Unravelling mental representations in aphantasia through unsupervised alignment

Project design and study simulation

Maël Delem

Abstract

Research on aphantasia is confronted with a long-standing conundrum of all research on consciousness and representations, namely the theoretical inaccessibility of subjective representations. Drawing on concepts from similarity and representation research, I endorse the view that the study of an individual's mental representations is made possible by exploiting second-order isomorphism. The concept of second-order isomorphism means that correspondence should not be sought in the first-order relation between (a) an external object and (b) the corresponding internal representation, but in the second-order relation between (a) the perceived similarities between various external objects and (b) the similarities between their corresponding internal representations. Building on this idea, this study project report was divided into four parts. **First**, I outline the central ideas underlying similarity research and its applicability to aphantasia research. **Second**, I present a complete paradigm with an experimental design and a data analysis plan. The design will be based on multi-arrangement and inverse multidimensional scaling, a protocol that can be implemented online to conduct such large-scale research with high efficiency. The analysis plan will present a state-of-the-art method for similarity analysis, unsupervised alignment with Gromov-Wasserstein optimal transport (GWOT). **Third**, I report a data simulation I've done of a potential outcome of this study, and the successful analysis of this synthetic data using GWOT alignment. **Fourth**, I analyse the feasability of such a project given the material constraints of my thesis. I conclude with the expected utility and benefits of this project.

Table of contents

Work-In-Progress everywhere, *so look away! Shoo!*

I just wanted to try Quarto's new manuscript format with this personal project. Conclusion: it's ***incredibly cool***.

Project inception

This project stems from several elements:

1. The long standing knowledge of the fact that internal representations seem impossible to reach due to their subjective nature.
2. The discovery of the article of Shepard and Chipman (1970) that expose the idea of "second-order isomorphism".
3. The discovery of state-of-the-art and accessible unsupervised analytic methods to study this principle in an astonishing way. The last two discoveries (and many more) are the fruit of amazing discussions and recommendations from Ladislas when he came to the lab on Jan. 26. These motivated me to try to implement GWOT in R on data that I wanted to create myself to emulate a study we could do.

I promise that I did this mostly on my spare time, we have too many other things to do elsewhere.
Note: This website may seem very fancy. I wanted to take advantage of this personal project to try Quarto's new manuscript format for scientific editing. Conclusion: it's ***awesome****. It is very likely that I'll end up writing my thesis using Quarto's book format (through RStudio). This will allow me to render the raw text and computations as beautifully formatted PDF and Word documents with low effort, and eventually port it as a self-contained website when I'm authorized to share it openly… All with a single command, just like I did for this website.* ***This also means that you can read the present report on a PDF or Word if you wish to do so, the links are in the header****. You'll freeze the nice interactive figures though. As a bonus for the curious (or the reviewer), the "MECA Bundle" contains absolutely everything tied to this manuscript, well sorted, from the code scripts and configuration files to the final documents in all formats.* ***Awesome, I tell you****.*

1. Theoretical context

1.1. From similarity to second-order isomorphism

When we try to compare our thoughts and representations with those of others, we quickly realize that the task will be really difficult, if not impossible, as we are of course incapable of "living in someone else's head". If we both try to imagine a dog, I can examine what goes on in my head, so can you, but apart from trying to describe our experiences verbally, we are up against a wall.

Now, what if I asked you to tell me how similar you think a dog and a panther look like? Let's say, in the context of the animal kingdom as a whole. Visualize them well. Well, I could tell you that, *in my opinion, a dog and a panther look no more alike than a dog and a whale*. You might tell me:

> "***What on earth do you imagine a dog and a panther look like? Do you also think that a dog looks nothing like a cat? What goes on in your head?***"

…And many people probably agree with you. They mentally "see" and compare certain items the same way you do… And just like that, *we are back on track*. We managed to better "compare our thoughts"! And we even felt we could dive a bit into the "weird" representations of someone else.

*For this thought experiment, let's imagine two things: (1) that someone could honestly say that, and (2) that people would be rating the animals purely on the basis of their mental images, and not on categorical features (number of legs, fur, etc.), which is unfortunately almost **never** the case in reality.*

The study of individual differences in the format of representations and the attempt at understanding those of others obviously has a very rich history. It has interested many fields, in philosophy, linguistics, sociology, biology, psychology, or neuroscience, to name but a few. A myriad of ideas, concepts, models, methods, and paradigms have tried to deepen our understanding of representations and find the "key" to objectifying them. The principle I tried to illustrate with the thought experiment above is at the heart of one of these methods trying to unravel representations that was born in psychophysics[1]: the study of ***similarity.***

1.2. Work-In-Progress

1.3. Psychological spaces and aphantasia

While attempting to demonstrate the uselessness of the concept of similarity as a philosophical and scientific notion, Goodman (1972) has inadvertently expressed an aspect of similarity judgements of primary importance to us aphantasia researchers:

> Comparative judgments of similarity often require not merely selection of relevant properties but a weighting of their relative importance, and variation in both relevance and importance can be rapid and enormous. Consider baggage at an airport checking station. The

[1] Most notably in the works of Fechner (1860) and Mach (1890); see also Roads and Love (2024) for an extended review.

