Individual Differences in Preferred Thought Formats Predict Features of Narrative Recall

Ana Chkhaidze (achkhaid@ucsd.edu)
Seana Coulson (scoulson@ucsd.edu)
Anastasia Kiyonaga (akiyonaga@ucsd.edu)
Department of Cognitive Science, University of California, San Diego
9500 Gilman Dr, La Jolla, CA 92093, USA

Abstract

Humans differ in how they experience their own thoughts. Some say they hear sentences in their *"mind's ear"*, others report seeing images in their *"mind's eye"*, and many struggle to describe their inner worlds. Here, we tested whether individual differences in thought formats predict accuracy and properties of verbal recall after listening to short podcasts about science. To assess the accuracy of recall, we measured the semantic similarity between embeddings of participant recall statements and the original podcasts. To characterize the properties of participants' recall language, we measured the perceptual strength of content words in their responses. Individual differences in thought formats were not associated with differences in the accuracy of verbal recall. By contrast, recall statements high in perceptual strength were more likely among participants who reported vivid visual imagery, while statements low in perceptual strength were more likely among those with higher verbal scores. Results highlight an intriguing connection between subjective reports about thought format and the attributes of naturalistic verbal memory recall.

Keywords: language; thought; memory; imagery

Introduction

Phenomenal experiences of thought greatly vary across individuals, ranging from abstract to perceptual, verbal to visual, dim to vivid (Hurlburt & Akhter, 2008; Richardson, 1977; Zeman et al., 2020).

Given their hidden nature, the format and content of our thoughts—internal representations—have proven to be difficult to quantify, sparking a long-standing debate in cognitive science. Some have argued that all information is stored in the human mind in a *descriptive*, *symbolic*, language-like format, and phenomenal differences in thought are merely epiphenomenal (Pylyshyn, 1973). Others have maintained that thoughts are *depictive* in nature; that our mental representations iconically *re-present* the world, recruiting neural resources used in perception (Pearson & Kosslyn, 2015).

Information about mental images can be extracted using decoders trained on neural activity recorded during visual perception (Harrison & Tong, 2009; Naselaris et al., 2015), suggesting that mental imagery recruits visual representations assumed to be depictive. Indeed, visual imagery also recruits peripheral responses in the eye, as pupils can respond to imagined luminance much like the response to retina-based light sources (Kay et al., 2022).

Moreover, humans seem to use those depictive representations functionally for processes beyond the visual domain. These include *false memory formation*, in that brain activity associated with visual imagery predicts falsely remembering objects that were only imagined, never perceived (Stephan-Otto et al., 2017); *reading comprehension*, in that humans are susceptible to the famous motion aftereffect illusion just from reading linguistic descriptions of motion (Dils & Boroditsky, 2010); as well as cognitive acts as complex as *moral decision-making* (Amit & Greene, 2012).

At the same time, we see recurring and pervasive reports of individual differences in subjective descriptions of the vividness of internal representations, yet this variability is rarely considered in theories and models of human cognition—even those that emphasize the involvement of imagery and thus require the assumption that people have one. Including the phenomenal experience of thought formats in the *equations of cognition* has the potential to reshape our understanding of cognition by challenging existing theories (Lupyan et al., 2023). Here, we examine the relationship between these phenomenal differences and recall of naturalistic stimuli.

Individual Differences in Thought Format

Subjective differences in thought formats have been formally quantified using various psychometric tools, including the Internal Representations Questionnaire (IRQ: Roebuck & Lupyan, 2020), the Vividness of Visual Imagery Questionnaire (VVIQ: Marks, 1973), and the Object-Spatial Imagery and Verbal Questionnaire (OSIVQ: Blazhenkova and Kozhenikov, 2009). However, there is no consensus on which of these measures best captures these phenomenological effects and their construct validity has been under question (see Roebuck & Lupyan, 2020 for a review).

