 2020; Murray, Thelen, et al., 2016, and see Driver & Noesselt, 2008; Ghazanfar & Schroeder, 2006 for reviews). Generally speaking, the effects of combining multiple senses, regardless of whether true integration has occurred, can manifest In a number of different ways, but are often clearest in tasks that display one of the following: a spatial component, a temporal component, or weak/noisy information in unisensory channels (Otto et al., 2013). Altogether, research has uncovered that multisensory processing is not simply the sum of unisensory processes, which implies that multisensory learning cannot be simplified to the sum of the constituent unisensory learning and memory. Indeed, researchers have begun investigating learning and memory under multisensory conditions, and these studies have revealed surprising phenomena that point to multisensory processing being a unique and powerful mechanism for learning and memory.

Research on multisensory processing in the domains of perceptual learning, sensory recalibration, and implicit associative learning has often shown that multisensory stimulus presentation can make learning faster, easier, or otherwise more effective (see Murray & Shams, 2023). However, it has only been more recently that multisensory researchers have attempted to expand these findings into the realm of memory. As such, while both the fields of multisensory perception and human episodic memory have significant history and literature, it is only in the last two decades that these have been brought together. We will attempt to bring these two areas of research together by investigating some remaining holes in the area of the influence of multisensory processing on memory processes.

To that end, we will briefly review some studies that investigate memory through a multisensory lens, with particular focus on audio-visual studies. We will discuss relevant theories in memory and learning research that examine the role of sensory inputs during

encoding. We will then highlight remaining questions and propose studies that could help to answer these remaining questions.

Multisensory Memory

While much work in multisensory processing has been dedicated to the topic of low-level learning, the benefits of multisensory processing are not limited to just the realm of learning. The memory systems of the brain must also, crucially, be able to store and represent information across senses in order for humans to make sense of our environment. In addition, our episodic memory, as well as being a useful guide on our environment, helps us to store information crucial to the events of our lives, which helps us to store information crucial to social interactions and aid in decision making critical for survival. Episodic memory is commonly defined as memories for events and experiences, rich in sensory and contextual details, rather than memories for facts (Tulving, 1993). Memories are rich in sensory detail and can typically be cured by many senses. Neuroimaging studies have revealed that the role of perception in memory was not unidirectional upon encoding: recall of visual and auditory stimuli reactivates sensory-specific cortices that were active at encoding. This is true within modality, where a sensory region active during encoding is reactivated upon recall (Nyberg et al., 2000) but has also been shown in multisensory conditions, where a visual probe for an audiovisually-encoded item reactivates auditory regions as well as visual ones (Wheeler et al., 2000). This highlights a clear link between sensory representations and mnemonic codes. Many studies of human memory have focused on individual senses (for examples, see Brady et al., 2008; Schurgin, 2018; Slotnick et al., 2012; Weinberger, 2004) or chosen to not view memory through a sensory lens at all. However, given that multisensory training has now been shown to benefit learning (Shams & Seitz, 2008), and that episodic memory ties together information across senses in a way that seems to naturally take advantage of crossmodal processing, work in the past two

decades has begun to explore the benefits of multisensory stimulus presentation for memory performance.

Some of the earliest work performed by multisensory researchers investigating the effects of crossmodal stimulus presentation on memory outcomes focused on investigating object recognition. In a continuous object recognition task, researchers showed that multisensory presentation of objects during the encoding phase seems to enhance later recognition of unisensory representation of the objects. Recognition performance for visual objects presented initially with congruent audio and visual cues was reported to be higher than that of objects initially presented only visually, or with an incongruent audio (Lehmann & Murray, 2005; Thelen et al., 2015). When the recognition test is auditory instead of visual, the pattern of results has been shown to be similar, where multisensory encoding produces higher recognition than audio-alone encoding (Moran et al., 2013).

The aforementioned studies all used a continuous recognition task in which the first and second presentations of the same object are presented within a stream of objects that are interleaved. Experiments that use a more traditional memory paradigm, with distinct encoding and retrieval phases separated by a delay interval, and also those attempting to study more naturalistic tasks have also found a benefit to multisensory encoding. Heikkilä et al. (2015) used such a paradigm to compare benefits in visual recognition to benefits in auditory recognition for stimuli encoded in a multisensory condition compared to stimuli encoded in a unisensory fashion. Contrary to some earlier studies, this study found no benefit to visual recognition between the two conditions, though there was a significant improvement to recognition for auditory memory for items encoded with a visual compared to those encoded as audio only. This study also looked for improvement in recognition of spoken and written words and found that adding audio to written words and vice versa improved recognition, so the benefits seen in previous studies may not be limited to perceptual representations and appear to extend to semantic information. This study noted an asymmetry in the effect of multisensory encoding on

recall: auditory representations benefit from multisensory training whereas visual representations do not. Given that auditory recognition memory is typically noted for being worse than its visual counterpart (M. A. Cohen et al., 2009; Gloede & Gregg, 2019), the representations supporting auditory memory may be more ambiguous, and thus may particularly benefit from multisensory encoding.

However, such findings are not ubiquitous. A study that attempted to replicate the findings of Thelen et al. (2015), but made a few changes to the paradigm reduce potential sources of bias, including using signal detection theory sensitivity (d'; Snodgrass & Corwin, 1988) as a measure of performance. Across four experiments, only one showed a weak confirmation of the previous results (Pecher & Zeelenberg, 2022). This highlights a few important needs in this area of research. Firstly, and of more interest methodologically, obtaining a reliable and unbiased measure of participant performance is crucial. Using sensitivity measures helps reduce the influence of participants' different response biases when assessing performance, but other factors are known to also influence participant accuracy. For example, there is a known interaction between speed and accuracy in decision making, termed the speed-accuracy tradeoff (Fitts, 1966), which is present in a number of other multisensory paradigms (e.g. Arieh & Marks, 2008; Diederich & Colonius, 2009).

Secondly, these findings indicate that memory performance benefits from multisensory stimulus presentation may not ubiquitous and utilize a limited variety of paradigms. Existing research investigating multisensory benefits for human episodic memory processes is somewhat limited. Continuous recognition experiments used a set-up wherein old-new judgments were made throughout the task, rather than by explicitly separating encoding and retrieval phases, which could result in some items showing up very close in time to other items, limiting the ability to draw conclusions about long-term memory performance. Indeed, Heikkilä et al. (2015) report average duration between the first and second presentations of a stimulus in the Lehmann and Murray (2005) study was only 25 seconds. A few studies exist in multisensory

literature outside of the continuous recognition paradigm, but this should be expanded to better capture whether multisensory stimulus presentation can benefit long-term memory performance.

These findings also primarily investigate recognition memory, and thus also leave limited our understanding of what types of memory retrieval multisensory stimulus presentation can support. Only one study has investigated recall instead of recognition, finding that recall for visual objects was better when those objects were initially presented with congruent auditory information, even if participants were explicitly told to ignore that auditory information (Duarte et al., 2022). However, whether this generalizes to other recall tasks, or other more complicated memory tasks (associative memory, learning of concepts, etc.) remains unclear. As such, further experiments should be performed to explore under what conditions it may be possible to receive a multisensory benefit to memory retrieval.

Relevant Memory Theories

Memory studies focusing on multisensory approaches make up a small portion of the relevant literature. It is thus important to integrate broader theories explaining human memory and highlight in what ways the multisensory approach is distinct from these existing theories. To that end, we will now discuss a few relevant theories that allow for sensory inputs to aid human memory. While much has been written about the relative merits of these theories, the current goal is not necessarily to show one is more empirically in line with multisensory memory findings than others. Instead, it is to highlight the difference between the multisensory approach and such existing theories. A few theories of memory function, both historical and active, that may be particularly relevant to addressing how multisensory benefits may arise include depth of processing, context-reinstatement, and dual processing theory.

The *depth of processing* framework, popularized in the 1970s, is one such framework through which multisensory benefits may be expected. Proposed by Craik and Lockhart (1972) and further explored by Craik and Tulving (1975), this framework posits that memory is strongly affected by how deeply individuals interact with stimuli at encoding. The framework indicates that, in general, the more one is required to engage with semantic or other high-level properties of a stimulus, compared to very basic perceptual features, the better their encoding of the object will be. This would tie in to then the quality of the memory for this object, altering how likely it is to be remembered later, such that deeper encoding correlated with better memory traces and superior ability to retrieve the memory at test. Neuroimaging studies can support the depth of processing idea with activation: fMRI studies have found that, at encoding, frontal and medial temporal regions of the brain show greater activation for deeper, semantic judgments of words compared to judgments about alphabetical properties of the words (Otten et al., 2001).

While many studies within this part of the literature focus specifically on how processing of words could change the memory for them—for example, asking phonological questions about a word compared to asking if it would work in the syntax of a specified sentence—the framework itself allowed for broadening into any type of deeper perceptual processing, and findings do appear to translate outside of processing of words. Recognition of images of human faces has been improved when participants were asked to make judgments about the character of the person shown (for example, if they were honest) rather than a basic judgment about the gender of the face (Bower & Karlin, 1974; Strnad & Mueller, 1977). While such visual judgments are hard to fit explicitly into the “semantic processing” idea common to word stimuli, the generalizability of this effect does indicate that generally being asked to make more effortful interactions with stimuli leads to superior encoding, which will correlate with generally superior memory for those items. From a sensory level, this could imply that multisensory stimuli are simply creating a deeper level of processing than a unisensory item. While adding a sound to a visual may not greatly deepen the processing of a stimulus, it could still require an assessment

of the congruency of the sound and the image, or a superior activation of an existing schema for an item and encourage participants to engage more deeply with the stimulus.

Given that Lehmann and Murray (2005) and Moran et al. (2013) only found an effect for semantically congruent audiovisual combinations, it is possible that multisensory stimulus presentation would fit within this framework. The simultaneous presentation of audiovisual stimuli may prime participants to encode information more deeply by providing extra information about a stimulus, compared to unisensory processing. For example, an image of a dog may provide a sense of the color and shape of the animal, but a bark may give additional information, such as a sense of the overall size of the dog this image is representing. However, this interpretation is weakened somewhat by some existing evidence that semantic relationships between audio and visual stimuli may not be necessary to see improved memory performance. Previous unpublished work in our lab indicates that, with sufficient training in a novel audiovisual association (randomized per participant, to prevent meaningful audiovisual correspondences across individuals), participants will see slight changes in their signal detection d' between stimuli presented originally with a sound from those presented without.

Another theory that could help to explain some of the observed effects from multisensory stimuli could include the idea of context reinstatement, and *context-dependent memory*. Studies on the impact of context on human memory indicate, overall, that a shared context for study and test tends to boost performance. In a classic and dramatic example, Godden and Baddeley (1975) showed that divers who learned a word list either on land or while diving remembered more items from the list in a recall test when they were tested in the same environment as they had learned the list. While such drastic changes could make this effect seem difficult to replicate, further experiments have shown that the environments need not be so extremely different to obtain this result. Changing classroom environments between single-session study and test locations, even if the rooms are similar lecture halls, can also obtain this effect (Metzger et al., 1979). The cues also need not be visual to create this effect; auditory cues such

as playing background music or white noise to a room during both study and test can also elevate memory performance in a free recall task (Balch et al., 1992; Smith, 1985). Olfactory cues can also be used: Herz (1997) found that the ambient scent of peppermint or osmanthus plants boosted performance on a surprise word recall task when present at both study and test.

These would seem somewhat at odds with the existing multisensory findings where encoding with different sensory conditions from those used at test could be interpreted as a difference in context between encoding and test. In that framework, the recognition improvements seem counter intuitive. However, there are several factors that could be used to explain this supposed contradiction. Redintegration in human memory—the retrieval of a complete, rich memory episode from a subset of the cues or a single cue (Horowitz & Prytulak, 1969)—has been observed, so providing full sensory cues are not necessary for memory retrieval, merely a way of improving the chance of retrieval. To this end, we propose that such considerations will need to be taken into account when assessing multisensory benefits, but these concerns may be reduced through careful experimental design, given that many factors are known to alter the importance of these context- and state-dependent effects. For example, increasing the study sessions and the time delay between them (see, for examples, Kornell et al., 2010; Smith et al., 1978) or adding more testing events (for example, see Roediger & Karpicke, 2006) have been shown to reduce the context effect. Indeed, solutions as simple as mentally recreating the context for the original learning can help overcome the drop in performance observed when switching contexts at test (Smith & Vela, 2001). The effects of state-dependent retrieval cues are not always reliable, either, though the efficacy of this effect seems to also depend on the availability of additional cues beyond state-dependent ones upon retrieval (Eich, 1980). Indeed, studies of alcohol-induced state-dependent memory have shown that recall, which provides fewer cues to guide participants' memory, shows stronger impact of state changes than cued recall tasks (Petersen, 1977). Certainly, these results indicate that, while context- and state-dependent effects must be considered, some experimental design

choices may mitigate their impact on participant memory performance. This does not, however, remove the importance of context from many memory tasks, especially as multisensory stimuli will provide richer encoding, they may provide extra contextual information that could be used to reinstate context, but it may imply that careful experimental design will allow for the differentiation of these effects from multisensory effects.

Among the most compatible theory with multisensory stimulus presentation, however, is the *dual coding theory* of memory, which posits that, as the number of traces for a memory increases, the likelihood of it being remembered increases (Clark & Paivio, 1991; Paivio, 1991). The two codes in this theory correspond to two distinct methods of processing and handling information, which are commonly conceptualized as being verbal and nonverbal, more conceptual processing. As such, there is a clear similarity in conceptualization to Baddeley's model of working memory, with its independent verbal and visuospatial working memory resources (Baddeley, 2000; Baddeley & Hitch, 1974). The resources needed to process information in the two streams are seen to be relatively independent, and dual activation of both streams during encoding can improve later memory retrieval. Of course, the Clark & Paivio model is not the only one—dual processing models exist in a number of forms and specifications, including closely related forms used to characterize human learning in audiovisual multimedia environments (R. E. Mayer, 2014).

In a similar timeframe, other models began to explain memory traces—the pieces of information that encapsulate memories—as associated features that together encapsulate a memory and can be used to retrieve it (Tulving & Bower, 1974). It would not be out of the ordinary to conceptualize some of the memory features to be the sensory information available at encoding—relevant sights and smells that help to identify an object or location. It has thus also been suggested that memory retrieval can be conceptualized as checking the retrieval prove to existing traces, where the speed and success of this process is related to how similar the probe is to the existing traces (Ratcliff, 1978). Some similar models are even more fine-

grained with their approach, suggesting perceptual information could be further broken down into features to be compared to those stored in memory (Norman & Rumelhart, 1970).

This theory, perhaps more than the others, is specifically interesting from a multisensory point of view, as it directly addresses the idea of sensory cues providing a way of improving memory performance. More sensory details, improved grouping of sensory details, and details that are a better match to retrieval probes would potentially all benefit retrieval. However, it does not appear to explicitly handle sensory interactions, and how integrated information may play a role in memory.

Defining the Multisensory Approach

This primary difference between the existing models of memory common to memory researchers and the point of view held by multisensory researchers is the purported role of sensory integration in the process. Are memory traces created from unimodal representations of stimuli, or are multimodal representations specifically useful for processing? Dual coding theory draws its processing distinctions between verbal and non-verbal modes. Thus, while providing images and sounds in a way that encourages sensory integration can be helpful in this framework, the framework is relatively agnostic to any differences between mere sensory combination and an integrated representation of an object or event. Conceptualizations of memory traces as associations of related features likewise do not draw a distinction between having multiple senses available at encoding, thus providing more routes back to the information and improving retrieval, from any particular benefit available from multisensory stimulus presentation.

There is an acknowledged possibility that items in different modalities can interact in memory (Logie et al., 1990), but these hypotheses appear to be more common in short-term or working memory studies than in investigations of long-term memory performance. Within

studies of long-term memory, reactivation of auditory cortex during visual recognition has been observed when the original stimulus presentation was multisensory (Nyberg et al., 2000), indicating that, even when information across senses is not required for a task, the brain may encode and retrieve a multisensory representation. Similarly, EEG analysis of participant performance during a visual recognition memory task indicates that memories encoded under multisensory conditions can be discriminated from those encoded under unisensory conditions in the brain, in regions as early in processing as the lateral occipital cortex (Murray et al., 2004).

Multisensory processing has been shown in the related field of learning to improve learning, and some of the mechanisms at play may also explain how multisensory encoding can improve later memory retrieval. A recent review by Mathias and von Kriegstein (2023), focusing the many facets of multisensory learning, came to the conclusion that multisensory mechanisms provide a better explanation for the observed benefits from multisensory learning as opposed to unisensory learning mechanisms. Many imaging and neurostimulation studies report that functional connectivity between sensory-specific areas is altered after crossmodal learning (as in K. M. Mayer et al., 2015; Thelen et al., 2012; von Kriegstein & Giraud, 2006). It has also been suggested via simulation studies that both crossmodal connectivity and connections between unisensory regions and higher-level association areas could be strengthened simultaneously during multisensory learning (Cuppini et al., 2017). Proposed Hebbian learning model, following the principle of “fire together, wire together” for the unisensory and multisensory regions (Hebb, 1949; Magee & Grienberger, 2020) have also been proposed to explain benefits in multisensory learning. Multisensory learning under this model takes place in part because the two senses contributing to a multisensory signal are co-occurring, which encourages these regions to become more strongly connected. This stronger connection will allow for activation of one region to recruit a larger population of neurons post-training more easily, due to stronger crossmodal connections.

While single-trial multisensory memory presentation may not be able to utilize the exact mechanisms that improve memory performance (though rapid recalibration is possible with multisensory stimulus presentation, as in Wozny & Shams, 2011, so not all learning mechanisms require longer exposure, and would thus not be out of the realm of possibility), most learning theories posit that multisensory regions will be activated during multisensory stimulus processing. This would allow multisensory stimulus presentation to activate a larger population of neurons, and to produce representations that are more refined relative to unisensory representations. Thus, sensory integration, specifically, has been put forward as a potential explanation for the benefit observed from multisensory encoding (Quintero et al., 2022; Shams & Seitz, 2008). To date, few empirical studies have attempted to tease these interpretations apart, however. To that end, we seek to explore this gap, and investigate if the presence of many senses is sufficient for benefit, or if integration itself is also able to support better memory retrieval.

Remaining Questions & Overview of the Current Studies

The literature leaves several questions open, that could all benefit from further research. In brief, these are the following:

a. Given that, partially through altering the means of analysis, Pecher and Zeelenberg (2022) failed to replicate the findings of Thelen et al. (2015), it seems there may be a need for additional analytical tools that can tease apart what components of task performance appear to be receiving the most benefit. To address this concern, we will investigate using measures that can simultaneously address speeded reactions and response accuracy, as those features of a response often interplay in multisensory research, as well as general decision-making. To that end, Study 1 will investigate drift diffusion modeling (DDM) as an

approach to analyze benefit from integration. We seek to investigate if this methodology provides an accurate sense of multisensory benefit in basic perceptual tasks, and if this is as or more sensitive than traditionally used measures of task performance (Study 1). We will also attempt to use similar measures that consider speed-accuracy tradeoffs in analyzing results in specifically multisensory memory research (Study 3).

b. Given the limited number of studies and variability in memory tasks used to assess claims that multisensory stimulus presentation is beneficial for memory (as well as the failure to replicate previous findings reported by Pecher & Zeelenberg, 2022), there is a need to see if multisensory stimulus presentation can improve long term memory performance. We additionally propose expanding the tasks used from relatively simple object and word memory to more challenging associative tasks. We additionally will build on the findings of Duarte et al. (2022) and explore if multisensory stimulus presentation can aid in recall tasks. To this end, we will propose two studies. In the first, we will explore if multisensory stimulus presentation during encoding will improve memory for face-name associations (Study 2). In the second, we will attempt to generalize these findings to multisensory encoding in learning vocabulary in a foreign language (Study 3).

c. As there is only limited empirical work exploring the claim that multisensory representations are distinctly beneficial for retrieval, rather than benefit arising from merely having multiple senses available at encoding, we would seek to explore this claim. This will require that we are able to explore if multisensory representations are encoded into memory, or if unisensory information is encoded instead. To this end, we suggest using audiovisual illusory stimuli, which will allow for us to probe if unfused or fused—and, thus, multisensory—

representations are available at retrieval. We will additionally investigate to what extent these findings appear to be universal, or ruled by individual differences in how participants encode information (Study 4).

Chapter 2: Revealing Multisensory Benefit with Diffusion Modeling

Abstract

Multisensory information can benefit perceptual, memory, and decision-making processes. These benefits commonly manifest in superior detection and discrimination of multisensory stimuli, as well as improved perception and subsequent memory of unisensory representation of an object previously encoded in a multisensory context. However, the vast majority of studies to date analyze accuracy, sensitivity and/or reaction time data independently to compare multisensory and unisensory conditions. Considering the well-established speed-accuracy trade-off, we asked whether some multisensory benefits go unnoticed when measured using traditional methods that do not take both reaction time and accuracy into account simultaneously, and whether an approach combining them can more reliably characterize and quantify the broad extent of multisensory interactions across perception and cognition. While drift diffusion models have been previously shown to be effective in addressing the speed-accuracy trade-off and providing a reliable and accurate measure of multisensory benefits, one impediment of this approach is the requirement of a large number of trials to estimate model parameters and to characterize effects. This may be prohibitive in many experimental paradigms. Several model variants attempt to reduce the required number of trials, either by averaging across participants or limiting the search space for the parameters. Here, we employed a hierarchical drift diffusion model, that utilizes Bayesian priors, allowing parameter estimation with smaller sample sizes while still making subject-specific parameter estimates. We

analyzed data in perceptual detection and discrimination tasks across multiple sensory combinations, to investigate if the diffusion model would provide a sensitive and reliable measure of multisensory benefits. Results indicate that across visual, auditory and tactile modality combinations, the diffusion model was either as or more sensitive than traditional accuracy, sensitivity, or reaction time measures, and was the only measure that consistently detected multisensory benefits in a statistically significant fashion. We recommend the use of diffusion modeling approaches when assessing the outcomes of multisensory experiments, especially as they become more computationally efficient.

Introduction

The extent to which we are able to integrate information across the senses to reach a behavioral decision and the nature of this integration has been the focus of many studies across the sensory modalities (for some reviews see Driver & Spence, 2000; Shams & Seitz, 2008; Rosenblum et al., 2017). More often than not, the world around us concentrates contingent information about objects that can be picked up by our distinct senses (Shinn-Cunningham, 2008). The multisensory integration that can be achieved by our neural system in this object-oriented context ultimately serves to provide a more accurate picture of what is it that we are experiencing, especially in the face of uncertainty regarding the origins and reliability of information (Stein et al., 2004; Gold & Shadlen, 2007; Raposo et al., 2012). Multisensory integration has been defined differently by different authors, without a clear consensus about the behavioral signature of multisensory integration. Regardless of the criterion used for multisensory integration, it is clear that even when senses do not fully integrate (for example, they do not meet the criterion of optimal integration in a given dimension such as accuracy), there is often evidence for a crossmodal interaction, whereby information from multiple modalities benefits performance compared to the unisensory conditions. Multisensory benefits

can also take the form of enhanced attention to both senses, enhanced processing of a dimension in one modality due to a congruent dimension in a second modality, and more.

Typically, inferences regarding the presence or absence of a multisensory benefit are guided by behavioral data based on accuracies and/or reaction times. However, these measures are known to interact. The speed-accuracy tradeoff is one such example, where pressure to complete a task in a short time frame causes participants to respond less accurately (Wickelgren, 1977; Fitts, 1966). Of perhaps greater interest to the area of multisensory research, is the different behavior participants display in different sensory conditions. For example, responses to auditory stimuli are typically faster than those to visual information (Shelton & Kumar, 2010; Brebner & Welford, 1980), and many spatial tasks, such as localization tasks, show higher response accuracy in unisensory visual condition than unisensory auditory condition (Bushara et al., 1999). This makes a comparison between multisensory conditions and the unisensory conditions more difficult to quantify without a single, combined measure that takes these factors into account. Many studies lack such a combined measure, leaving a reliance on separate measures of response time and accuracy, potentially obscuring the interpretation of the results. To help solve this problem, we propose a methodological approach to combine these measures, allowing for an improved ability to uncover these interactions even when traditional approaches fail to do so.

One such model that has made its way into the psychological literature is the *diffusion decision model* (Ratcliff, 1978), sometimes also called the *drift diffusion model*. The model utilizes trial-by-trial accuracy and reaction time data to estimate parameters that capture a dynamic decision-making process, wherein decision-making is a process of evidence accumulation over time. Under this model, the decision-making process is a dynamic and noisy process of evidence accumulation from some starting point towards one of two decision boundaries. When the accumulated evidence reaches a boundary threshold, sufficient evidence has been gathered to make a decision. The parameters of the model capture key aspects of this

process, including the distance between the decision thresholds, called the boundary separation (*a*), the starting point for evidence accumulation (*z*), and the non-decision time in this process (*t;* see Figure 1 for a graphical representation). Of greatest interest for the current study, however, is the *drift rate (v)*, which can be thought of as the average rate of evidence accumulation (Ratcliff & McKoon, 2008). Higher drift rates typically lead to reaching a decision boundary more rapidly and are considered an indication that participants are better at extracting information from the evidence than if the drift rate was lower. As the effects of multisensory stimulus presentation are often modeled in terms of reducing variance in sensory representations (Gingras et al., 2009), which in turn alters the signal-to-noise ratio, this could be a key parameter to uncover multisensory benefits in this model.

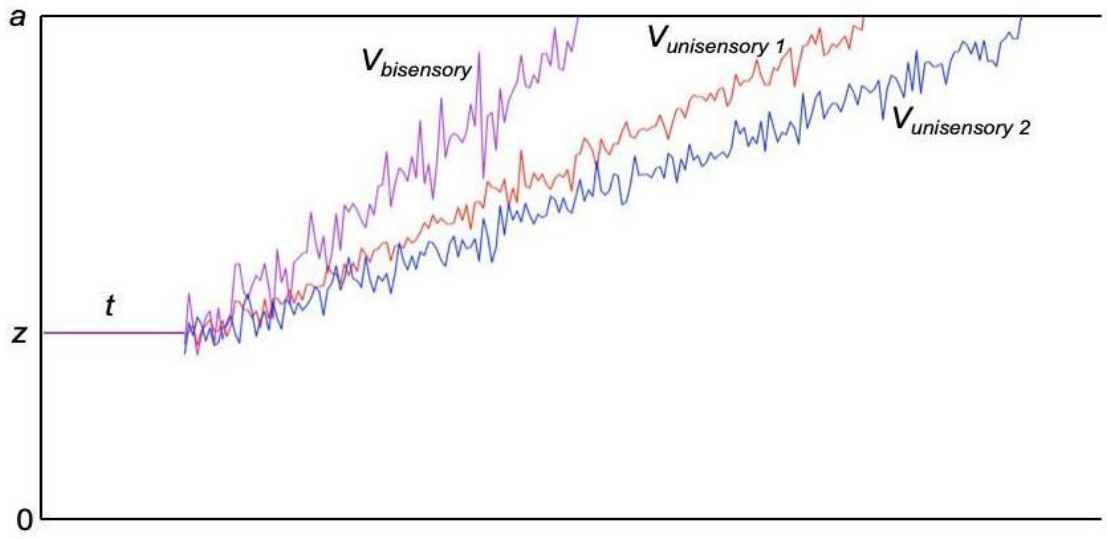


Figure 1: A schematic of the basic components of a drift diffusion model.

The hypothesis that the drift rate in the bisensory condition is faster than those of unisensory conditions is displayed. The four major components included in models used to fit the current experimental data are shown. Drift rate (v) is the average slope of the noisy evidence schematically portrayed here.

Crucially, for our interests, this model has been used previously in modeling memory decisions. When introduced by Ratcliff in 1978, this framework was proposed as a way to model speed-accuracy tradeoffs in memory retrieval. As a non-exhaustive list, It has since been used to show that older adults accumulate evidence more slowly and are more conservative in setting their accumulation boundaries than younger adults in episodic memory tasks (Ratcliff et al., 2004; Spaniol et al., 2006), that interference in prospective memory tasks in adults show increased boundary separation & slower processing of information (Boywitt & Rummel, 2012; Horn et al., 2013), and that drift rate in working memory tasks can predict variance in working memory capacity (Weigard & Huang-Pollock, 2017).

Previous studies have shown that parameters of the drift diffusion model measure can also track multisensory benefits (Drugowitsch et al., 2014, 2015; Diederich, 2008), but such modeling approaches have not yet become well-employed in either multisensory or memory experiments. We suspect this is, at least in part, due to the large number of trials that standard implementations of these models need to converge on a solution. Standard drift diffusion models commonly need several hundred trials in each experimental condition to fit model parameters reliably (see Drugowitsch et al., 2014; Leite & Ratcliff, 2010; Gomez et al., 2007; Van Zandt et al., 2000 for examples), which can be prohibitive for their use. Methods do exist that either combine trials across participants to obtain a large enough number of trials (Mahani et al., 2019), or constrain the search space for parameters to allow estimation to converge more quickly (Nidiffer et al., 2018), however these are not widely employed.

As a method to use fewer trials per condition while still managing to fit individual participant parameters, we utilized the Hierarchical Drift Diffusion Model (HDDM; Wiecki et al., 2013). This variant of the drift diffusion model uses Bayesian priors to begin a Markov Chain Monte Carlo (MCMC) process to converge on model estimates in fewer trials. The model also fits parameters per subject, and then creates an estimate of the population distribution from the participant parameters, allowing for estimates at both individual and population levels. This

particular variant of the model has previously been used to show that drift rate reflects the strength of audiovisual integration in a detection task (Regenbogen et al., 2016). The current study aims to expand this investigation to a systematic investigation of multiple perceptual tasks and sensory combinations, and to examine whether the HDDM can reliably and accurately detect multisensory interactions with moderate sample sizes at least as well as traditional accuracy, reaction time, and sensitivity measures. This would support the general use of such a modeling approach to more fully characterize the results of multisensory experiments.

General Methods

Participants

A total of 67 participants were recruited across 4 separate studies. All participants were undergraduate students at either the University of California, Los Angeles (UCLA), or at the University of California, Riverside (UCR). All reported having normal or corrected to normal vision and hearing, and no history of neurological issues that would reduce tactile sensitivity. Written informed consent was collected from each participant and experimental procedures were reviewed and approved by the UCLA and UCR Institutional Review Boards. Full details about the participants in each experiment have been included in the relevant experiments.

Task Overview

In each experiment, participants were presented with a combination of pseudo-randomly interleaved unisensory and bisensory stimuli. Across three experiments, audiovisual, visuotactile, and audiotacile bisensory combinations were used. Visual stimuli were squares of dynamic salt-and-pepper noise, and auditory and tactile stimuli were Gaussian white noise, delivered over headphones or a vibro-tactile stimulator, respectively. In each experiment, participants were asked to complete two separate 2AFC tasks (Figure 2). In what we will call the

detection task, participants were provided with a stimulus and were asked to determine if the stimulus “pulsed.” Pulses were rhythmic changes in contrast between the light and dark pixels in the stimuli in the visual condition and rhythmic changes in amplitude in the auditory and tactile conditions. In what we will call the *discrimination task*, the stimulus in a given trial would always pulse, and participants were asked to determine if the pulse had been a “slow” or “fast” pulse. All trials lasted a maximum of 3000 ms from stimulus onset.

Data was analyzed for each participant in terms of average accuracy and reaction time in each modality. Signal detection theory sensitivity (d’; Macmillan & Creelman, 2005) was also calculated for each participant, to provide a measure of performance that would not be affected by response biases, although, like accuracy measure, would not consider response times simultaneously. Drift rates were also estimated per-participant for each sensory condition. In each of these measures, bisensory performance was compared to the best unisensory performance to investigate if a multisensory benefit could be detected in each case.

Materials

In the audiovisual task, adapted from Raposo, Sheppard, Schrater, and Churchland (2012), participants were presented with visual, auditory, or audiovisual stimuli. Visual stimuli were squares of dynamic salt-and-pepper noise, half the width and height of the 18’’ CRT monitors used, lasting 500 ms. The arrangement of light and dark pixels in the stimulus changed at a 100 Hz frequency. Participants were asked to detect or discriminate *pulses*, which were changes in contrast polarity of each pixel, relative to the gray background, at either 8 or 12 Hz (which were “slow” and “fast” pulses, respectively). Overall contrast between the visual stimulus and the background was adjusted following each detection or discrimination mini-block for each participant, to keep participant accuracy on detection and discrimination between 60 and 80%. In experiment 1a, if mini-block performance was above or equal to 80% stimulus intensity was decreased 10%, else if mini-block performance was less or equal to 60% stimulus intensity was increased 10%. Auditory stimuli were Gaussian white noise lasting 500 ms. Pulses in this

modality were fluctuations in sound amplitude occurring at 8 and 12 Hz (which were, again, considered slow and fast pulses). Overall difference in sound amplitude was adjusted for each participant in a similar manner than the visual stimuli in experiment 1a to keep accuracy between 60 and 80%, with the additional rule of a 20% decrease in stimulus intensity when mini-block performance was 100%. In experiment 1b, the accuracy cutoffs were the same, but the change in stimulus intensity was altered to 1.2% in order to more precisely match task difficulty across different sensory modalities. Audiovisual stimuli were constructed by combining auditory and visual stimuli on screen, in synchrony, with a visual contrast and auditory amplitude modulation that matched unisensory stimuli in intensity. All stimuli were created and presented using MATLAB (Mathworks Inc., Natick MA), with the use of Psychophysics Toolbox (Brainard, 1997).

Visuotactile used identical stimuli to the audiovisual experiment, but instead of playing through audio-headphones the Gaussian white noise and its associated amplitude fluctuations, the stimuli were presented to participants through vibro-tactile electromagnetic solenoid-type stimulators powered by a vibro-tactile amplifier tactamp 4.2 (Dancer Design, 2017). Participants were presented with visual, tactile, or visuotactile stimuli, and asked to perform the detection and discrimination tasks identical to those in the audiovisual experiments. Audiotactile stimuli were identical to the one generated for audiovisual and visuotactile experiments, but no visual stimuli were presented on screen. Both headphones and vibro-tactile stimulators were used to deliver the stimuli. Participants were presented with audio, tactile, or audiotactile stimuli during detection and discrimination tasks. During both visuotactile and audiotactile tasks, the sound from the vibro-tactile stimulators was masked by an external speaker playing the white noise at an individual comfort level.

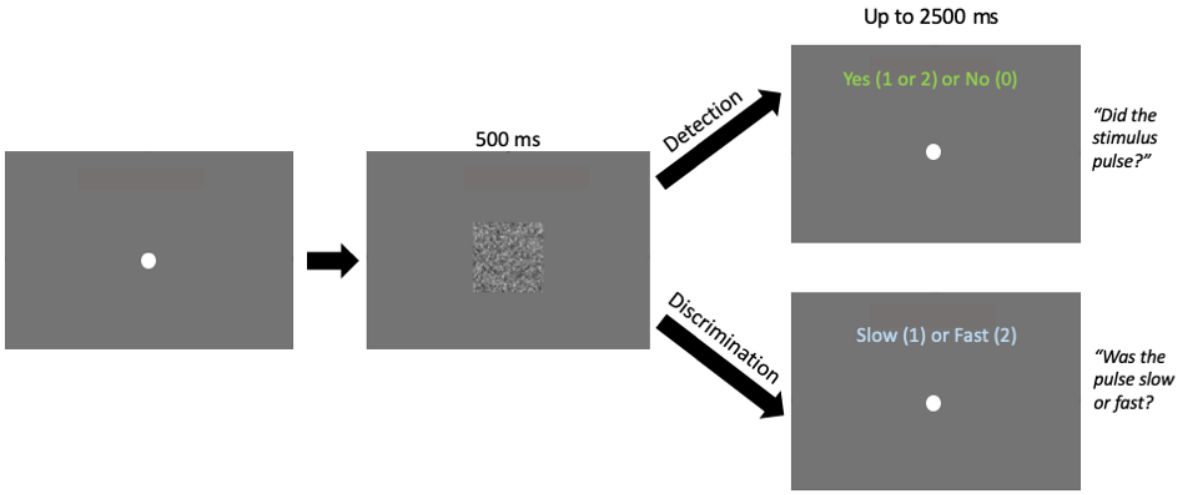


Figure 2: Experimental procedure for general tasks

The stimulus shown here is an example of one of our visual stimuli. Participants were made aware whether the block was a detection or discrimination block and were then presented with a unisensory or bisensory stimulus for 500 ms. For up to 2500 ms after the stimulus disappeared from the screen, the participant was able to respond to the stimulus for that trial.