> spectator may notice shape, size, color, material, and even make of luggage; the pilot is more concerned with weight, and the passenger with destination and ownership. Which pieces are more alike than others depends not only upon what properties they share, but upon who makes the comparison, and when. . . . Circumstances alter similarities.

Goodman's claim was dismissed since then by propositions of robust mathematical models of similarity, e.g. Gardenfors (2004), Decock and Douven (2011).

This can be easily reversed as an argument in favor of the **potential of similarity analyses to highlight the inter-individual differences in sensory mental representations**. For example, should we ask individuals to judge the similarities in shape or color between various objects, the *differences between the similarity structures* of individuals will be precisely the most important phenomenon for us, far less than the constancy between these structures. If we can account for the context dependence, as we will propose here with explicit instructions, clever task design, and hypothesis-neutral analysis, we could overcome the limitations of the inherently subjective nature of similarity judgements.

This idea of a difference in similarity judgements in aphantasia seems to transpire in the results of Bainbridge et al. (2021) on their drawing study. They have shown that aphantasics had more schematic representations during recall, accurate in their spatial positioning, but with less sensory details. This difference can be seen from two perspectives: (1) a memory deficit for sensory properties; (2) a different representational structure of the items in their psychological spaces. In the latter case, aphantasics would have greater/faster abstraction of their representation of a perceived scene, reducing the amount of encoded sensory details unconsciously considered to be relevant. Both (1) and (2) can theoretically explain the same behavioural response, i.e. less sensory elements and correct spatial recall accuracy in aphantasic drawings, but **the two have drastically different consequences on how we define, characterize, and judge aphantasia.**

The dominant hypothesis seems to be that aphantasics simply have an episodic or general memory deficit. Conversely, I hypothesize that aphantasics have different representational structures than phantasics in certain dimensions of their psychological spaces (notably sensory, but potentially abstract too). More generally, I hypothesize that the concept of visual imagery evaluates in reality the continuous spectrum of representational structures in *sensory* dimensions of psychological spaces. Mirroring visual imagery, spatial imagery could also be a rough psychometric evaluation of the continuous spectrum of structural differences in *conceptual/abstract* dimensions of psychological spaces. In this view, the psychological space of aphantasics would constrain internal representations to particularly abstract forms from a very early stage, thus selectively limiting the item properties thereafter encoded in long-term memory. In other terms, **I hypothesize that aphantasia would not be characterized by an episodic memory deficit, but by an episodic memory *selectivity* caused by the specific characteristics of their representational structures and psychological spaces.** This selectivity would have, as we already

hypothesized several times, benefits and drawbacks.

Gardenfors (2004) proposed that differences in psychological (in his terms, conceptual) spaces could arise from various sources, whether innate, due to learning, or broader cultural or social differences. All these hypotheses could be coherent to explain the sources of aphantasia. Nevertheless, the study of these sources should be the subject of very large-scale or longitudinal studies, which are out of the scope of this project.

Here, we shall rather attempt to **develop a method to characterize the differences in aphantasics' representational structures and psychological spaces.**

1.4. The present project

2. Methods

Roads and Love (2024), in a recent review on the state and perspectives of similarity research, highlighted two challenges that studies in this field had to face: (1) The high cost of collecting behavioral data on a large number of stimuli; (2) The lack of software packages being a high barrier to entry, making the task of coding models difficult for the uninitiated.

To solve these problems, we present here two solutions, respectively for (1) experimental design and (2) data analysis:

1. A recent method to efficiently acquire similarity judgements, the "multiple arrangement of items" and "inverse multidimensional scaling" developed by Kriegeskorte and Mur (2012).
2. An accessible and robust Python toolbox provided by Sasaki et al. (2023) to conduct unsupervised alignment analysis using Gromov-Wasserstein optimal transport.

2.1. Experimental design

Multi-arrangement and inverse multidimensional scaling

Assuming a geometric model of representational similarities, Kriegeskorte and Mur (2012) developed a multi-arrangement (MA) method to efficiently acquire (dis)similarity judgments for large sets of objects. The subject has to perform multiple arrangements of item subsets adaptively designed for optimal measurement efficiency and for estimating the representational dissimilarity matrix (RDM) by combining the evidence from the subset arrangements.