Here, we aimed to (1) compare several instruments that have been put forward as assessments of thought format and (2) test whether they predict features of participants' free recall responses for verbal content. Accordingly, we administered several previously validated instruments—the IRQ, the VVIQ, and the OSIVQ—to the same participants to establish whether they all capture participants' subjective assessment of their own thought format. If these instruments accurately measure individual differences in thought profiles,

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

we would expect participants' scores on the visual dimensions of these instruments to be correlated with one another, and likewise for their verbal dimensions. Further, if these measures index the way participants think they think, the assessment scores should predict whether a given individual considers themselves to think in words or in pictures.

Beyond their subjective adequacy, is the question of whether these instruments tap into objective aspects of cognitive function. We hypothesized that differences in thought format would be related to different components of language comprehension characterized as embodied and symbolic. Whereas embodied meanings involve the activation of the sensorimotor cortex implicated in experiences with the referent, symbolic meaning recruits a network of polymodal association areas with more abstract representational content (Pulvermüller, 2013).

Because we expect that individual differences in thought format might arise due to asymmetries in the activation of the sensorimotor versus association cortex, these differences might be manifested in the kinds of detail people attend to in language or in the characteristics of the language they produce, for instance, in how much perceptual detail they mention—further referred as *perceptual strength*. Language researchers have noted that speakers display a remarkable degree of consistency in their ratings of the sensory and motor origins for a vast range of concepts and that these intuitions are captured in the *perceptual strength* ratings participants assign to individual words (Lynott & Connell, 2013).

To test for objective consequences of thought format, we asked participants to listen to two short podcasts and then complete a free (verbal) recall task after each. We asked whether individual differences in thought format were associated with how much perceptual language they produced on the free recall tasks. We hypothesized that participants who scored high on the visual dimensions of our instruments might produce recalls high in perceptual strength reflecting their attention to perceptual information conveyed by the language during the encoding process. Likewise, participants who scored high on the verbal dimensions might produce recalls lower in perceptual strength, reflecting their attention to more abstract language.

Methods

Participants

We recruited 59 participants through the online experiment participation platform *Prolific*. All participants gave informed consent and were compensated at the rate of $10 (US) per hour. Participation was restricted to adults (over 18 years of age) whose first language was English, and who lived in countries where the official language is English. All participants successfully answered at least 2/3 of catch trial attention checks. It took participants 25-30 minutes to complete the study, depending on the duration of free recall.

Materials & Procedure

Each participant listened to two short popular science podcast episodes from *Scientific America's 60-Second Science,* the order of which was counterbalanced. One podcast was on *autophagy* (*Ep 31: Nobel in physiology or medicine to Yoshinori Ohsumi* (01:52)) and the other—*pigmentation* (*Ep 47: Color-changing skin aids climate control and communication* (02:06)). Each trial began by playing a podcast, and participants were not allowed to pause, stop, or replay the recording. Next, participants were instructed to retell the content of the podcast they had just listened to: *"Retell the podcast as if you were conveying the content of the story to a friend. Recall as much detail as possible."* Participants had to spend at least two minutes retelling each podcast.

Following each recall, participants were asked to rate a set of statements concerning their experience with listening to and recalling the podcast. These statements asked about their level of *confidence* in their recollection, their ability to maintain *focus*, their *familiarity* with the podcast's topic, and their overall *enjoyment* of the podcast. As none of these factors were found to predict participants' responses, they are not discussed further.

After the podcast and recall portion of the session, participants were asked to complete three previously published questionnaires measuring individual differences in the format of thought. These included: The *Internal Representation Questionnaire* (IRQ: Roebuck & Lupyan, 2020) to assess verbal, visual, and manipulation modes of thought; the *Object-Spatial Imagery and Verbal Questionnaire* (OSIVQ: Blazhenkova and Kozhenikov, 2009) to differentiate between visual object and visual spatial imageries; and the *Vividness of Visual Imagery Questionnaire* (VVIQ: Marks, 1973) to assess the vividness of visual imagery. Each statement from the OSIVQ and IRQ was presented together with a five-point Likert scale from Strongly Disagree to Strongly Agree. For each VVIQ statement, participants were asked to visually imagine an object or a scene and then rate their imagery using one of these statements: "No image at all, I only "know" I am thinking of the object," "Dim and vague image," "Moderately realistic and vivid," "Realistic and reasonably vivid," "Perfectly realistic, as vivid as real seeing." Following the formal questionnaires, participants were posed with an informal inquiry regarding their phenomenal experiences. They were asked the question: "How do you think?" and were given the response options: "In pictures," "In written words," "In spoken words," "In abstract symbols," "Can't describe," or "Other." An optional field was also provided for participants to offer a description of their thought formats. We refer to this below as the *Thought Format* assessment.