Procedure

The task was organized such that blocks of detection and discrimination trials were alternated, and participants completed two blocks of each task per session. In each block, participants experienced 40 mini-blocks of 5 trials each for a total of 200 trials per block and 400 trials per task. In the first audiovisual experiment, 25% of the trials were visual only, 25% were auditory only, and 50% were audio-visual, all interleaved pseudorandomly. In all other experiments, unisensory trials made up 40% of the total trials, and the remaining 60% of trials were bisensory, all interwoven pseudorandomly. Before each block, instructions were presented to let participants know if they were supposed to respond for the detection or discrimination task. Task difficulty was adjusted by increasing or decreasing the contrast of a pulsation based on the criteria reported above on participant performance.

In each discrimination block, participants were asked to judge if the stimulus presented to them was pulsing at a relatively fast or slow rate. Fast pulsations occurred at 12 Hz, and slow pulsations occurred at 8 Hz. Participants were instructed to press a button “1” for slow or “2” for fast pulsations. There was a 50% chance that a stimulus presented to them was pulsing at either rate and occurred equally in each sensory modality. In the case of detection blocks, participants were asked to judge if the stimulus presented to them was pulsing or not. Participants were instructed to press either button “1” or “2” if the stimulus was pulsing or a “0” if there was no pulsation. We used different keys for each type of response to avoid interference of inter-block stimulus-response mappings. There was a 50% chance that a stimulus presented to them was pulsing. Prior to completing the main task, participants completed a practice run on each task for at least 10 trials in each unisensory modality. In some cases, additional verbal instruction and a second practice block was delivered to ensure the tasks were adequately understood.

Hierarchical Drift Diffusion Model fitting

Drift diffusion modeling utilized the Hierarchical Drift Diffusion Model (HDDM) toolbox for Python (Wiecki et al., 2013). Data was split by task (detection vs discrimination) for modeling, and two different models were compared for each task. Reaction times (RT) were trimmed such that any RT that was 20 ms or fewer were removed before analysis, accounting for fewer than 0.5% of trials in any experiment. Additionally, only trials from the second half of each session were used in the model and behavioral analyses, to ensure participant thresholds and task difficulty were largely stable throughout the data.

Accuracy-coded drift diffusion models were fit to the data, such that the upper boundary of the model represented correct responses and the lower boundary was incorrect responses, and included terms boundary separation (a), non-decision time (t), drift rate (v), and an outlier term for any extreme reaction times in the data. Drift rate was allowed to vary with sensory

condition, as we predicted any difference in performance should emerge as a result of a change in evidence accumulation. We did not predict significant differences in boundary separation as we did not predict these would vary with sensory condition, given participants had the same speed and accuracy instructions for all conditions. We also did not expect non-response time to vary with condition, as the motor responses to hit buttons were not linked to sensory condition but rather the "yes/no" or "slow/fast" distinction. The modelling process started with priors on all parameters as set by the toolbox (Wiecki et al., 2013), which, for our free parameters, were as follows:

$$a \sim G(1.5, 0.75)$$

$$t \sim G(0.4, 0.2)$$

$$v \sim N(2, 3)$$

where *G* represents a gamma distribution and *N* represents a normal distribution. Additionally, the starting point for the drift process (z), was not allowed to vary freely, and instead used the prior values from the model, such that z ~ N(0.5, 0.5), placing z halfway between 0 and *a*.

The model conducted a 6000-sample Markov Chain Monte Carlo (MCMC) simulation by running 8000 samples with a burn-in of 2000 samples. Comparison of these models was conducted using the deviance information criterion (DIC; Spiegelhalter et al., 2002), which functions similarly to other information criteria but is specialized for hierarchical models. DIC results for all models were compared to single-drift versions of the same model, to ensure the additional complexity of separate drift rates provided better fit to the data, and, in all of the models assessed, this was the case. In addition, we performed a posterior predictive check, which assessed the ability of the model to recapture the behavioral data using its own parameters across the 10th through 90th quantiles of the data, which allowed us to check the fit of the model to the data across the entire distribution of reaction times. This posterior predictive

check indicated that 95% confidence intervals around the model's estimates did recapture the observed reaction times in the data. Parameter recovery on the model indicated that the parameters output by the model could be reliably recovered using synthesized data created using the parameters from the model. These tests indicated a good fit of the model to our data, with less than 9% deviation from the input on all model parameters. Model convergence was assessed with a Gelman-Rubin statistic (<1.02; Gelman & Rubin, 1992) calculated across 5 models. This statistic compares within- and between-model variability in the estimates, and provided evidence for convergence of the models on stable solutions.

Data Analysis

Multisensory benefit ("MSB") in accuracy, signal detection theory sensitivity (d'), and drift rate was calculated for each participant as the proportion change in performance in the bisensory condition above that of the best unisensory condition (Rach et al., 2011). As such, the best unisensory performance was subtracted from the multisensory performance in the same experiment, and this difference was divided by the best unisensory performance (Eqn 1).

$$MSB = \frac{(bisensory) - max\ (unisensory)}{max\ (unisensory)} \qquad (1)$$

For reaction time, the equation was changed somewhat to reflect fast reaction times as superior performance, such that the average bisensory reaction time was subtracted from the faster of the average unisensory reaction times, and divided by the faster of the average unisensory reaction times (Eqn 2).

$$MSB = \frac{min(unisensory) - bisensory}{min\ (unisensory)} \qquad (2)$$

In addition to analyzing average reaction time for each participant per condition, we also assessed *inverse efficiency scores* (Rach et al., 2011), where the average reaction time (RT) is adjusted by the average detection rate for a stimulus. This adjusted RT measure, hereafter shortened as RT*, helps to separate improved performance due to speed accuracy tradeoffs from accuracy changes that reflect improved performance (Rach et al., 2011). Benefit to the adjusted RT was also assessed for multisensory benefit, using the formula in Equation 2.

The MSB for accuracy, d', reaction time, RT*, and drift rate was calculated per participant and averaged across individuals for analysis. Average MSB for accuracy, d', and reaction time were analyzed with t-tests with p-values adjusted for multiple comparisons using the Holm procedure (Holm, 1979). Because the HDDM procedure violated the independence of observations assumption necessary for a t-test, we instead used Bayesian hypothesis testing (conducted using the BEST package in R; Kruschke, 2013) to assess via 10000-sample MCMC simulation if the MSB for drift rate was significant.

In addition to the diffusion model, the data was investigated using a race model (Miller, 1982). This model compares observed bimodal reaction times to those that would be predicted by optimally combining the unisensory reaction times. Violations of this model such that reaction times for bimodal conditions are significantly faster than those predicted by the probability sum of the unimodal components. Such violations would indicate RT facilitation above that expected with fully independent unisensory inputs, and would indicate crosstalk between these senses (see Colonius & Diederich, 2017 for an overview in multisensory contexts). In a similar fashion to the drift diffusion model, this model examines deviations across the entire distribution of reaction times, though it only utilizes reaction time information. Comparison to the race model helps establish if simultaneous use of reaction time and accuracy data are important for establishing multisensory benefit. We have specifically chosen to use an extension of the race model which uses a permutation test of the model to control for Type I error rate (Gondan, 2010; Gondan & Minakata, 2016). For each experiment, the race model was assessed at every 5th

quantile in the data, with 10,001 permutations computed to create an estimate of the distribution of the probability sum per participant. The results of this test are normed such that negative t_{max} values indicate multisensory performance regularly below the probability sum of the unisensory components, and sufficiently large positive values indicate a violation of the race model inequality.

Experiment 1a

Audiovisual Integration

Participants

Experiment 1 was split into two halves based on different staircasing procedures for thresholding used for the participants and the proportions of unisensory and multisensory trials. In experiment 1a, participants were 15 undergraduate students (11 female) from the University of California, Los Angeles, with an average age of 20.20 years (*SD* = 1.15 years). In experiment 1b, participants were 18 undergraduate students (13 female) from the University of California, Los Angeles, with an average age of 19.56 years (*SD* = 1.04 years). Participants in both experiments had normal or corrected-to-normal vision and reported no hearing issues. Participants in both experiments were compensated with course credit for their participation. Prior to the start of the experiment, participants signed an informed consent and were presented with written and verbal instructions of both tasks.

Results

In the detection task portion of experiment 1a, participants showed no significant MSB in mean accuracy (*M* = 0.011, *SD* = 0.067, *t*(14) = 0.640, *p* = 0.99), reaction time (*M* = 0.007, *SD* = 0.076 , *t*(14) = 0.347, *p* = 0.99), or signal detection sensitivity (*M* = -0.019, *SD* = 0.212, *t*(14) = -0.339, *p* = 0.99). RT*, where RT was adjusted by detection rate to assess intersensory

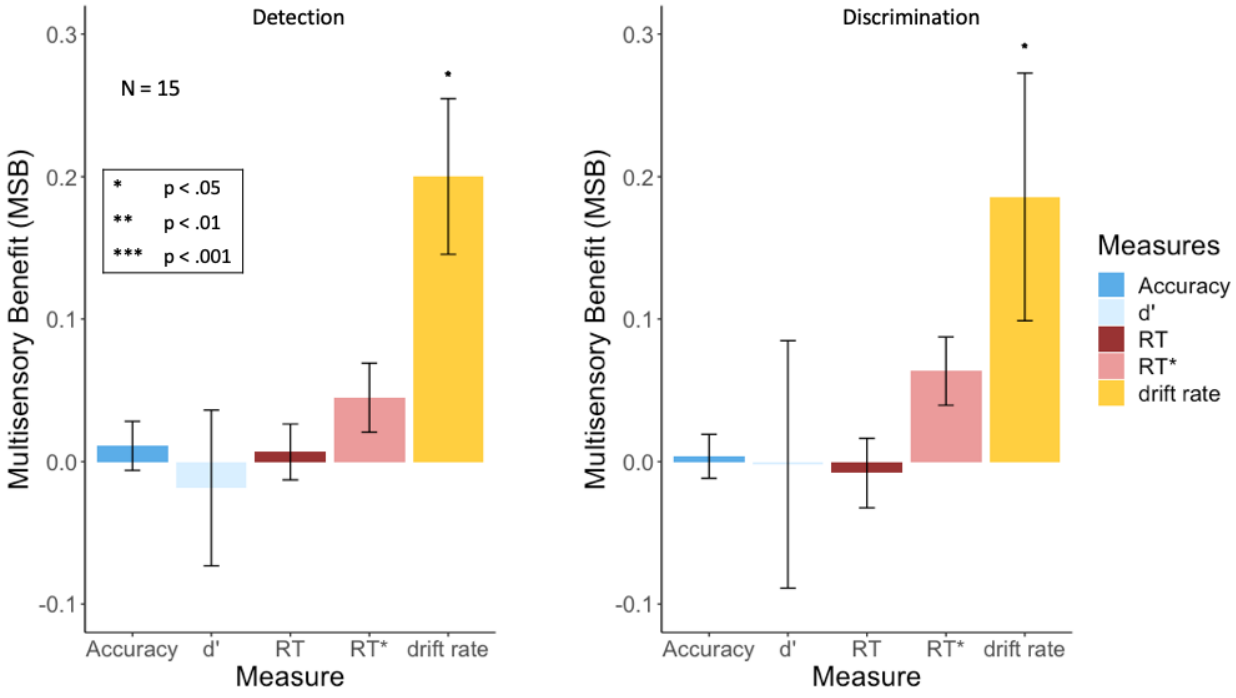
facilitation, likewise showed a non-significant benefit (M = 0.045, SD = 0.094, t(14) = 1.850, p = .342). However, there was a significant MSB for audiovisual drift rates over the unisensory drift rates (M = 0.200, SD = 0.078, 95% credible interval = [0.044, 0.355]), indicating a multisensory advantage was present in the data (Figure 3a). The discrimination task in experiment 1a showed a similar pattern of results (Figure 3b), such that there was no significant MSB in accuracy (M = 0.004, SD = 0.060, t(14) = 0.241, p = 0.99), sensitivity (M = -0.002, SD = 0.336, t(14) = -0.023, p = 0.99), reaction time (M = -0.008, SD = 0.095, t(14) = -0.331, p = 0.99), or RT* (M = 0.063, SD = 0.093, t(14) = 2.648, p = .076). However, drift rate did show a significant MSB in the performance (M = 0.186, SD = 0.072, 95% credible interval = [0.042, 0.327]).

Experiment 1b (Figure 3c and d) showed a somewhat different pattern of results. In the detection task, a significant multisensory benefit was observed in response accuracy (M =0.043, SD = 0.049, t(17) = 3.771, p = .006) and RT* (M = 0.040, SD = 0.045, t(17) = 3.761, p = .006). Unadjusted reaction time (M = -0.006, SD = 0.069, t(17) = -0.383, p = 0.707) and signal detection sensitivity (M = 0.111, SD = 0.206, t(17) = 2.295, p = 0.070) did not show a multisensory benefit. A benefit of multisensory presentation of stimuli on drift rate was also apparent in the detection task (M = 0.135, SD = 0.024, 95% credible interval = [0.087, 0.181]). The discrimination portion of this experiment showed no significant advantages in response accuracy (M = 0.046, SD = 0.079, t(17) = 2.451, p = 0.051) or sensitivity (M = 0.158, SD = 0.363, t(17) = 1.845, p = 0.083), but there was a significant benefit in reaction time (M = 0.047, SD = 0.058, t(17) = 3.423, p = .010) and in RT* (M = 0.104, SD = 0.088, t(17) = 5.037, p < .001). Further, drift rate, still showed a significant benefit for multisensory stimuli in the discrimination task (M = 0.403, SD = 0.092, 95% credible interval = [0.247, 0.611]) tasks.

Tests of the race model inequality for both portions of experiment 1a indicate that there was no significant deviance from predicted sensory facilitation for either detection (t_{max} = -2.898, p > 0.99) or discrimination tasks (t_{max} = -1.493, p > .99). The same was found for detection (t_{max} = -3.522, p > .99) and discrimination tasks (t_{max} = 0.319, p = .716) in experiment 1b. As such,

we do not observe multisensory benefit in reaction time that exceeds the expectation of statistical facilitation between multiple sensory signals.

(a) (b)



(c) (d)

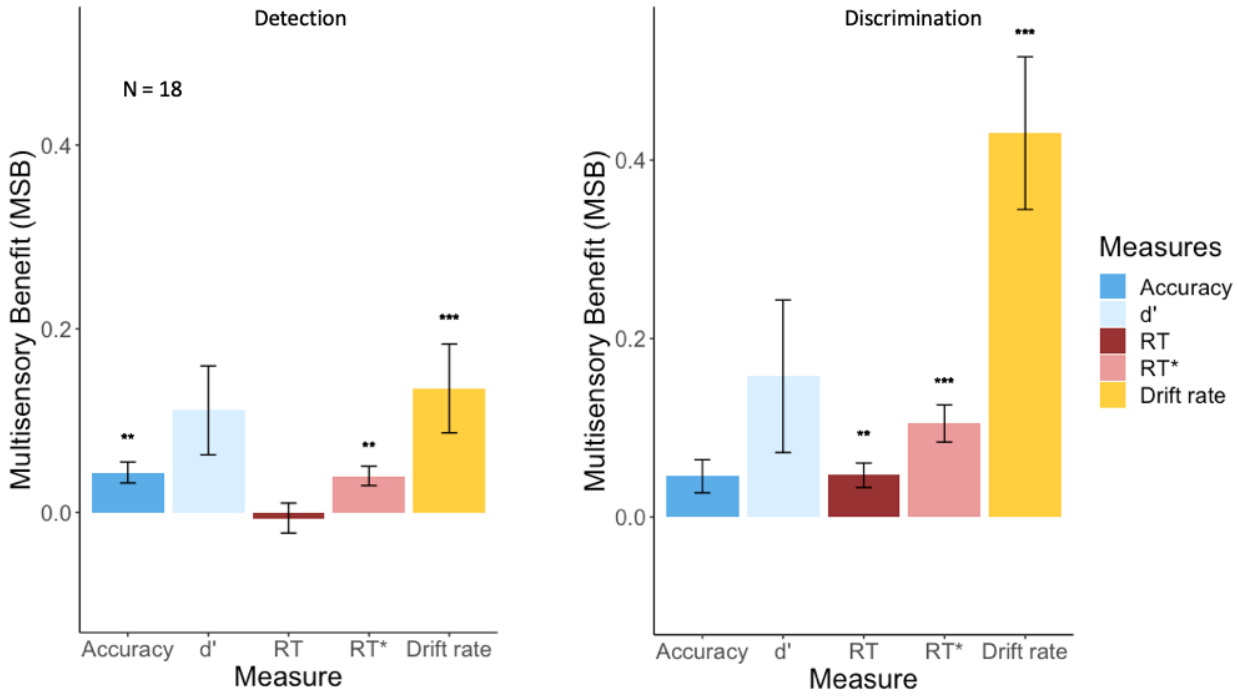


Figure 3: Results from Experiment 1a

Advantage in accuracy, d', RT, and RT* for (a) the detection and (b) discrimination portions of experiment 1a, as well as the(c) detection and (d) discrimination tasks in experiment 1b.

Experiment 1b

Visuotactile integration

Participants

Participants in experiment 2 were 17 undergraduate students (7 female) from the University of California, Riverside, with an average age of years 22.75 years (*SD* = 5.43 years). All had normal or corrected-to-normal vision and reported no tactile issues. Participants were compensated with 10 dollars per hour of their participation. Prior to the start of the experiment, participants signed an informed consent and were presented with written and verbal instructions of both tasks.

Results

Results for the visuotactile detection task (Figure 4a) showed a significant multisensory benefit for response accuracy (M = 0.097, SD = 0.072, t(16) = 5.578, p < .001) and sensitivity (M = 0.274, SD = 0.269, t(16) = 4.190, p = 0.001). There was no significant advantage in reaction time (M = 0.021, SD = 0.055, t(16) = 1.557, p = 0.139), however, RT* did show a significant multisensory benefit (M = 0.085, SD = 0.063, t(16) = 5.520, p < .001). Drift rate also revealed a multisensory performance benefit (M = 0.510, SD = 0.058, 95% credible interval = [0.395, 0.623]). The discrimination task showed a similar pattern of results (Figure 4b), where significant multisensory benefits was observed in sensitivity (M = 0.134, SD = 0.190, t(16) = 2.912, p = .041), but not in reaction time data (M = 0.020, SD = 0.056, t(16) = 1.442, p = .168). Accuracy was marginally significant (M = 0.033, SD = 0.051, t(16) = 2.668, p = .051), as was RT* (M = 0.030, SD = 0.047, t(16) = 2.594, p = .051). Drift rates, again, indicated there was a significant benefit obtained from multisensory stimulus presentation (M = 0.279, SD= 0.048, 95% credible interval = [0.183, 0.376]).

Race model inequality tests indicated no significant difference between observed and predicted multisensory reaction time in either detection (t_{max} = -2.564, p > .99) or discrimination (t_{max} = -2.495, p > .99).

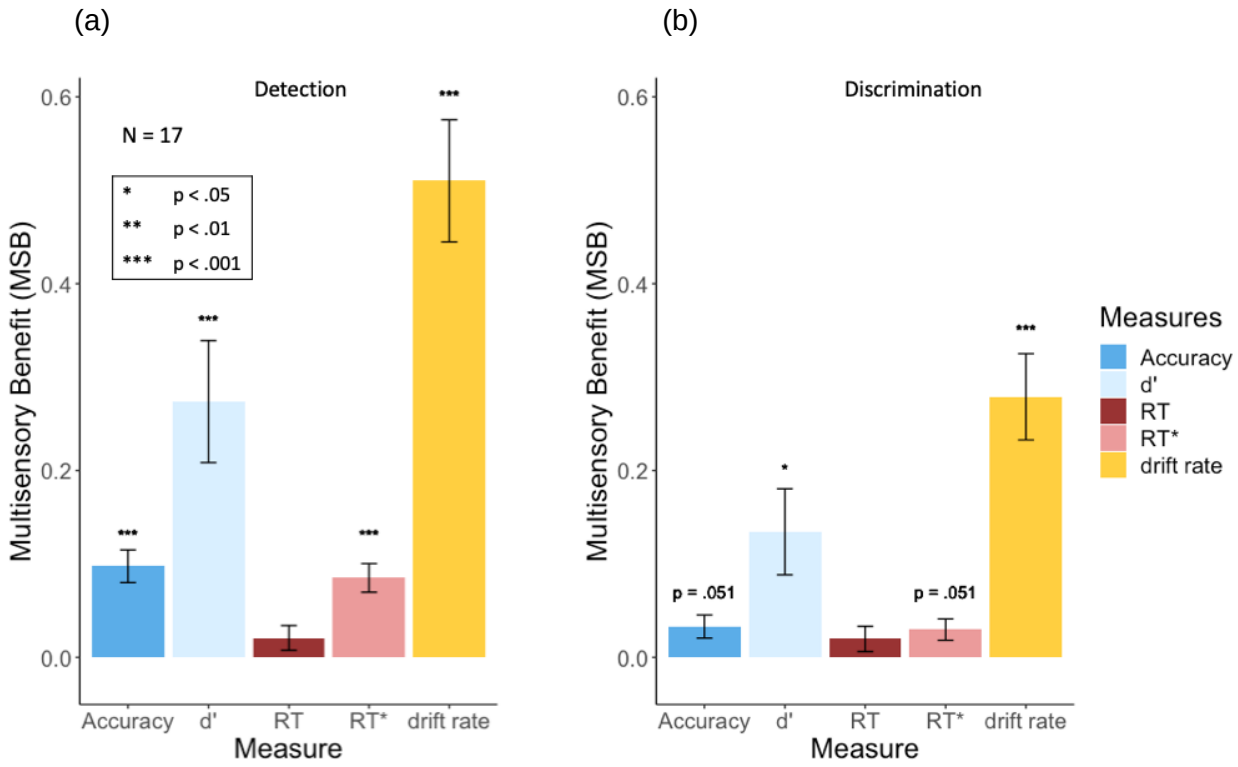


Figure 4: Results from Experiment 1b

For the visuotactile experiment, the average advantage in accuracy, d', RT, RT*, and drift rate for (a) detection and (b) discrimination tasks. Visuotactile data showed advantages in accuracy, sensitivity, inverse efficiency score, and drift rate, while also showing no advantage for participant RT.

Experiment 1c

Audiotactile integration

Participants

Participants in experiment 3 were 17 undergraduate students (11 female) from the University of California, Riverside, with an average age of 21.18 years (SD = 2.34 years). All

reported normal hearing and reported no tactile issues. Participants were compensated with 10 dollars per hour of their participation. Prior to the start of the experiment, participants signed an informed consent and were presented with written and verbal instructions of both tasks.

Results

In the audiotactile detection task (Figure 5a), the data revealed no significant benefit for multisensory stimuli compared to unisensory stimuli in accuracy (M = -0.052, SD = 0.135, t(16) = -1.600, p = 0.388), sensitivity (M = -.102, SD = 0.415, t(16) = -1.013, p = 0.652), reaction time (M = -0.010, SD = 0.043, t(16) = -0.987, p = .652), or RT* (M = -0.010, SD = 0.043, t(16) = -1.897, p = .304). Analysis of model drift rates also showed no significant MSB (M = -0.015, SD = 0.092, 95% credible interval = [-0.120, 0.165]).

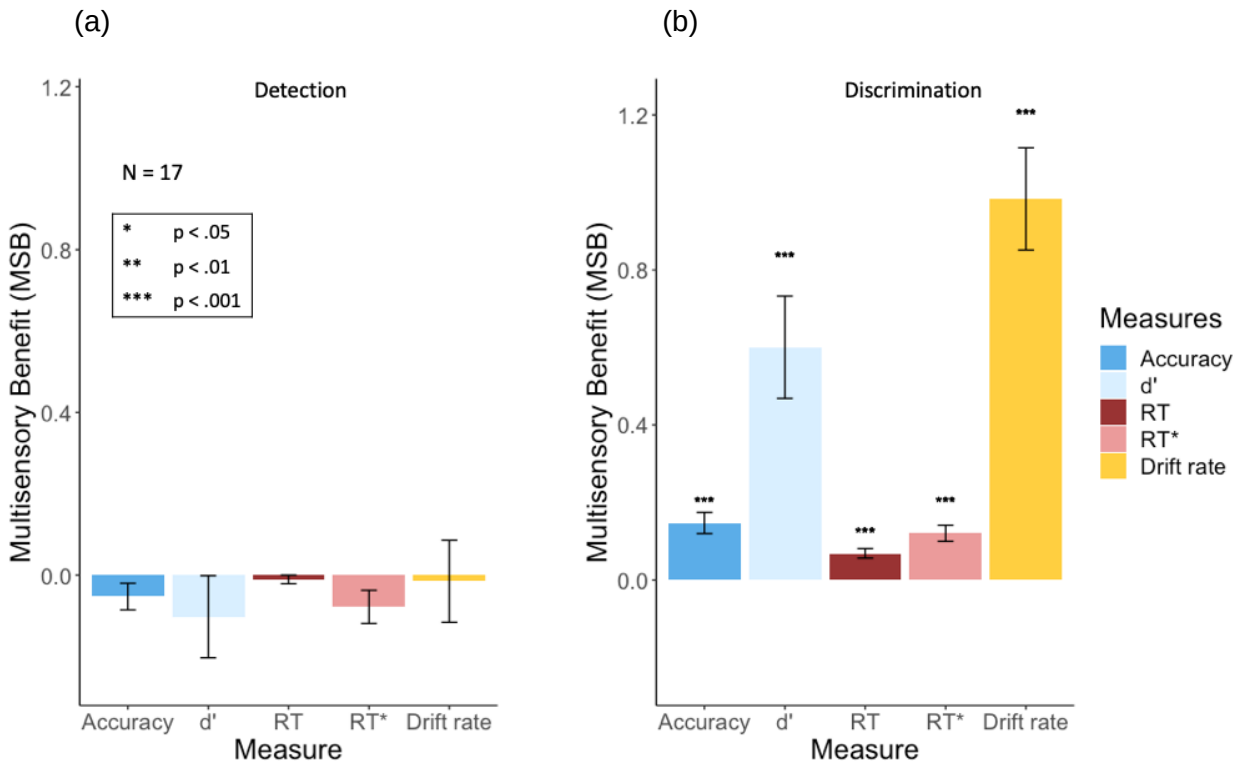


Figure 5: Results of experiment 1c

For the audiotactile experiment, the average advantage in accuracy, d', RT, RT*, and drift rate for (a) detection and (b) discrimination tasks. Audiotactile detection showed no multisensory advantage in any of the measures used, and audiotactile discrimination showed an advantage in all of the measures used.

Audiotactile discrimination results (Figure 5b) showed a significant MSB for multisensory presentation of stimuli in accuracy (M = 0.148, SD = 0.112, $t(16) = 5.419$, $p < .001$), sensitivity (M = 0.601, SD = 0.543, $t(16) = 4.449$, $p < .001$), reaction time (M = 0.069, SD = 0.051, $t(16) = 5.580$, $p < .001$), and RT* (M = -0.077, SD = 0.168, $t(16) = 5.828$, $p < .001$). The model drift rates also indicated a significant MSB (M = 0.983, SD = 0.123, 95% credible interval = [0.741, 1.229]).

Race model results for experiment 3 indicated no significant deviation from probability sum predictions in either detection (t_{max} = -6.771, p > .99) or discrimination tasks (t_{max} = 1.066, p = .462).

Discussion

The results presented here suggest that diffusion models are sensitive to multisensory benefits, across both detection and discrimination tasks, and across different multiple sensory combinations. In each of the presented experiments, drift rate was found to be at least as sensitive as measures of accuracy, sensitivity, reaction time, inverse efficiency scores, and the race model. Across all of these experiments, drift rate was consistently the most reliable measure of multisensory benefit. The next most consistent measure was RT*, the only other measure that combined RT and accuracy, highlighting the need for a combined measure. However, drift rate in experiment 1a did pick up a benefit even when RT* did not, indicating a benefit of the diffusion model. While multisensory benefit may manifest in a variety of different

measures of performance, here, we show that in the majority of cases, the benefit is manifested in the change in drift rate, and less consistently in other measures commonly used in the literature such as accuracy, reaction time, or sensitivity.

Is it possible that drift rate shows a benefit when it should not (i.e., when in reality there is not a true multisensory benefit to processing)? In principle, this possibility cannot be ruled out, the same way that any other measure (such as accuracy and reaction time) can by random chance exhibit a benefit. In the absence of a ground truth about the presence of enhanced neural processing, one cannot rule out the possibility of false positives in any measure. Strictly speaking, ground truth about multisensory benefit is never available, even if one could track the activity of all neurons in the brain in different conditions. However, the literature on multisensory processing in the last two decades has established that multisensory redundant signals generally result in enhanced processing compared to unisensory conditions, as shown in a variety of perceptual tasks, settings, and sensory combinations (for examples of reviews and overviews, see Groh, 2014; Murray & Wallace, 2011; Trommershauser et al., 2011; Shams & Kim, 2010; Calvert et al., 2004; Ernst & Bülthoff, 2004). Further, the results shown here were replicated across multiple experiments and sensory combinations. Therefore, it is reasonable to expect that the redundant multisensory signals in the current experiments also result in superior processing over those of unisensory conditions. Therefore, we consider it reasonable to conclude that the drift rate is reflective of a true benefit in processing compared to unisensory conditions. Still, future work should examine how the relative false positive rates of diffusion models compare to alternative methods.

Interestingly, in the experiment 1a discrimination task, no benefit was seen in either task for any measure but the drift rate. The pattern of these results appears to reveal a difference in the more accurate and the faster unisensory conditions, and a multisensory condition that captures the best of these separate components. Of the 15 participants in the discrimination task, 14 of those were, on average, more accurate in the auditory condition. Of those 14

participants, 11 responded, on average, more quickly in the visual condition. This means that the majority of our participants showed higher accuracy in their slower modality. As drift rate combines these into a single measure, we believe that this reflects that the multisensory condition resembles the speed of the faster unisensory modality and the accuracy of the more accurate unisensory modality, and only then in combination does this benefit emerge.

The significant benefit to RT and RT* without a parallel race model inequality violation, observed across multiple experiments, may also initially appear somewhat contradictory. However, it should be noted that these measures are quantifying benefit against different baselines. RT and RT* are comparing multisensory performance to a single unisensory performance, whereas the race model inequality compares multisensory performance to an additive combination of the unisensory components. Thus, such results can occur when performance in the multisensory condition exceeds those of unisensory conditions but not the probability sum (indicating no coactivation between these senses).

That the model can robustly detect multisensory benefits, across a number of sensory combinations or patterns of benefits, makes drift diffusion models a potentially consistent tool for analyzing the results of multisensory studies, as well as memory studies. Furthermore, this particular variant of the model, as it can create estimates in a relatively small number of trials, may more easily allow this to be done on experiments with a reasonable memory load, without the need to include hundreds of trials. The current experiments had between 100 and 200 trials per condition, though the model has been used with fewer (as in Regenbogen et al., 2016). Drift diffusion models have also, traditionally, imposed design restrictions on experiments, as they are typically only used to analyze tasks with binary decisions and are limited to one-step decisions (Ratcliff & McKoon, 2008; Voss et al., 2013). While an issue for use, then, in free recall tasks, this should be employable in recognition tasks, where participants are asked to make judgments more easily compressed into a decision binary, such as new/old distinctions. This is not to say it would be impossible, however, to employ diffusion modeling on more

continuous free recall tasks; Extensions of the model that address these issues do exist that allow for more alternatives in the decision, up to and including continuous response scales (Ratcliff, 2018; Smith, 2016; Krajbich & Rangel, 2011), allow for multiple steps in a decision-making process (Pleskac & Busemeyer, 2010; Resulaj et al., 2009), and allow for participant bias to change through sequential trials (Nguyen et al., 2019), though such variants are less commonly used.

Also, we would caution that this particular model is not the only option for use in analysis of multisensory stimuli or memory performance. While the HDDM uses Bayesian priors on its parameters to allow the model to converge more rapidly on parameter values, it is not the only Bayesian drift diffusion model available and drift diffusion modeling is not the only technique available to simultaneously investigate reaction time and accuracy. Previous studies on this topic have, in fact, shown that this particular Bayesian hierarchical method does not necessarily outperform alternative models in terms of efficiency (Lerche et al., 2017). Other groups have created similar hierarchical models (Vandekerckhove et al., 2011) that can be implemented using similar Bayesian optimization techniques via Gibbs sampling (Wabersich & Vandekerckhove, 2014). Other non-hierarchical models have also attempted to improve efficiency by limiting the range in which model parameters can fall (Diederich & Busemeyer, 2003; Nidiffer et al., 2018). Indeed, the hierarchical solution provided here is only one of the possible solutions to making this model more computationally efficient, and such methods are being worked into traditional DDMs as well as hierarchical variants. More traditional variants of sequential sampling models are also beginning to become more efficient in their processing, including the *Ornstein-Uhlenbeck model*, a variant of a sequential sampling model that combines evidence accumulation with a decay term that acts to bring the evidence accumulation back towards the starting point (Ratcliff & Smith, 2004; Diederich, 1995), or the *compatibility bias model* (Yu et al., 2009), that investigates decision-making processes through the framework of Bayesian causal inference modeling.

A significant benefit that can be reaped from any of these models, though, beyond just the ability to investigate a speed-accuracy tradeoff in multisensory experiments, is the ability to break down the decision-making process to investigate what portion of the decision-making process is influenced by the presentation of multisensory information. The current experiment focused on drift rate as a parameter of interest because we hypothesized that the evidence accumulation process would be most affected by the inclusion of a second sensory modality. However, the parameters available through this type of modeling allow a large number of features of the decision-making process to be investigated. The compatibility bias model has been used to investigate how multisensory information may be differentially used by older and younger adults and found that reduced reaction times in older adults were caused in part by more conservative decision making, in terms of larger boundary separation, and slower non-decision response time (Jones et al., 2019). This allows this type of decision-making model to additionally provide more insight into where in the decision-making process multisensory stimuli may exert influence, allowing us to better categorize where the benefits of multisensory stimuli may arise from. This may be helpful for forming hypotheses about how multisensory stimulus presentation may improve memory performance.

Given the potential benefits of this type of modeling, then, we would generally advocate for greater use of such sequential sampling models, especially those that allow for a smaller number of trials to be used effectively. This could provide greater insight into both when and how multisensory stimuli benefit performance on a large number of tasks, with a reliability that has the ability to surpass current methods that independently investigate reaction speed and response accuracy.

Chapter 3: Multisensory encoding of names via name tags facilitates remembering

Abstract

Associating names to faces can be challenging, in part because this task lacks an inherent semantic relationship between a face and name. The current study seeks to understand whether bolstering names with cross-modal cues—specifically, name tags—may aid memory for face and name pairings. In a series of five experiments, we investigated whether the presentation of congruent vocalized and written names at encoding might benefit subsequent cued recall and recognition memory tasks. The results showed that participants, cued with a picture of a face, were more likely to recall the associated name when those names were encoded with a name tag (a congruent visual cue) compared to when no supporting cross-modal cue was available. The findings were consistent with a benefit of multisensory encoding, above any effect from the availability of independent unisensory traces, extending previous findings of multisensory learning and memory benefits to a naturalistic associative memory task.

Introduction

As any individual who has been to a large gathering can attest, remembering the association between names and faces is a challenge. While associative memory tasks tend to be among the most challenging in a laboratory setting, memory for names is one that is considered especially so, in part due to the lack of an inherent, semantic relationship between a face and name (e.g., there isn't anything about one face that makes it seem more "Hannah" than another). This makes learning names a challenge, which has been the topic of much past research (for examples, see McWeeny et al., 1987; Cohen & Faulkner, 1986; Brooks et al., 1993).

Previous studies have tested a number of different approaches to improve recall of names associated with particular faces (we will provide a brief overview, but see Brédart, 2019 for a more comprehensive review). Spacing the learning of the names has been shown to improve recall performance (Carpenter & DeLosh, 2005), as has retrieval practice of particular names (Morris et al., 2005). However, these previously studied approaches may be difficult to use in the real world; for example, one probably cannot control how many people they meet at a conference, let alone the spacing between these meetings. Semantic associations or mental imagery devices, such as creating a mnemonic around the name and associating this with a physical feature or fact about the person, have also been shown to improve how well names are remembered (e.g., McCarty, 1980). Comparisons of these techniques, however, show that mnemonic techniques are less effective than spacing (Morris et al., 2005; Neuschatz et al., 2005), and so may benefit from additional supporting cues.

More recent research has begun to tap into the connection between the sensory content available at the time of encoding and the conditions present during later retrieval. Previous work has shown that encoding an audiovisual of a person talking is more effective for subsequent recognition of their voice than encoding the voice alone, showing the superiority of audiovisual

encoding over auditory encoding in auditory recognition (von Kriegstein & Giraud, 2006). Learned congruence between a face and voice has been reported to speed recognition of a familiar face-voice pair, compared to an incongruent audiovisual pairing (O'Mahony & Newell, 2012). Interestingly, it has also been found that regions of the brain involved in audiovisual integration—for creating an association between congruent audio and visual cues—are activated more strongly during encoding for faces that will later be remembered than for those that are forgotten (Lee et al., 2017), so perhaps multisensory stimuli can support recall of face-name associations.