The procedure is illustrated in Figure 1.

A key strength of this method that sets it as particularly effective is the "adaptive" part. The goal of the process is to acquire similarity judgements as precisely as possible while minimizing the total amount of trials. To do so, starting from the second trial, selected subsets of the items to be compared are presented to the subject: these items are the ones that were very close on-screen in previous trials and thus had their distance evaluated with lower accuracy by

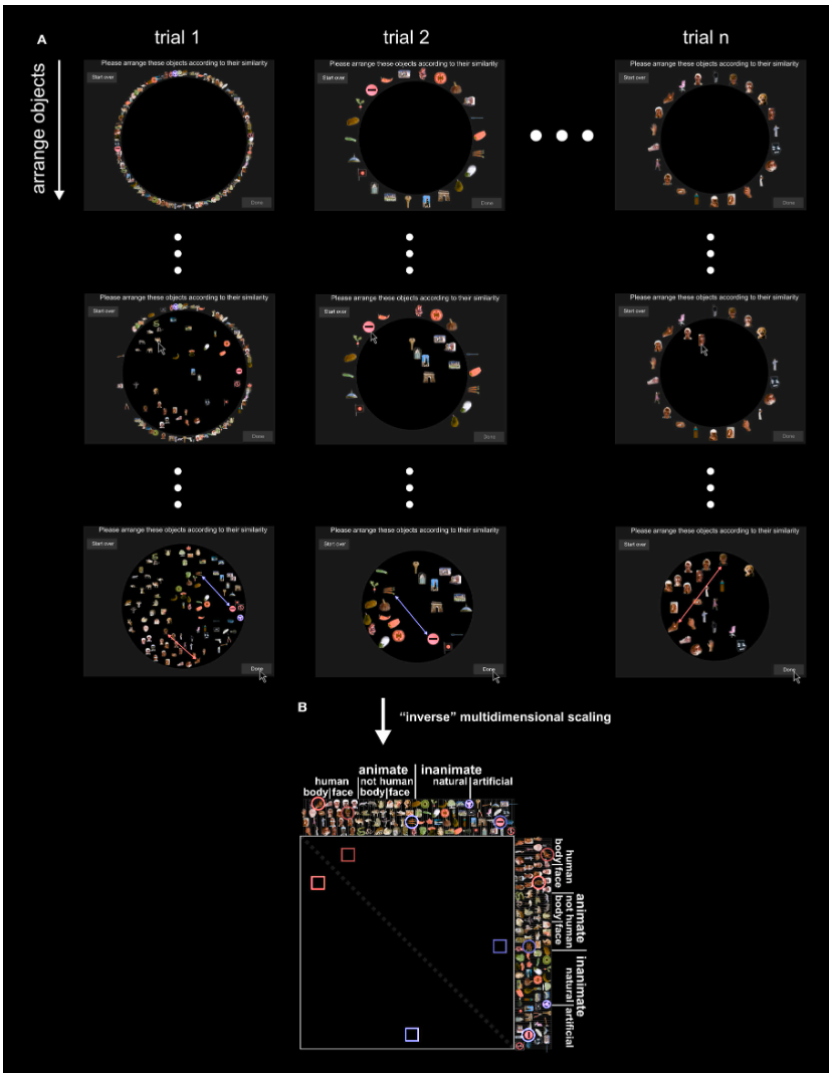


Figure 1: **Acquiring similarity judgements with the multi-arrangement method. (A)** Subjects are asked to arrange items according to their similarity, using mouse drag-and-drop on a computer. The similarity measure is taken as the distances between the items: similar items are closer, while dissimilar items are further apart. The upper part of the figure shows screenshots at different moments of the acquisition for one subject. Columns are trials and rows show the object arrangements over time, running from the start (top row) to the end (last row). The first trial contains all items; subsequent trials contain subsets of items that are adaptively selected to optimally estimate judged similarity for each subject. **(B)** Once acquisition of the final judgements is completed, inter-item distances in the final trial arrangements are combined over trials by rescaling and averaging to yield a single dissimilarity estimate for each object pair. The process is illustrated in this figure for two example item pairs: a boy's face and a hand (red), and carrots and a stop sign (blue). Their single-trial dissimilarity estimates (arrows) are combined into a single dissimilarity estimate, which is placed at the corresponding entry of the RDM (lower panel). Mirror-symmetric entries are indicated by lighter colors. Figure from Mur et al. (2013).

the subject. As the subject has to fill the entire "arena" with the items, these subsequent trials will necessarily increase the level of precision in the similarity judgement between pairs of items. The second key benefit of this method is the time and effort gain compared to others. For example, to compare every pair of items among 64 different items would require $\frac{64\times(64-1)}{2} = 2016$ comparisons (i.e. trials). This would be extremely time-consuming, while also losing the *context-independence* afforded by the MA method due to the presence of other items around every time the subject mentally performs a pairwise comparison.