Analysis

Individual Differences We scored responses on different dimensions of the IRQ, OSIVQ, and VVIQ following the guidelines described in Roebuck and Lupyan (2020), Blazhenkova and Kozhevnikov (2009), and Marks (1973),

respectively. Responses to the Thought Format assessment were divided into three categories: participants who think in words (those who responded either "In written words" or "In spoken words"), participants who think in pictures (those who responded "In pictures"), and a third catch-all category for those who responded either "In abstract symbols," "Can't describe," or "Other."

Text Preprocessing Recall statements and podcast transcripts were initially run through a basic Natural Language Processing (NLP) pipeline. This pipeline included steps such as correcting typos, standardizing cases, removing punctuation, and tokenizing and lemmatizing words. However, when calculating similarities between free recalls and podcasts, the original punctuation and untokenized text were preserved.

Podcasts We quantified perceptual richness of each podcast using the *maximum perceptual strength* variable from the Lancaster Sensorimotor Norms (Lynnot et al., 2020). In this database, 40,000 English words are rated on six perceptual and five action modalities. Maximum perceptual strength (hereafter referred to as perceptual strength) is a single composite variable representing the rating of the dominant perceptual dimension for the word. We calculated the perceptual strength of a given podcast by averaging together the perceptual strength measures for every word in the podcast.

Free Recalls To quantify the perceptual characteristics of the free responses, we assessed the perceptual richness of each response in a similar manner as we did for podcasts. That is, we retrieved the perceptual strength for every word in the recall statements from the Lancaster norms and computed the average.

To quantify how accurately each participant remembered the podcasts, we used Semantic Textual Similarity (STS) approaches from the area of NLP to compare the similarity between participant recall and the actual podcast content. We implemented similarity analysis using Transformers—a powerful NLP architecture that has been used to measure semantic similarity at the sentence, paragraph, and text levels (Ormerod et al., 2021). We used *SentenceTransformers* Python framework to compute embeddings for podcasts and recall responses and then calculated cosine similarity between the two. A recent study has shown that NLP models of similarity analysis can be good metrics for measuring free recall accuracies (Shen et al., 2022).

Results

Are Different Subjective Measures of Thought Related to One Other?

Overview The visual dimensions on all three questionnaires—IRQ, OSIVQ, and VVIQ—were correlated with one another, except for the visuospatial dimension on the OSIVQ. The latter was correlated only with the verbal dimension on the OSIVQ and the manipulation scale on the IRQ. Notably, the verbal dimensions of the IRQ and the OSIVQ were not correlated with each other. As described below, three dimensions of the two instruments (VVIQ, OSIVQ Object, and OSIVQ Verbal) predicted participants' responses on the Thought Format assessment, suggesting these measures capture participants' own assessment as to whether they think in words or pictures and quantify their subjective impressions in a continuous manner.

Correlations Between Instruments To assess the relationships between different instruments, we computed Pearson correlation coefficients between all relevant scores (Bonferroni-corrected for multiple comparisons; see Figure 1). Reliable correlations were found between multiple pairs, even after Bonferroni correction. In the IRQ, visual and verbal factors were positively correlated (r = .38), and so were manipulation and visual factors (r = .38). The verbal and manipulation dimensions of IRQ were not correlated (r = .16). The verbal and spatial factors on the OSIVQ were correlated (r = .31). However, there was no reliable relationship between the object dimension of the OSIVQ and the spatial and verbal factors of this assessment (r = .05 and r = -.3, respectively). All three visual factors of these questionnaires (VVIQ, OSIVQ object, and IRQ visual) except for the visuospatial dimension on the OSIVQ were correlated with each other. The object dimension on the OSIVQ was strongly correlated both with the visual dimension of the IRQ and with VVIQ scores (r = .8, r = .65, respectively). There was also a positive correlation between visual IRQ and VVIQ scores (r = .53). Finally, the correlation between verbal dimensions of OSIVQ and IRQ was not reliable (r = .27).