Facilitation of memory by utilizing multiple sensory cues would be consistent with a few memory models, most notably with dual-coding theory (see Clark & Paivio, 1991; Paivio, 1991 for reviews), wherein providing verbal and non-verbal representations (that often occur across different senses) can facilitate memory. Another similar model is the cognitive theory of multimedia learning (see Mayer, 2014 for overview), wherein presentation of stimuli across verbal and pictorial working memory channels allows for better learning. In general, encoding information across different channels can provide more routes by which a memory can be accessed. This would seemingly support findings in the multisensory research literature that multisensory information can improve memory, as multisensory information provides information through at least two senses, while unisensory information can provide only one sensory route to a memory. However, this particular framework fails to make a distinction between having information available across multiple senses and unified multisensory experiences, where congruency (temporal, spatial, structural, semantic, etc.) between stimuli can lead to the creation of integrated multisensory representations (e.g., Spence, 2007; Laurienti et al., 2004; Lacey et al., 2009; Butler et al., 2012; Ernst & Bülthoff, 2004; Shams & Kim, 2010). The integration of cues from multiple sensory modalities can result in overall improvement of the sensory signals, by, for example, uncertainty reduction leading to improved precision and/or accuracy. We seek to investigate if multisensory integration mechanisms, in particular, are able

to support remembering face-name associations, beyond any benefit provided by multiple unisensory traces.

Previous work indicates there may be a memory benefit to presenting stimuli with meaningful and congruent cross-modal sensory inputs (see Matusz et al., 2017; Shams & Seitz, 2008 for an overview). For instance, studies have shown that object images are recognized better when they are originally presented with their iconic sound compared to when they are presented without sound, even when only the visual cue is presented at test (Lehmann & Murray, 2005). Similarly, auditory recognition is better for objects originally presented together with congruent images compared to audio-alone encoding (Moran et al., 2013), or to presenting the sound with a meaningless visual stimulus (Thelen et al., 2015). Improvements to recognition memory performance were also shown to extend to written words accompanied by audio of those words (Heikkilä et al., 2015; Heikkilä & Tiippana, 2016). While the exact mechanisms by which multisensory encoding benefits recall or recognition remain unexplained (though see proposed mechanisms in Shams & Seitz, 2008), electroencephalographic (EEG) signals measured during memory retrieval begin to diverge at a relatively early stage of processing for visual versus audiovisual information (Murray et al., 2004), indicating that multisensory stimulus encoding may involve distinct processes not triggered by unisensory encoding. This would suggest that there is a distinct benefit to using multisensory cues as opposed to multiple unisensory ones, which could provide an avenue to boost memory performance in everyday tasks.

The present research seeks to expand upon these findings in a number of ways, by exploring how such mechanisms could be translated into benefitting naturalistic memories for face-name associations. Of particular note in the case of name memory, where the face and name share no semantic information, providing a visual cue that is semantically congruent with the auditory cue may prove to be beneficial. In-person introductions inherently engage multiple senses in a cross-modal associative learning task, the association between a name and a face.

However, each component of the association is presented in only one modality– the face is visual, and the name is auditory. To bolster memory performance for the association between a face and name, it could be beneficial to enhance each of those components by making it multisensory, and thus creating a multisensory representation for each component. While there is not a simple way to transform seeing a face into a multisensory experience (short of touching a face, which is seldom socially acceptable), the auditory presentation of the name (ie., the spoken name) could be augmented with a visual representation, by for example, the addition of a name tag. When name tags are presented in one's native language, they provide a visual component to an introduction that is congruent specifically with the auditory information being given. Name tags thus provide a natural correspondence with the spoken name and are an ideal cue for testing whether multisensory stimulus presentation can aid with associative memory tasks.

Here, in a series of experiments, we systematically investigate the role of multisensory presentations in associative memory. We present a multisensory representation of a name through the use of vocalized names and congruent name tags, to see if a multisensory stimulus presentation would aid face-name memory. In experiment 1, we test if presenting a name tag during an introduction will improve memory for names when participants are later probed with previously encountered faces. In experiment 2-4, we alter initial stimulus presentation to rule out the influence of visual text guiding attention, lip reading, and duration of time spent with the name on cued recall improvement when a name tag is provided. In experiment 5, we test whether the synchrony of the auditory name and the visual tag– with synchrony being an important factor in multisensory integration– is useful above merely providing more information, to investigate if the memory improvement is mediated by multisensory integration. If multisensory stimulus presentation is generally helpful for this process, then performance with congruent and synchronous name tag presentation with the name should improve memory performance above and beyond the baseline in each experiment.

Experiment 1

In this experiment, we examined whether name tags can improve memory of names using a within-subject design in which during the encoding phase half of the trials included a name tag and half of them did not. We hypothesized that the addition of this visual information would improve the recall of names.

Methods

Participants

Participants were 38 undergraduate students (22 females) at the University of California, Los Angeles. Average participant age was 19.49 years (*SD* = 1.07), and all reported normal or corrected-to-normal sight and hearing, except for one participant who reported that they did not have corrected-to-normal sight, but reported no difficulty observing the stimuli on the computer screen and were thus included in the analyses. Additionally, 30 of these participants were native English speakers. The remaining 8 were fluent in English. Initial analyses indicated that the results did not differ if non-native speakers were excluded, so those participants were kept in the analyses for this and the follow-up experiments. Two participants were excluded from analyses due to computer errors resulting in incomplete session data.

Written informed consent was obtained from each participant and experimental procedures were reviewed and approved by the UCLA Institutional Review Board.

Materials

Experimental stimuli were 60 brief video clips (1-2 s duration) of young adults (age 18-22; half male, half female) captured from the chest up against a white background. In each

video, the speaker introduced themself with the phrase “Hello, my name is [name].” Names presented during these videos were selected from the most common first names given to male and female children in the United States between 1990 and 1999 as reported by the United States Social Security Administration, so all of the names would have similar familiarity to participants.

A white rectangle acting as a name tag was placed over the chest and neck of each individual video but did not obscure the mouth. This remained in the same location for the duration of the experiment. During half of the trials, this rectangle remained blank, presenting no additional name information to the participants (the “no tag” level of the name tag condition). In the other half of the trials, black text spelling the name given in the video was presented for the duration of the video in this white rectangle (the “tag” level of name tag condition). See Figure 6 for examples of both conditions.

Experimental stimuli were presented using PsychoPy software (Peirce et al., 2019) on a Mac Mini computer.

Procedure

The 60 videos were presented to participants across 4 blocks. During each block, participants were shown 15 videos with a mix of genders. As this meant that there were an uneven number of trials in each block, the first and third block had 8 tag trials, and the second and fourth blocks had 7 tag trials. Individuals within each block were presented in a random order, and the order of these blocks of individuals were also randomized.

In each block, participants were first given an encoding phase (Figure 6a), where they were presented with each of the 15 videos in that block. To ensure participants were attending to the videos, they were asked to make a button-press response to report the gender of the speaker after each video, using ‘1’ to indicate a male speaker and ‘2’ to indicate female speaker. Participants were not informed that they would be tested later on their memory of the

names. After seeing and reporting the gender for all 15 videos in the block, participants were given a 3-min break during which they were asked to close their eyes and relax. At the end of this delay, participants were given a cued recall test of the name. They were presented with a still image from each of the videos they had seen before the delay, in a randomized order, and prompted to type in the name they remembered being associated with that person. Still images were created from the final frames of each video, and were selected such that the faces had closed lips, to remove any facial cues for sounds in the name. Participants were given 10 s to recall and type the name; after the 10 s, the experiment would advance to the next question.

After each cued recall attempt, participants were asked to rate their confidence in their memory for the name on a scale from 1 (low confidence) to 4 (high confidence). After being tested on all 15 names and providing confidence ratings, participants were given a 1-min break before moving on to the next block.

Analysis

Participant responses were rated by three blind raters for correctness, as well as by computer test-matching. The human- and computer-based scoring did not qualitatively alter the results, so human ratings were used to allow for spelling errors and alternative name spellings. Human raters were instructed to rate a response as correct if the typed response was an alternative spelling of a name, if an answer was cut off by the response time limit and could not reasonably be mistaken for another name, or if the name typed was a shortened version of the correct name that could not be mistaken for another name. If any of the raters judged a participant's attempt as correct, the response was marked as correct for final analysis. Raters largely agreed with one another, such that all three raters matched their judgments on 98% of responses.

Reaction time (RT) was collected for each key press in the typed response, and was analyzed using the first key input from the participant. As reaction times were non-normal in

their distribution, median values on correct trials were used in the analyses. In the case where participants did not have any correct responses in one condition, they were removed from pair-wise analyses.

Data used in these analyses have been made available in a GitHub repository (https://github.com/murray-carolynA/Data_MultisensoryNametagStudy).

Results

Results from the attention check (i.e., the gender judgment task) during the encoding phase showed high accuracy across all participants (*M* = 98.6%, *SD* = 2.4%), indicating they were attending to the stimuli at encoding. As such, all trials were included in the final analyses.

Initial analyses showed that performance differences between name tag conditions persisted across blocks, regardless of whether participants did not know their memory would be tested (as in block 1) or if they did (all subsequent blocks), so analyses collapsed performance across blocks (Fig 6b). Descriptive statistics of participant performance are printed in Table 1. Pairwise one-way t-test comparison of accuracy between the two conditions showed superior recall performance in the tag condition over the no-tag condition ($t(36) = 3.59$, $p < .001$, Cohen's $d = 0.44$; calculated as in Cohen, 1988, such that $d = \frac{M_{tag} - M_{no\ tag}}{\sqrt{\frac{sd_{tag}^2 + sd_{no\ tag}^2}{2}}}$

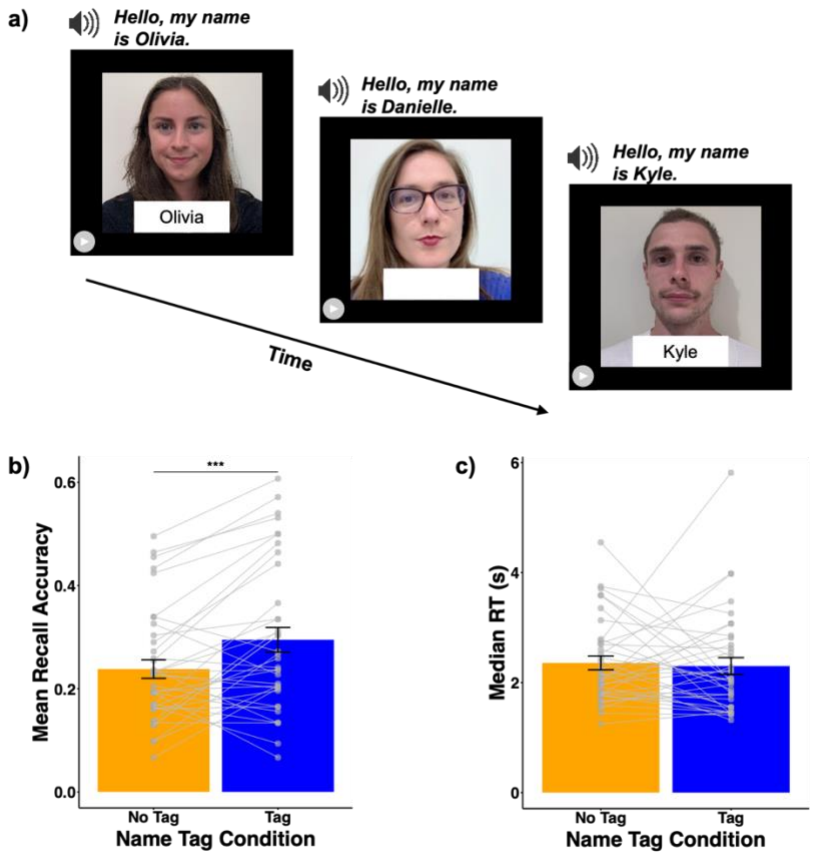


Figure 6: Methods & Results of Study 2, Experiment 1

(a) A diagram of the encoding procedure. Participants were presented with 15 videos per block, half with a name tag and half without. (b) Recall performance. There was a significant main effect such that participants recalled a higher proportion of names presented with a name tag compared to those presented without. The overlaid scatter plot represents individual participant scores. (c) Reaction time (for correct responses, in seconds) for the recall task measured as the first keystroke made in the response. Error bars are standard errors.

Pairwise one-way t-test comparison of RT for correct responses showed no significant effect of name tag condition ($t(36) = 0.32$, $p = .75$; Fig. 6c). The results were the same when including all trials.

Confidence results generally tracked the accuracy data across experiments, and can be found in the Supplementary material.

Table 1: Descriptive Statistics for Accuracy and Reaction Time, Study 2, Experiments 1-5

Means for accuracy are reported as a proportion correct, and reaction times (RT) is reported as the average of the median response times for correct responses in seconds. Standard deviations for each measure are provided in parentheses.

		Recall Accuracy	Recall RT	Recognition Accuracy	Recognition RT
Experiment 1	**Name tag**	0.29 *(0.15)*	2.30 *(0.93)*	N/A	N/A
	No tag	0.24 *(0.11)*	2.36 *(0.76)*	N/A	N/A
Experiment 2	**English Tag**	0.25 *(0.13)*	2.21 *(0.59)*	N/A	N/A
	Armenian Tag	0.19 *(0.11)*	2.67 *(1.00)*	N/A	N/A
Experiment 3	**Name tag**	0.27 *(0.15)*	2.11 *(0.79)*	N/A	N/A
	No tag	0.19 *(0.13)*	2.71 *(1.33)*	N/A	N/A
Experiment 4	**Name tag**	0.31 *(0.13)*	2.25 *(0.45)*	0.54 *(0.15)*	2.96 *(0.81)*
	No tag	0.27 *(0.14)*	2.34 *(1.03)*	0.51 *(0.15)*	3.04 *(0.91)*
Experiment 5	**Synchronous**	0.43 *(0.17)*	3.09 *(0.98)*	0.64 *(0.17)*	2.40 *(0.76)*
	Asynchronous	0.39 *(0.14)*	3.24 *(1.12)*	0.62 *(0.13)*	2.67 *(0.85)*

Interim Discussion:

Results from this experiment indicate that participants do perform better when they are given a semantic visual cue congruent with the auditory stimuli, even if these stimuli were not

available at the time of retrieval. This would seem to generally support the utility of multisensory stimulus presentation for this type of recall task.

However, there are alternative explanations for the observed superiority of the tag condition. For example, it has been shown that objects presented with an accompanying irrelevant stimulus in a different modality can improve memory for that object relative to objects presented alone (Matusz et al., 2017). Alternatively, the presence of the name tag may have increased the salience of the visual stimuli, and therefore led to higher arousal in the tag condition, compared to the no tag condition that contained a blank rectangle. To investigate if the mere presence of an additional visual in the form of a name tag could explain improved performance in the tag condition, we conducted a second experiment.

Experiment 2

In this experiment we investigated whether the superior memory performance in the previous experiment was due to the difference in visual salience in the two conditions. We compared the performance between two name tag conditions that had equal visual salience and only differed in the semantic content. In one condition, the name tag could be read and understood by the participants (Latin alphabet, hereafter called the English condition), and in the control condition it was written in an unfamiliar alphabet (Armenian alphabet, hereafter called the Armenian condition) that was unfamiliar and incomprehensible to the participants. If the difference in performance observed in the previous experiment was due to visual salience of the tag, then that difference should disappear in this experiment. On the other hand, if the superiority of the tag condition was due to the additional visual semantic cue, then we should observe a superior performance of the English name tag over the Armenian name tag.

Methods

Participants

Participants were 41 undergraduate students at the University of California, Los Angeles, with an average age of 19.66 years (*SD* = 1.86). Two participants were excluded from analyses because they knew individuals in the videos from everyday life by different names, leaving 39 participants in the analysis (30 female). All participants except one reported normal or corrected-to-normal vision, and all reported having normal hearing and being unable to read Armenian. The participant who reported not having corrected-to-normal vision reported no difficulty seeing the stimuli on the computer screen, and was kept in the analyses.

Materials & Procedure

Videos and the name tag format matched those in experiment 1 except in the no tag condition. In this experiment, to control for the visual saliency of having a name tag, the no tag condition was replaced by a condition with a name tag written in an alphabet unfamiliar to the participants. In this half of trials, the name was written in the Armenian alphabet, so the size and shape of the letters would be similar to the names written in a familiar alphabet (see Fig. 7a), but the congruency between the visual and audio signals would not be present for participants.

The procedure and data analysis matched that of Experiment 1.

Results

Pairwise t-test comparison of cued recall accuracy showed superior performance in the English name-tag condition compared to the Armenian name tag condition , $t(38) = 4.21$, $p < .001$, Cohen's $d = 0.54$, (Fig 7b). Median reaction time for correct responses also showed an effect of tag, $t(38) = 3.16$, $p = .003$, Cohen's $d = 0.57$, such that median response time was shorter when recalling names originally presented with an English tag than an Armenian one

(Fig 7c). The results were qualitatively the same when including all trials (including incorrect responses).

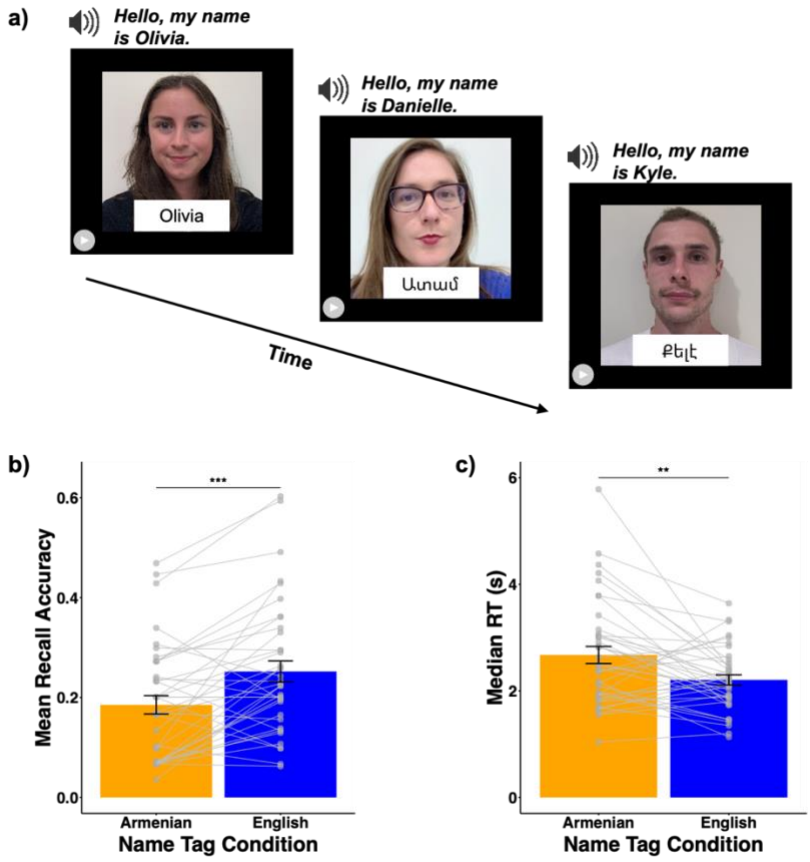


Figure 7: Methods & Results for Study 2, Experiment 2

(a) A diagram of the encoding procedure for Experiment 2. Armenian names are presented on the trials lacking a congruent name tag to control for the visual saliency of having a name tag with writing. (b) Recall performance. Participants showed significantly higher recall for names when the face was initially presented with an English name tag than with an Armenian name tag. *** = $p < .001$. The overlaid scatter plot represents individual participant scores. (c) Reaction time (for correct responses, in seconds) for the recall task measured as the first keystroke made in the response. Error bars are standard errors.

Interim Discussion

Results from this experiment show that semantically congruent visual stimuli can facilitate remembering names and rules out the role of visual salience and arousal as the underlying mechanism for this facilitation. These findings are consistent with the hypothesis of multisensory integration as an underlying mechanism for improved memory.

It is important to note that in the control conditions for the previous two experiments (no name tag or Armenian name tag) there are two cues available about the name: the acoustic cue (the voice) and the lip movement cue. Previous work has shown that lip reading may provide important multisensory cues, and can assist with disambiguating sounds (Bernstein et al., 2004). The written name (name tag) information can help encoding the face-name association in two different ways: by disambiguating (reducing the uncertainty) of the lip movement cue or by disambiguating the auditory cue. In order to gain insight into which process is occurring, we conducted the following experiment.

Experiment 3

The goal of this experiment was to gain insight into the role of the lip movement cue in the facilitation effect of name tags observed in previous experiments. To investigate whether the observed facilitation effect stems primarily from the disambiguation of lip movements by the name tag (both visual cues, but one perceptual and the other semantic) videos were replaced by still images in this experiment , to remove lip reading cues. To the extent that the benefit of the name tag cue stems from its interaction with the lip movement cue, in this experiment the effect should disappear or be weakened. Conversely, if the benefit of name tag stems primarily

from interaction with the auditory cue or just by providing an additional source of information without interacting with the other cues, then the effect should remain the same here.

Methods

Participants

A total of 44 participants (37 female), who were all undergraduate students at the University of California, Los Angeles, were enrolled. Participants had an average age of 19.58 years (*SD* = 2.11), and all reported normal or corrected-to-normal sight and hearing. Thirty-five reported being native English speakers, and all reported being fluent in the language. Two participants were excluded from analysis due to computer issues interrupting the experiment.

Materials & Procedure

Materials had one major change from the preceding experiments: the video of the speaker was replaced by a still image of the individual from the video, to remove the ability of participants to use lipreading to help with the task. The images were taken from the end of each video, selected so the speakers' lips were closed and provided no cues for what sounds the individuals may have been speaking. Each image was presented for the duration of the video it was replacing, and the audio that accompanied it was taken from the original video.

The procedure and data analysis matched that of the first experiment, where the tag was either blank or had an English tag.

Results

Accuracy results for the recall task were very similar to those of the previous experiments (Figure 8). Pairwise on-way t-test analyses showed higher recall for names originally presented with a name tag compared to those presented with no tag ($t(41) = 4.64$, $p <$

.001, Cohen's *d* = 0.57). Median correct reaction time also showed an effect of name tag (*t*(37) = 2.82, *p* = .008, Cohen's *d* =0.55), such that recall responses to names originally presented with a name tag were faster than those for names originally presented with no tag. The results were qualitatively the same when all reaction times were used.

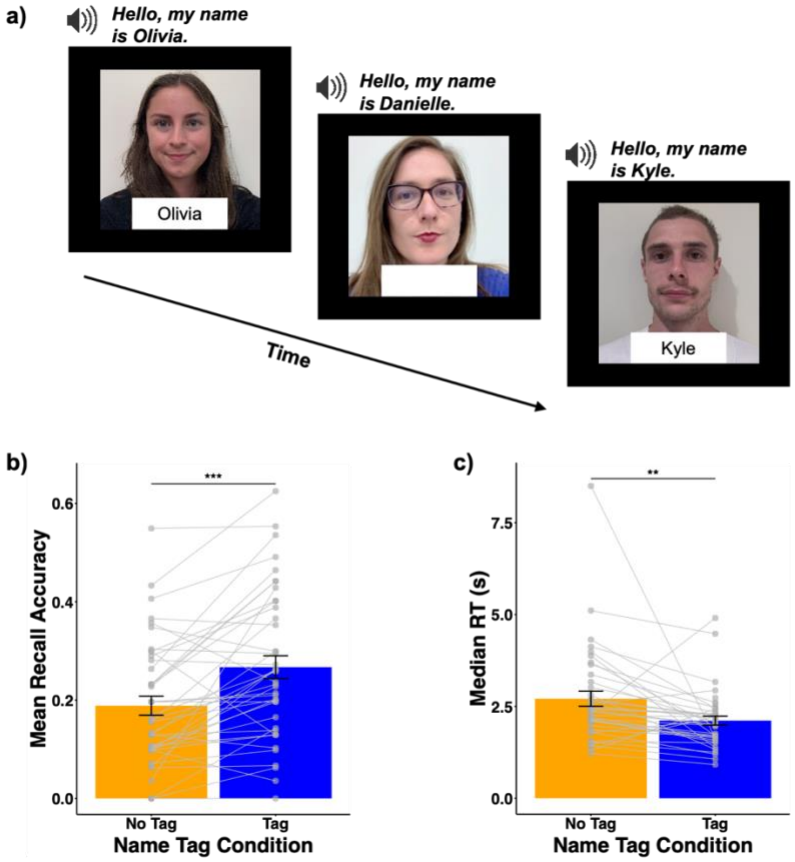


Figure 8: Methods & Results for Study 2, Experiment 3

(a) A diagram of the modified encoding procedure for Experiment 3 (still pictures instead of videos). (b) Recall performance. Participants showed higher average recall performance for names encoded with a name tag than for those encoded without. The overlaid scatter plot represents individual participant scores. (c) Reaction time (for correct responses, in seconds) for

the recall task. Measured as the first keystroke made in the response. Error bars are standard errors.

Interim Discussion

Results for Experiment 3 indicate that the observed superiority of the tag condition over no tag is not due (at least entirely) to interaction with the lip-reading cue. The findings were consistent with those of experiments 1 and 2, supporting the interpretation that multisensory mechanisms may be able to explain the improved recall performance when name tags are presented at encoding. However, it should be noted that the amount of time participants were exposed to each name differed between the two conditions: when name tags were presented, participants were aware of the name much earlier than in the no tag condition. This difference in duration could lead to improved performance from longer exposure to the visual cue, rather than any multisensory mechanisms. As such, Experiment 4 was designed to keep name exposure times equal between the tag and no tag conditions.

Experiment 4

The objective of this experiment was to equate the duration of time in which the name of the speaker is available to the participant across conditions to test whether the observed superiority of the tag condition was due to the longer duration of the name information being available in the tag condition.

Methods

Participants

Participants for this experiment were 49 undergraduate students (39 female) at the University of California, Los Angeles, with a mean age of 19.06 years (*SD* = 0.87). All participants reported normal or corrected-to-normal vision and normal hearing, and 7 reported being non-native speakers of English, but were fluent and so kept in for analyses.

Materials & Procedure

Videos and the name tag format matched those in Experiment 1, except all videos were cut such that the introduction (“Hello, my name is”) was removed, leaving only the name. This meant that the tag and video were on screen for only the duration of the stated name.

The experimental procedure was similar to those of Experiment 1, with a few notable changes. As task performance had been, overall, somewhat low in previous experiments, the number of names participants were asked to learn per block was reduced from 15 to 10, and the number of blocks increased from 4 to 6. Moreover, at test, the confidence rating task was replaced by a recognition memory task, to probe if recognition would benefit from multisensory encoding as well as recall. Participants were given the same 10 s to type a response to the recall prompt as in Experiment 1, and then given a 5-alternative multiple-choice recognition test for the name, using the same image as a prompt (Fig 9a). The 5 names selected for the recognition test included the correct name and 4 alternatives that had been presented in the same block. To ensure this task would not be trivial and 5 probable names would exist, blocks now consisted of the same gender of speaker in all videos, resulting in 3 blocks of female and 3 blocks of male speakers. The assignment of male or female speakers to blocks was pseudorandom between participants, as was block order. As all of the speakers within one block were of the same gender, the encoding task of recognizing the gender was removed for this experiment.

Results

Recall results in this experiment (Fig. 9 b-c) largely follow those of the previous experiments: pairwise one-way t-test results showed that average cued recall performance was higher for names originally presented with a name tag than for names presented without a tag ($t(49) = 3.11$, $p = .003$, Cohen's d = 0.32). Median correct reaction time in the recall task showed no significant effect of name tag ($t(49) = 0.59$, $p = .56$). Results were qualitatively the same when using all trials.

On the recognition task, there was a marginal effect of name tag condition ($t(49) = 1.83$, $p = .07$, Cohen's $d = 0.22$), such that names originally presented with a tag were remembered more often than names originally presented with no tag. There was no significant difference between name tag conditions on median correct response time in the recognition task ($t(49) = 0.75$, $p = .45$). These results were qualitatively the same when analyzed using response times from both correct and incorrect trials.

Interim Discussion

Experiment 4 further supports that the addition of a visual stimulus congruent with auditory stimulus improves performance in cued recall for names, even if the presentation of the congruent visual stimulus matches the length of the auditory stimulus. Interestingly, recognition of the names does not show a similar benefit in accuracy, though the data trends such that recognition accuracy is somewhat higher for faces originally presented with a tag compared with those that were not.

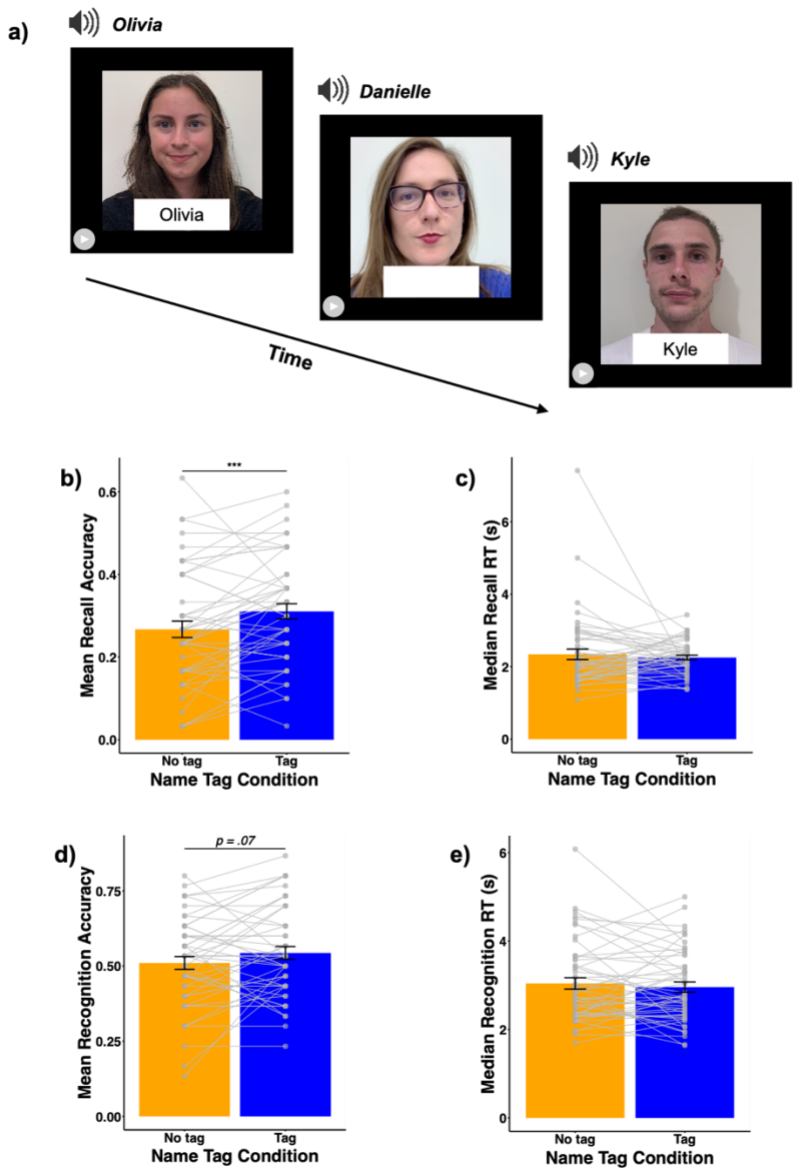


Figure 9: Methods & Results for Study 2, Experiment 4

(a) A diagram of the modified encoding procedure for Experiment 4. This design ensures the onset time, and subsequent presentation duration of the name tag and stated name are matched. (b) Recall performance. There was a main effect of name tag, such that presenting a name tag produced higher average recall than if no tag was present during encoding. The overlaid scatter plot represents individual participant scores. (c) Reaction time (for correct responses, in seconds) for the recall task measured as the first keystroke made in the response.

(e) Recognition performance. Participants showed no significant difference in recognition performance based on the tag condition. (f) Reaction time (for correct responses, in seconds) for the recognition task. Error bars are standard errors.

Experiment 5

Experiments 1-4 establish that addition of a name tag improves the recall of names. However, two distinct underlying mechanisms could mediate this facilitation (see Fig. 10). One possibility is that the name tag could be serving as an additional memory trace that would aid recall by providing a second redundant route to the desired information (i.e., the name), Fig. 10a. Alternatively, the tag cue provides a multisensory representation of name by combining with the audio (and maybe also lip movements; Fig. 10b) and a richer encoding of name-face association. To tease apart these two potential mechanisms, in this fifth experiment we compared two conditions that were equal in the number of "traces" during encoding, but one condition allows for multisensory integration to occur, whereas the other condition does not. This was achieved by manipulating the relative timing of the cues, because it is well established that temporal congruency between cues plays an important role in integration of cues (see Calvert et al., 2004; Shams & Kim, 2010). In both conditions, the same cues (video, audio, name tag) were presented, however, in one condition the audio and tag were presented simultaneously, and in the other condition the tag followed the audio with a delay that is expected to disrupt integration. If the benefits of name tags derive exclusively from their provision of an additional memory trace, then we would expect to see equal performance across conditions. In contrast, a multisensory framework would predict that simultaneity between the audio and congruent visual stimuli would be necessary to receive a memory benefit, and therefore we should see better performance in the synchronous condition.

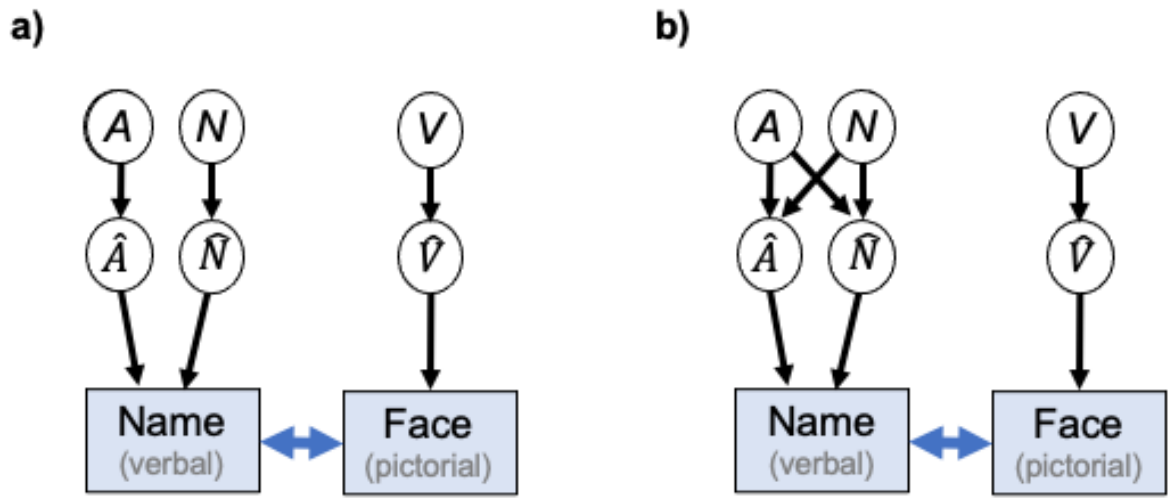


Figure 10: Schematic showing two possible mechanisms underlying the name tag facilitation of face-name associative memory

A denotes auditory input provided by the spoken name. N denotes the visual input (text) provided by the name tag. *V* denotes the visual input provided by the video of the individual. A, N, and V are noisy sensory inputs, and $\hat{A}$, $\hat{N}$, and $\hat{V}$ represent the perceptual estimates. a) If multisensory mechanisms are not utilized, $\hat{A}$ (spoken name) and $\hat{N}$ (written name) independently provide information about name (verbal representation), without interacting with each other perceptually. b) If multisensory mechanisms are utilized, both $\hat{A}$ and $\hat{N}$ provide information about name and each now provide an improved estimate of name due to integration with the other sensory stimulus, as depicted by arrows from both sensory stimuli to each of the perceptual estimates.

Methods

A total of 38 participants (24 female), who were all undergraduate students at the University of California, Los Angeles, were enrolled. Participants had a mean age of 20.89 years (*SD* = 3.36), and all reported normal or corrected-to-normal sight and hearing. Thirty-five

reported being native English speakers, and all reported being fluent in the language. Four participants were excluded from analysis due to remembering zero names in either condition during any block of the experiment.

Materials & Procedure

Materials were the still images from Experiment 3, as these reduce the influence of congruency between lipreading and the visual name tag from playing a role in participants performance. Still images were presented for 5.5 s with the audio from the original videos played starting at the visual stimulus onset. At the bottom of the image, placed over the neck and torso for the duration of the stimulus, as in experiments 1-4, was a white rectangle. Both name tag conditions in this experiment present a name tag and differ in when the tag is displayed: synchronously with the name, or asynchronously. In the synchronous condition, the name is visible starting simultaneously with the still image and audio, and, in the asynchronous condition, the name is visible beginning 2.5 s after the start of the presentation of the still image. In both cases, the visual name will be presented for 2.5 s.

Blocks are organized as in Experiment 4: a total of 6 blocks containing 10 same-gender speakers and names to remember in each block, with tests of both cued recall and recognition for the names given after a 3-min delay.