Historically, when referring to the projection of the representations of stimuli (e.g., coordinates in geometric space) from a high-dimensional space into a lower-dimensional space, inference algorithms were commonly called multidimensional scaling (Roads and Love, 2024). By analogy, the process of combining several lower-dimensional (2D) similarity judgements on-screen to form one higher dimensional similarity representation (in the RDM) can be conceptually seen as "inverse" multidimensional scaling, hence the name given to the method by Kriegeskorte and Mur (2012).

Principle

The idea is simple: for a given set of items that have distinct and very pictorial visual properties, we would ask a wide range of aphantasics, phantasics or hyperphantasics to imagine, mentally compare and make similarity judgements between the items. To compare these representations with actual perceptual representations, the subjects would also perform the same task afterwards, this time with actual pictures to compare. Subjects would also fill our usual psychometric imagery questionnaires.

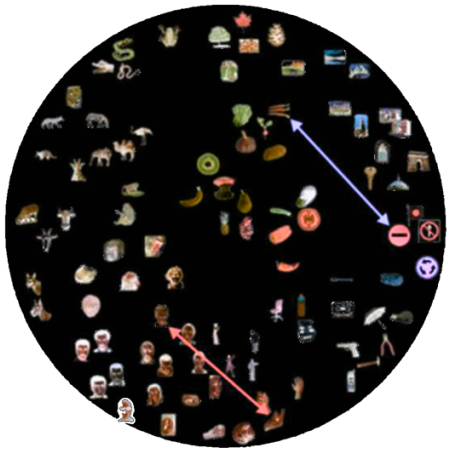
To "compare imagined items", we could use a "word" version of the MA paradigm. An example from Majewska et al. (2020) - *who used the method to build large-scale semantic similarity resources for Natural Language Processing systems* - is represented in Figure 2b.

We could have the stimuli rated by another set of participants on several features.

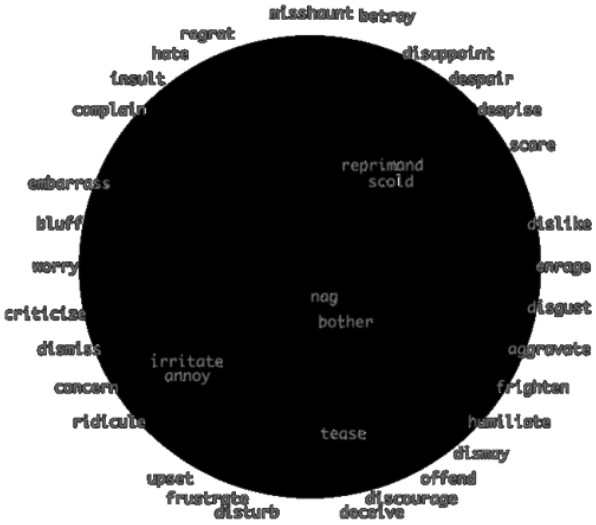
> « *We deliberately did not specify which object properties to focus on, to avoid biasing participants' spontaneous mental representation of the similarities between objects. Our aim was to obtain similarity judgments that reflect the natural representation of objects without forcing participants to rely on one given dimension. However, participants were asked after having performed the task, what dimension(s) they used in judging object similarity.* » (Jozwik et al., 2016)

> « ***All but one of the 16 participants reported arranging the images according to a categorical structure.*** » (Jozwik et al., 2017)

This result of Jozwik et al. (2017) suggests that we should give an explicit instruction about the features to focus on, otherwise everyone might bypass



(a) Arena layout of the MA protocol used by Mur et al. (2013) to acquire perceptual similarity judgements on natural images.



(b) Arena layout of the MA protocol used by Majewska et al. (2020) to acquire similarity judgements on word pairs.

Figure 2: Examples of arena layouts for the multi-arrangement (MA) paradigm.

visual features and mental images in favour of concepts and categories, regardless of their mental imagery profile.

In contrast, if we ask to focus specifically on the visual features, then ask subjects about the strategy they used to evaluate the similarities, then on the subjectively felt mental format of these strategies, we might grasp better insight on the sensory representations of subjects.

We could even go for several comparisons - even though this would increase quadratically the number of trials - e.g. :

- Evaluate to what extent the **shape** *of these animals are* ***similar*** **at rest, ignoring size differences.**
- Evaluate to what extent these animals **sound like each other.**
- Etc.