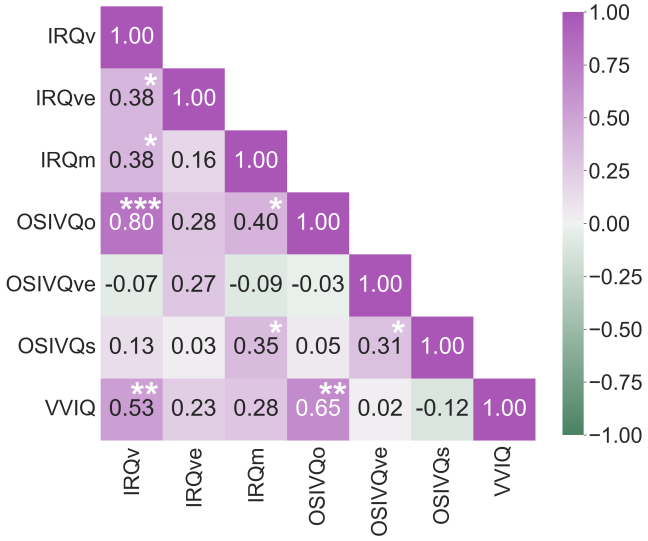

Figure 1: Heatmap of correlation coefficients between assessment measures (v: visual, ve: verbal, m: manipulation, o: object, s: spatial)

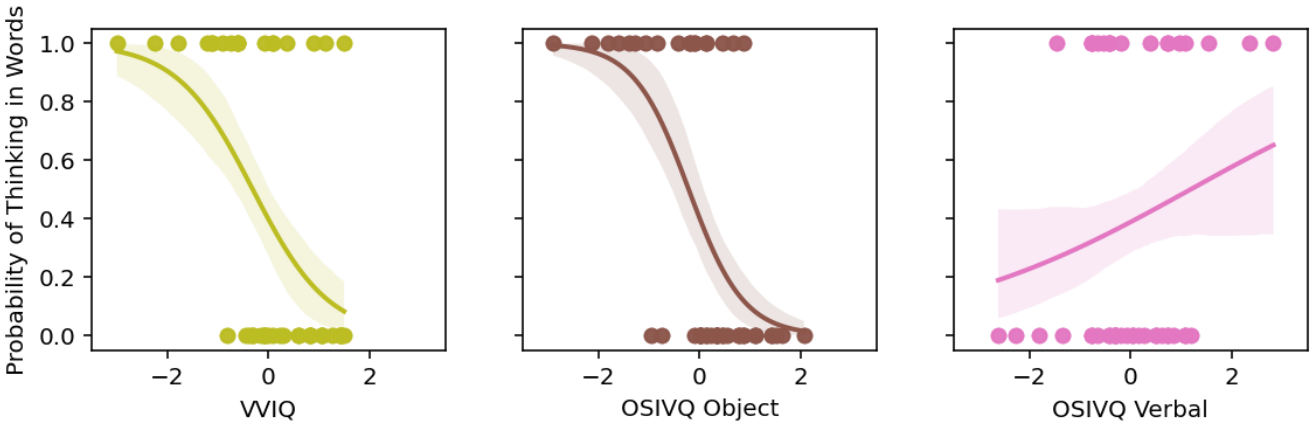

Figure 2: Probability of thinking in words as a function of VVIQ (left), OSIVQ Object (middle), and OSIVQ Verbal (right)