Results

Recall results in this experiment (Fig. 11) largely follow those of the previous experiments: a pairwise one-way t-test showed that recall performance was higher in the multisensory synchronous condition compared to performance in the asynchronous condition ($t(33) = 2.27$, $p = .03$, Cohen's $d = 0.23$). There was no significant effect of name tag condition

on median correct recall time, $t(32) = 0.82$, $p = .42$. This effect was qualitatively the same when all response times were included.

Recognition results showed no significant effect of name tag condition, $t(33) = 0.90$, $p = 0.37$. However, there was a significant effect of name tag condition on recognition response time ($t(33) = 2.27$, $p = .03$, Cohen's d = 0.35), such that participants on average had faster median responses to names originally presented synchronously than asynchronously. These results were qualitatively different when all response trials were included, such that there was no significant difference between name tag conditions when correct and incorrect responses were used ($t(33) = 1.31$, $p = .20$).

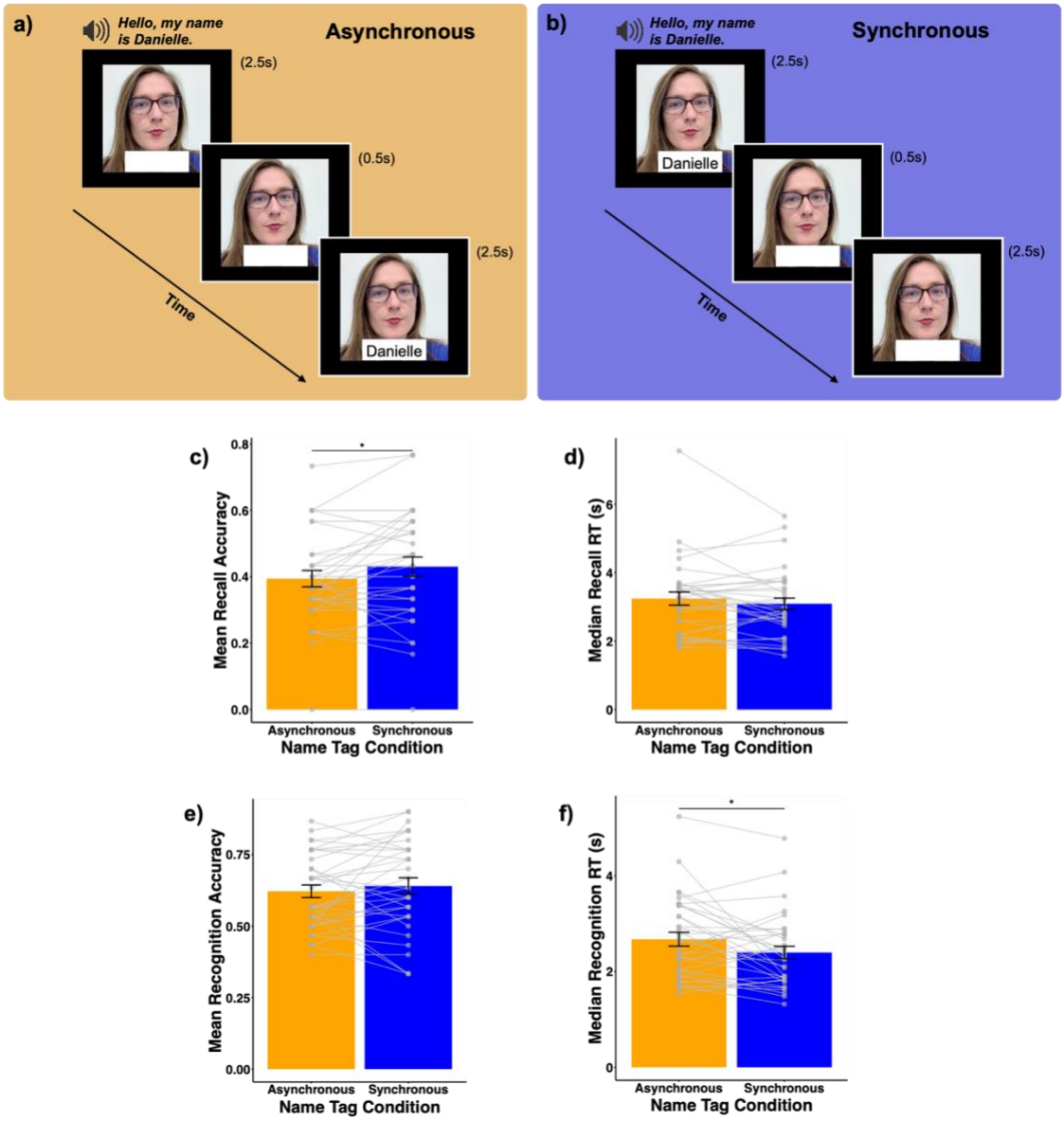


Figure 11: Methods & Results for Study 2, Experiment 5

(a) Schematic of asynchronous condition. The asynchronous condition presented the name tag after a 0.5 second delay. (b) Schematic of synchronous condition. The synchronous condition presented the name tag and audio synchronously. (c) Recall performance. There was a significant main effect of condition on recall. The overlaid scatter plot represents individual participant scores. (d) Reaction time (for correct responses, in seconds) for the recall task, measured as the first keystroke made in the response. (e) Recognition performance. There was

no main effect of condition on recognition. (g) Reaction time (for correct responses, in seconds) for the recognition task. Error bars are standard errors.

Discussion

In this study, we investigated memory of people's names using naturalistic stimuli of videos in which speakers introduced themselves, as is often the case in daily life. Remembering people's names in this context amounts to an associative memory task in which the brain encodes an association between a visual face/body and an auditory presentation of a name (although the lip movements of the speaker may also contribute to this encoding). Because there is no inherent relationship between one's name and one's face, the learning and retention of this association is non-trivial, especially when tasked with the learning of multiple face-name pairs within a short period of time, which is often the case when we attend a party or a professional event.

A few previous studies have shown that multisensory encoding of objects or object features (e.g., motion, or voice) facilitates learning (e.g., Seitz et al., 2006; von Kriegstein & Giraud, 2006; Kim et al., 2008; Shams et al., 2011) and episodic memory (e.g., Lehmann & Murray, 2005; Moran et al., 2013; Heikkilä et al., 2015). However, these learning and memory tasks involved processing of a single feature or recognition of an object or object feature and did not involve memory of an association. Here, we examined whether the benefit of multisensory encoding extends to associative memory. Specifically, we asked whether a multisensory encoding of a name can aid people's ability to bind that name to a face. To render the encoding of a name multisensory, we added a written representation of the name in the form of a name tag in addition to the auditory introduction given by the speaker. We then compared the memory of names in the presence and absence of name tags.

Across a series of five experiments, we found that participants, when cued with a face, were more likely to remember the associated name when that name had been encoded with a name tag, compared to when no name tag was provided. Experiment 1 showed a robust superiority of the tag condition (effect size 0.44) over the no tag condition. Experiment 2 examined whether the observed effect in Experiment 1 was due to the difference in visual saliency of the two conditions (blank vs. text below the face) by controlling for the visual saliency. In both conditions name tags were presented, but in one condition they were in English and congruent with the spoken name, and in the other condition they were in Armenian (a language that could not be understood by participants) and not congruent with the spoken English name. The English tag condition resulted in superior cued recall performance compared to the unintelligible name tag (effect size 0.54), ruling out that the difference in performance was due to visual saliency. Experiment 4 examined whether the observed effect in the earlier experiments was due to the fact that name information was available to the observers throughout the trial whereas the name information conveyed by voice was only available for a portion of trial duration. In that experiment, the presentation of the name tag during the trial was cut and matched the duration of the vocalization of the name. The name tag advantage effect persisted, ruling out the role of the difference in duration of name information as the underlying factor. These experiments collectively establish that the presentation of name tag aids memory of names by providing an additional cue for name. However, the mechanism by which this additional cue facilitates face-name memory remains unclear. Experiments 3 and 5 aimed to shed light on this question.

The name tag cue is a visual semantic cue. It can interact and disambiguate (reduce the uncertainty of) the other semantic cues, namely the vocal cue and the lip-reading cue. The lip-reading cue is an impoverished cue and, as such, could benefit from disambiguation in a within-modality (vision) manner when a name tag is added, bypassing multisensory mechanisms. Therefore, we asked if the interaction between name tag and the lip movements is the primary

factor underlying the observed facilitation of memory. In Experiment 3, lip movement cues were eliminated by replacing videos with static images during the encoding phase. The superiority of name tag condition over no-name tag persisted with a similar effect size (effect size 0.44 with the lip movement vs. 0.55 without lip movements), suggesting that the putative enhancement of lip-reading cue by name tag cannot account for the observed effect.

Finally, we aimed to gain insight into the underlying mechanism of the name tag benefit by teasing apart the role of multiple independent memory traces (Fig. 5a) vs. the role of integration of multisensory cues (Fig. 5b). The name tag provides an additional memory trace, which can facilitate recall by providing an alternative retrieval route to access the name when cued with the face. That is, the face might trigger the retrieval of the auditory memory of the spoken name *or* the visual memory of the written name, essentially giving participants an extra chance to succeed at recalling the name. In this framework (Fig. 5a), the mere existence of an additional cue is sufficient for improved recall. On the other hand, in the multisensory encoding framework (Fig. 5b), the interaction between the cues and the integrated representation of the feature/object can play a key role in the richness of the encoding, thus increasing the likelihood of later recall (Shams & Seitz, 2008). More specifically, the name tag cue can be integrated with the vocal cue, resulting in a more accurate and/or more precise representation of the name. This improved name representation can strengthen the encoding of the face-name association and lead to improved memory performance.

In order to tease apart these two possible accounts, in Experiment 5, we compared two conditions in which the number of traces were equivalent, but one condition lends itself to integration of the name tag cue with other cues, whereas the other condition does not. It is well established that temporal congruency is key in integration of sensory cues, and the lower the temporal congruency the lower the probability of integration (e.g., Shams et al., 2002; Shams & Kim, 2010; Calvert et al., 2004; see Ernst & Bülthoff, 2004). Therefore, by manipulating the relative timing of the name tag and the vocal (and lip movement) cues, we can influence their

probability of integration. It has been shown that introducing audio and visual stimulus onset asynchronies of between 150 and 250 ms reduces audiovisual speech fusion and alters brain activity in speech-processing regions of the brain (Macaluso et al., 2004; van Atteveldt et al., 2006; Miller & D'Esposito, 2005). Therefore, it is to be expected that the name tag cue would get integrated with the other name cues when it is presented synchronously and not integrated when it is presented with a delay of 500 ms. On the other hand, in both of these conditions all of the cues are available in each trial, and by delaying the name tag relative to the video, the performance may even be expected to improve according to multiple independent memory trace account: the name information, which is initially encoded by voice and lip-reading, gets reinforced by the later presentation of the name tag. The results of Experiment 5 showed that the synchronous presentation of the name tag leads to better memory performance than the asynchronous presentation. This would support the multisensory integration hypothesis, that multisensory object representation itself can be helpful to memory above what would be predicted by having multiple independent sensory traces. Future research will need to probe this question further by examining the nature of multisensory interactions that promote facilitation of memory, including which sensory combinations can facilitate memory performance, and what kinds of memory tasks will benefit from multisensory integration.

The present results cannot be accounted for by the dual-coding theory or the cognitive theory of multimedia learning, according to which the combined verbal and pictorial presentation of words facilitates memory compared to verbal-alone presentations. In the present study, in all conditions, including the baseline no tag condition both verbal (name) and non-verbal (video/image) representations are available (see Fig. 5). The only difference between the experimental and control conditions is the availability of *additional* verbal information (name tag), or, in the case of Experiment 5, the relative timing of the additional verbal (name tag) information.

The improved memory accuracy under multisensory stimulus presentation conditions does not seem to be as robust in the multiple-choice recognition task compared to the recall task. Experiments 4 and 5 evaluated both name recall and name recognition in response to face cues. In Experiment 4 there was a trend for a multisensory benefit in recognition, whereas in both experiments the benefit of multisensory presentation in recall was statistically significant. While previous experiments have shown multisensory benefits in recognition tasks, those experiments were structured quite differently from the current experiment. This experiment, unlike many previous multisensory memory studies, used an associative memory task. Previous multisensory research has probed memory for single items, while the current study investigated memory for an association between a name and a face. Moreover, additional experimental power may be needed to uncover statistically significant effects in the recognition task.

Also of note are that the brain mechanisms by which multisensory stimuli benefit recall performance are unclear. The current results can speak to a few different behavioral theories, but cannot distinguish between them decisively. Previous work has indicated that cross-modal interactions allow for information distributed across multiple senses to be combined into meaningful representations. This combination of senses has been found to allow for optimal processing of sensory information and can help disambiguate noisy stimulus presentation via uncertainty reduction (one signal can disambiguate another signal, leading to the improvement in precision and/or accuracy) (see Ernst & Bülthoff, 2004; Shams & Beierholm, 2010 for overviews). Multisensory stimulus presentation may also change how attention is directed and multisensory scenes are segregated (Lewkowicz et al., 2021) at the time of encoding. Which mechanism, if any of these, supports the current findings is currently unclear, and future neuroimaging research may be able to identify the neural mechanisms supporting multisensory memory benefits.

It should also be noted that, while these experiments do provide evidence for multisensory memory benefits, further research could help directly rule out the possibility that

participants were using strategies during encoding that would selectively benefit the name tag conditions. For example, it is possible that participants preferentially encode the visual tag by default, and only use the auditory information when the tag is unavailable. This would mean that participants may have a switching cost as they change strategies between trials where a tag is synchronously presented with the audio as opposed to when the tag is asynchronous or absent, and the cognitive cost of this visual-to-auditory attentional switching on no-tag trials could potentially explain the benefit of tags seen in all experiments. We believe this explanation is very unlikely, particularly given the results of experiment 4. In that experiment, the average duration of the videos was greatly reduced, such that none were longer than 2 seconds, and the audio presentation of the name began immediately at the start of the trial (see videos in the supplemental materials for an example). If the aforementioned task switching strategy explained the full set of results, one might expect to see that any benefit would disappear if participants were denied the time needed to assess which strategy they should use, but a difference still existed between the tag and no tag conditions in experiment 4. However, participants were not asked explicitly to describe any strategies they were using, and therefore we cannot entirely rule out that strategizing could play some role in the observed benefit. Further research using a between-subjects design (where some participants have only tag trials and others have only no-tag trials) to prevent task switching could investigate this further. Such an investigation could also probe how each level of synchrony in experiment 5 compares to a unisensory baseline. This has been left out of experiment 5 to maximize experimental power in testing the underlying mechanism of the observed benefit to remembering names when a tag is present– i.e., whether it was due to having more sensory cues available or if multisensory integration was specifically helpful– but future experiments could investigate this relationship.

Our findings are generally in line with previous multisensory findings, and expand those results to associative memory, and to a more naturalistic memory task. The current experiments also suggest that multisensory mechanisms can be leveraged in daily, difficult tasks to improve

memory performance. These findings do not contradict previous memory theories, but rather can function as an additional tool that can be used to improve human memory in difficult situations. Traditional techniques for improving memory—including mnemonics and spatial mapping—are effective but do require a relatively high level of sophistication and intent to employ. Using basic sensory information could be more easily and passively implemented to improve memory. This could lead to the development of new strategies, techniques, and technologies to improve everyday life and learning, even for relatively difficult associative memory tasks.

Chapter 4: Effective Language Learning: The Impact of Sensory and Cognitive Cues

Abstract

Multisensory explanations for memory benefit suggest that crossmodal integration is uniquely poised to support retrieval of information from memory. Existing theories allow that crossmodal stimuli can be helpful, but often draw distinctions between higher-level cognitive processes, claiming that providing information across varied cognitive domains is helpful for learning, as in the Cognitive Theory of Multimedia Learning (CTML; Mayer, 2014). To investigate the roles of sensory and cognitive factors in memorization, as well as to explore if the findings of Study 2 will generalize to other associative memory tasks, we asked participants to learn vocabulary in a foreign language, namely Swahili. Guided by the CTML as well as previous research in multisensory memory benefits, we expected to see that congruent crossmodal word encoding with audio and text—here termed *sensory enrichment*— would improve word recall relative to unimodal, text-based word encoding, in keeping with predictions from the CTML's modality principle. We additionally expected that adding a different cognitive stream of processing—images, in addition to audiovisual verbal presentation, which we have termed *cognitive enrichment*—would improve performance as well. The first two experiments show that cognitive, but not sensory, enrichment improves memory for words in a foreign language. We then investigated if the memory benefit observed in those experiments could be due to crossmodal stimulus presentation, as cognitive and sensory enrichment had always co-occurred. In experiments 3 and 4, we found that recall was highest when images accompanied text, in a

violation of the expectations of the modality principle. We additionally found that benefit received from simultaneously providing crossmodal and cognitively complementary stimuli, through audio with images, did not provide a larger benefit to unisensory processing than cognitive enrichment alone. We discuss this seeming violation of the modality principle and previous multisensory findings and suggest that foreign language learning provides an interesting border case that helps to explore where multisensory benefit to memory is more limited.

Introduction

The ability to speak a foreign language has many cognitive, professional, social and intercultural advantages. For example, research on the cognitive benefits suggest that multilingual individuals have improved brain executive functioning compared to monolinguals (Costa & Sebastián-Gallés, 2014). Moreover, multilingualism has some neuroprotective effects, delaying Alzheimer's disease (Chertkow et al., 2010). In the professional setting, this ability also confers an advantage as it increases job opportunities (Gándara, 2018). In particular, with the rise of technology and social media, boundaries are being diminished and people worldwide are given the opportunity to interact with each other and exchange ideas, cultures, and thoughts. This has made novel scientific findings more accessible. If one is able to understand multiple languages, this access to information can be easier, increasing communication on a global scale. Thus, learning a foreign language can have many benefits. However, many people might not seek to learn a foreign language because of the difficulty associated with it, and thus, they cannot take advantage of these benefits. One demanding aspect of learning a foreign language is memorizing a large number of foreign words and their meanings. As such, it is important to study how to make learning foreign language more effective at the early word-learning stage to make language learning easier for all individuals.

The traditional classroom or individual strategy to learn new vocabulary was by looking at written words and reading their meanings in a textbook. More modern approaches have moved this to audiovisual multimedia learning, which combines text, audio, and pictorial representations of the concepts being taught. As such, it is important to consider aspects of effective multimedia learning in assessing how to make learning new words easier for language learners. These multimedia frameworks largely build off of traditional text-based learning in a few ways: what we will call *sensory enrichment,* where additional low-level sensory input from a different sense (e.g., audio in addition to text) is included in learning, and *cognitive enrichment*,

wherein higher-level modes of cognition are added that require different processing from the existing modes (e.g., an image in addition to text). In many modern learning tools, these are both present, either separately or together (e.g., using text with audio and an image), and distinguishing which is more helpful for learning can be challenging. As such, the current experiment seeks to make this distinction clearer by investigating which of these factors is more important.

Principles of multimedia learning

One of the best-known frameworks for understanding learning in multimedia contexts Is the Cognitive Theory of Multimedia Learning (Moreno & Mayer, 1999; Mayer, 2014). This theory makes a distinction between two important cognitive modes of processing, verbal and pictorial, wherein activating both processing pathways will improve learning outcomes. This creates two important principles to consider when evaluating and predicting the most effective combination of stimuli on learning: the redundancy and modality principles.

The redundancy principle is closely tied to cognitive load theory, and suggests that redundancy of stimuli in different forms, when each form can be understood alone, can increase the strain on working memory beyond its limited capacity, and impair learning (Sweller, 2005). For example, the results of one study suggested that college students were able to perform better on a retention test when they viewed an educational animation with its narration, compared to when they viewed the animation, narration, and written subtitles all together (Mayer et al., 2001). Definitions of redundancy, however, depend often on individuals' proficiency with the information presented. For beginners who want to learn a complex task, the repeated information may in fact aid learners and deepen their understanding, improving learning (Sorden, 2005). For upper intermediate learners, the results might be different as they might experience the negative aspects of the redundancy principle. The concept of cognitive

enrichment ties into this idea—utilizing complementary cognitive processes would improve learning and retention of new words in a foreign language.

The modality principle suggests that we have limited capacity within a single sense to process information, and so presenting information across different sensory modalities is beneficial to learning new information. This is particularly of note in the cognitive theory of multimedia learning in the case of adding audio to an animation instead of adding text to an animation (Mayer, 2014), but will also apply to still images. This also ties in with the concept of a limited processing ability but is perhaps more closely related to the idea of limited processing ability and working memory stores; showing too much information in one sensory modality can overwhelm the limited processing power in that sense, impairing leaning compared to dispersing information across multiple processing modalities (Mayer, 2014).

In light of both of these principles, it would be expected that sensory enrichment and cognitive enrichment could improve learning. Sensory enrichment– adding senses that are complementary to one another, using different processing pathways but both providing information about the stimulus– will reduce processing demands in one sense, aligning with the modality principle. Cognitive enrichment– providing information across two higher-level processing domains, such as verbal and visual processing– will help reduce redundancy at the cognitive level. Additionally, these can be combined– as in the case of auditory processing being accompanied by a visual image of the item being described– to meet both criteria at once. Previous research has investigated these different types and combinations of enrichment in language learning, however, and the results are often mixed, making it difficult to determine what type of enrichment may be most effective for supporting multimedia learning of foreign vocabulary.

Cognitive enrichment in language learning

There have been multiple studies which compare learning vocabulary in a foreign language, with image and text versus text alone. Overall, most of the previous experiments suggest that the image and text condition results in better foreign language learning compared to text alone. For example, Carpenter and Olson conducted a series of experiments on Swahili word learning (2012). They compared vocabulary learning, when Swahili words were presented with their English translations to when they were presented with simple pictures. In this study, when participants believed that pictures facilitated learning, an overconfidence bias developed, and consequently the foreign words were not learned better when accompanied with pictures. However, when researchers included instructions to minimize overconfidence bias, the picture and text condition did result in better learning of the foreign word compared to the text and translation condition (Carpenter & Olson, 2012). Another experiment done by Webber on elementary school students also suggested that learning foreign words is more effective in a picture-word pair compared to a word-translation pair (1978). The target language examined in this study was Indonesian, the pictures used were simple line-drawings and the participants were 42 fourth graders (Webber, 1978). This study assessed whether learning was more effective for native English-speaking participants in Japanese and English word pairs compared to Japanese word and picture pairs. The results suggested that learning foreign words is facilitated by pictures (Deno, 1968). This phenomenon was also investigated in a comprehensive study on Spanish word learning in 1967 by Lado. The results again suggested that compared to other conditions when no picture was included in learning, accompanying foreign words and their translations with pictures, further improves learning (Lado, 1967). All these experiments were done on native English speakers. However, Lotto and de Groot studied Italian vocabulary learning in Dutch learners and discovered different results. They reported that participants learnt Italian words worse when they were presented with images compared to when they were presented with only their Dutch translations (Lotto & de Groot, 1998). In addition, Boers et al. studied vocabulary learning in L2 English learners and compared text only

and text-image conditions (2017). In this study, adding pictures did not significantly improve meaning recognition but led to lower word form recall, possibly due to reduced attention to target words (Boers et al., 2017). However, in 2001, Al-Seghayer, investigated vocabulary acquisition in ESL students. The results of this study suggested that learning a foreign language is more effective when it is presented as still-image text compared to text alone (Al-Seghayer, 2001). Overall, it appears that there is an advantage for providing images in addition to text when English-speaking participants seek to learn a foreign language.

Sensory enrichment in language learning

Aside from presenting texts with images, some researchers have investigated the effects of adding auditory stimuli to text on novel word learning. A survey done on 100 girls in Pakistan suggests that people believe it is easier to learn a foreign language when the verbal stimuli is accompanied by an auditory factor (Kausar, 2013). However, to understand whether this positive effect of auditory stimuli is actually present in language learning, different experiments were conducted. A few studies compared word learning when presented as audio and text compared to text alone, suggesting that using multiple senses might facilitate learning. For example, a study conducted on sight-word learning of 4-year-old children suggests that new words are better learned when they are presented as audio and text compared to when there is text alone (Arlin et al., 1978). To study foreign language word learning specifically, some researchers studied German vocabulary learning and compared the effectiveness of different presentation formats on foreign word memory. These researchers compared text-only presentations to text and audio presentations and concluded that for individuals who scored below the median score, the multisensory audio and text condition had a significant impact on learning, improving scores compared to the text condition (Macedonia & Repetto, 2016).

Comparisons of Cognitive and Sensory Enrichment

As both sensory and cognitive enrichment appear to positively affect learning and memory outcomes in language learning, it would seem natural to combine these features. Several studies have previously studied the effects of presenting audio, text and images together on foreign language learning. However, the results are not clear, and the understanding of this condition compared to other sensory combinations is limited. For example, in one experiment, English speaking participants were asked to read a German text with audio and then had an option of viewing verbal annotations (written translations in English) or visual annotations (pictures or videos) or both (Plass et al., 1998). The results suggested that participants memorized foreign words better when they had both verbal and visual annotations (audio, image, and text). Between the verbal only and visual only annotation conditions, participants performed stronger if they could choose their preferred method of learning. These results suggest that the most effective sensory combination for foreign language learning may differ for visual and verbal learners (Plass et al., 1998). In a more recent experiment however, on 15 students with disabilities, foreign words were taught over a course of 7 weeks with multisensory enrichment (text, audio and image). For these students, learning foreign vocabulary did not significantly improve with the multisensory condition compared to the control group which received text only instruction. This may be due to the small sample size or that the students had disabilities (Ciccarone, 2019).

It is also unclear whether sensory or cognitive enrichment is more useful for supporting memory for new words in a foreign language. Other researchers investigated the effects of adding auditory stimuli when audio plus text is compared to image and text and the results were mixed. For example, when Arlin et al. compared presenting text with its image to presenting text with its audio, the auditory condition resulted in less effective word learning (1978). However, in this experiment the words were not from a foreign language, as they studied the effects of audio in general words (Arlin et al., 1978). To study foreign vocabulary learning specifically, Kozan et al. divided participants to two groups based on their working memory level (high vs low) and

compared the effects of audiovisual presentation to visual-only presentation (2015). The experiment was done on 29 Turkish speaking English learners and the English text needed to be learned was about tornado formation (Kozan et al., 2015). In the audiovisual condition, static pictures were presented with an audio. in the visual-only condition, the same text was presented in a written form with the image, instead of the auditory stimulus. Results suggest that for individuals with a high working memory, the audiovisual condition resulted in better retention of the foreign text compared to the visual-only condition. For the low working memory condition however, there was no significant effect of modality for low working memory participants (Kozan et al., 2015). It should be noted however that the small sample size may have affected the results. However, the results for the high-working memory participants suggests that activating multiple senses enhances learning.

The current studies

Based on the body of evidence presented by literature, the results regarding the most effective sensory and cognitive combinations for foreign language word learning have been mixed and/or inconclusive. Thus, our knowledge of the effectiveness of different learning schemas on foreign language word learning remains limited. To clarify confusions and investigate the most effective sensory combination to present novel foreign language words in, the current research has been done. In our first two experiments, we compared sensory enrichment with cognitive enrichment, to investigate which is more effective for retaining new words in a foreign language. In the next two experiments, we investigated the interplay of modality and cognitive enrichment by investigating if cognitive enrichment was equally effective in improving learning across two different sensory modalities.

Experiment 1

In experiment 1, we investigated the differences between three different conditions: text alone (T), test plus audio (AT), and image plus audio plus text (IAT). Based on previous experiments, the hypothesis is that foreign language vocabulary learning can be improved when novel words are presented in multisensory and multimedia combinations. In particular, it is expected that learning in the IAT condition would have the best results in terms of the recall and retention of foreign language words--because of the complexity of the task and the novelty of the language--while learning in the T condition would have the worst results. According to the modality principle, we expect that the AT condition would have better learning outcomes than the T condition.

Methods

Participants

Participants were 65 undergraduate students recruited from the University of California, Los Angeles (UCLA) psychology department subject pool. The average age of these participants was 22.2 years. Thirty-one identified as female, 14 identified as male, and 20 declined to provide a gender identity. All participants reported having normal or corrected-to-normal sight, normal hearing, being fluent in English, and not being familiar with Swahili. One participant was fluent in another language from the Niger-Congo language family but was kept in the analyses because the language they spoke was from a different branch of this language family than Swahili.

Materials

Stimuli were drawn from a list of Snodgrass-type images (Rossion & Pourtois, 2004), including color images of common body parts, animals, and household items. The names for these items were translated into Swahili using Google translate. In the case where a word started with an uncommon letter combination in English (e.g., *Ny* or *Ng*), the word was simplified to be more consistent with standard English pronunciation, to make the words easier for English speakers to understand. From here, words were then selected from the full list of items such that no Swahili translation of the word with fewer than 4 or more than 6 letters would be selected. Length of the English translation of the word was not restricted. This left 90 words and their translations and images available for the final experiment.

Audio stimuli were created using Amazon Web Services Polly text-to-speech software. The voice selected was a computer imitation of a female speaker with a standard American accent. We obtained recordings of the English word, the Swahili word, and a phrase linking the two words, always phrased as "[Swahili word] means [English word]." All files were downloaded as mp3s for use in the experiment.

Procedure

Participants began in the encoding phase of the experiment, where participants were instructed that they would be taught new words in Swahili, and that they would later be tested on these translations. In each block of the encoding, they were given 10 words to learn in Swahili. In this block, learning could take place in one of three conditions. In the Text condition, the Swahili word and its English translation were presented in text format as "[Swahili word] means [English word]." In the audio plus text condition, the text was accompanied by an audio track that simultaneously said "[Swahili word] means [English word]." In the image plus audio plus text condition, a Snodgrass-type color drawing of the item was presented simultaneously with the text and audio (Figure 12). Each Swahili word and its corresponding English translation

was presented to participants once, with a fixation cross lasting 1.2 seconds between each stimulus.

After hearing the full list of 10 words with their translations, participants were given a 3-minute break, during which they were encouraged to meditate. After 3 minutes, the participants began the test phase of the task. For each word, participants were first given a Cued Recall phase, where they were presented with a text form of the Swahili word and were prompted to write the English translation of that word. Participants were given 10 seconds in which to finish typing their responses. After this, participants entered a Recognition phase, wherein they were again prompted with the Swahili word and were then asked to select among 5 multiple choice options which was the corresponding English word. These English words were presented below the Swahili word with a number, 1 to 5, that corresponded to the number key the participant would need to use to select that word. Each correct option corresponded to a number, 1 through 5, that correct and equal number of times within the block (i.e., 2 times/block). The other 4 words were randomly selected from the remaining 9 words that had been seen in the preceding encoding block. The Cued Recall followed by Recognition phases were repeated for each word learned in the preceding Encoding block, in a random order. Participants then had a one-minute break before moving on to the next encoding phase.

There was a total of 3 encoding-retrieval blocks for each condition, for a total of 9 blocks. Blocks were semi-randomly organized, such that all three conditions were seen before any condition was repeated (e.g., participants had to experience text, audio plus text, and image plus audio plus text conditions before they could see another block of any one of those conditions again). In each block, words were presented in a random order in both encoding and retrieval phases. Words were assigned to a block and condition within a version of the experiment, and a total of 9 different versions of the experiment with different word assignments were created, to limit the influence of item effects.

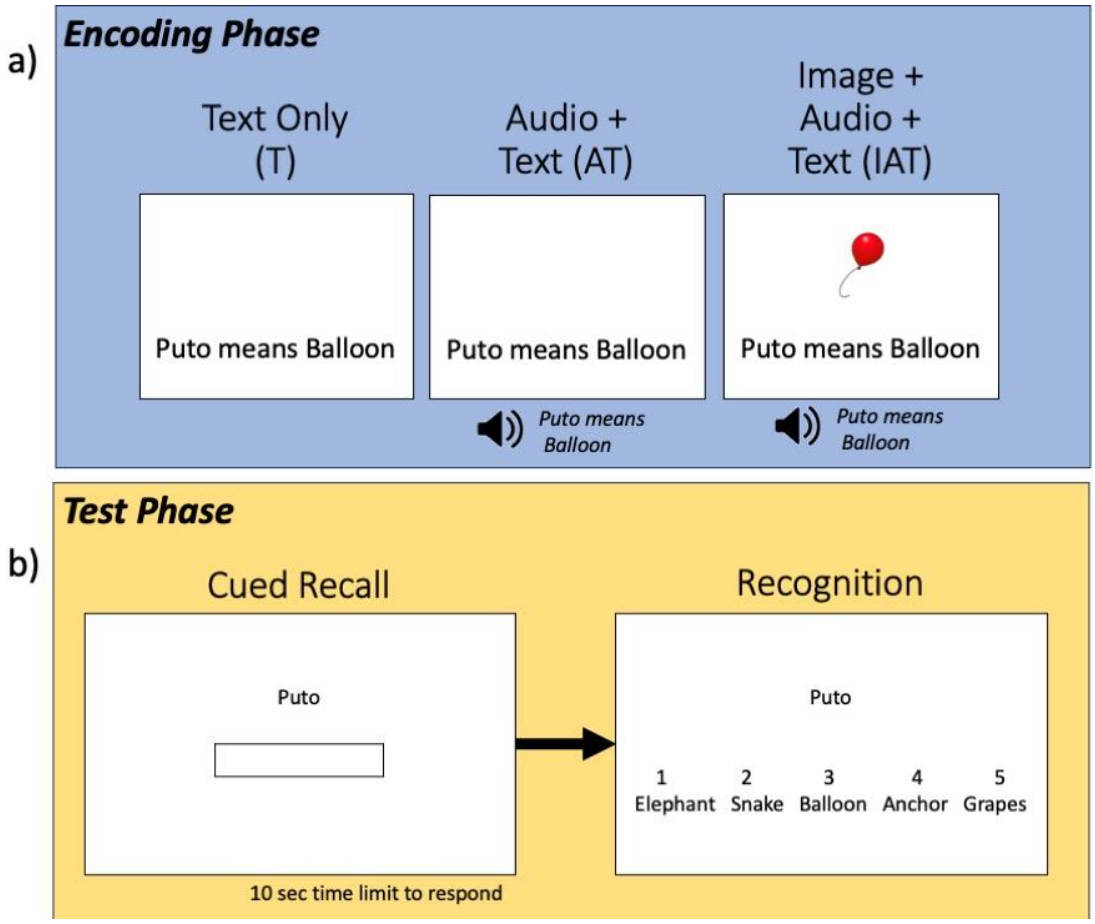


Figure 12: Encoding Condition and Test Phase for Study 3, Experiment 1

All examples in this figure use the translation “Puto means Balloon,” though each participant would only encounter the word in one encoding condition. (a) Examples of the stimulus “Puto means Balloon” in all 3 different conditions present in the encoding phase of the experiment. (b) Example trials for Cued Recall and Recognition for the word Puto, where participants would respond with the English translation. The other translations presented in the recognition phase would be randomly selected from words presented in the same block as the target translation.

Statistical Analysis

Participant responses were recorded for both the cued recall and recognition tasks. For cued recall, participants’ typed responses were scored. This was done by computer, scoring participants responses as a match to the correct word, ignoring capitalization and any spacing

or other punctuation. Recognition was scored as participants selecting the correct English translation out of their five options.

Response time in the recall task was measured as the first key stroke given during the participant's response period. Trials where no response was attempted were not included in the calculation in the mean. For cases where participants had no response attempts for a particular encoding condition—which only occurred twice across all 4 experiments—participants were given an average RT of 10, the maximum possible RT, for that condition. Recognition RTs were scored as the time after the stimulus presentation began when participants pressed a key to make their recognition choice.

In addition to these measures, we also calculated a measure of speed-accuracy tradeoff for the recognition task. While drift diffusion modeling (see Chapter 2) would provide a stringent measure of speed-accuracy tradeoff, the trial count per condition is too low (~30 trials/condition) to obtain reliable estimates, even with the hierarchical variant we have previously used. However, there are other analytical methods available. Our measure was inspired by *inverse efficiency scores* (Rach et al., 2011; Townsend & Ashby, 1983), which can be similarly effective to DDMs for uncovering benefit to processing in multisensory experiments (Rach et al., 2011). Inverse efficiency (IE) looks at RT divided by accuracy, thus penalizing low accuracy responses by inflating IE relative to the base RT. We instead chose to look at the inverse of this measure, what we will be terming the *efficiency score* (ES), which was accuracy divided by RT. Thus, RTs that are longer, leading to higher accuracy, will decrease ES relative to accuracy scores.