  Note to be added: if you do not know the animal, just guess its placement, as this situation is quite unlikely to happen (animals chosen are fairly common knowledge).

Kawakita et al. (2023): To assess whether the color dissimilarity structures from different participants can be aligned in an unsupervised manner, we divided color pair similarity data from a large pool of 426 participants into five participant groups (85 or 86 participants per group) to obtain five independent and complete sets of pairwise dissimilarity ratings for 93 color stimuli (Fig. 3a). Each participant provided a pairwise dissimilarity judgment for a randomly allocated subset of the 4371 possible color pairs. We computed the mean of all judgments for each color pair in each group, generating five full dissimilarity matrices referred to as Group 1 to Group 5.

Stimuli

We would have a list of animal items, that would have several characteristics:

- A name
- A category
- A shape

We need orthogonal data:

- Each class of animal should include each shape (roughly)
- Each shape should have an animal

This would imply that category cannot be derived from shape, and vice-versa. Thus, a **sorting by shape would reveal to be innately visual** (or maybe spatial, if shape concerns this type of imagery), and a **sorting by category would reveal an abstraction** from these shapes. We expect that the two will be mixed to some degree in every subject, but that low-imagery would rather

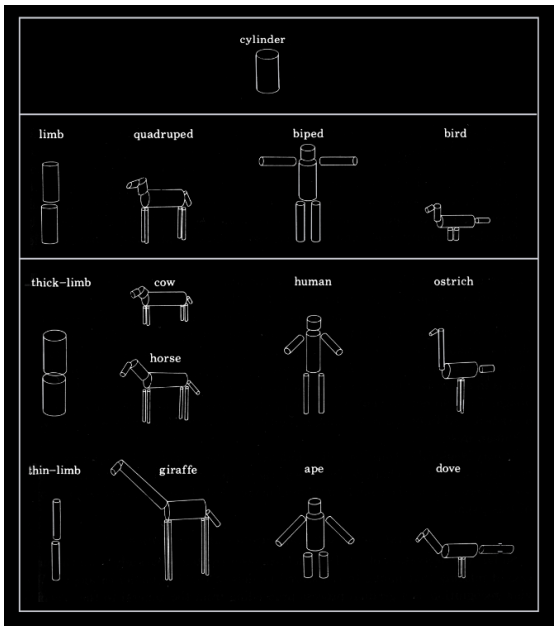


Figure 3: Representing the characteristics of shapes with cylinders. Figure from Marr et al. (1997). *Click to expand.*

tend towards category sorting, while high-imagery would tend towards shape sorting.

Shapes could be very tricky stimuli to discuss. Gardenfors (2004) noted that we only have a very sketchy understanding of how we perceive and conceptualize things according to their shapes. The works of Marr et al. (1997) highlight this difficulty when analysing the complexity of the hierarchical judgements of shapes and volumes, as shown in Figure 3.

2.2. Data analysis plan

Unsupervised alignment rationale

Visual images can be represented as points in a multidimensional psychological space. Embedding algorithms can be used to infer latent representations from human similarity judgments. While there are an infinite number of potential visual features, an embedding algorithm can be used to identify the subset of salient features that accurately model human-perceived similarity. (*From Roads' CV*)

Using an optimization algorithm, the free parameters of a psychological space are found by maximizing goodness of fit (i.e., the loss function) to the observed data. Historically, when referring specifically to the free parameters that correspond to the representation of stimuli (e.g., coordinates in geometric space),

inference algorithms were commonly called multidimensional scaling (MDS), or simply scaling, algorithms.

In the machine learning literature, analogous inference algorithms are often called embedding algorithms. The term "embedding" denotes a higher-dimensional representation that is embedded in a lower-dimensional space. For that reason, the inferred mental representations of a psychological space could also be called a psychological embedding.

Numerous techniques exist, and each has limitations. Popular techniques for comparing representations include RSA Kriegeskorte et al. (2008) and canonical correlation analysis (CCA) (Hotelling 1936). Briefly, RSA is a method for comparing two representations that assesses the correlation between the implied pairwise similarity matrices. CCA is a method that compares two representations by finding a pair of latent variables (one for each domain) that are maximally correlated.