Relationship Between Instruments and the Thought Format Assessment To assess the relationship between the scores on formalized instruments and how people think they think (as measured by our Thought Format Assessment), we performed logistic regression using the *glm* package in R. The Hosmer and Lemeshow chi-square test indicates the data adequately fit the model ($\chi^2(7) = 3.54$, $p = .83$). The model explained 58% (Nagelkerke R^2) of the variance in participants' self-reported thought format, as it discriminated between the two categories—words and pictures (AUC = .90). The odds of thinking in words increased by a factor of 1.02 for every unit increase in standardized scores on the OSIVQ verbal dimension ($p = .03$). By contrast, as scores increased on the object dimension of OSIVQ and on the VVIQ, the odds of thinking in words changed by a factor of -1.60 and -1.18 ($p = .02$ and $p = .007$), respectively (Figure 2). None of the IRQ dimensions predicted differences in participants' subjective descriptions of their thoughts ($p_{visual} = .55$, $p_{verbal} = .58$, $p_{manipulation} = .29$). As IRQ factors were not predictive of participants' thought formats, they are not discussed further.

Do Subjective Measures of Thought Reflect Differences in Verbal Memory Recall?

Overview Self-reported differences in thought formats were not associated with our measure of recall accuracy. However, those same differences were related to the type of language participants used in their recall responses. Those with higher VVIQ scores produced recall statements richer in *perceptual language* than those with less vivid visual imagery. Likewise, relative to those with lower scores on the verbal dimension of the OSIVQ, participants with higher verbal scores produced recall statements that were lower in perceptual strength. Thus, two of the instruments assessing participants' subjective experience were reliably associated with objective features of their free recall statements.

For all the analyses reported below, we used mixed effects models with participant and podcast as random intercepts using the *lmer* package in R.

Stimuli-Response Similarity We first examined whether OSIVQ and VVIQ predicted how well participants recall information from the podcasts. We operationalized recall accuracy using cosine similarity distances between each podcast and the corresponding free response on the recall task. None of the questionnaire predictors in this model were significant, suggesting that individual differences in thought profiles were not reliably associated with the accuracy of their recall of the podcasts. However, longer responses were predictive of stronger similarity to the podcasts, $b = .0.05$, (SE = .01), $t = 4.66$ (Figure 3, left).

Perceptual Strength of Responses Next, we looked at the relationship between IRQ, OSIVQ, and VVIQ and the type of language used in participants' free recalls. We built a mixed effects model with the perceptual strength of the responses as the predicted variable. Since the perceptual strength of the recall task is expected to be greater when participants recall a podcast that is itself high in perceptual strength, we also included the perceptual strength for each podcast as an additional predictor in our mixed effect models. The full model (using z-scored predictors) was:

Recall_perceptual_strength ~
*OSIVQ_factor*podcast_strength + VVIQ*podcast _strength + (1|participant) + (1|podcast)*

As expected, the perceptual strength of the podcast predicted greater perceptual strength in participants' recall responses, $b = .15$, (SE = .01), $t = 11.78$. More germane to our study goals, higher scores on the VVIQ also predicted increased perceptual strength in recall language, $b = .06$ (SE = .02), $t = 2.77$. Finally, there was also an interaction between the verbal dimension of the OSIVQ and the perceptual strength of the podcast: For the perceptually stronger podcast, participants with greater verbal scores produced language with lower perceptual strength, $b = -.03$, (SE = .01), $t = -2.33$.