Results

Figure 13 shows the overall trend in recall and recognition performance in experiment 1. A repeated-measures ANOVA was run on the recall data and revealed a significant difference in recall performance based on what condition the participant had been given during encoding

($F(2,128) = 5.71$, $p = .004$, generalized $\eta^2 = 0.011$). Post-hoc paired Student's t-tests with Holm corrections on the p-values showed that participants did not differ significantly in their performance between the text only and audio plus text conditions ($t(64) = 0.39$, $p = 0.70$). However, participants showed higher recall accuracy when given an image, audio, and text compared with text alone ($t(64) = 2.66$, $p = 0.02$), or compared with audio plus text ($t(64) = 3.02$, $p = 0.01$). Response times did not differ significantly between the encoding conditions ($F(2,128) = 0.12$, $p = .89$, generalized $\eta^2 < .001$).

Recognition data followed a somewhat different pattern. The repeated measures ANOVA indicated that encoding condition had a significant effect on recognition score ($F(2,128) = 4.49$, $p = .01$, generalized $\eta^2 = 0.013$). Post-hoc tests indicated that participants had higher recognition accuracy for translations presented as text alone compared to audio alone ($t(64) = 2.43$, $p = 0.04$). Accuracy in the image, audio, and text condition was also significantly higher than in audio plus text ($t(64) = 2.83$, $p = 0.02$), but did not differ significantly from accuracy observed for items encoded with text alone ($t(64) = 0.52$, $p = 0.61$). Response times did not differ significantly between the encoding conditions ($F(2,128) = 0.18$, $p = .84$, generalized $\eta^2 < .001$). We additionally looked at efficiency scores for this recognition task. Efficiency scores showed no significant effect between the conditions ($F(2,128) = 2.32$, $p = .10$, generalized $\eta^2 = 0.006$). The difference in significance between ES and accuracy may indicate the presence of a speed-accuracy in the recognition responses, explaining some of the differences in performance.

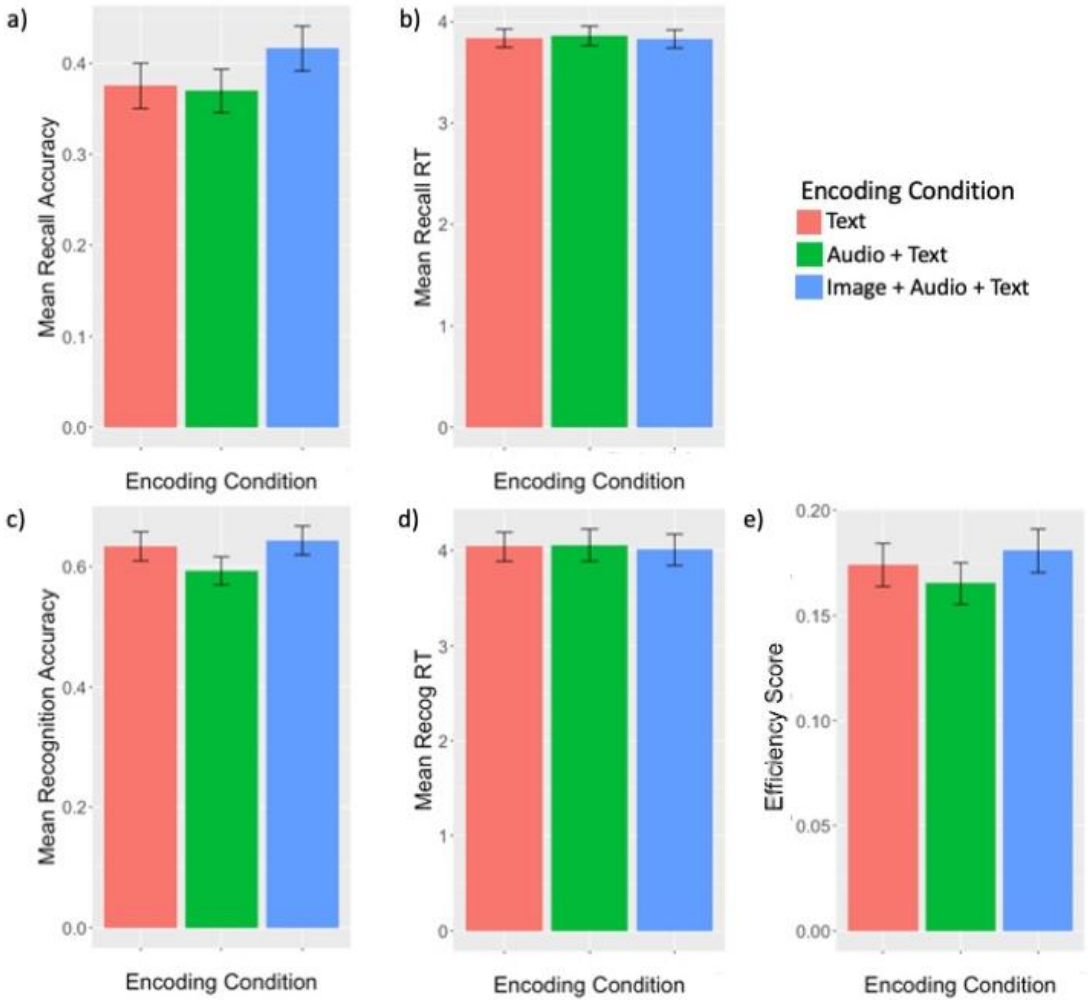


Figure 13: Results for Study 3, Experiment 1

Performance in (a) recall accuracy and (b) recall RT. Participants show higher accuracy in the IAT condition than the other two conditions, and show no significant difference in RT between the three conditions. In (c) recognition accuracy, performance in T and IAT conditions outperforms the AT condition, and (d) no significant difference observed in average recognition RT. We also analyzed performance using (e) a measure of speed-accuracy tradeoff, recognition accuracy/average response time. This pattern of results, while it trends similarly to the accuracy score, has no significant differences between the encoding conditions.

Interim Discussion

Results from experiment 1 in both memory tasks violate the modality principle, as adding multiple pieces of visual information improves performance, but adding information in the auditory domain does not improve performance. Sensory enrichment, here included as adding audio to the text, does not improve performance, and in fact reduces accuracy in recognition relative to text-alone encoding. Such findings are out of line with the original hypotheses of this experiment, which predicted that sensory enrichment would improve performance relative to encoding in a single sense.

Experiment 2

Experiment 2 was a replication of experiment 1 in a separate group of participants, where the test phase was changed so participants needed to retrieve the new Swahili word they had learned when prompted with English, instead of producing the English translation from the Swahili word. Results were expected to be similar to those of experiment 1.

Methods

Participants

Participants were 70 undergraduate students recruited from the UCLA psychology department subject pool, similar to the first experiment. The average age of the participants was 20.4 years. Fifty-six identified as female, 13 identified as male, and 1 identified as nonbinary. All participants reported having normal or corrected-to-normal sight and hearing, and being fluent in English. Only one participant reported knowing limited Swahili.

Materials

Stimuli used were the same as those described in Experiment 1 materials.

Procedure

Similar to Experiment 1, participants began in the encoding phase and were instructed that they would be taught new words in Swahili, and that they would later be tested on these translations. Each block contained 10 words to learn in Swahili. Learning took place in the same three conditions described in Experiment 1 (text, audio plus text, image plus audio plus text), and timing of the words and fixation cross being presented was the same as Experiment 1 as well.

This was followed by a three-minute break where the participant was encouraged to meditate until the test phase began. The test phase consisted of cued recall and recognition questions for each of the 10 words shown during encoding. The order of questions was randomized so as not to match the order of words shown during the encoding phase. The participant was shown the English translation as a cue during the recall test, while being given 10 seconds to respond by typing in the corresponding Swahili translation. This was followed immediately by the recognition test. In the recognition portion, they were given five options of Swahili translations for the English cue and were prompted to type a number ‘1’ through ‘5’ corresponding to what they believed was the correct translation. This cued recall-recognition test pairing repeated for all 10 words from the associated encoding phase. Once the encoding and test phase were completed for one round, the study advanced to the next round within that block (Fig. 14).

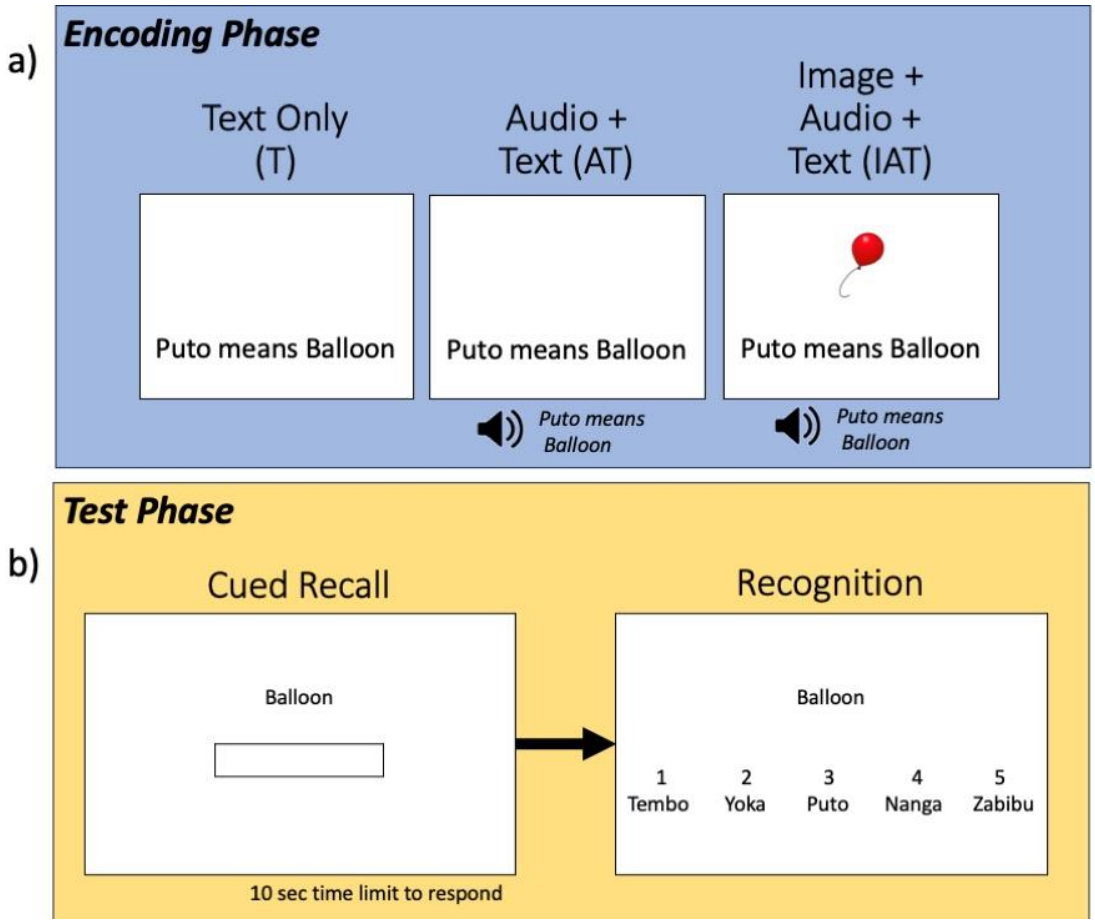


Figure 14: Encoding Condition and Test Phase for Study 3, Experiment 2

All examples in this figure use the translation “Puto means Balloon,” though each participant would only encounter the word in one encoding condition. (a) Examples of the stimulus “Puto means Balloon” in all 3 different conditions present in the encoding phase of the experiment. (b) Example trials for Cued Recall and Recognition for the word Balloon, where participants would respond with the Swahili translation. The other translations presented in the recognition phase would be randomly selected from words presented in the same block as the target translation.

Statistical Analysis

Data was scored identically to that of experiment 1. Response times and ES were measured similarly, with the only change being that recall RTs reflected participants' final keystrokes rather than their first.

Results

Figure 15 shows the overall trend in recall and recognition performance in experiment 2. A repeated-measures ANOVA was run on the recall data and revealed no significant difference in recall performance based on what condition the participant had been given during encoding ($F(2,132) = 0.88$, $p = .42$, generalized $\eta^2 = 0.011$). RT likewise showed no significant differences between the conditions (($F(2,132) = 1.29$, $p = .28$, generalized $\eta^2 = 0.006$).

Recognition data followed a different pattern. The repeated measures ANOVA indicated that encoding condition had a significant effect on recognition score ($F(2,132) = 7.93$, $p < .001$, generalized $\eta^2 = 0.02$). Post-hoc tests indicated that participants had no significant difference in recognition accuracy if items were encoded as text alone or as audio and text ($t(66) = 0.54$, $p = 0.59$). Accuracy in the image, audio, and text condition was significantly higher than in audio plus text ($t(66) = 3.45$, $p = .003$), and was higher than text alone ($t(66) = 3.03$, $p = 0.006$). RT for recognition choices across the different encoding conditions did not differ significantly ($F(2,132) = 1.78$, $p = .17$, generalized $\eta^2 = 0.005$).

Efficiency scores for the data showed a similar trend to the recognition accuracy scores, with a significant difference between the groups ($F(2,132) = 6.79$, $p = .002$, generalized $\eta^2 = 0.021$). As in recognition, this reflected no significant difference in recognition accuracy if items were encoded as text alone or as audio and text ($t(66) = 0.13$, $p = 0.90$), but superior efficiency in image plus audio plus text compared to audio plus text $t(66) = 2.99$, $p = 0.008$) or text alone $t(66) = 3.26$, $p = 0.005$).

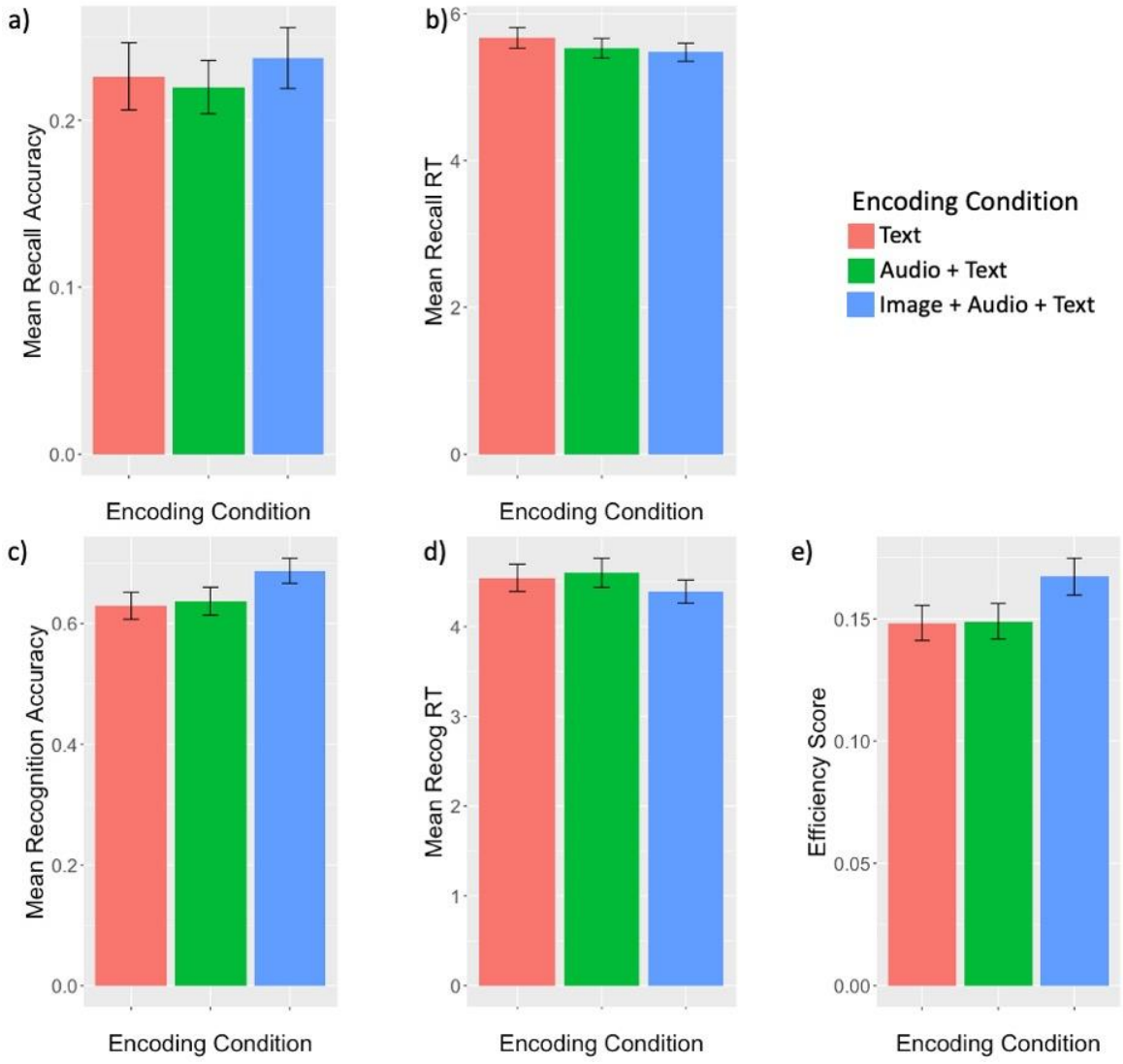


Figure 15: Results for Study 3, Experiment 2

Performance in (a) recall accuracy and (b) recall RT. Participants show no differences in recall accuracy or RT between the three conditions. In (c) recognition accuracy, we that IAT accuracy is higher than AT or T accuracy, and (d) no significant difference observed in average recognition RT. (e) Efficiency scores showed a significant effect, such that participants performed higher in the IAT condition compared with the AT or T conditions, in line with recognition accuracy.

Interim Discussion

Experiment 2 had several key differences from experiment 1. Accuracy in cued recall showed no significant differences, and accuracy in recognition showed was higher when image, audio, and text were presented at encoding when compared to text alone or audio and text. However, while quantitatively different, theoretically these findings share many similarities. Both seem to violate the modality principle, as the condition with a higher visual processing load shows equal or higher accuracy than the condition using two different senses, and adding information across senses in audio plus text does not improve performance on recall or recognition above encoding with text alone. This, again, indicates that sensory enrichment is not facilitating retention of word translations for these participants.

Experiment 3

Experiment 3 sought to more directly investigate whether cognitive enrichment was equally effective in different sensory contexts. Based on the modality principle, it is expected that the AT condition yields better results compared to the IT condition as it would avoid sensory overload.

Methods

Participants

Participants were 52 undergraduate students recruited from the University of California, Los Angeles (UCLA) psychology department subject pool. The average age of these participants was 20.4 years. Thirty-six of these participants were female, eight were male, and eight declined to provide their gender identity. All participants reported having normal or corrected-to-normal sight, normal hearing, being fluent in English, and not being familiar with Swahili. One participant was fluent in another language from the Niger-Congo language family,

but was kept in the analyses because the language they spoke was from a different branch of this language family than Swahili.

Materials

Materials were identical to those used in experiment 1, though a subset of 80 of the 90 original words were used for this experiment.

Procedure

The procedure was similar to that used in experiments 1 and 2, but with the following changes (Fig. 16).

During encoding, participants were exposed to four different conditions. In two of these conditions, participants were exposed to audio recordings of the translation. These were the Audio condition, where participants listened to just the audio track stating “[Swahili Word] means [English Word],” and the Audio plus Image condition, where the audio track was accompanied by a Snodgrass-type color image depicting the item being named. In the other two conditions, text was presented with no accompanying audio. These included the Text condition, where a line of text reading “[Swahili Word] means [English Word]” was shown, and the Text plus Image condition, where this text was accompanied by a Snodgrass-type image depicting the item being named.

During the cued recall phase, participants were presented with an image representing the word they were trying to recall. Participants were asked to type the Swahili word corresponding to the word represented by that image.

Blocks still contained 10 words each for participants to learn, but each condition was only repeated two times, meaning that there were 8 blocks total in the experiment. Words were assigned to a block and condition within a version of the experiment, and a total of 4 different

versions of the experiment with different word assignments were created, to limit the influence of item effects.

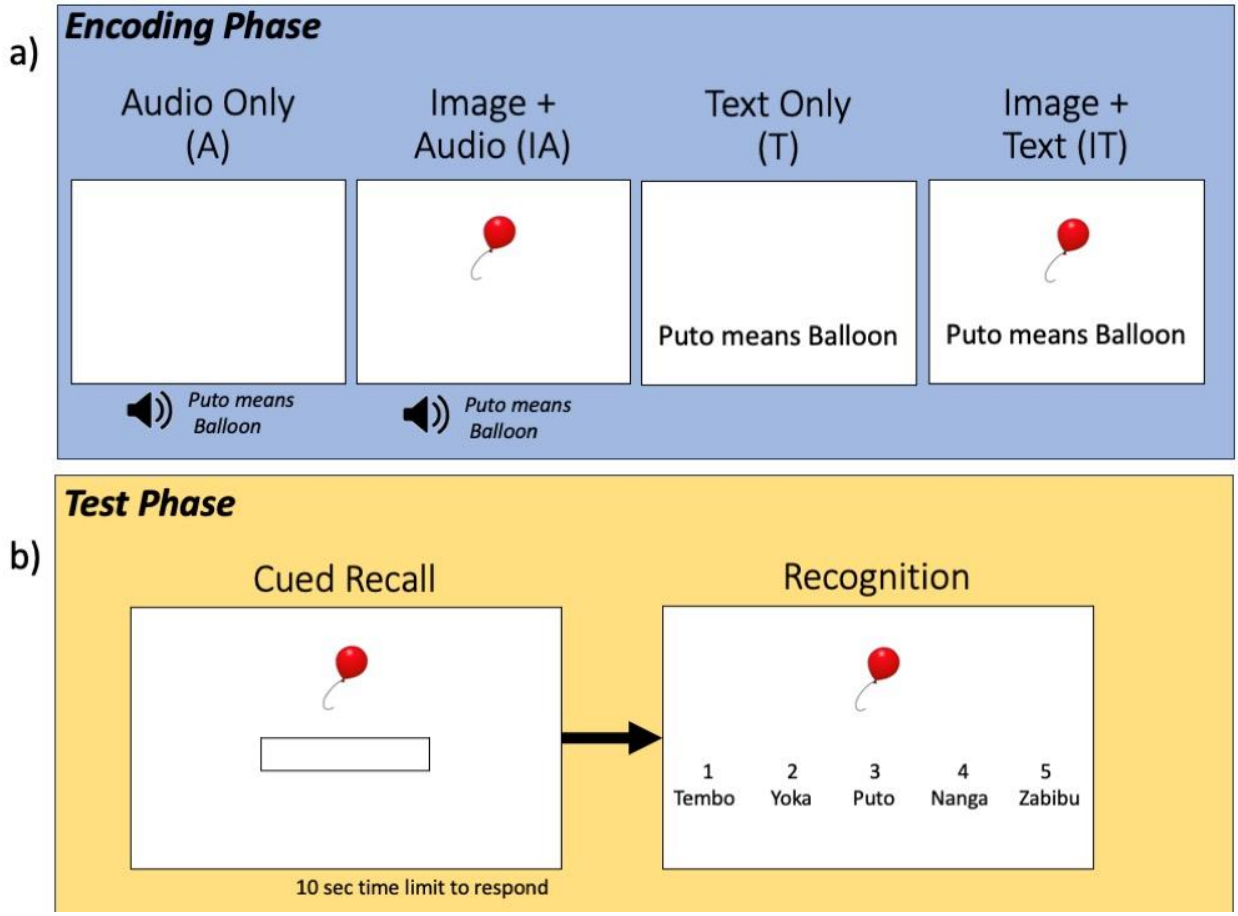


Figure 16: Encoding Condition and Test Phase for Study 3, Experiment 3

All examples in this figure use the translation “Puto means Balloon,” though each participant would only encounter the word in one encoding condition. (a) Examples of the stimulus “Puto means Balloon” in all 4 different conditions present in the encoding phase of the experiment. (b) Example trials for Cued Recall and Recognition for the word Balloon, where participants would respond with the English translation. The other translations presented in the recognition phase would be randomly selected from words presented in the same block as the target translation.

Statistical Analysis

Participant responses for the cued recall phase were scored by computer, ignoring capitalization and any spacing or punctuation, as in experiment 1. Recognition was scored as in the previous two experiments.

Results

Figure 17a-c shows the overall trend in recall performance in experiment 3. A repeated-measures ANOVA was run on the recall data, and revealed a significant difference in recall performance based on encoding condition ($F(3,153) = 50.62$, $p < .001$, generalized $\eta^2 = 0.31$). Post-hoc paired Student's t-tests with Holm corrections on the p-values showed that participants remembered more words on average when encoded as text alone when compared to audio alone ($t(51) = 5.96$, $p < .001$) or when compared to encoding with an image and audio ($t(51) = 2.98$ $p = 0.004$), but was less effective than encoding with an image accompanying the text ($t(51) = 6.51$, $p < .001$). Audio encoding produced lower recall accuracy than encoding with images accompanied by audio ($t(51) = 3.27$, $p = 0.003$) or by encoding images with text ($t(51) = 9.53$, $p < .001$). Images that accompanied text produced higher recall accuracy than images accompanied by audio ($t(51) = 8.21$, $p < .001$; Fig. 17a). Response time for recall showed no significant differences between groups ($F(3,153) = 1.92$, $p = .13$, generalized $\eta^2 = 0.02$; Fig 17b).

To investigate whether cognitive enrichment produced a larger effect when the modality principle was being observed, we additionally analyzed difference scores between our conditions with images and those without the images included (Fig. 17c). This allowed us to compare the difference between the benefit received from using an image in conjunction with audio to that received when using an image in conjunction with text. Participants scored an average of 4.90% (SD = 1.50%) higher in auditory conditions when given an image and scored

an average of 13.37% (SD = 2.05%) higher in text-based conditions when given an image. The difference between these improvement scores was significant ($t(51) = 3.92$, $p < .001$).

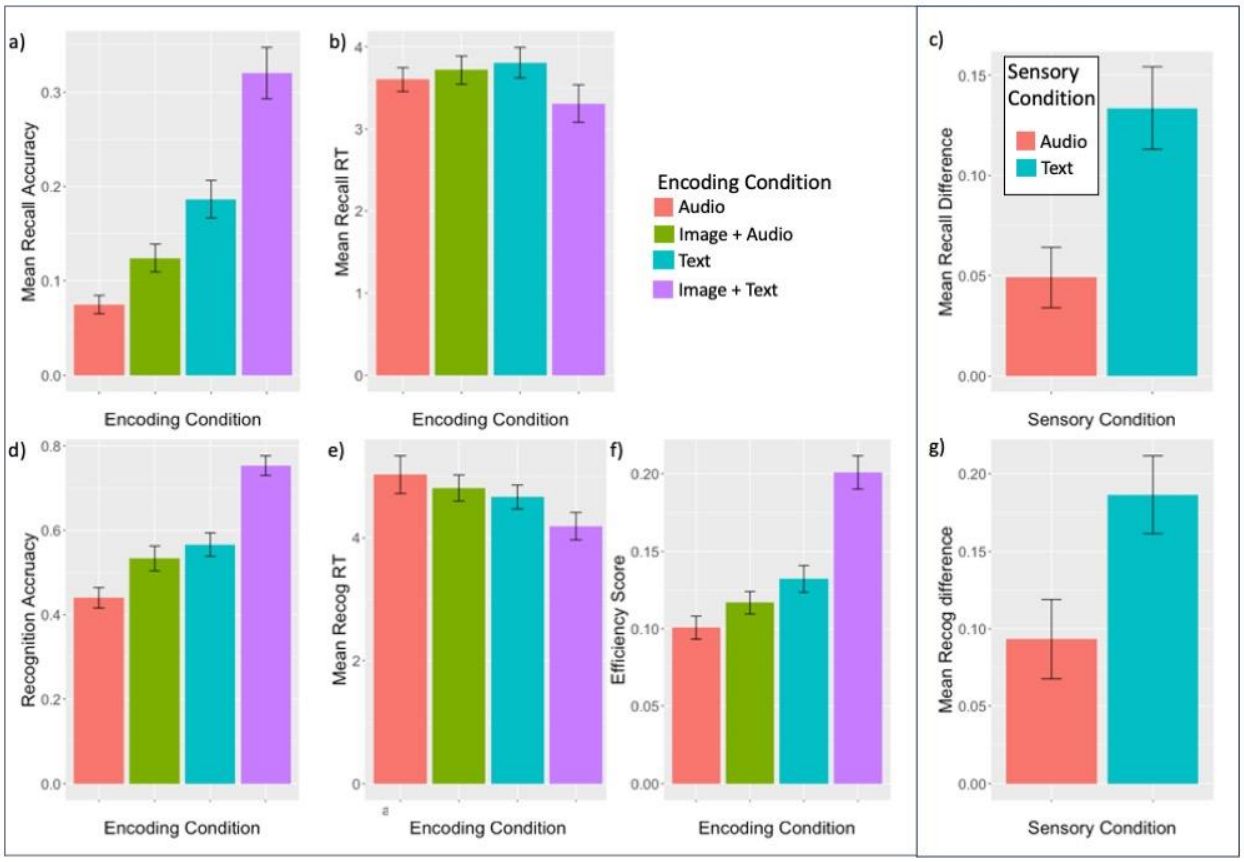


Figure 17: Results for Study 3, Experiment 3

Performance in (a) recall accuracy and (b) recall RT. Participants show significant differences in accuracy between all encoding conditions, with highest performance in IT, and no difference in response time. Additionally, (c) differences between audio-based conditions anf text-based conditions show that text received a larger benefit from adding images than audio, even though these utilize the same sensory modality. In (d) recognition accuracy, all differences are significant except that between IA and T. (e) Response times did significantly differ, with participants responding faster in the IT condition as compared to the A and IA. (f) Efficiency scores showed a significant effect, in line with recognition accuracy except in that IA is no longer

different from A. (g) Difference scores in recognition showed a similar trend to those in recall, where text received a larger benefit from adding images than did audio.

Recognition data for this experiment showed a similar pattern of results (Fig 17e-g). A repeated-measures ANOVA showed a significant difference in recall performance based on encoding condition ($F(3,153) = 48.98$, $p < .001$, generalized $\eta^2 = 0.27$). Post-hoc paired Student's t-tests with Holm corrections on the p-values showed that participants remembered more words on average when encoded as text alone when compared to audio alone ($t(51) = 4.45$, $p < 0.001$), but not when compared to encoding with an image and audio ($t(51) = 1.22$, $p = 0.23$). Text was also less effective than encoding with an image accompanying the text ($t(51) = 7.43$, $p < .001$). Audio encoding produced lower recall accuracy than encoding with images accompanied the audio ($t(51) = 3.63$, $p = 0.001$) or by encoding images with text ($t(51) = 12.26$, $p < .001$). Images that accompanied text produced higher recall accuracy than images accompanied by audio ($t(51) = 7.98$, $p < .001$).

Response time showed a significant difference across the different conditions in the recall task ($F(3,153) = 5.60$, $p = .002$, generalized $\eta^2 = 0.03$). Post-hoc tests shows that average response time was faster for words presented as image plus text than for those presented as audio ($t(51) = 3.24$, $p = .01$) and those presented as image plus audio ($t(51) = 3,34$, $p = .009$). There was also a trend for responses to be faster in image plus text than text alone, though this was not significant ($t(51) = 2.37$, $p = .086$). All other comparisons were non-significant ($p \geq .41$ for all comparisons).

Efficiency scores showed a similar trend to the recognition accuracy, with an overall significant difference existing between the efficiency score in different encoding conditions ($F(3,153) = 52.50$ $p < .001$, generalized $\eta^2 = 0.28$). Significant differences emerged in most of the same post hoc comparisons as in the recognition accuracy. Participants remembered more words on average when encoded as text alone when compared to audio alone ($t(51) = 4.82$, $p <$

0.001), but not when compared to encoding with an image and audio ($t(51) = 1.85$, $p = 0.13$). Text was also less effective than encoding with an image accompanying the text ($t(51) = 8.00$, $p < .001$). Audio encoding showed equal recall accuracy as encoding when images accompanied the audio ($t(51) = 1.87$, $p = 0.13$), but showed worse efficiency than encoding images with text ($t(51) = 11.16$, $p < .001$). Images that accompanied text produced higher recall accuracy than images accompanied by audio ($t(51) = 8.27$, $p < .001$).

We also investigated the difference in benefit received in audio- and text-based conditions when an image was added in the recognition task (Fig. 17g). Participants scored an average of 9.33% (SD = 2.57%) higher in auditory conditions when given an image and scored an average of 18.65% (SD = 2.51%) higher in text-based conditions when given an image. The difference between these improvement scores was significant ($t(51) = 2.42$, $p = 0.02$).

Interim Discussion

Data from experiment 3 show that participants show the best memory for new words when these are encoded as text with image and perform the worst when these words are encoded as audio. We also see, contrary to our hypothesis, that participants performed significantly better when text was accompanied by an image than when audio was accompanied by an image, again violating the modality principle. Indeed, the benefit received from adding an image to text was even greater than the benefit received from adding the image to audio, highlighting this violation. While some of the benefit observed in the image-included conditions may derive from the nature of the cueing in this task with an image, as these conditions are more closely replicated at test, they do not explain the observed violation of the modality principle.

Experiment 4

Experiment 4 was a replication of experiment 3, where the crucial difference was the means through which participants provided their response. Participants here provided their response by speaking the English translation of the Swahili word aloud, in case the text-based method of response created a bias in responding that benefitted text-based encoding. Participants were also prompted with audio in the cued recall instead of an image, to ensure the presentation of the image at retrieval was not responsible for the improved memory performance in the image-present conditions.

Methods

Participants

Participants were 54 undergraduate students recruited from the University of California, Los Angeles (UCLA) psychology department subject pool. The average age of these participants was 20.4 years. Thirty-seven of these participants were female, 15 were male, and two declined to provide their gender identity. All participants reported having normal or corrected-to-normal sight, normal hearing, being fluent in English, and not being familiar with Swahili or any related languages.

Materials

Materials were identical to those used in experiment 3.

Procedure

The procedure of this experiment was identical to experiment 3 in the encoding phase, but differed in the test phase.

During cued recall, participants were prompted with audio that repeated the Swahili word 3 times. Participants were then asked to say aloud the English translation of that word. Participants' auditory responses were recorded via zoom. The recognition phase was also removed from this experiment (Fig. 18).

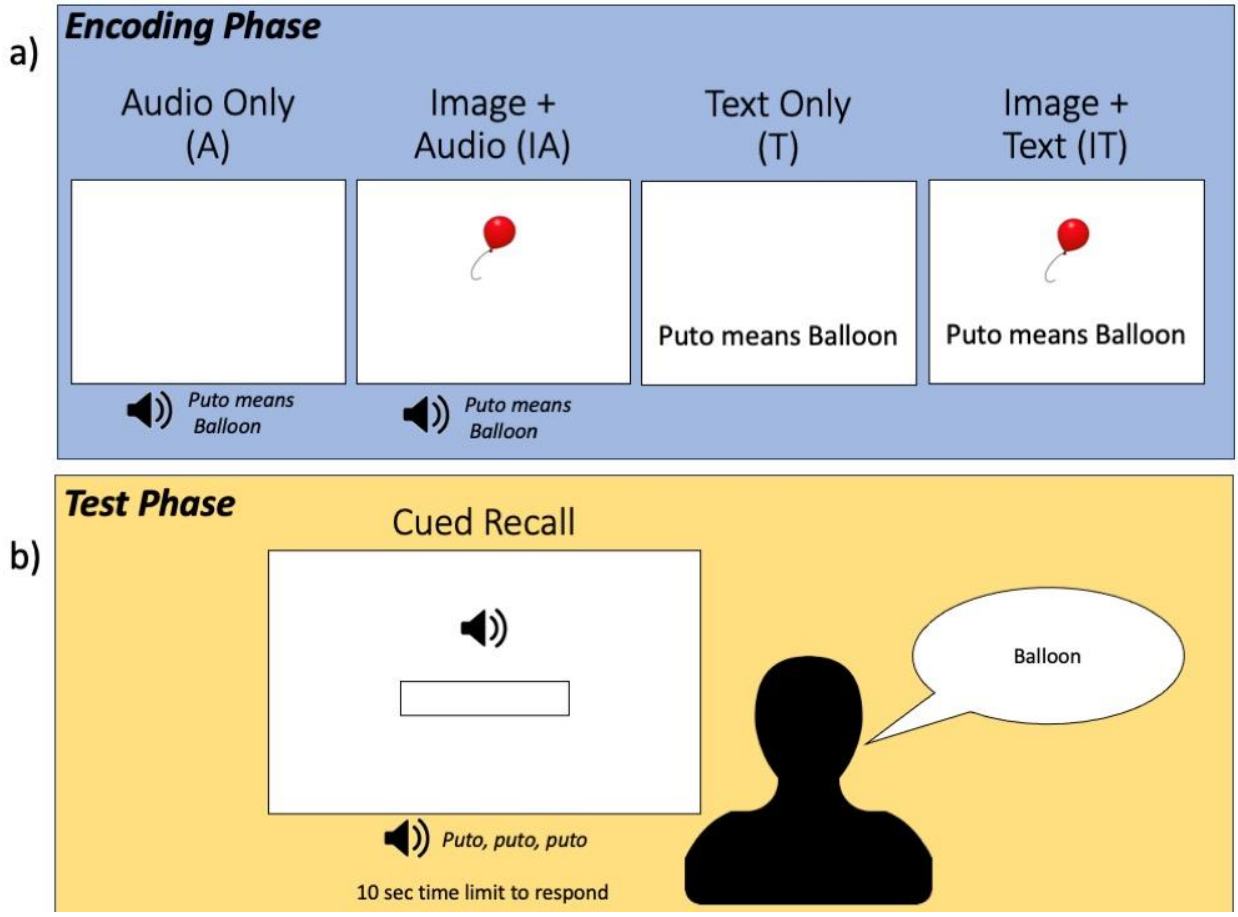


Figure 18: Encoding Condition and Test Phase for Study 3, Experiment 4

All examples in this figure use the translation "Puto means Balloon," though each participant would only encounter the word in one encoding condition. (a) Examples of the stimulus "Puto means Balloon" in all 4 different conditions present in the encoding phase of the experiment. (b) Example trials for Cued Recall for the word Puto, where participants would respond by speaking the English translation aloud.