One might be tempted to compare two dissimilarity matrices assuming stimulus-level "external" correspondence: my "red" corresponds to your "red"(Fig. 1d). This type of supervised comparison between dissimilarity matrices, known as Representational Similarity Analysis (RSA), has been widely used in neuroscience to compare various similarity matrices obtained from behavioural and neural data. However, there is no guarantee that the same stimulus will necessarily evoke the same subjective experience across different participants. Accordingly, when considering which stimuli evoke which qualia for different individuals, we need to consider all possibilities of correspondence: my "red" might correspond to your "red", "green", "purple", or might lie somewhere between your "orange" and "pink"(Fig. 1e). Thus, we compare qualia structures in a purely unsupervised manner, without assuming any correspondence between individual qualia across participants.

Gromov-Wasserstein optimal transport

To account for all possible correspondences, we use an unsupervised alignment method for quantifying the degree of similarity between qualia structures. As shown in Fig. 2a, in unsupervised alignment, we do not attach any external (stimuli) labels to the qualia embeddings. Instead, we try to find the best matching between qualia structures based only on their internal relationships (see Methods). After finding the optimal alignment, we can use external labels, such as the identity of a color stimulus (Fig. 2b), to evaluate how the embeddings of different individuals relate to each other. This allows us to determine which color embeddings correspond to the same color embeddings across individuals or which do not. Checking the assumption that these external labels are consistent across individuals allows us to assess the plausibility of determining accurate inter-individual correspondences between qualia structures of different participants.

To this end, we used the Gromov-Wasserstein optimal transport (GWOT) method, which has been applied with great success in various fields. GWOT

aims to find the optimal mapping between two point clouds in different domains based on the distance between points within each domain. Importantly, the distances (or correspondences) between points "across" different domains are not given while those "within" the same domain are given. GWOT aligns the point clouds according to the principle that a point in one domain should correspond to another point in the other domain that has a similar relationship to other points. The principle of the method is illustrated in Figure 4

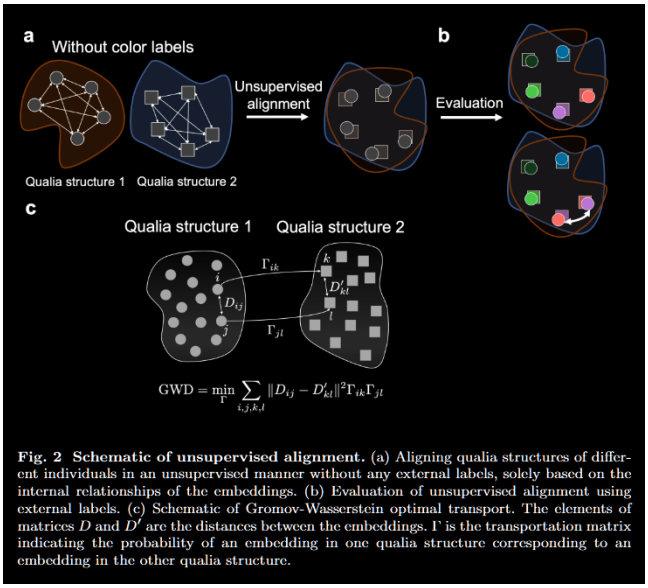


Fig. 2 Schematic of unsupervised alignment. (a) Aligning qualia structures of different individuals in an unsupervised manner without any external labels, solely based on the internal relationships of the embeddings. (b) Evaluation of unsupervised alignment using external labels. (c) Schematic of Gromov-Wasserstein optimal transport. The elements of matrices D and D' are the distances between the embeddings. Γ is the transportation matrix indicating the probability of an embedding in one qualia structure corresponding to an embedding in the other qualia structure.

Figure 4: Gromov-Wassertein optimal transport principle. Figure from Kawakita et al. (2023). *Click to expand.*

We first computed the GWD for all pairs of the dissimilarity matrices of the 5 groups (Group 1-5) using the optimized ϵ. In Fig. 3b, we show the optimized mapping $\Gamma*$ between Group 1 and Groups 2-5 (see Supplementary Figure S1 for the other pairs). As shown in Fig. 3b, most of the diagonal elements in $\Gamma*$ show high values, indicating that most colors in one group correspond to the same colors in the other groups with high probability. We next performed unsupervised alignment of the vector embeddings of qualia structures. Although $\Gamma*$ provides the rough correspondence between the embeddings of qualia structures, we should find a more precise mathematical mapping between qualia structures in terms of their vector embeddings to more accurately assess the similarity between the qualia structures. Here, we consider aligning the embeddings of all the groups in a common space.