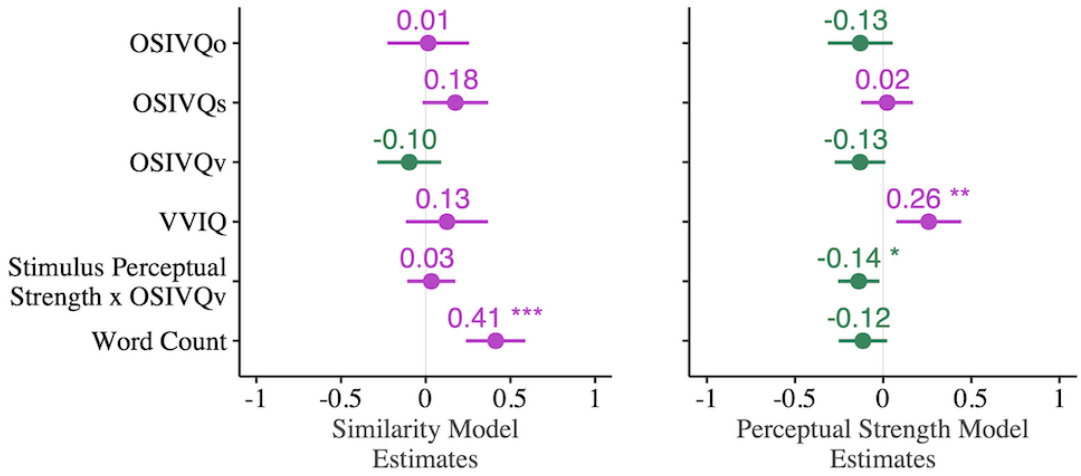

Figure 3: Regression coefficients from mixed effects models predicting *similarity* between podcast and recall (left) and *perceptual strength* of recall (right), showing the main effects of OSIVQ, VVIQ, and length of recall

Other dimensions of OSIVQ did not predict participants' responses (Figure 3, right).

Discussion

Some psychometric measures of thought format reliably captured subjective experiences of how people think they think, while others did not. When asked a simple question—*How do you think?*—people scoring high on OSIVQ's object imagery dimension and those with higher vividness scores on VVIQ said they thought *in pictures*, while high OSIVQ verbal scores reliably predicted that participants would report thinking *in words*. The verbal and visual dimensions of IRQ did not capture these phenomenological differences across participants.

Of course, the relationship, or lack thereof between an instrument and a single-word introspective report cannot serve as the only test of the instrument's utility and construct validity. Here, we also examined the relationship between these questionnaires and an objective measure of naturalistic verbal recall.

Preferred Thought Format and Verbal Recall

We asked whether different thought profiles predict how well participants recalled rich naturalistic linguistic stimuli—short podcasts—by assessing similarities between their written summaries and the podcasts themselves. To quantify similarity, we utilized natural language processing (NLP) models of semantic text similarity (STS) which have been found to be valid and reliable to automatically score the quality of narrative recalls (Shen et al., 2022). We found no relationship between preferred thought formats and the quality of participants' recall, which, taken at face value, suggests that neither format is inherently superior for remembering verbally-presented content. Alternatively, this null result may reflect that our measure of semantic similarity is inadequate to capture recall accuracy. This could be remedied in future work by supplementing the NLP analysis with content coding by trained human annotators.

A more intriguing question is whether thought profiles can predict the perceptual properties of how people remember and report naturalistic stimuli—more specifically, the type of language participants produce when they recall podcasts. These data suggest the answer is *yes*. Higher visual vividness scores from the VVIQ were associated with more perceptual information in participants' recalls. Likewise, higher scores on the verbal dimension of the OSIVQ were associated with less perceptual information in the recalls.

Here we show that scores on instruments that measure how people think they think—in words or in pictures—were associated with differences in the type of language that characterized their verbal recall. While this result should be replicated before drawing strong conclusions, it raises intriguing questions regarding the underlying neural substrates of preferred thought formats and how they relate to the processes involved in verbal encoding and recall.

We suggest that participants with a visualiser thought profile are more likely to engage in visual mental imagery processes when they listen to verbal materials (such as the podcasts used in this study), as compared to those with a verbaliser profile. Consequently, when they are asked to recall what they heard, their memory is marked by more perceptual details, which results in verbal recall statements with higher perceptual strength.

Likewise, when verbalisers listen to the same podcasts, they are less likely to activate "grounded" sensorimotor features associated with the material, utilizing abstract language-based codes instead. Their verbal recall would thus be characterized by words with lower perceptual strength.

This work represents a first step toward linking various subjective and objective measures of thought. The interpretations here are speculative and will require further work to assess their validity, but overall, the results suggest that (1) the imagery instruments may capture consistent cognitive differences, and (2) differences in representational formats may influence how naturalistic language stimuli are remembered.