Statistical Analysis

Participant responses in cued recall were scored by human raters. One human rater scored each participant's files, blinded to the condition presented in each block. Raters were informed to score items as correct if the participant produced the word correctly, ignoring issues of plurality of the word or slight mispronunciation that did not make the word difficult to discern. Additionally, RT could not be reliably measured for this experiment, so analyses involving RT have not been included.

Results

Figure 19 shows the overall trend in recall performance in experiment 4. A repeated-measures ANOVA was run on the recall data and revealed a significant difference in recall performance based on encoding condition ($F(3,153) = 16.02$, $p < .001$, generalized $\eta^2 = 0.10$). Post-hoc tests showed that participants remembered more words on average when encoded as text alone when compared to audio alone ($t(51) = 2.77$, $p < 0.016$), but not when compared to encoding with an image and audio ($t(51) = 0.09$ $p = 0.93$).Text alone encoding was less effective than encoding with an image accompanying the text ($t(51) = 3.99$, $p = .001$). Audio encoding produced lower recall accuracy than encoding with images and audio ($t(51) = 3.47$, $p = 0.003$) or by encoding images with text ($t(51) = 7.01$, $p < 0.001$). Images that accompanied text produced higher recall accuracy than images accompanied by audio ($t(51) = 3.74$, $p = 0.002$; Fig 19a).

To investigate whether cognitive enrichment produced a larger effect when the modality principle was being observed, we additionally analyzed difference scores between our conditions with images and those without the images included (Fig. 19b). This allowed us to compare the difference between the benefit received from using an image in conjunction with audio to that received when using an image in conjunction with text. Participants scored an

average of 5.77% (SD = 1.66%) higher in auditory conditions when given an image and scored an average of 7.43% (SD = 1.86%) higher in text-based conditions when given an image. The difference between these improvement scores was non-significant ($t(51) = 0.64$, $p = 0.53$).

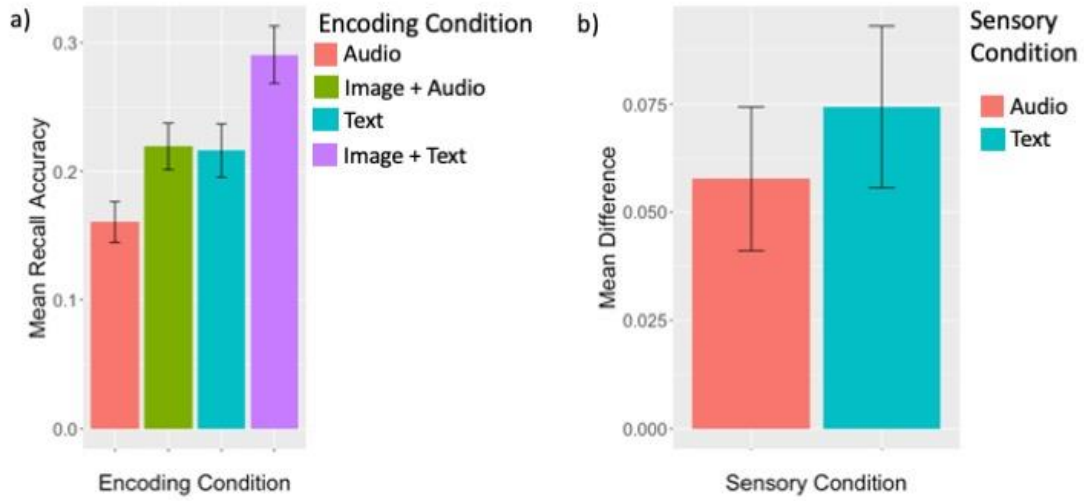


Figure 19: Results for Study 3, Experiment 4

(a) There was a significant difference in recognition accuracy based on encoding score. All methods differ significantly from one another except text alone and image plus audio, which are not significantly different from one another. (b) Difference scores in audio- and text-based conditions between when images are presented in addition to that sensory information or not. These do not differ significantly from each other, though both are significantly greater than zero, indicating a general benefit to recall from using images. Error bars represent standard errors.

Discussion

Across four experiments, we observed several different trends of performance, with a few important similarities across the experiments. In experiments 1 and 2, we directly tested the relative memory improvement with sensory enrichment to the benefit obtained from adding cognitive enrichment. Neither experiment showed a significant improvement in cued recall or recognition for Swahili-English translations when audio was presented with text and audio, providing sensory enrichment, beyond what was seen when text was presented alone. However, adding an image, and thus creating cognitive enrichment, did improve performance in experiment 1. This result was not observed in experiment 2. The differences in these experiments primarily lay in the language participants were asked to respond in, which may have caused the differences in the outcomes of these experiments. These experiments were also underpowered, as estimated in a post-hoc power analysis, which could also explain why their results differ. Future experiments could replicate these procedures with higher participant enrollment to test these outcomes with higher power.

Cognitive enrichment was also beneficial for learning in experiments 3 and 4, where adding images to audio or text improved learning. Additionally, in these experiments, adding images to text provided as much (exp 4) or a greater (exp 3) benefit to memory as adding images to audio, even though adding visuals to text was predicted to be less effective than adding images to audio, under the modality principle.

Taken all together, these experiments seem to indicate that, in many foreign-language learning tasks, cognitive enrichment is more effective for improving retention of new words learned than sensory enrichment. Indeed, this study showed no benefit to sensory enrichment between audio and visual information– there was no observed benefit for adding audio to text, and adding an image to audio during encoding was not more effective for improving word retention than adding an image to text. Such findings may seem at odds with multimedia

learning frameworks, given that the modality principle in the cognitive theory of multimedia learning promotes utilization of multiple sensory modalities to avoid overloading the processing capacity of an individual sense. Even outside the framework of the cognitive theory of multimedia learning, providing information across multiple senses is expected to improve memory for items. Paivio's Dual Coding theory (1991) suggests that providing information across multiple sensory stores is helpful (though, as the cognitive theory of multimedia learning is also a dual-coding model, perhaps this shared prediction is not shocking). Empirical evidence looking at simpler memory paradigms for objects suggests that memory for images can be improved by providing audio that is congruent in meaning with the image during encoding (Lehmann & Murray, 2005; Duarte et al., 2022), and that the inverse of providing images to audio will improve ability to recognize that audio later (Moran et al., 2013). Findings also suggest that providing audio accompanying text in one's native language can improve memory for those words compared to text presented with white noise (Heikkilä & Tiippana, 2016), implying that these findings can generalize to verbal stimuli.

Previous research in multimedia language learning, however, sometimes indicates that the redundancy and modality principles do not always appear to predict performance in language-learning experiments. When looking at vocabulary learning in a foreign language, a recent review suggests that empirical research finds that text with audio and graphics is most often conducive to participant learning, though reports of text and image-based learning being superior to text- or audio-only learning for retention of vocabulary in a foreign language were also prevalent (Zhang & Zou, 2022). Such findings are often attributed to a few different reasons. Among these are that reading text as opposed to listening to a foreign language can be less cognitively demanding, decreasing the load of learning and making it easier (Zhang & Zou, 2022). Indeed, high comprehension and high working memory load, which would lessen the extent to which a foreign language task is cognitively demanding, have been correlated with decreased reliance on captions upon second viewing of information, indicating captions (and

text information in general) may be of greatest assistance to individuals who are less fluent with the task or have low working memory capacity (Gass et al., 2019).

Providing images was also hypothesized in many cases to activate both semantic representations for words, in addition to the lexical activation from text or auditory word presentation, which allowed participants to more effectively process the information across multiple levels, improving their memory for vocabulary (Zhang & Zou, 2022). Violations of the redundancy principle are also observed in previous work looking specifically in foreign language learning, where retention of information given in a speaker's second language is better for text accompanying a video image than for audio accompanying a video image (Lee & Mayer, 2018). However, where present, the benefit for including images is often suggested to be due to the increased cognitive load of processing information in a second language. This creates a case where the transient nature of speech may not allow adequate time for participants to process what is being said, whereas text is often present for longer and thus provides more time for participants to process the information they are learning (Lee & Mayer, 2018).

Generally, these violations seem to indicate that cognitive enrichment may be more useful for improving performance for high-cognitive load tasks than sensory enrichment. This would imply that previous findings supporting a benefit from adding more sensory information at encoding leading to better memory outcomes may only generalize to tasks with low or middling cognitive demands. Future research will be required to investigate at what level of cognitive load and for what kinds of tasks this may be true. As split attention is also a hypothesized reason why adding auditory information to visual processing for words may not be helpful for learning those words, it may also be that sensory enrichment is most helpful in cases where it helps to guide attention. Multisensory stimulus presentation can help direct attention towards crossmodal events, but this may work best when competition for attentional resources is low (Talsma et al., 2010). Learning vocabulary in a new language may have higher attentional demands (e.g., associating the word and the meaning, learning new phoneme frequencies, etc.) than previous

multisensory memory studies whose stimuli were common objects or familiar words, and thus not benefit from sensory enrichment. A more systematic study of the relationship between attentional demand and memory benefits from multisensory stimulus presentation would help to investigate this possibility. Additionally, sensory enrichment may be useful in cases where cognitive enrichment provides less semantic information. For example, face-name associations seldom have a deep semantic relationship between the visual face and spoken name, but providing written information supporting the auditory name has been shown to improve memory for the face-name pairing (Murray et al., 2022). More research would be needed however, across a variety of tasks where semantic information is relatively limited, to test this hypothesis.

Overall, the current research indicates that cognitive enrichment is more effective for promoting memory of new vocabulary in a foreign language than sensory enrichment. Additionally, in violation of the modality principle, we found that providing images with text produced better memory for new words and their translations than providing images with audio. While the literature in the area of multimedia theory in language learning is overall mixed, the current results are in line with previous work that indicates that the redundancy and modality principles may not adequately predict outcomes in second-language learning, due to the high cognitive load and attentional demands associated with the task. This may also suggest that multisensory presentation and sensory enrichment during encoding is most effective in cases with low cognitive load and attentional demand, whereas cognitive enrichment may be helpful in those cases.

Chapter 5: Sensory Inputs are Encoded in Memory After Integration with Other Senses for Most Individuals

Abstract

At almost every moment, we are bombarded with sensory information across different sensory modalities which we encode and use for later memory retrieval. It is well established that perception of the objects and events around us involves interaction between the senses, at times leading to modifications of perception in one modality by signals from another modality. However, it remains unclear whether memory encoding utilizes sensations prior to crossmodal integration, the fusion across the senses, or both. To tease apart these possibilities, across two experiments, we presented participants (N = 121 and N = 125) with speech stimuli, wherein a simultaneously presented video of an incongruent syllable changes the perception of the auditory syllable. Thus, the perception of the syllable after integration with vision differs qualitatively from the perception of the syllable prior to integration. The participants' memory of auditory syllables was then tested after a delay in a recognition task. The majority of participants rated the integrated syllable as being old more often on average than the unfused syllable. However, a minority of participants found the auditory syllable more familiar than the integrated syllable. Individual variability in the results suggests varying styles of memory

encoding across individuals, with a majority primarily encoding multisensory fused representations, a minority primarily encoding independent unisensory representations, and some both.

Introduction

Events in daily life are almost always multisensory. At any given moment, the human brain is processing sights, sounds, smells, and other sensory information to create a coherent understanding of the world around us. To do this, we must be able to understand which information in a scene goes together, and which does not. Crafting multisensory representations--those that combine information that is congruent in space and/or time across multiple sensory inputs to create a unique cross-modal representation-- are crucial to helping us parse this barrage of sensory information efficiently. Furthermore, we do not just parse scenes in the present, but we also store information about the present in memory for use in the future. Human episodic memory is noted for being rich in sensory details (Gillund, 2012), which reflects that remembering events must utilize memory traces from across different senses.

Indeed, human memory appears to be improved by the presence of multiple sensory sources of information during encoding. Audiovisual encoding of objects has been shown to improve later visual (Lehmann & Murray, 2005) and auditory (Moran et al., 2013) recognition for previously seen objects compared to those initially encountered in a unisensory fashion, even though the final memory test was unisensory. Recall performance for visual objects has also been shown to be improved if those stimuli are encoded with congruent auditory

information (Duarte et al., 2022). Similar findings have been reported for other types of memory , including remembering words (Heikkilä et al., 2015; Heikkilä & Tiippana, 2016), or remembering associations between names and faces (Murray et al., 2022). These findings can be viewed as largely consistent with existing abstract models of human memory, notably dual-coding theory (Clark & Paivio, 1991; Paivio, 1991), in which the presence of multiple routes to a memory will improve the chance of later retrieval. Neural evidence also supports a special role for multisensory stimulus presentation in memory retrieval, such that information encoded in a multisensory manner is differentiable at retrieval from information that was encoded in a unisensory manner (Thelen et al., 2012). However, theories such as dual-coding theory do not make a distinction between providing *more* sensory information and providing *multisensory information*, and thus fail to consider interactions between representations prior to encoding. The mere presence of two or more sensory sources of information is believed to be the basis for the facilitation of retrieval. On the other hand, accounting for behavioral findings on multisensory learning (e.g. Kim et al., 2008; Mathias & von Kriegstein, 2023; Seitz et al., 2006) and memory (e.g. Duarte et al., 2022; Murray et al., 2022; Lehmann & Murray, 2005; Heikkilä et al., 2015), others have proposed neural mechanisms that involve integration and interaction between the sensory representations as the basis for the benefits of multisensory encoding (Quintero et al., 2022; Shams & Seitz, 2008). In these postulated models, it is not the mere presence of multiple sensory signals during encoding that is helpful, but rather the integration of the sensory signals that benefits memory and learning. While there are experimental findings supporting both models, there remain important questions about the mechanisms of human memory. Namely, it is unclear at what stage of processing sensory representations are encoded

into memory. It is conceivable from existing evidence that information is encoded into memory prior to sensory interaction and integration, that encoding occurs after sensory integration, or that both pre- and post-integration representations are encoded in memory. While one of these scenarios may appear more plausible than others, in fact arguments can be made in favor of each of these three schemas.

There are three distinct ways that concurrent representations across sensory modalities could be encoded into memory (Figure 20). One possible schema would be that unisensory information is all that is encoded for memory, with multisensory representations existing only for other purposes. This encoding scheme would allow for relatively fast encoding of sensory information as less processing would need to be performed prior to storage. This encoding method would allow for a better match between unisensory information and memory traces, potentially allowing for easier activation of a memory using a unisensory probe. A challenge of this schema, however, would be recognition of multisensory experiences if they are very different from the unisensory representations. However, multisensory experiences can be reconstituted from unisensory memory representations upon retrieval, albeit this could slow down the retrieval process.

A second possible encoding schema would be to encode only integrated multisensory representations into memory. This would provide relatively high reliability between the lived experience and the memory representation, as both would reflect the full sensory experience. This would also be a highly efficient way to represent the information, as it would provide only one sensory representation of our lived experience, whereas schemes with unisensory representations would necessitate more traces be stored to represent the same experience as a

single unit sensory representation. However, this encoding mechanism would have the disadvantage of making it challenging to retrieve the memory using only one sensory cue, especially in cases where the unisensory information would be qualitatively different from the multisensory representation. This hypothesis would be in line with existing studies showing that familiarity ratings for images to be more accurate for images originally presented with congruent sound than those presented with incongruent sound (e.g. Lehmann & Murray, 2005).

A third possibility would be that both unisensory and multisensory representations are encoded into memory. This would combine the relative strengths of the previous two schemas, and would provide easier recognition in all cases. Whether or not the unisensory and multisensory experiences are a close match to one another would not matter, and either input could be used to retrieve the memory. However, this encoding method has the disadvantage of requiring more processing and memory storage than the other encoding schemas. In cases where unisensory and multisensory representations would be similar to one another, it would be redundant to have both types of information encoded.

Distinguishing which of these three schemas are used in human memory processes is challenging with existing paradigms, due to the close match between the multisensory and unisensory representation of items; the representations of the dog presented as just a picture compared to a dog presented as the same picture with a bark does will be highly similar, and those slight differences will not be enough to distinguish which trace is encoded and available for retrieval. Addressing this question necessitates using a novel paradigm, where unisensory and multisensory representations can be more easily distinguished from one another. To that end, the current research study used examples of the *McGurk illusion* (Alsius et al., 2018;

McGurk & McDonald, 1976) where a visual representation of a syllable is played simultaneously with an incongruent auditory representation, creating a distinct fused multisensory representation of the sound (see Fig. 20 text for an example). This paradigm allows us to test participants' memory for both the unisensory auditory representation, using the auditory syllable, and the multisensory representation, using the fused syllable. This test allows the current study to clarify which of these patterns are used in human memory processes.

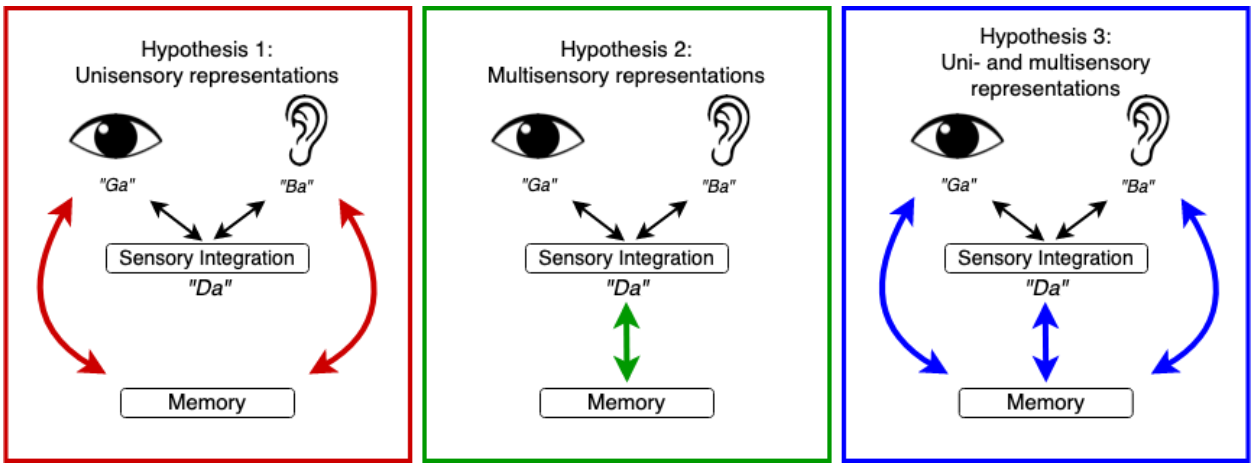


Figure 20: Schematic diagram of hypotheses for Study 5

Competing hypotheses for how humans may encode and retrieve multisensory and unisensory representations of sensory stimuli. Hypothesis 1 poses that individuals are able to integrate multisensory information, but do not encode that in memory storage, instead encoding only unisensory representations. Hypothesis 2 posits that individuals only store crossmodal information in memory, and do not store unisensory representations. Hypothesis 3 proposes that individuals store both kinds of representations and have access to both during memory retrieval.

Experiment 1

Methods

Participants

Participants in the study were 121 volunteers (78 female, 38 male, 4 declined to answer) recruited from the University of California, Los Angeles psychology subject pool. The average participant age was 20.81 years. All reported normal or corrected-to-normal sight, normal hearing, and all were fluent English speakers. Informed consent was obtained from each participant and experimental procedures were reviewed and approved by the UCLA Institutional Review Board.

Materials

A total of 21 videos, each lasting between four and five seconds, were created in-lab for the experiment. The subject of all videos was a volunteer who was recorded standing against a white background. The volunteer was provided an earbud that played a sound from a metronome so they would provide syllables at the same pace across recordings. In each video, the participant repeated the same syllable three times. In the 18 control videos, the participant's lip movements and auditory track were consistent. The remaining three videos were then further processed to create McGurk stimuli: audio and video of different syllables that, when combined, produce a percept different from the original audio. We used three different McGurk stimuli: *Ba* (A) + *Ga* (V) → *Da* (AV), *Pa* (A) + *Ka* (V) → *Ta* (AV), and *Ga* (A) + *Ma* (V)--> *Na* (AV). These were selected because they were used in previous studies (e.g., Brown et

al., 2018), whose auditory and multisensory representations would not overlap with one another. Video editing software was used to pair separate audio and visual recordings to make the McGurk stimuli, with audiovisual alignment adjusted to bring out the McGurk illusion. Stimuli were piloted on lab personnel who did not partake in the final experiment to ensure the McGurk illusion was present upon viewing the stimuli. Examples of congruent and McGurk videos can be found in the supplementary materials. Videos were presented in a size of 960 by 540 pixels against a dark gray background using Qualtrics online experimental software with custom JavaScript code to randomize the video order.

Procedure

The basic procedure is outlined in Figure 21. Participants started in the encoding phase, where they were instructed to view videos of an individual speaking syllables and were told they would later be tested on the syllables they heard. During this phase participants were shown four videos in a random order: one McGurk stimulus and four randomly selected congruent syllable stimuli. Before each video affixation cross was presented in the middle of the screen for 1.5 seconds. Following this, videos were presented on a white background and at the center of the screen.

After this, participants were given a one-minute delay, during which they played a game of snake—a silent version of the classic computer game where players are asked to capture squares with their lengthening “snake” without crossing the snake’s own body or colliding with a wall. The game was intended to prevent participants from rehearsing the audio while not providing any interfering auditory information.

Participants were then given a recognition task. During the recognition task participants were given four auditory stimuli to respond to. These auditory stimuli were split into four categories:

1) An old audio: a previously presented audio from a congruent video
2) A new audio: a syllable which the participant has not head during encoding
3) A *fused audio*: the McGurk percept from a previously observed video (so, for auditory *ba* and visual *ga*, this would be *da*), reflecting a multisensory representation of the sound heard
4) An *unfused* audio: the "true" audio heard in the McGurk trials (so, for auditory *ba* and visual *ga*, this would be *ba*), reflecting a unisensory representation of the sound heard

Participants were presented with one example of each category per recognition block, in a random order. For each trial, participants were presented with audio with a prompt asking them to judge if the audio was new or old by using the number keys. The response and reaction time to input the response was recorded for each trial period there were three repetitions of the encoding-delay-recognition blocks.

After this task, participants were given an abbreviated Raven's Progressive Matrices. After that, participants were given a perception task, where a syllable video was played, and the participant was asked to report which syllable they heard via a multiple-choice response. Participants were first given the three McGurk stimuli they had previously seen, in a random order. For each question, the options given to participants always included the fused audio, the

unfused audio, a third syllable chosen randomly from the full list of congruent syllables in our dataset, and "other." The first three of these multiple-choice options were ordered randomly. Participants who chose "other" were prompted to type what sound they heard. After these three McGurk stimuli participants were asked to also report on three randomly selected congruent audiovisual stimuli. The multiple-choice options for these included the actual sound heard, two randomly selected syllables from those heard previously in the experiment, and an "other" category that prompted further response, as in the McGurk trials. The first three of these multiple-choice options were ordered randomly. After each report of what syllable the participant heard, participants were asked to rate their confidence in their decision on a scale from 1 (low confidence) to five (high confidence).

To ensure participant compliance with experimental protocol, experimental sessions were monitored by an experimenter via Zoom software.

Statistical analyses

Data was collected from the Recognition portion of the memory task, the Raven's progressive matrices, and the Perception Task. The recognition task was a New/Old judgment, where "New" judgments were scored as zero and "Old" judgments were scored as a 1. Response time (RT) was measured as the mean time to respond to the prompt of the New/Old judgment in the Memory Task. These RTs were then trimmed, such that values exceeding the 99th percentile were removed (RT > 15.61 seconds), as these likely represented attentional lapses. For the perception task, only responses of the traditional McGurk percept (e.g., the multiple-choice option *da* for the video representing were rated as *Ba* (A) + *Ga* (V) → *Da* (AV))

were rated as participants experiencing the fusion. All other responses, including “other” were rated as not perceiving the fusion.

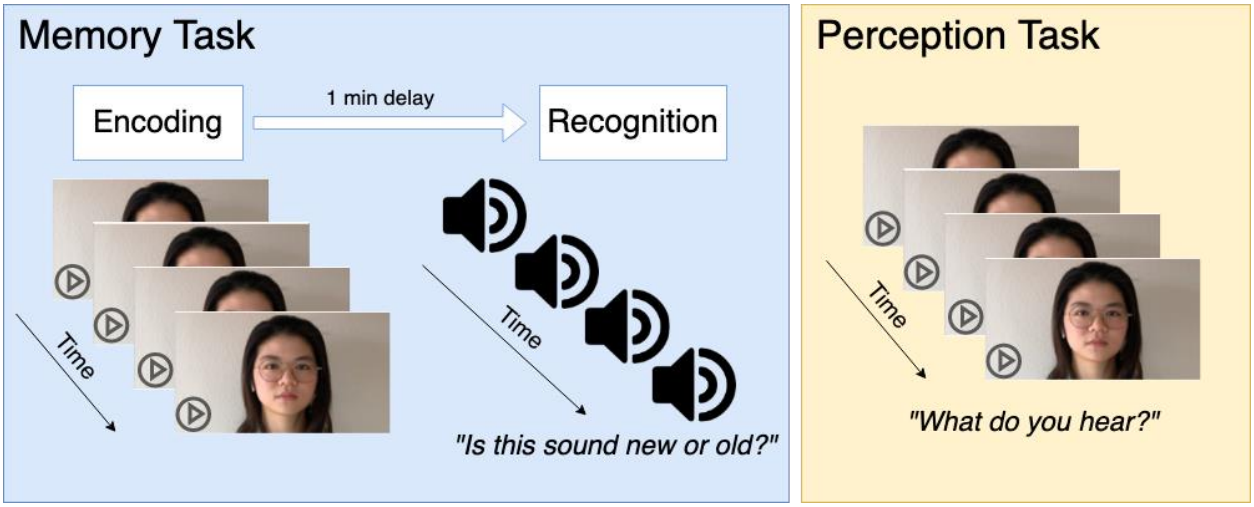


Figure 21: Methods for Study 4

Participants were first given a Memory Task, during which they were presented with videos of a speaker who repeated a syllable 3 times, where the stimulus was either a congruent syllable or a McGurk stimulus. Participants were then given a 1-minute delay and then asked to rate audio-only syllables (also repeated 3 times) as new or old. After the memory task was completed, participants were given another delay and then a Perception Task, where what syllable they perceived was assessed.

Statistical analyses were performed in R (Version 4.0.2). Comparisons between all four auditory memory tests conditions were made using a Friedman test, and post-hoc tests were two-tailed Wilcoxon sign rank tests with Holm-corrected p-values. All correlations performed

use the Spearman's method as our predictor variables were often range-limited, requiring the use of a non-parametric test.

Results

An overview of the memory results can be found in Figure 22. Participants, on average, recognized old stimuli as "old" on 87.22% of trials (95% CI = [83.38%, 91.06%]). New audio was recognized as "old" on an average of 25.00% of trials (95% CI = [20.36%, 29.64%]). Fused audio was rated "old" on an average of 61.38% of trials (95% CI = [56.21%, 66.57%]) and unfused audio were rated as "old" in 39.72% of trials (95% CI = [34.13%, 45.32%]).

A Friedman's test indicated that there were significant differences between recognition scores of the different stimulus types, $\chi^2(3) = 185.15$, $p < .001$, Kendall's $W = 0.51$ (Fig. 22a). As such, post hoc Wilcoxon sign rank tests with a Holm correction were conducted. Participants were generally able to distinguish previously-heard congruent and new syllables from one another, such that previously encountered syllables were more often rated as old than new syllables ($V = 5878$, $p < .001$, effect size $r = 0.86$). This indicates the participants did remember the syllables they had heard in general. The fused audio was less often rated as old, on average, than an old congruent syllable ($V = 381.5$, $p < .001$, $r = 0.64$), but were also rated as old more often than new audio ($V = 4634$, $p < .001$, $r = 0.70$). The unfused audio was also rated as "old" more often than the new sounds ($V = 2164$, $p < .001$, $r = 0.34$), but less often than a congruent old sound ($V = 67$, $p < .001$, $r = 0.80$). Of greatest interest, however, the comparison between the unfused and fused audio showed that participants were more likely to rate the fused audio as previously heard compared to the unfused audio ($V = 3212.5$, $p < .001$, $r = 0.43$).

To help differentiate between hypotheses one and three from Figure 20, it is also important to investigate whether fused audio appear to be constructed from unfused audio. If this is the case, it should be expected that remembering a fused audio will take longer than remembering an unfused audio, as the multisensory percept would need to be reconstructed from retrieved unfused audio. To investigate this, we examined response times for participant familiarity ratings (Fig. 22b). An ANOVA on response times indicates that there are significant differences between the conditions ($\chi^2(3) = 44.27$, $p < .001$, Kendall's $W = 0.12$). However, the response time for the fused and unfused audio did not differ significantly, indicating that there was no significant difference in the time it took to make a recognition judgement about an multisensory representation relative to a unisensory representation ($V = 4040.5$, $p = 0.28$, $r = 0.10$). All other comparisons are significant, $p = .001$ or smaller.

These findings represent an overall trend for an average across subjects, but leave unexplored whether this pattern of results is present in all participants. To investigate if this pattern of results generalized across all participants, we looked at the relative recognition ratings of the fused and unfused audio for each participant (Figure 23). Of the 120 participants included in the analysis, 65 gave a higher recognition score to fused audio than to unfused audio. Thirty of the remaining participants provided equal recognition scores to the multisensory representation and the unfused audio. Only 25 of our participants reported higher recognition scores for the unfused audio than the multisensory representation. Thus, we do observe individual variability, though the majority of participants seem to have access to fused audio of some form (Fig. 23a).

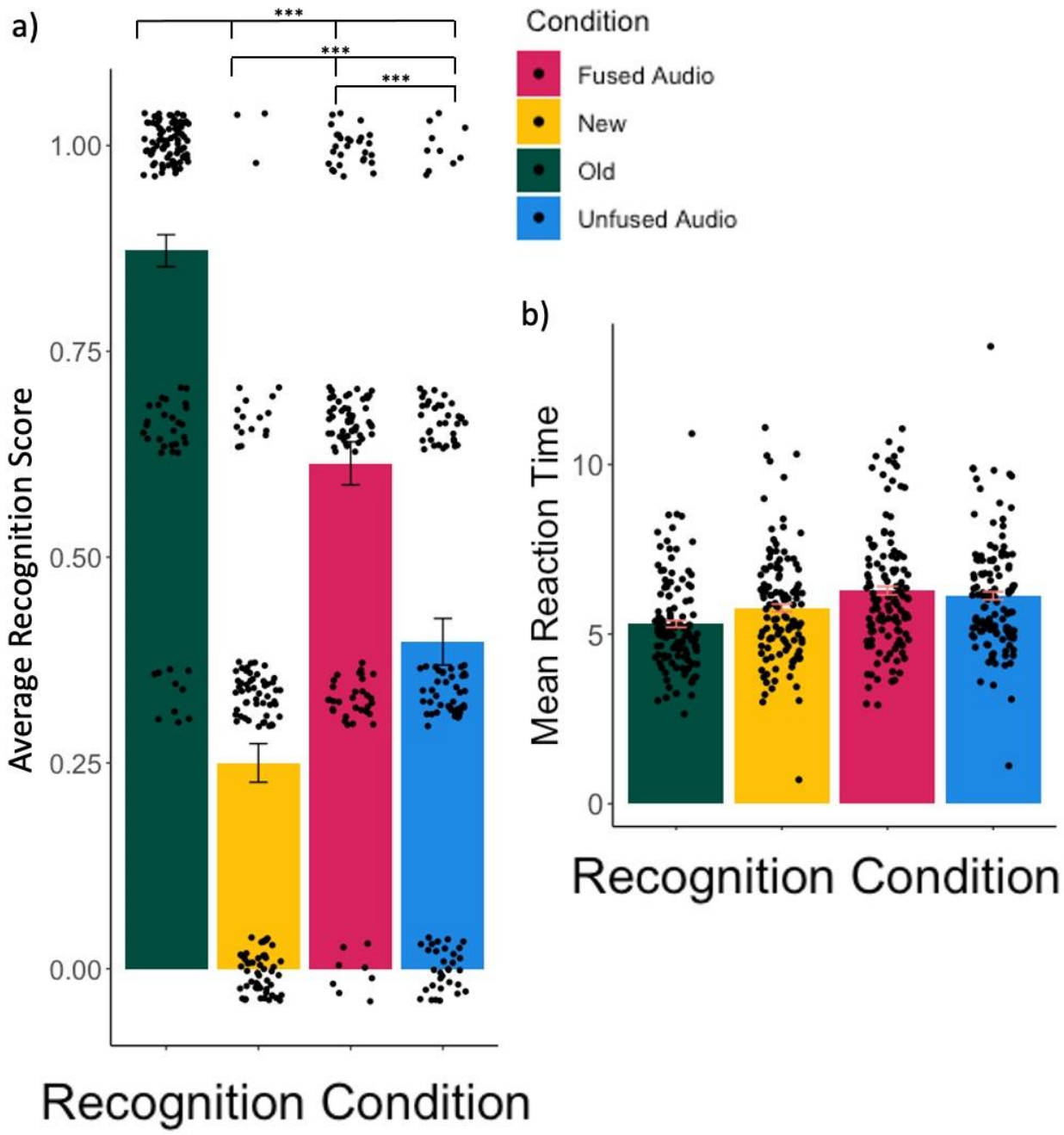


Figure 22: Averaged participant recognition scores for auditory syllables in Study 4, Experiment 1

a) Overall participant averages for recognition scores of old, new, multisensory, and unfused audio. (b) Mean reaction time for recognition judgments for all stimulus types. All error bars represent standard errors.

There is of course some expectation that individual differences in perceiving the McGurk illusion could explain these results. It would be expected that individuals who perceive the illusion would be more likely to rate the fused audio as "old" than those who do not perceive the illusion. To investigate this possibility, we performed a Spearman's correlation looking at the percent of McGurk illusions that were perceived and reported in the perception test with the recognition score difference between fused and unfused audio. This correlation was significant ($r(118)= .18$, $p = .044$; Fig. 23b), indicating that perceiving the McGurk illusion was a predictor of the relative recognition score for the fused audio compared to the unfused audio. Participants who perceived the McGurk illusion more often were more likely to rate the fusion audio as "old" than the unfused audio. However, it should be noted that even in this high-fusion group, there were individuals who provided a higher recognition score to the unfused audio than the fused audio (Fig. 23c), indicating that even among individuals who are likely to fuse, there may not be a single preferred method for encoding stimuli.

Results of the Raven's progressive matrices showed a mean score of 7.55 correct out of 12 questions (95% CI = [7.01, 8.09]), indicating they were generally engaged in the task, and likely not thinking about the illusions during the delay between the memory blocks and the perception test. Descriptive statistics regarding the confidence rating given during the perception test are included in the supplementary materials.