By applying MDS, we obtained the 3-dimensional embeddings of Group 1 and Groups 2-5, referred to as X and Yi, where i = 2, ..., 5 (Fig. 3c). We then aligned Yi to X with the orthogonal rotation matrix Qi, which was obtained by solving a Procrustes-type problem using the optimized transportation plan

$\Gamma*$ obtained through GWOT (see Methods). Fig. 3d shows the aligned embeddings of Group 2-5 (QiYi) and the embedding of Group 1 (X) plotted in the embedded space of X. Each color represents the label of a corresponding external color stimulus. Note that even though the color labels are shown in Fig. 3d, this is only for the visualization purpose and the whole alignment procedure is performed in a purely unsupervised manner without relying on the color labels. As can be seen in Fig. 3d, the embeddings of similar colors from the five groups are located close to each other, indicating that similar colors are 'correctly' aligned by the unsupervised alignment method.

To evaluate the performance of the unsupervised alignment, we computed the k-nearest color matching rate in the aligned space. If the same colors from two groups are within the k-nearest colors in the aligned space, we consider that the colors are correctly matched. We evaluated the matching rates between all the pairs of Groups 1-5. The averaged matching rates are 51% when k = 1, 83% when k = 3, and 92% when k = 5, respectively. This demonstrates the effectiveness of the GW alignment for correctly aligning the qualia structures of different participants in an unsupervised manner.

However, as can be seen in Fig. 4b, the optimized mapping $\Gamma*$ is not lined up diagonally unlike the optimized map- pings between color-neurotypical participants groups shown in Fig. 3b (see Supplementary Figure S1 for the other pairs). Accordingly, top k matching rate between Group 1-5 and Group 6 is 3.0% when k = 1 (Fig. 4c), which is only slightly above chance ($\approx$ 1%). The matching rate did not improve even when we relaxed the criterion (6.9% and 11% for k = 3 and k = 5, respectively). Moreover, all of the GWD values between Group 1-5 and Group 6 are larger than any of the GWD values between color-neurotypical participant groups (Fig. 4d).

These results indicate that the difference between the qualia structures of neuro-typical and atypical participants is significantly larger than the difference between the qualia structures of neuro-typical participants.

2.3. Paradigm summary

The experimental design and data analysis plans are succinctly summarised in Figure 5.