References

Amit, E., & Greene, J. D. (2012). You See, the Ends Don't Justify the Means: Visual Imagery and Moral Judgment. *Psychological Science, 23*(8), 861–868.

Blazhenkova, O., & Kozhevnikov, M. (2009). The new object-spatial-verbal cognitive style model: Theory and measurement. Applied Cognitive Psychology, 23(5), 638–663.

Dils, A. T., & Boroditsky, L. (2010). Visual motion aftereffect from understanding motion language. *Proceedings of the National Academy of Sciences of the United States of America, 107*(37), 16396–16400.

Harrison, S. A., & Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. *Nature, 458(*7238).

Hurlburt, R. T., & Akhter, S. A. (2008). Unsymbolized Thinking. *Consciousness and Cognition, 17(*4), 1364–1374.

Kay, L., Keogh, R., Andrillon, T., & Pearson, J. (2022). The pupillary light response as a physiological index of aphantasia, sensory and phenomenological imagery strength. *ELife*, *11*, e72484.

Lupyan, G., Uchiyama, R., Thompson, B., & Casasanto, D. (2023). Hidden Differences in Phenomenal Experience. *Cognitive Science, 47(*1), e13239.

Lynott, D., & Connell, L. (2013). Modality exclusivity norms for 400 nouns: The relationship between perceptual experience and surface word form. Behavior Research Methods, 45(2)

Lynott, D., Connell, L., Brysbaert, M., Brand, J., & Carney, J. (2020). The Lancaster Sensorimotor Norms: Multidimensional measures of perceptual and action strength for 40,000 English words. Behavior Research Methods, 52(3), 1271–1291.

Marks, D. F. (1973). Visual Imagery Differences in the Recall of Pictures. British Journal of Psychology, 64(1), 17–24.

Naselaris, T., Olman, C. A., Stansbury, D. E., Ugurbil, K., & Gallant, J. L. (2015). A voxel-wise encoding model for early visual areas decodes mental images of remembered scenes. *NeuroImage, 105,* 215-228.

Ormerod, M., Rincón, J. M. del, & Devereux, B. (2021). Predicting Semantic Similarity Between Clinical Sentence Pairs Using Transformer Models: Evaluation and Representational Analysis. JMIR Medical Informatics,

Pearson, J., & Kosslyn, S. M. (2015). The heterogeneity of mental representation: Ending the imagery debate. Proceedings of the National Academy of Sciences, 112(33), 10089–10092.

Pulvermüller, F. (2013). How neurons make meaning: Brain mechanisms for embodied and abstract-symbolic semantics. Trends in Cognitive Sciences, 17(9), 458–470.

Pylyshyn, Z. W. (1973). What the mind's eye tells the mind's brain: A critique of mental imagery. Psychological Bulletin, 80, 1–24.

Richardson, A. (1977). Verbalizer-visualizer: A cognitive style dimension. *Journal of Mental Imagery, 1*, 109–125.

Roebuck, H., & Lupyan, G. (2020). The Internal Representations Questionnaire: Measuring modes of thinking. Behavior Research Methods, 52(5), 2053–2070.

Shen, X., Houser, T., Smith, D. V., & Murty, V. P. (2022). Machine-learning as a validated tool to characterize individual differences in free recall of naturalistic events. Psychonomic Bulletin & Review.

Stephan-Otto, C., Siddi, S., Senior, C., Muñoz-Samons, D., Ochoa, S., Sánchez-Laforga, A. M., & Brébion, G. (2017). Visual Imagery and False Memory for Pictures: A Functional Magnetic Resonance Imaging Study in Healthy Participants. *PLoS ONE, 12*(1), e0169551.

Zeman, A., Milton, F., Della Sala, S., Dewar, M., Frayling, T., Gaddum, J., Hattersley, A., Heurman-Williamson, B., Jones, K., MacKisack, M., & Winlove, C. (2020). Phantasia–The psychological significance of lifelong visual imagery vividness extremes. *Cortex, 130,* 426–440.