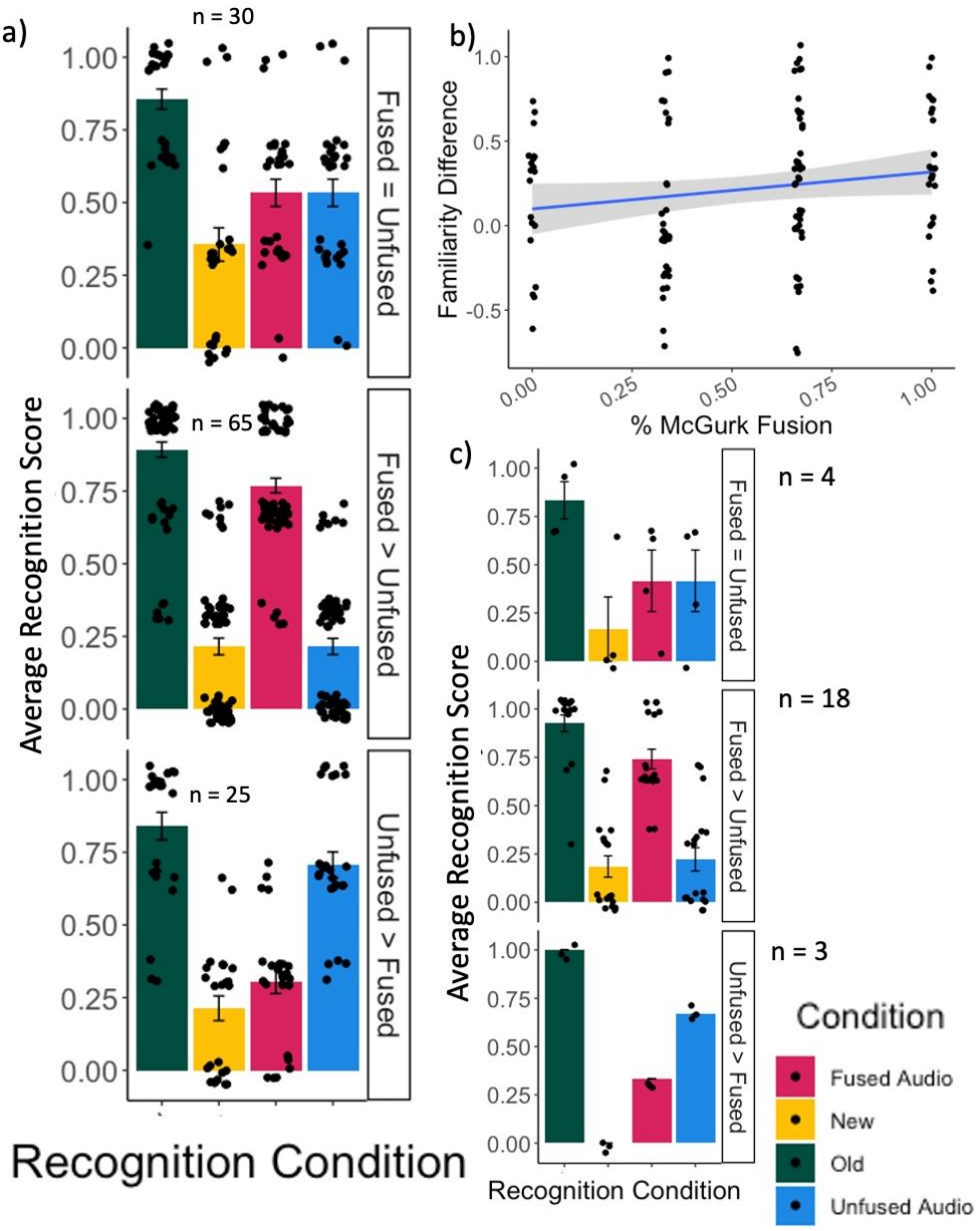


Figure 23: Participant recognition scores, split by relative score for fusion audio vs unfused audio, Study 4, Experiment 1

(a) Participants, split into three groups by relative recognition scores (fused - unfused audio recognition score), and the participant counts within these groups. While most participants rate the fused audio as more familiar than the unfused audio, approximately 20% of participants rate more of the unfused audio as "old" than the fused audio. (b) Performance on the McGurk

perception task, as percentage of McGurk stimuli where the fusion audio was identified as the syllable perceived, is a weak but significant predictor of the recognition score difference between multisensory and unfused audio. (c) Within individuals who fused 100% of the McGurk stimuli, all three patterns of relative recognition score still exist.

Experiment 2

Experiment 2 was conducted to act as a replication of experiment 1 in a separate group of participants. Additionally, this experiment was conducted in a laboratory setting instead of online, for better control over presentation of audiovisual stimuli.

Methods

Participants

Participants in the study were 125 volunteers (83 female, 37 male, 2 nonbinary, and 3 declined to answer) recruited from the University of California, Los Angeles psychology subject pool. The average participant age was 20.38 years. All reported normal or corrected-to-normal sight, normal hearing, and all were fluent English speakers. Informed consent was obtained from each participant and experimental procedures were reviewed and approved by the UCLA Institutional Review Board.

Materials

Videos used in this experiment were identical to those used in Experiment 1. Videos were presented in a size of 800 by 450 pixels against a black background using PsychoPy software.

Procedure

The procedure in Experiment 2 was similar to experiment 1. Participants started in the encoding phase, where 3 videos were shown: one McGurk stimulus and two congruent syllables. Participants were then given a one-minute break where they played a silent game, where they attempted to catch falling objects by moving the mouse. After one minute, participants were tested on their memory for the syllables, and were given four stimuli to remember, as outlined in the procedure for experiment 1. This encoding-break-retrieval procedure was completed 3 times.

After the memory portion of the task was completed, participants went immediately to the perception test, where their perception for McGurk stimuli was tested. As in experiment 1, participants were shown the three McGurk stimuli they had previously seen in a random order, followed by three randomly selected congruent stimuli. Multiple choice options were identical to those in experiment 1, but participants were not given the option to type what they had heard if they selected “other” in this experiment.

Statistical analysis

The analyses conducted on this data matched the methodology used in experiment 1.

Results

A Friedman's test indicated that there were significant differences between recognition scores of the different stimulus types, $\chi^2(3) = 241.65$, $p < .001$, Kendall's $W = 0.64$ (Fig. 24a). As such, post hoc Wilcoxon sign rank tests with a Holm correction were conducted. Participants were again able to distinguish previously heard congruent and new syllables from one another, such that previously encountered syllables were given higher recognition scores than new syllables ($V = 7037.5$, $p < .001$, effect size $r = 0.85$). This indicates the participants did remember the syllables they had heard in general. The fusion audio was given a lower recognition score, on average, than an old congruent syllable ($V = 4431.5$, $p < .001$, $r = 0.74$), but were also more often scored as "old" than new audio ($V = 310.5$, $p < .001$, $r = 0.65$). The unfused audio was also rated as "old" less often than the old sounds ($V = 7550.5$, $p < .001$, $r = 0.86$), but were not rated as significantly different from new sounds ($V = 1799$, $p = 0.07$, $r = 0.17$). The comparison between the unisensory and fusion audio showed that participants were more likely to rate the multisensory stimulus as "old" compared to the unisensory stimulus ($V = 4944$, $p < .001$, $r = 0.73$).

A Friedman test on average response times was also conducted, and showed a trend towards significance ($\chi^2(3) = 7.65$, $p = .053$, Kendall's $W = 0.03$; Fig. 24b). Post-hoc Wilcoxon signed rank tests indicated that RTs did not differ significantly between any of the individual conditions ($p \geq .07$ for all comparisons).

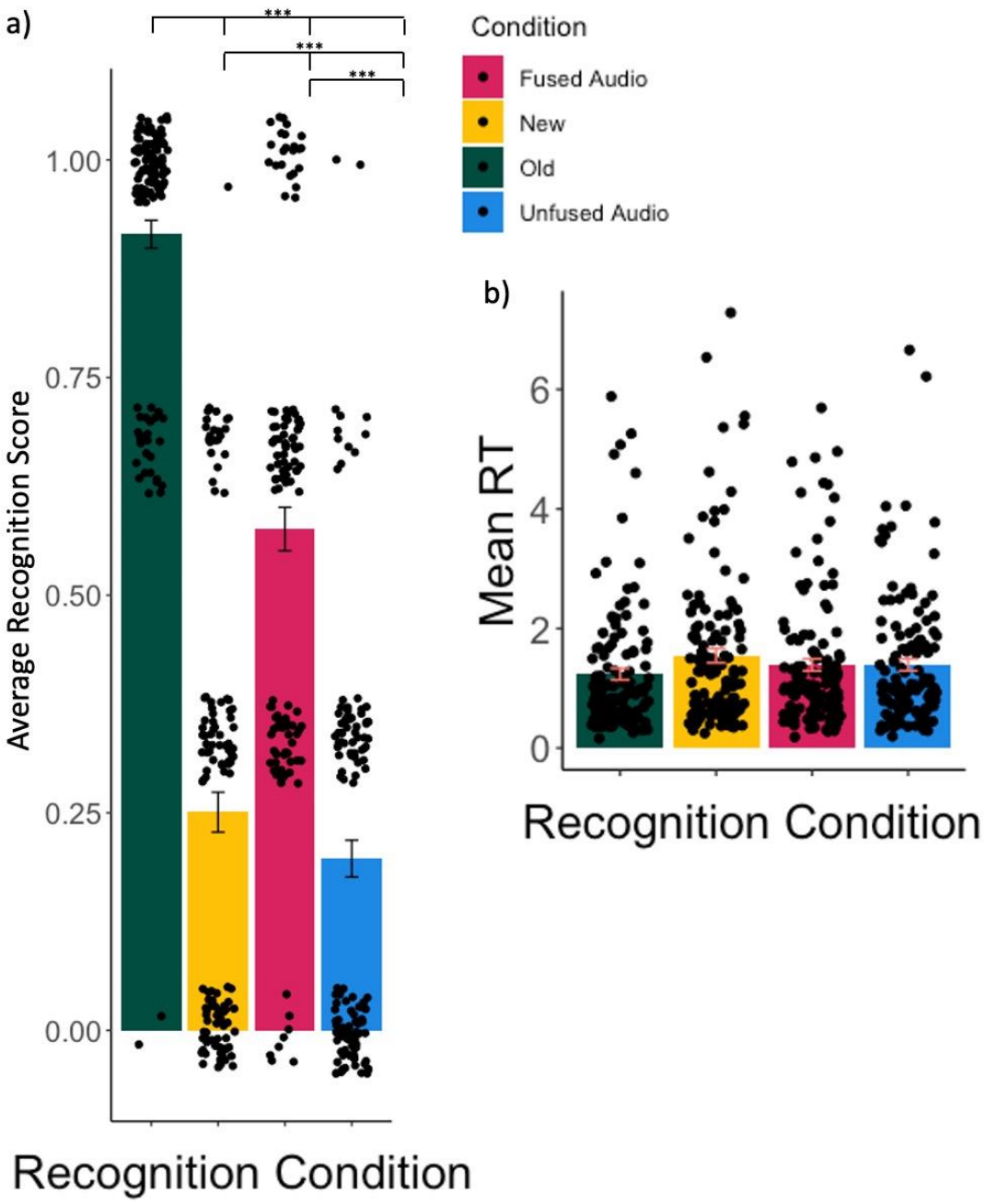


Figure 24: Averaged participant recognition scores for auditory syllables in Study 4, Experiment 2

(a) Overall participant averages for recognition scores of old, new, multisensory, and unfused audio. (b) Mean reaction time for recognition judgments for all stimulus types. All error bars represent standard errors.

To investigate if this pattern of results generalized across all participants, we looked at the relative recognition rating of the fused and unfused audio for each participant (Fig. 25a). Of the 125 participants included in the analysis, 92 rated that the fused audio with a higher recognition score than the unfused audio. Twenty-three rated the multisensory representation and the unfused audio equally, and 10 participants provided a higher recognition score for the unfused audio than the fused audio. However, it is also worth noting that, in this experiment, participants who rated the multisensory and unfused audio as equally familiar, on the whole, rated these items as very unfamiliar, indicating that they did not particularly remember either item.

We additionally looked at how performance in the fusion test could predict the difference in fused and unfused audio recognition scores. A Spearman's correlation indicated that there was no significant correlation between the proportion of McGurk stimuli fused during the perception task and the difference between recognition scores for fused and unfused audio ($r(123) = 0.04$, $p = .66$; Fig. 25b). As in experiment 1, however, we see all patterns of response even among individuals who perceived all McGurk stimuli as fusions (Fig. 25c), indicating that perceptual fusion of the McGurk stimulus does not predict memory patterns for all individuals.

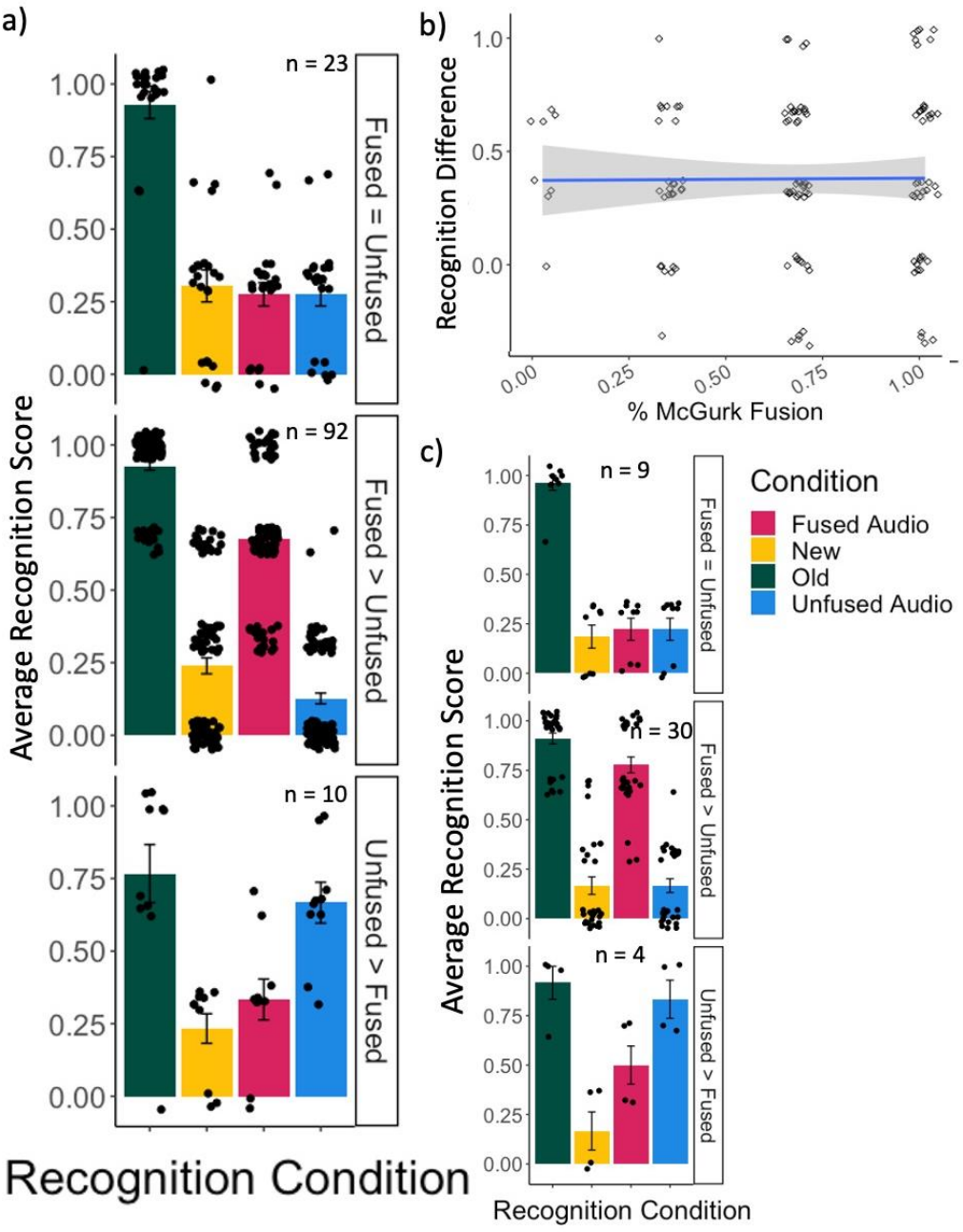


Figure 25: Participant recognition scores, split by relative score for fusion audio vs unfused audio, Study 4, Experiment 2

(a) Participants, split into three groups by relative recognition, and the participant counts within these groups. Most participants provide a higher recognition score to the fused audio than the unfused audio, with only 10 of participants providing the unfused audio with a higher

recognition score. (b) Performance on the McGurk perception task, as percentage of McGurk stimuli where the fusion audio was identified as the syllable perceived, is a nonsignificant predictor of the recognition score difference (fused - unfused audio recognition scores) between fused and unfused audio. (c) Within individuals who fused 100% of the McGurk stimuli, all three patterns of fused vs unfused recognition scores exist.

Discussion

We utilized a novel paradigm, looking at memory for illusory stimuli to differentiate unisensory and multisensory representations in memory, to investigate what type of sensory information is stored in memory. We found, across two experiments, that the majority of participants rated more of the fused, multisensory audio as "old" than unfused, unisensory representations. Participants did not take additional time to create these multisensory representations compared to unisensory representations. This indicates that participants do have access to multisensory representations, and that these multisensory representations are distinct from unisensory representations. The percentage of McGurk stimuli participants perceived did not reliably predict how much participants rated fused audio as more familiar than unisensory representations– this was marginally significant in experiment one and non-significant in experiment two. While non-intuitive, this indicates that low-level sensory processing does not predict memory outcomes, indicating some independence between sensory and higher cognitive processes. Further research, with more trials and a wider variety of McGurk stimuli could help to flesh out to what extent this would be true.

These findings are consistent with the idea that the average participant does have access to multisensory representations during retrieval, and that sensory representations are combined prior to encoding in memory. This would support an encoding scheme where the majority of participants store multisensory representations for later retrieval. However, not all participants showed this trend in their memory. The vast majority of participants (210 of the 245 enrolled) did show equal or higher recognition scores for multisensory representations compared to unisensory representations. However, a non-negligible proportion of our sample showed higher recognition scores for unisensory representations. As such, it appears that some combination must exist of multisensory and unisensory representations, which participants are able to use in individualized ways. As perception is not itself a reliable predictor of the familiarity difference, this would indicate that ability to fuse stimuli is not itself a strong predictor and leaves open for future research what might be a predictor of what representation is most familiar to participants.

One possibility is that there are distinct memory "styles," wherein individuals prefer to rely on unisensory or multisensory representations to different extents when encoding and retrieving memories. This style may influence or reflect which representation you have easiest access to, and thus influence what is encoded and remembered about life events. The investigation of such styles would require further research with more rigorous measures of multisensory binding; the current study has relatively few trials per subject, so while results qualitatively suggest such styles could exist, the stochasticity of memory and perceptual processes make it difficult to state to what extent and how strong such styles would be. However, if such styles do exist, they would provide another means of understanding individual

differences in memory and would supply a new method for evaluating how to most effectively encode information we want to remember in our everyday lives.

On the average, however, participants do appear to have access to multisensory representations that are separate from unisensory representations of life events, and are able to pull these from memory. This implies that the organization and interactions at the level of the sensory regions may cascade up to have significant effects in memory encoding. Such a relationship between perceptual processing and higher-level memory outcomes could be utilized to benefit human memory performance, as it provides a unique tool that can be used to improve human memory in day-to-day experiences.

Chapter 6: Discussion and Future Directions

Across four experiments, we have investigated open questions that connected multisensory stimulus presentation to ease of retrieval in human memory. The first study investigated analytical means that could be used to identify multisensory benefits considering speed-accuracy tradeoffs. We investigated the use of hierarchical drift diffusion models and associated measures that assess speed-accuracy tradeoff both detection and discrimination tasks and found that drift rate—the rate of evidence accumulation in the drift process— was a reliable and sensitive measure for assessing multisensory benefit. In light of this, where the trial count is sufficiently high to allow such modeling, we recommend using these models in assessing multisensory benefit. In cases where this is computationally challenging, due to low trial counts for a given stimulus, we have investigates a less computationally intensive measure of accuracy controlled for response time, guided by the conceptualization of inverse efficiency scores (Rach et al., 2011; Townsend & Ashby, 1983). In applying this model to our recognition memory for foreign vocabulary and their translations in Study 3, we found that participants did experience a speed-accuracy tradeoff in some conditions that altered the results in recognition memory performance. As such, this analytical approach of assessing multisensory performance while controlling for any speed-accuracy tradeoff in performance appears to generalize broadly to multisensory tasks, making it a valuable addition to the tools available to multisensory researchers.

In addition to these more methodological questions, we also sought to investigate empirically if multisensory memory benefit existed. In Study 2, we extended previous experiments, that looked primarily at recognition (though see Duarte et al., 2022) for objects, to investigate potential memory benefits in cued recall of face-name associations. Across five experiments, we found that participants had higher recall accuracy for face-name pairs when given an audiovisual nametag cue as opposed to when the name was presented as audio only.

Crucially, the final experiment in that group showed that recall was higher specifically when there was temporal co-occurrence of the audio with the nametag, as opposed to if these were presented with a delay between them. This suggests that the multisensory presentation, specifically, was helpful for the observed improvement in recall for this condition. This suggests that multisensory integration does provide unique benefits to memory, beyond what the presence of multiple sensory cues without integration can provide. This supports previous multisensory findings for memory benefit likely reflect a real benefit,

However, this finding leaves open the neural mechanisms explaining such a benefit. While it has been suggested that the increased recruitment of neural populations that occurs with multisensory processing, through cross-modal activation of sensory-specific cortices and the activation of multisensory-specific neurons (Shams & Seitz, 2008), this has not been empirically shown in the case of multisensory memory. Indeed, the neural structures that support memory benefits are unclear. Encoding information for memory is often ascribed to the hippocampus, and this structure is suggested to be crucial for associative binding of related objects (Shohamy & Turk-Browne, 2013), which could support the benefit stemming from sensory associations. However, regions earlier in the processing stream for stimuli have also been noted to show different activity for items encoded in multisensory conditions compared to those encoded as unisensory objects (M. M. Murray et al., 2004; Thelen & Murray, 2013).

Regions as early in the processing steam as the auditory cortex have been suggested to be capable of storing information crucial for memory traces (Weinberger, 2004). As perceptual processes in sensory-specific regions are able to influence processing in other sensory specific regions (e.g. Kayser et al., 2008; Watkins et al., 2006), and this crossmodal co-activation can re-occur during memory retrieval with a unisensory probe (e.g. Nyberg et al., 2000; Wheeler et al., 2000), it is distinctly possible that multisensory benefit emerges earlier in the neural pathways forming memory traces than the hippocampus. Hypothesized neural models in perceptual learning with multisensory stimuli highlight that changes in multisensory regions or

the connections between unisensory regions, modulated by multisensory activity, best explain observed benefits from multisensory learning (Mathias & von Kriegstein, 2023). This approach highlights that several levels of neural processing may be important for receiving multisensory benefit, and it is possible that this is also true for observed multisensory memory benefits. However, more rigorous study using neuroimaging techniques will be required to better explain how multisensory encoding can lead to improved retrieval performance.

Study 3 investigated multisensory stimulus presentation in another challenging associative task: learning vocabulary in a foreign language. In this study, across four experiments, we showed that providing verbal information with pictorial information was helpful, as would be expected from existing dual-processing and multimedia learning frameworks (e.g., Clark & Paivio, 1991; Mayer, 2014). However, we did not see a benefit from multisensory stimulus presentation—providing audio in addition to text, or images in addition to audio did not produce higher learning that using text alone, or than adding text to images. Thus, in this type of memory task, it appears multisensory encoding is not helpful for improving recall or recognition performance. This provides an interesting boundary condition, wherein we must explore the differences between study 2 and study 3 that can explain the seeming discrepancy in these results.

Foreign language learning, in the framework of multimedia learning, is a somewhat unusual case. While much learning does benefit from providing information across different modalities, to avoid overwhelming processing resources in visual or auditory streams—a guideline termed the modality principle (Mayer, 2014)—much research looking at learning foreign vocabulary finds that memory for new words is most accurate when participants are given images, audio, and text, or text with images (Zhang & Zou, 2022). As such, foreign language learning is itself an interesting case where the usual expectations for best learning are challenged. Many hypotheses exist to explain this, and we will discuss some that are especially relevant for discovering the differences between the findings of studies 2 and 3.

First, the differences in the design of these studies should be mentioned, which could be expected to produce different results. In existing multisensory memory findings, there is a noted asymmetry in when participants receive benefits for audiovisual stimulus presentation, such that auditory memory benefits more from the addition of a visual than visual memory benefits from the addition of auditory information (Heikkilä et al., 2015; Heikkilä & Tiippana, 2016; Pecher & Zeelenberg, 2022). In multisensory processing, the strongest benefits to using crossmodal stimulus presentation are often observed when unisensory processing would be particularly noisy or ineffective on its own, a principle known as *inverse effectiveness* (e.g. Senkowski et al., 2011; Stein & Meredith, 1993). Auditory memory is noted for being worse than visual memory, in the case of recognition (Cohen et al., 2009; Gloede et al., 2017), which could mean that a principle like inverse effectiveness may explain this asymmetry in memory results. Study 2 utilized an auditory performance baseline, and thus a stronger multisensory benefit may have been possible when text was added. Study 3, by contrast, used text as baseline performance in experiments 1 and 2, and showed that this did not benefit from audio, which could be because memory for text—a visual stimulus—would be higher than for audio (as observed in experiments 3 and 4 of study 3), and thus benefit less from the addition of a different sensory cue. Further research exploring this expansion of the idea of inverse effectiveness to multisensory memory could help to discover if such mechanisms are able to explain this seeming asymmetry of results, or if another factor is at play.

Second, and also a note based in the methodological differences between these experiments, is the difference in how transient auditory and text representation of words are. Text is relatively less transient than auditory representations of word. Presenting text on a screen for the translations in study 3 would leave the new Swahili word on screen for the duration of the translated phrase, where in the auditory representation of the new Swahili word only lasts as long as that word is being said. It has been suggested that text presentation of audio thus allows for a longer exposure to the stimulus, and can allow participants to revisit the

word or phrase, improving their memory for the item (Lee & Mayer, 2018). Study 3 did not attempt to control for this discrepancy in duration across its 4 experiments, and so this is a notable limitation of that design. Study 2, by contrast, did have experiments that attempted to decrease the temporal discrepancy in stimulus duration, by trimming down presentation of names and faces to just include the name in experiment 4. Future experiments looking at multisensory contributions to foreign language learning should seek to control for this discrepancy to better explore if this factor contributes to the observed retrieval benefit for text-based encoding conditions.

Finally, it is also important to consider the different cognitive demands of the tasks used in studies 2 and 3. While both are challenging associative tasks, there could be a distinct difference in how challenging these are. Study 2 looked at face-name associations, but the names were very commonly used and thus relatively familiar to participants. Study 3 used common English words, but the Swahili translations were unfamiliar to almost every participant. This likely added an additional cognitive challenge to this task—not only did participants have to build an association between two words, but they also had to memorize a completely new word. This could factor into the results of study 3 in a few ways. It is suggested that reading text may be less cognitively demanding than listening to audio in foreign language contexts, which would mean that text-based learning decreases cognitive load, making learning easier in this context (Zhang & Zou, 2022). Such a suggestion is supported by empirical findings that high comprehension in a foreign language and high working memory capacity correspond to reduced reliance on captions during learning in one's second language (Zhang & Zou, 2022). It is also possible that participants felt less sure of their crossmodal associations between text and sounds in a foreign language, where phoneme-to-grapheme mappings may differ from in their primary language. Multisensory stimulus presentation is suggested to guide attention effectively to crossmodal events, but this is suggested to be most effective when attentional demands are low (Talsma et al., 2010). The attentional demands of this task are relatively high, and

participants may have a sense that their intersensory reliability is not as high as it would be in their native language. This may lead to attention being divided between modalities in the multisensory condition, as opposed to being guided to important features of the stimulus being presented.

Both of these options provide insight into cases where multisensory perception is helpful for memory and learning outcomes. All together, these differences could suggest that principles important for multisensory learning may also apply to receiving benefits from multisensory encoding. Semantic congruence and spatiotemporal congruence of audiovisual stimuli, which are crucial for engaging neural mechanisms of sensory integration and guiding attention to multisensory events, appear to play a role in receiving multisensory benefit from learning. Principle such as inverse effectiveness may play a role in explaining the strength of a multisensory benefit in memory paradigms. This presents an exciting possibility that the neural mechanisms underpinning memory benefits may resemble those that explain multisensory learning benefit, though to our knowledge this has not been rigorously tested through empirical of computational means. Further research investigating this through rigorous computational modeling could greatly help to elucidate the extent to which these neural mechanisms overlap. This could also help to predict what kinds of tasks would show a benefit in retrieval when given multisensory stimulus presentation.

In our final study, we additionally investigated if multisensory representations are stored in memory. In two separate studies, it was shown that participants rate multisensory representations of their experiences as more familiar than unisensory representations. Additionally, they do not need extra time to build these representations relative to unisensory representations, indicating that they do store the multisensory representation and are not reconstructing the multisensory experience from unisensory memory traces. Interestingly, one's likelihood to fuse stimuli was not a strong predictor of how they rated the relative familiarity of the unisensory and multisensory representations of events, indicating that likelihood of

integration does not necessarily predict the pattern of memory results. This is interesting in combination with the results that a subset of our participants remembers the unisensory representation as more familiar. This suggests that, while most of our participants do store multisensory representations, this is not the case for all individuals. As this is not predicted strongly by perceptual processes, this may indicate the existence of a memory "styles," wherein participants, regardless of their likelihood to integrate information, may store this information differently. This raises the question of what, if not perceptual integration, may explain the existence of such styles. Future studies should investigate possible predictors of this difference in order to explain these styles. If the difference is not at the level of integration, it is possible this emerges based on differences in cognitive processing between these low- and higher-level cognitive processes, though that leaves many candidate steps. Understanding more aspects of cognitive processing, and perhaps investigating with a more rigorous measure of integration than fusion in the McGurk illusion, could help to explore the existence of these styles.

Overall, the series of studies introduced here, investigating the role of multisensory stimulus presentation on memory retrieval outcomes, suggest that, in many cases, multisensory representations are helpful for supporting memory retrieval. It also suggests that these multisensory representations are not just, in memory, the sum of two different sensory inputs; rather the integration process produces a representation that is uniquely helpful in retrieval. We also observe interesting border cases, such that under high cognitive load, and when inverse effectiveness does not apply, a multisensory benefit will not be observed. Together, this suggests that, if used in the correct situations, providing multisensory cues could provide an easy-to-implement means of benefitting human memory performance. This could lead to the development of audiovisual tools that could improve human memory performance, that would be easier to implement than mnemonics or other cognitively demanding methods. Such tools could be used broadly in everyday life and education. Additionally, this could provide a way to bolster memory performance in groups that struggle with memory. Visual-somatosensory

integration has been shown to predict cognitive decline (Mahoney & Verghese, 2020), and older adults also can show greater response time benefit from visuo-tactile stimulus presentation than younger adults (Mahoney et al., 2011). This could imply that older adults would likewise benefit from audiovisual integration, and thus this could prove a useful tool to improve memory with age. Additionally, while tactile/somatosensory stimulation requies specialized equipment to implement, audiovisual stimulus presentation can be done by any number of modern devices, making it a potentially more accessible route to improve memory. Further research would be necessary to investigate to what extent this would be helpful for older adults—particularly given that sensory integration processes operate differently in older adults than in younger adults (e.g. Hirst et al., 2019; McGovern et al., 2014)—but this could provide an easy-to-implement means of improving memory in older adults. Overall, this opens up opportunities for future research and development of tools using multisensory principles to improve memory and, thus, the quality of everyday life.

References

Al-Seghayer, K. (2001). “The Effect of Multimedia Annotation Modes on L2 Vocabulary Acquisition: A Comparative Study.” *Language Learning & Technology, 5*(1): 202–232.

Alsius, A., Paré, M., & Munhall, K. G. (2018). Forty Years After Hearing Lips and Seeing Voices: The McGurk Effect Revisited. *Multisensory Research, 31*(1–2), 111–144. https://doi.org/10.1163/22134808-00002565

Arieh, Y., & Marks, L. E. (2008). Cross-modal interaction between vision and hearing: A speed-accuracy analysis. *Perception & Psychophysics*, *70*(3), 412–421. https://doi.org/10.3758/PP.70.3.412

Arlin, M., Scott, M., & Webster, J. (1978). The Effects of Pictures on Rate of Learning Sight Words: A Critique of the Focal Attention Hypothesis. *Reading Research Quarterly, 14*(4), 645-660. doi:10.2307/747266

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences, 4*(11), 417–423. https://doi.org/10.1016/S1364-6613(00)01538-2

Baddeley, A. D., & Hitch, G. (1974). Working Memory. In G. H. Bower (Ed.), *Psychology of Learning and Motivation* (Vol. 8, pp. 47–89). Academic Press. https://doi.org/10.1016/S0079-7421(08)60452-1

Balch, W. R., Bowman, K., & Mohler, L. A. (1992). Music-dependent memory in immediate and delayed word recall. *Memory & Cognition, 20*(1), 21–28. https://doi.org/10.3758/BF03208250

Bernstein, L. E., Auer, E. T., & Takayanagi, S. (2004). Auditory speech detection in noise enhanced by lipreading. *Speech Communication, 44*(1), 5–18. https://doi.org/10.1016/j.specom.2004.10.011

Boers, F., Warren, P., He, L., & Deconinck, J. (2017). Does adding pictures to glosses enhance vocabulary uptake from reading? *System*, 66, 113–129. https://doi.org/10.1016/j.system.2017.03.017

Botvinick, M., & Cohen, J. (1998). Rubber hands 'feel' touch that eyes see. *Nature, 391*, 756. https://doi.org/10.1038/35784

Bower, G. H., & Karlin, M. B. (1974). Depth of processing pictures of faces and recognition memory. Journal of Experimental Psychology, 103(4), 751–757. https://doi.org/10.1037/h0037190

Boywitt, C. D., & Rummel, J. (2012). A diffusion model analysis of task interference effects in prospective memory. *Memory & Cognition*, *40*, 70-82. https://doi.org/10.3758/s13421-011-0128-6

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*(38), 14325–14329. https://doi.org/10.1073/pnas.0803390105

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

Brandwein, A. B., Foxe, J. J., Russo, N. N., Altschuler, T. S., Gomes, H., & Molholm, S. (2011). The Development of Audiovisual Multisensory Integration Across Childhood and Early Adolescence: A High-Density Electrical Mapping Study. *Cerebral Cortex, 21*(5), 1042–1055. https://doi.org/10.1093/cercor/bhq170

Brebner, J. T., & Welford, A. T. (1980). Introduction: An historical background sketch. In A. T. Welford (Ed.), *Reaction Times*. Academic Press, New York, pp. 1-23.

Brédart, S. (2019). Strategies to improve name learning: A review. *European Psychologist, 24*(4), 349–358. https://doi.org/10.1027/1016-9040/a000363

Brooks, J. O., Friedman, L., Gibson, J. M., & Yesavage, J. A. (1993). Spontaneous mnemonic strategies used by older and younger adults to remember proper names. *Memory, 1*(4), 393–407. https://doi.org/10.1080/09658219308258245

Brown, V. A., Hedayati, M., Zanger, A., Mayn, S., Ray, L., Dillman-Hasso, N., & Strand, J. F. (2018). What accounts for individual differences in susceptibility to the McGurk effect? *PLOS ONE, 13*(11), e0207160. https://doi.org/10.1371/journal.pone.0207160

Bruns, P. (2019). The ventriloquist illusion as a tool to study multisensory processing: An update. *Frontiers in Integrative Neuroscience*, *13*, 51. https://doi.org/10.3389/fnint.2019.00051

Burr, D., & Gori, M. (2012). Multisensory integration develops late in humans. In: *The Neural Bases of Multisensory Processes.* CRC Press/Taylor & Francis, Boca Raton (FL).

Bushara, K. O., Weeks, R. A., Ishii, K., Catalan, M. J., Tian, B., Rauschecker, J. P., & Hallett, M. (1999). Modality-specific frontal and parietal areas for auditory and visual spatial localization in humans. *Nature Neuroscience, 2*(8), 759-766.

Butler, J. S., Foxe, J. J., Fiebelkorn, I. C., Mercier, M. R., & Molholm, S. (2012). Multisensory Representation of Frequency across Audition and Touch: High Density Electrical Mapping Reveals Early Sensory-Perceptual Coupling. *Journal of Neuroscience, 32*(44), 15338–15344. https://doi.org/10.1523/JNEUROSCI.1796-12.2012

Calford, M. B. (2002). Dynamic representational plasticity in sensory cortex. *Neuroscience*, *111*(4), 709-738. https://doi.org/10.1016/S0306-4522(02)00022-2

Calvert, G., Spence, C., & Stein, B. E. (Eds.). (2004). *The Handbook of Multisensory Processes*. Cambridge, MA: MIT press.

Carpenter, S. K., & DeLosh, E. L. (2005). Application of the testing and spacing effects to name learning. *Applied Cognitive Psychology, 19*(5), 619–636. https://doi.org/10.1002/acp.1101

Carpenter, S. K., & Olson, K. M. (2012). Are pictures good for learning new vocabulary in a foreign language? Only if you think they are not. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(1), 92–101. https://doi.org/10.1037/a0024828

Chertkow, H., Whitehead, V., Phillips, N., Wolfson, C., Atherton, J., & Bergman, H. (2010). Multilingualism (but not always bilingualism) delays the onset of Alzheimer disease: evidence from a bilingual community. *Alzheimer disease and associated disorders*, *24*(2), 118–125. https://doi.org/10.1097/WAD.0b013e3181ca1221

Ciccarone, S. (2019). The effects of multisensory learning on second language acquisition for students with learning disabilities (Master's thesis). Retrieved from Goucher Special Collections & Archives. (11603/13842).

Clark, J. M., & Paivio, A. (1991). Dual coding theory and education. *Educational Psychology Review, 3*(3), 149–210. https://doi.org/10.1007/BF01320076

Cohen, G., & Faulkner, D. (1986). Memory for proper names: Age differences in retrieval. British *Journal of Developmental Psychology, 4*(2), 187–197. https://doi.org/10.1111/j.2044-835X.1986.tb01010.x

Cohen, J. (1988). The t Test for Means. In *Statistical Power Analysis for the Behavioral Sciences* (2nd ed., pp. 19–74). L. Erlbaum Associates.

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

Colonius, H., & Diederich, A. (2017). Measuring multisensory integration: from reaction times to spike counts. *Scientific reports*, *7*(1), 3023. https://doi.org/10.1038/s41598-017-03219-5

Costa, A., & Sebastián-Gallés, N. (2014). How does the bilingual experience sculpt the brain?. Nature reviews. *Neuroscience, 15*(5), 336–345. https://doi.org/10.1038/nrn3709

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*(6), 671–684. https://doi.org/10.1016/S0022-5371(72)80001-X

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 104*(3), 268–294. https://doi.org/10.1037/0096-3445.104.3.268

Cuppini, C., Ursino, M., Magosso, E., Ross, L. A., Foxe, J. J., & Molholm, S. (2017). A computational analysis of neural mechanisms underlying the maturation of multisensory

speech integration in neurotypical children and those on the autism spectrum. *Frontiers in Human Neuroscience*, *11*, 518. https://doi.org/10.3389/fnhum.2017.00518

Deno, S. L. (1968). Effects of words and pictures as stimuli in learning language equivalents. *Journal of Educational Psychology, 59*(3), 202–206. https://doi.org/10.1037/h0025772

Diederich, A. (1995). Intersensory facilitation of reaction time: Evaluation of counter and diffusion coactivation models. *Journal of Mathematical Psychology, 39*(2), 197–215. https://doi.org/10.1006/jmps.1995.1020

Diederich, A. (2008). A further test of sequential-sampling models that account for payoff effects on response bias in perceptual decision tasks. *Perception & Psychophysics, 70*(2), 229–256. https://doi.org/10.3758/PP.70.2.229

Diederich, A., & Busemeyer, J. R. (2003). Simple matrix methods for analyzing diffusion models of choice probability, choice response time, and simple response time. Journal of *Mathematical Psychology, 47*(3), 304-322. https://doi.org/10.1016/S0022-2496(03)00003-8

Diederich, A., & Colonius, H. (2009). Crossmodal interaction in speeded responses: Time window of integration model. In *Progress in Brain Research* (Vol. 174, pp. 119–135). Elsevier. https://doi.org/10.1016/S0079-6123(09)01311-9

Driver, J., & Noesselt, T. (2008). Multisensory interplay reveals crossmodal influences on 'sensory-specific' brain regions, neural responses, and judgments. *Neuron*, *57*(1), 11–23. https://doi.org/10.1016/j.neuron.2007.12.013

Driver, J., & Spence, C. (2000). Multisensory perception: Beyond modularity and convergence. *Current Biology, 10*(20), 10–12. https://doi.org/10.1016/S0960-9822(00)00740-5

Drugowitsch, J., DeAngelis, G. C., Angelaki, D. E., & Pouget, A. (2015). Tuning the speed-accuracy trade-off to maximize reward rate in multisensory decision-making. *Elife, 4*, e06678.

Drugowitsch, J., DeAngelis, G. C., Klier, E. M., Angelaki, D. E., & Pouget, A. (2014). Optimal multisensory decision-making in a reaction-time task. *Elife, 3*, e03005.

Duarte, S. E., Ghetti, S., & Geng, J. J. (2022). Object memory is multisensory: Task-irrelevant sounds improve recollection. *Psychonomic Bulletin & Review, 30*, 652–665. https://doi.org/10.3758/s13423-022-02182-1

Eich, J. E. (1980). The cue-dependent nature of state-dependent retrieval. *Memory & Cognition, 8*(2), 157–173. https://doi.org/10.3758/BF03213419

Ernst, M. O., & Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Sciences, 8*(4), 162–169. https://doi.org/10.1016/j.tics.2004.02.002

Ferraro, S., Van Ackeren, M. J., Mai, R., Tassi, L., Cardinale, F., Nigri, A., Bruzzone, M. G., D'Incerti, L., Hartmann, T., Weisz, N., & Collignon, O. (2020). Stereotactic electroencephalography in humans reveals multisensory signal in early visual and auditory cortices. *Cortex*, *126*, 253-264. https://doi.org/10.1016/j.cortex.2019.12.032

Fiser, J., & Lengyel, G. (2022). Statistical learning in vision. *Annual Review of Vision Science*, *8*, 265-290. https://doi.org/10.1146/annurev-vision-100720-103343

Fitts, P. M. (1966). Cognitive aspects of information processing: III. Set for speed versus accuracy. *Journal of Experimental Psychology, 71*(6), 849-857.

Gándara, P. (2018). The economic value of bilingualism in the United States. *Bilingual Research Journal, 41*(4), 334-343. https://doi.org/10.1080/15235882.2018.1532469

Gass, S., Winke, P., Isbell, D. R., & Ahn, J. (2019). How captions help people learn languages: A working-memory, eye-tracking study. *Language Learning & Technology, 23*(2), 84–104. https://doi.org/10125/44684

Gau, R., Bazin, P. L., Trampel, R., Turner, R., & Noppeney, U. (2020). Resolving multisensory and attentional influences across cortical depth in sensory cortices. *Elife*, *9*, e46856. https://doi.org/10.7554/eLife.46856

Gelman, A., & Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science, 7*(4). https://doi.org/10.1214/ss/1177011136

Ghazanfar, A. A., & Schroeder, C. E. (2006). Is neocortex essentially multisensory?. *Trends in cognitive sciences*, *10*(6), 278-285. https://doi.org/10.1016/j.tics.2006.04.008

Gillund, G. (2012). Episodic Memory. *Encyclopedia of Human Behavior*, 68-72. https://doi.org/10.1016/b978-0-12-375000-6.00152-x

Gingras, G., Rowland, B. A., & Stein, B. E. (2009). The Differing Impact of Multisensory and Unisensory Integration on Behavior. *The Journal of Neuroscience, 29*(15), 4897-4902. https://doi.org/10.1523/jneurosci.4120-08.2009

Gloede, M. E., & Gregg, M. K. (2019). The fidelity of visual and auditory memory. *Psychonomic Bulletin & Review, 26*(4), 1325-1332. https://doi.org/10.3758/s13423-019-01597-7

Gloede, M. E., Paulauskas, E. E., & Gregg, M. K. (2017). Experience and information loss in auditory and visual memory. *Quarterly Journal of Experimental Psychology, 70*(7), 1344-1352. https://doi.org/10.1080/17470218.2016.1183686

Godden, D. R., & Baddeley, A. D. (1975). Context-dependent memory in two natural environments: on land and underwater. *British Journal of Psychology, 66*(3), 325–331. https://doi.org/10.1111/j.2044-8295.1975.tb01468.x

Gold, J. I., & Shadlen, M. N. (2007). The Neural Basis of Decision Making. *Annual Review of Neuroscience, 30*(1), 535-574. https://doi.org/10.1146/annurev.neuro.29.051605.113038

Goldstone, R. L. (1998). Perceptual learning. *Annual review of psychology*, *49*(1), 585-612. https://doi.org/10.1146/annurev.psych.49.1.585

Gomez, P., Ratcliff, R., & Perea, M. (2007). A model of the go/no-go task. *Journal of Experimental Psychology: General, 136*(3), 389-413. https://doi.org/10.1037/0096-3445.136.3.389

Gondan, M. (2010). A permutation test for the race model inequality. *Behavior Research Methods, 42*(1), 23-28. https://doi.org/10.3758/brm.42.1.23

Gondan, M., & Minakata, K. (2016). A tutorial on testing the race model inequality. *Attention, Perception, & Psychophysics, 78*(3), 723-735. https://doi.org/10.3758/s13414-015-1018-y

Groh, J. M. (2014). *Making space: How the brain knows where things are.* Harvard University Press.

Hebb, D. O. (1949). *The organization of behavior: A neuropsychological theory*. Wiley.

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

Heikkilä, J., & Tiippana, K. (2016). School-aged children can benefit from audiovisual semantic congruency during memory encoding. *Experimental Brain Research, 234*(5), 1199-1207. https://doi.org/10.1007/s00221-015-4341-6

Herz, R. S. (1997). The effects of cue distinctiveness on odor-based context-dependent memory. *Memory & Cognition, 25*(3), 375-380. https://doi.org/10.3758/bf03211293

Hirst, R. J., McGovern, D. P., Setti, A., Shams, L., & Newell, F. N. (2020). What you see is what you hear: Twenty years of research using the Sound-Induced Flash Illusion. *Neuroscience & Biobehavioral Reviews*, *118*, 759-774. https://doi.org/10.1016/j.neubiorev.2020.09.006

Hirst, R. J., Setti, A., Kenny, R. A., & Newell, F. N. (2019). Age-related sensory decline mediates the Sound-Induced Flash Illusion: Evidence for reliability weighting models of multisensory perception. *Scientific Reports, 9*(1), Article 1. https://doi.org/10.1038/s41598-019-55901-5

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6(*2), 65-70. https://www.jstor.org/stable/4615733

Horn, S. S., Bayen, U. J., & Smith, R. E. (2013). Adult age differences in interference from a prospective-memory task: A diffusion model analysis. *Psychonomic Bulletin & Review*, *20*, 1266-1273. https://doi.org/10.3758/s13423-013-0451-y

Horowitz, L. M., & Prytulak, L. S. (1969). Redintegrative memory. *Psychological Review, 76*(6), 519-531. https://doi.org/10.1037/h0028139

Jones, S. A., Beierholm, U., Meijer, D., & Noppeney, U. (2019). Older adults sacrifice response speed to preserve multisensory integration performance. *Neurobiology of Aging, 84*, 148-157. https://doi.org/10.1016/j.neurobiolaging.2019.08.017

Kausar, G. (2013). Students' perspective of the use of audio visual aids in Pakistan. *International Proceedings of Economics Development and Research*, *68*(3), 11-13.

Kayser, C., Petkov, C. I., & Logothetis, N. K. (2008). Visual Modulation of Neurons in Auditory Cortex. *Cerebral Cortex, 18*(7), 1560-1574. https://doi.org/10.1093/cercor/bhm187

Kim, R. S., Seitz, A. R., & Shams, L. (2008). Benefits of Stimulus Congruency for Multisensory Facilitation of Visual Learning. *PLoS ONE, 3*(1), e1532. https://doi.org/10.1371/journal.pone.0001532

Knudsen, E. I. (1994). Supervised learning in the brain. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, *14*(7), 3985–3997. https://doi.org/10.1523/JNEUROSCI.14-07-03985.1994

Kolb, B., & Whishaw, I. Q. (1998). Brain plasticity and behavior. *Annual Review of Psychology*, *49*, 43–64. https://doi.org/10.1146/annurev.psych.49.1.43

Kornell, N., Castel, A. D., Eich, T. S., & Bjork, R. A. (2010). Spacing as the friend of both memory and induction in young and older adults. *Psychology and Aging, 25*(2), 498-503. https://doi.org/10.1037/a0017807

Kozan, K., Erçetin, G., & Richardson, J. C. (2015). Input modality and working memory: Effects on second language text comprehension in a multimedia learning environment. *System, 55*, 63-73. https://doi.org/10.1016/j.system.2015.09.001

Krajbich, I., & Rangel, A. (2011). Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proceedings of the National Academy of Sciences, 108*(33), 13852-13857. https://doi.org/10.1073/pnas.1101328108

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General, 142*(2), 573-603. https://doi.org/10.1037/a0029146

Lacey, S., Tal, N., Amedi, A., & Sathian, K. (2009). A Putative Model of Multisensory Object Representation. *Brain Topography, 21*(3-4), 269-274. https://doi.org/10.1007/s10548-009-0087-4

Lado, R. (1967). Massive Vocabulary Expansion in a Foreign Language beyond the Basic Course: The Effects of Stimuli, Timing and Order of Presentation. Report prepared for U. S. Department of Health, Education and Welfare, Georgetown University.

Laurienti, P. J., Kraft, R. A., Maldjian, J. A., Burdette, J. H., & Wallace, M. T. (2004). Semantic congruence is a critical factor in multisensory behavioral performance. *Experimental Brain Research, 158*(4), 405-414. https://doi.org/10.1007/s00221-004-1913-2

Lee, H., & Mayer, R. E. (2018). Fostering learning from instructional video in a second language. *Applied Cognitive Psychology, 32*(5), 648-654. https://doi.org/10.1002/acp.3436

Lee, H., Stirnberg, R., Stöcker, T., & Axmacher, N. (2017). Audiovisual integration supports face-name associative memory formation. *Cognitive Neuroscience, 8*(4), 177-192. https://doi.org/10.1080/17588928.2017.1327426

Lehmann, S., & Murray, M. M. (2005). The role of multisensory memories in unisensory object discrimination. Cognitive Brain Research, 24(2), 326–334. https://doi.org/10.1016/j.cogbrainres.2005.02.005

Leite, F. P., & Ratcliff, R. (2010). Modeling reaction time and accuracy of multiple-alternative decisions. *Attention, Perception, & Psychophysics, 72*(1), 246-273. https://doi.org/10.3758/app.72.1.246

Lerche, V., Voss, A., & Nagler, M. (2017). How many trials are required for parameter estimation in diffusion modeling? A comparison of different optimization criteria. *Behavior Research Methods, 49*(2), 513-537. https://doi.org/10.3758/s13428-016-0740-2

Lewkowicz, D. J., Schmuckler, M., & Agrawal, V. (2021). The multisensory cocktail party problem in adults: Perceptual segregation of talking faces on the basis of audiovisual temporal synchrony. *Cognition, 214*, 104743. https://doi.org/10.1016/j.cognition.2021.104743

Logie, R. H., Zucco, G. M., & Baddeley, A. D. (1990). Interference with visual short-term memory. *Acta Psychologica, 75*(1), 55-74. https://doi.org/10.1016/0001-6918(90)90066-o

Lotto, L., & de Groot, A. M. B. (1998). Effects of Learning Method and Word Type on Acquiring Vocabulary in an Unfamiliar Language. *Language Learning, 48*, 31-69. https://doi.org/10.1111/1467-9922.00032

Macaluso, E., George, N., Dolan, R., Spence, C., & Driver, J. (2004). Spatial and temporal factors during processing of audiovisual speech: a PET study. *NeuroImage, 21*(2), 725-732. https://doi.org/10.1016/j.neuroimage.2003.09.049

Macedonia, M., & Repetto, C. (2016). Brief Multisensory Training Enhances Second Language Vocabulary Acquisition in Both High and Low Performers. *International Journal of Learning, Teaching and Educational Research, 15*, 42-53.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection Theory: A User's Guide* (2nd ed.). New Jersey: Lawrence Erlbaum Associates.

Magee, J. C., & Grienberger, C. (2020). Synaptic Plasticity Forms and Functions. Annual Review of Neuroscience, 43(1), 95–117. https://doi.org/10.1146/annurev-neuro-090919-022842

Mahani, M.-A. N., Bausenhart, K. M., Ahmadabadi, M. N., & Ulrich, R. (2019). Multimodal Simon Effect: A Multimodal Extension of the Diffusion Model for Conflict Tasks. *Frontiers in Human Neuroscience, 12*. https://doi.org/10.3389/fnhum.2018.00507

Mahoney, J. R., Li, P. C. C., Oh-Park, M., Verghese, J., & Holtzer, R. (2011). Multisensory integration across the senses in young and old adults. *Brain Research, 1426*, 43-53. https://doi.org/10.1016/j.brainres.2011.09.017

Mahoney, J. R., & Verghese, J. (2020). Does Cognitive Impairment Influence Visual-Somatosensory Integration and Mobility in Older Adults? *The Journals of Gerontology: Series A, 75*(3), 581-588. https://doi.org/10.1093/gerona/glz117

Mathias, B., & von Kriegstein, K. (2023). Enriched learning: behavior, brain, and computation. *Trends in Cognitive Sciences, 27*(1), 81-97. https://doi.org/10.1016/j.tics.2022.10.007

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

Mayer, K. M., Yildiz, I. B., Macedonia, M., & von Kriegstein, K. (2015). Visual and Motor Cortices Differentially Support the Translation of Foreign Language Words. *Current Biology, 25*(4), 530-535. https://doi.org/10.1016/j.cub.2014.11.068

Mayer, R. E. (2014). Cognitive Theory of Multimedia Learning. *The Cambridge Handbook of Multimedia Learning*, 43-71. https://doi.org/10.1017/cbo9781139547369.005

Mayer, R. E., Heiser, J., & Lonn, S. (2001). Cognitive constraints on multimedia learning: When presenting more material results in less understanding. *Journal of Educational Psychology, 93*(1), 187-198. https://doi.org/10.1037/0022-0663.93.1.187

McCarty, D. L. (1980). Investigation of a visual imagery mnemonic device for acquiring face-name associations. *Journal of Experimental Psychology: Human Learning and Memory, 6*(2), 145-155. https://doi.org/10.1037/0278-7393.6.2.145

McGovern, D. P., Burns, S., Hirst, R. J., & Newell, F. N. (2022). Perceptual training narrows the temporal binding window of audiovisual integration in both younger and older adults. *Neuropsychologia*, *173*, 108309. https://doi.org/10.1016/j.neuropsychologia.2022.108309

McGovern, D. P., Roudaia, E., Stapleton, J., McGinnity, T. M., & Newell, F. N. (2014). The sound-induced flash illusion reveals dissociable age-related effects in multisensory integration. *Frontiers in Aging Neuroscience, 6*. https://doi.org/10.3389/fnagi.2014.00250

McGurk, H., & McDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746–748. https://doi.org/10.1080/00357529.1980.11764651

McWeeny, K. H., Young, A. W., Hay, D. C., & Ellis, A. W. (1987). Putting names to faces. *British Journal of Psychology, 78*(2), 143-149. https://doi.org/10.1111/j.2044-8295.1987.tb02235.x

Metzger, R. L., Boschee, P. F., Haugen, T., & Schnobrich, B. L. (1979). The classroom as learning context: Changing rooms affects performance. *Journal of Educational Psychology, 71*(4), 440–442. http://dx.doi.org/10.1037/0022-0663.71.4.440

Miller, J. (1982). Divided attention: Evidence for coactivation with redundant signals. *Cognitive Psychology, 14*(2), 247-279. https://doi.org/10.1016/0010-0285(82)90010-x

Miller, L. M., & D'Esposito, M. (2005). Perceptual Fusion and Stimulus Coincidence in the Cross-Modal Integration of Speech. *The Journal of Neuroscience, 25*(25), 5884-5893. https://doi.org/10.1523/jneurosci.0896-05.2005

Moran, Z. D., Bachman, P., Pham, P., Hah Cho, S., Cannon, T. D., & Shams, L. (2013). Multisensory Encoding Improves Auditory Recognition. *Multisensory Research, 26*(6), 581-592. https://doi.org/10.1163/22134808-00002436

Moreno, R., & Mayer, R. E. (1999). Cognitive principles of multimedia learning: The role of modality and contiguity. *Journal of Educational Psychology, 91*(2), 358–368. https://doi.org/10.1037/0022-0663.91.2.358

Morris, P. E., Fritz, C. O., Jackson, L., Nichol, E., & Roberts, E. (2005). Strategies for learning proper names: Expanding retrieval practice, meaning and imagery. *Applied Cognitive Psychology, 19*(6), 779–798. https://doi.org/10.1002/acp.1115

Murray, C. A., & Shams, L. (2023). Crossmodal interactions in human learning and memory. *Frontiers in Human Neuroscience*, *17*, 1181760. https://doi.org/10.3389/fnhum.2023.1181760

Murray, C. A., Tarlow, M., Rissman, J., & Shams, L. (2022). Multisensory encoding of names via name tags facilitates remembering. *Applied Cognitive Psychology, 36*(6), 1277–1291. https://doi.org/10.1002/acp.4012

Murray, M. M., Lewkowicz, D. J., Amedi, A., & Wallace, M. T. (2016). Multisensory processes: a balancing act across the lifespan. *Trends in Neurosciences*, *39*(8), 567-579. https://doi.org/10.1016/j.tins.2016.05.003

Murray, M. M., Michel, C. M., Grave de Peralta, R., Ortigue, S., Brunet, D., Gonzalez Andino, S., & Schnider, A. (2004). Rapid discrimination of visual and multisensory memories revealed by electrical neuroimaging. *NeuroImage, 21*(1), 125–135. https://doi.org/10.1016/j.neuroimage.2003.09.035

Murray, M. M., & Wallace, M. T. (Eds.). (2011). The neural bases of multisensory processes. CRC Press.

Nardini, M., Cowie, D., Bremner, A. J., Lewkowicz, D. J., & Spence, C. (2012). The development of multisensory balance, locomotion, orientation, and navigation. *Multisensory development*, 137-158.

Neil, P. A., Chee-Ruiter, C., Scheier, C., Lewkowicz, D. J., & Shimojo, S. (2006). Development of multisensory spatial integration and perception in humans. *Developmental Science, 9*(5), 454–464. https://doi.org/10.1111/j.1467-7687.2006.00512.x

Neuschatz, J. S., Preston, E. L., Toglia, M. P., & Neuschatz, J. S. (2005). Comparison of the Efficacy of Two Name-Learning Techniques: Expanding Rehearsal and Name-Face Imagery. *The American Journal of Psychology, 118*(1), 79–102. JSTOR.

Nguyen, K. P., Josić, K., & Kilpatrick, Z. P. (2019). Optimizing sequential decisions in the drift–diffusion model. *Journal of Mathematical Psychology, 88*, 32-47.

Nidiffer, A. R., Diederich, A., Ramachandran, R., & Wallace, M. T. (2018). Multisensory perception reflects individual differences in processing temporal correlations. *Scientific Reports, 8*(1). https://doi.org/10.1038/s41598-018-32673-y

Norman, D. A., & Rumelhart, D. E. (1970). A System for Memory and Perception. In D. A. Norman (Ed.), Models of Human Memory (pp. 19–64). Academic Press.

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*(20), 11120-11124. https://doi.org/10.1073/pnas.97.20.11120

O'Mahony, C., & Newell, F. N. (2012). Integration of faces and voices, but not faces and names, in person recognition: Integration of faces and voices in person recognition. *British Journal of Psychology, 103*(1), 73-82. https://doi.org/10.1111/j.2044-8295.2011.02044.x

Otten, L. J., Henson, R. N. A., & Rugg, M. D. (2001). Depth of processing effects on neural correlates of memory encoding: relationship between findings from across- and within-task comparisons. *Brain, 124*(2), 399–412. https://doi.org/10.1093/brain/124.2.399

Otto, T. U., Dassy, B., & Mamassian, P. (2013). Principles of Multisensory Behavior. *The Journal of Neuroscience, 33*(17), 7463-7474. https://doi.org/10.1523/jneurosci.4678-12.2013

Paivio, A. (1991). Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology / Revue canadienne de psychologie, 45*(3), 255-287. https://doi.org/10.1037/h0084295

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

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods, 51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Peters, M. A., Balzer, J., & Shams, L. (2015). Smaller= denser, and the brain knows it: natural statistics of object density shape weight expectations. *PloS one*, *10*(3), e0119794. https://doi.org/10.1371/journal.pone.0119794

Petersen, R. C. (1977). Retrieval failures in alcohol state-dependent learning. *Psychopharmacology, 55*(2), 141–146. https://doi.org/10.1007/BF01457849

Plass, J. L., Chun, D. M., Mayer, R. E., & Leutner, D. (1998). Supporting visual and verbal learning preferences in a second-language multimedia learning environment. *Journal of Educational Psychology, 90*(1), 25–36. https://doi.org/10.1037/0022-0663.90.1.25

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review, 117*(3), 864-901. https://doi.org/10.1037/a0019737

Quintero, S. I., Shams, L., & Kamal, K. (2022). Changing the Tendency to Integrate the Senses. *Brain Sciences, 12*(10), Article 10. https://doi.org/10.3390/brainsci12101384

Rach, S., Diederich, A., & Colonius, H. (2011). On quantifying multisensory interaction effects in reaction time and detection rate. *Psychological Research, 75*(2), 77–94. https://doi.org/10.1007/s00426-010-0289-0

Raposo, D., Sheppard, J. P., Schrater, J. P., & Churchland, A. K. (2012). Multisensory decision-making in rats and humans. *Journal of Neuroscience, 32*(11), 3726-3735.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*(2), 59–108. https://doi.org/10.1037/0033-295X.85.2.59

Ratcliff, R. (2018). Decision making on spatially continuous scales. *Psychological Review, 125*(6), 888-935. https://doi.org/10.1037/rev0000117

Ratcliff, R., & McKoon, G. (2008). The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks. *Neural Computation, 20*(4), 873-922. https://doi.org/10.1162/neco.2008.12-06-420

Ratcliff, R., & Smith, P. L. (2004). A Comparison of Sequential Sampling Models for Two-Choice Reaction Time. *Psychological Review, 111*(2), 333-367. https://doi.org/10.1037/0033-295x.111.2.333

Ratcliff, R., Thapar, A., & McKoon, G. (2004). A diffusion model analysis of the effects of aging on recognition memory. *Journal of Memory and Language*, *50*(4), 408-424. https://doi.org/10.1016/j.jml.2003.11.002

Regenbogen, C., Johansson, E., Andersson, P., Olsson, M. J., & Lundström, J. N. (2016). Bayesian-based integration of multisensory naturalistic perithreshold stimuli. *Neuropsychologia, 88*, 123-130. https://doi.org/10.1016/j.neuropsychologia.2015.12.017

Resulaj, A., Kiani, R., Wolpert, D. M., & Shadlen, M. N. (2009). Changes of mind in decision-making. *Nature, 461*, 263–266. https://doi.org/10.1038/nature08275

Roediger, H. L., & Karpicke, J. D. (2006). Test-Enhanced Learning: Taking Memory Tests Improves Long-Term Retention. *Psychological Science, 17(*3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Rohlf, S., Li, L., Bruns, P., & Röder, B. (2020). Multisensory Integration Develops Prior to Crossmodal Recalibration. *Current Biology, 30*(9), 1726-1732.e7. https://doi.org/10.1016/j.cub.2020.02.048

Rosenblum, L. D., Dias, J. W., & Dorsi, J. (2017). The supramodal brain: Implications for auditory perception. *Journal of Cognitive Psychology, 29*(1), 65–87.

Rossion, B., & Pourtois, G. (2004). Revisiting Snodgrass and Vanderwart's Object Pictorial Set: The Role of Surface Detail in Basic-Level Object Recognition. *Perception, 33*(2), 217–236. https://doi.org/10.1068/p5117

Schurgin, M. W. (2018). Visual memory, the long and the short of it: A review of visual working memory and long-term memory. *Attention, Perception, & Psychophysics, 80*(5), 1035–1056. https://doi.org/10.3758/s13414-018-1522-y

Seitz, A. R., Kim, R., & Shams, L. (2006). Sound Facilitates Visual Learning. *Current Biology, 16*(14), 1422–1427. https://doi.org/10.1016/j.cub.2006.05.048

Senkowski, D., Saint-Amour, D., Höfle, M., & Foxe, J. J. (2011). Multisensory interactions in early evoked brain activity follow the principle of inverse effectiveness. *NeuroImage, 56*(4), 2200–2208. https://doi.org/10.1016/j.neuroimage.2011.03.075

Setti, A., Finnigan, S., Sobolewski, R., McLaren, L., Robertson, I. H., Reilly, R. B., Kenny, R. A., & Newell, F. N. (2011). Audiovisual temporal discrimination is less efficient with aging: an event-related potential study. *Neuroreport*, *22*(11), 554–558. https://doi.org/10.1097/WNR.0b013e328348c731

Shams, L., & Beierholm, U. R. (2010). Causal inference in perception. *Trends in Cognitive Sciences, 14*(9), 425–432. https://doi.org/10.1016/j.tics.2010.07.001

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

Shams, L., Kamitani, Y., & Shimojo, S. (2002). Visual illusion induced by sound. *Cognitive brain research*, *14*(1), 147-152. https://doi.org/10.1016/S0926-6410(02)00069-1

Shams, L., & Kim, R. (2010). Crossmodal influences on visual perception. *Physics of Life Reviews, 7*(3), 269–284. https://doi.org/10.1016/j.plrev.2010.04.006

Shams, L., & Seitz, A. R. (2008). Benefits of multisensory learning. *Trends in Cognitive Sciences, 12*(11), 411–417. https://doi.org/10.1016/j.tics.2008.07.006

Shams, L., Wozny, D. R., Kim, R. S., & Seitz, A. (2011). Influences of Multisensory Experience on Subsequent Unisensory Processing. *Frontiers in Psychology, 2*. https://doi.org/10.3389/fpsyg.2011.00264

Shelton, J., & Kumar, G. P. (2010). Comparison between Auditory and Visual Simple Reaction Times. *Neuroscience and Medicine, 01*(01), 30-32. https://doi.org/10.4236/nm.2010.11004

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in Cognitive Sciences, 12*(5), 182-186. https://doi.org/10.1016/j.tics.2008.02.003

Shohamy, D., & Turk-Browne, N. B. (2013). Mechanisms for widespread hippocampal involvement in cognition. *Journal of Experimental Psychology: General, 142*(4), 1159-1170. https://doi.org/10.1037/a0034461

Slotnick, S. D., Thompson, W. L., & Kosslyn, S. M. (2012). Visual memory and visual mental imagery recruit common control and sensory regions of the brain. *Cognitive Neuroscience, 3*(1), 14–20. https://doi.org/10.1080/17588928.2011.578210

Smith, P. L. (2016). Diffusion theory of decision making in continuous report. *Psychological Review, 123*(4), 425-451. https://doi.org/10.1037/rev0000023

Smith, S. M. (1985). Background Music and Context-Dependent Memory. *The American Journal of Psychology, 98*(4), 591–603. https://doi.org/10.2307/1422512

Smith, S. M., Glenberg, A., & Bjork, R. A. (1978). Environmental context and human memory. *Memory & Cognition, 6*(4), 342–353. https://doi.org/10.3758/BF03197465

Smith, S. M., & Vela, E. (2001). Environmental context-dependent memory: A review and meta-analysis. *Psychonomic Bulletin & Review, 8*(2), 203–220. https://doi.org/10.3758/BF03196157

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

Sorden, S. D. (2005). A cognitive approach to instructional design for multimedia learning. *Informing Science*, *8*, 263-279.

Spaniol, J., Madden, D. J., & Voss, A. (2006). A diffusion model analysis of adult age differences in episodic and semantic long-term memory retrieval. *Journal of*

experimental psychology. Learning, memory, and cognition, *32*(1), 101–117. https://doi.org/10.1037/0278-7393.32.1.101

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

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian Measures of Model Complexity and Fit. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 64*(4), 583-639. https://doi.org/10.1111/1467-9868.00353

Stein, B. E. (2012). The New Handbook of Multisensory Processing. MIT Press. https://doi.org/10.7551/mitpress/8466.001.0001

Stein, B. E., & Meredith, M. A. (1993). The Merging of the Senses. MIT Press.

Stein, B. E., Stanford, T. R., Wallace, M. T., Vaughan, J. W., & Jiang, W. (2004). Crossmodal spatial interactions in subcortical and cortical circuits. *Crossmodal space and crossmodal attention*, *2550*.

Strnad, B. N., & Mueller, J. H. (1977). Levels of processing in facial recognition memory. *Bulletin of the Psychonomic Society, 9*(1), 17–18. https://doi.org/10.3758/BF03336915

Sweller, J. (2005). The Redundancy Principle in Multimedia Learning. In R. Mayer (Ed.), The Cambridge Handbook of Multimedia Learning (pp. 159-168). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511816819.011

Talsma, D., Senkowski, D., Soto-Faraco, S., & Woldorff, M. G. (2010). The multifaceted interplay between attention and multisensory integration. *Trends in Cognitive Sciences, 14*(9), 400–410. https://doi.org/10.1016/j.tics.2010.06.008

Thelen, A., Cappe, C., & Murray, M. M. (2012). Electrical neuroimaging of memory discrimination based on single-trial multisensory learning. *NeuroImage, 62*(3), 1478–1488. https://doi.org/10.1016/j.neuroimage.2012.05.027

Thelen, A., & Murray, M. M. (2013). The efficacy of single-trial multisensory memories. *Multisensory research*, *26*(5), 483-502. https://doi.org/10.1163/22134808-00002426

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

Thurlow, W. R., & Jack, C. E. (1973). Certain determinants of the “ventriloquism effect”. *Perceptual and motor skills*, *36*(3_suppl), 1171-1184. https://doi.org/10.2466/pms.1973.36.3c.1171

Townsend, J. T., & Ashby, F. G. (1983). The stochastic modeling of elementary psychological processes. Cambridge University Press.

Trommershauser, J., Kording, K., & Landy, M. S. (Eds.). (2011). Sensory cue integration. Oxford University Press.

Tulving, E. (1993). What Is Episodic Memory? *Current Directions in Psychological Science, 2*(3), 67-70. https://doi.org/10.1111/1467-8721.ep10770899

Tulving, E., & Bower, G. H. (1974). The Logic of Memory Representations. *Psychology of Learning and Motivation*, 265-301. https://doi.org/10.1016/s0079-7421(08)60457-0

van Atteveldt, N. M., Formisano, E., Blomert, L., & Goebel, R. (2006). The Effect of Temporal Asynchrony on the Multisensory Integration of Letters and Speech Sounds. *Cerebral Cortex, 17*(4), 962-974. https://doi.org/10.1093/cercor/bhl007

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2011). Hierarchical diffusion models for two-choice response times. *Psychological Methods, 16*(1), 44-62. https://doi.org/10.1037/a0021765

Van Zandt, T., Colonius, H., & Proctor, R. W. (2000). A comparison of two response time models applied to perceptual matching. *Psychonomic Bulletin & Review, 7*(2), 208-256. https://doi.org/10.3758/bf03212980

von Kriegstein, K., & Giraud, A.-L. (2006). Implicit Multisensory Associations Influence Voice Recognition. *PLoS Biology, 4*(10), e326. https://doi.org/10.1371/journal.pbio.0040326

Voss, A., Nagler, M., & Lerche, V. (2013). Diffusion Models in Experimental Psychology: A Practical Introduction. *Experimental Psychology, 60*(6), 385-402. https://doi.org/10.1027/1618-3169/a000218

Wabersich, D., & Vandekerckhove, J. (2014). Extending JAGS: A tutorial on adding custom distributions to JAGS (with a diffusion model example). *Behavior research methods*, *46*, 15-28. https://doi.org/10.3758/s13428-013-0369-3

Watkins, S., Shams, L., Tanaka, S., Haynes, J. D., & Rees, G. (2006). Sound alters activity in human V1 in association with illusory visual perception. *NeuroImage, 31*(3), 1247-1256. https://doi.org/10.1016/j.neuroimage.2006.01.016

Webber, N. E. (1978). Pictures and words as stimuli in learning foreign language responses. The Journal of Psychology, 98(1), 57-63. https://doi.org/10.1080/00223980.1978.9915946

Weigard, A., & Huang-Pollock, C. (2017). The Role of Speed in ADHD-Related Working Memory Deficits: A Time-Based Resource-Sharing and Diffusion Model Account. *Clinical Psychological Science, 5*(2), 195–211. https://doi.org/10.1177/2167702616668320

Weinberger, N. M. (2004). Specific long-term memory traces in primary auditory cortex. *Nature Reviews Neuroscience, 5*(4), 279-290. https://doi.org/10.1038/nrn1366

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*(20), 11125-11129. https://doi.org/10.1073/pnas.97.20.11125

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica, 41*(1), 67-85. https://doi.org/10.1016/0001-6918(77)90012-9

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. *Frontiers in Neuroinformatics, 7*. https://doi.org/10.3389/fninf.2013.00014

Wozny, D. R., Beierholm, U. R., & Shams, L. (2008). Human trimodal perception follows optimal statistical inference. *Journal of Vision*, *8*(3), 1-11. https://doi.org/10.1167/8.3.24.

Wozny, D. R., & Shams, L. (2011). Computational Characterization of Visually Induced Auditory Spatial Adaptation. *Frontiers in Integrative Neuroscience, 5*. https://doi.org/10.3389/fnint.2011.00075

Yu, A. J., Dayan, P., & Cohen, J. D. (2009). Dynamics of attentional selection under conflict: Toward a rational Bayesian account. *Journal of Experimental Psychology: Human Perception and Performance, 35*(3), 700-717. https://doi.org/10.1037/a0013553

Zhang, R., & Zou, D. (2022). A state-of-the-art review of the modes and effectiveness of multimedia input for second and foreign language learning. *Computer Assisted Language Learning, 35*(9), 2790-2816. https://doi.org/10.1080/09588221.2021.1896555