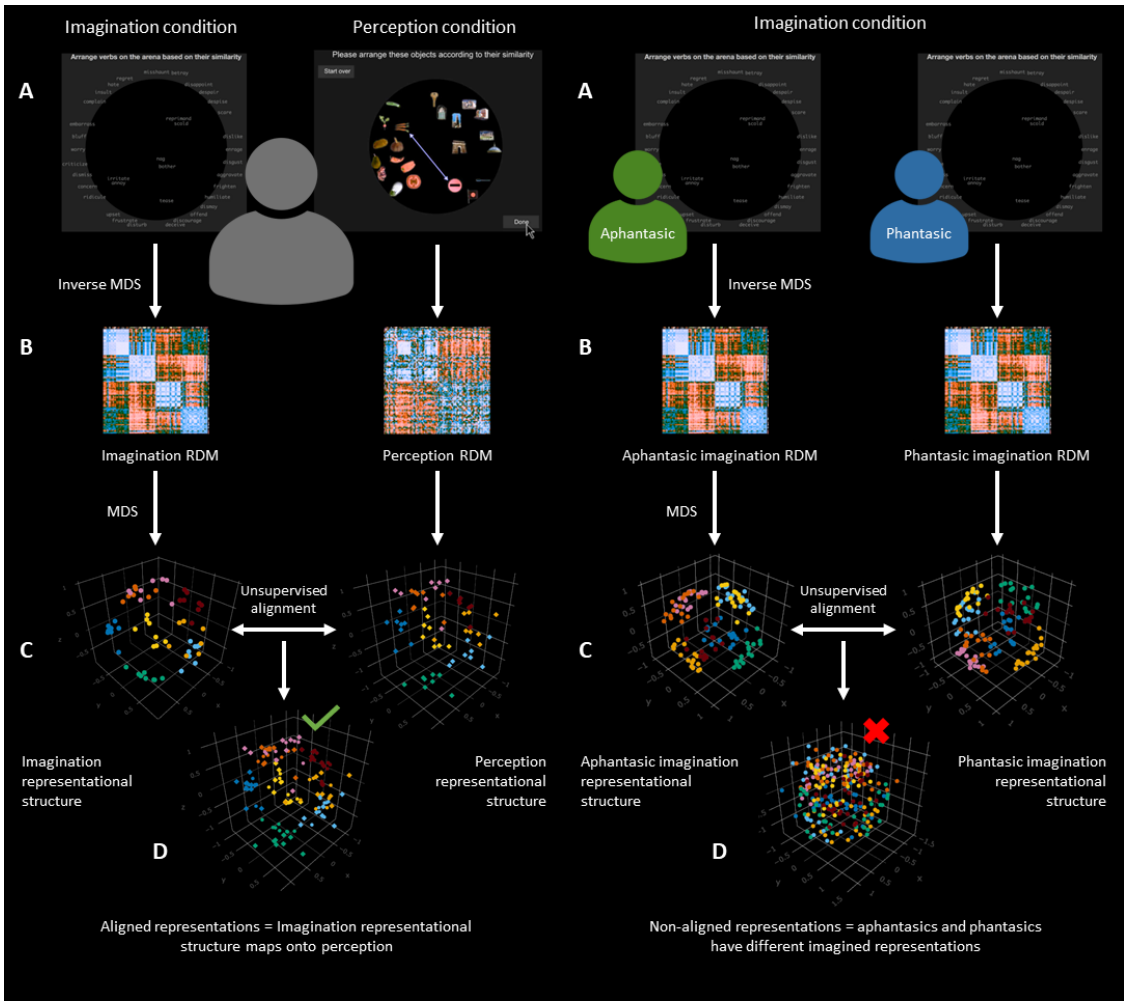
2.4. Hypotheses

Aphantasic and phantasic psychological spaces

The most representative members of a category are called prototypical members.

Prototype theory builds on the observation that among the instances of a property, some are more representative than others. The most representative one is the prototype of the property.

Thus, following the concepts illustrated by Gardenfors (2004), we would expect that aphantasics, when doing shape similarity judgements, would be more



(a) (b)

Figure 5: Summary schematics of the proposed experimental protocol and data analysis plan. *Click on the sub-figures to expand them.* Figure 5a represents the two conditions to be completed by each subject. These two conditions will allow to compute comparisons (alignments) within a subject's own perceptual and imaginal representational structures, but also between subjects (or groups) for each modality (see the next figure's description). **(A)** The subject performs two simililarity judgement tasks using the MA paradigm presented earlier. **(B)** The low-dimensional similarity judgements are converted to a high-dimensional Representational Dissimilarity Matrix (RDM) through inverse-MDS as a follow-up to extract the results of the MA. **(C)** The RDMs are then reduced in dimensionality once again to extract relevant dimensions reflecting inferred features of the items through MDS, yielding embeddings. Three-dimensional projections of these embeddings have been chosen here for visualization purposes. **(D)** These embeddings are compared through unsupervised alignment using GWOT, which results in an estimate of the degree of alignment of the two representational structures and in coordinates of aligned embeddings. These coordinates allow us to examine the 3D visualization shown here and judge by ourselves the "look" of the alignment. Here the perception representation aligns with the imagination one, from which we could infer that imagined representations are made of sensory (rather than abstract properties). We expect inter-individual variability in these perception-imagination alignments, as shown in the next figure. Figure 5b represents the comparison between the representational structure of different cognitive profiles. In practice, all pairs of subjects will be compared to assess their representational structure alignments, independently of arbitrary groups. This is computationally heavy, but analytically very powerful. This figure also tacitly shows an idea supporting the use of unsupervised alignment: it is possible that RDMs seem to be very correlated and similar, as shown in step **(B)**, but do not align when compared without supervision. This contrasts with several supervised alignment methods (such as RSA, see Kriegeskorte et al., 2008) which usually use the RDM as-is. This difference is due to the involvement of labels for items that are already known by the researcher to correlate the RDMs, whereas unsupervised algorithms such as GWOT are only concerned with the structures. This principle is eloquently illustrated by Figure 4 from Kawakita et al. (2023).

inclined to group items close to the prototypical items due to a lower definition of the mental image. In comparison, phantasics would have a much more distributed conceptual space of item shapes due to their higher-resolution mental images of said items.

Subjective imagery and psychological spaces

In the proposed view of visual imagery as the subjective expression of a given type of psychological space, we mentioned earlier that *spatial* imagery could also constitute a subjective expression of other dimensions of psychological spaces. Hence, the *verbal* dimension of the simplified model of imagery we outlined in my thesis project could also represent different dimensions.

This conception leads to the following theoretical hypothesis: provided that our visual-spatial-verbal model correctly fits subjective imagery, the imagery profile of individuals should map on their psychological spaces.

Operationally, this would be evaluated by the fact that **individuals with similar imagery profiles** (visual, spatial, verbal, or any combination of the three) **should have similar representations** in their given psychological space, **quantifiable by the degree of alignment between their similarity structures.**

3. Study simulation results

Source: Article Notebook

3.1. Visual-spatial-verbal model of cognitive profiles

One of the objectives of the study would be to link the subjective cognitive profiles of individuals with their representational structures. To evaluate these profiles, we are going to use psychometric questionnaires evaluating the visual-object, spatial, and verbal dimensions of imagery which will yield three scores, one for each dimension.

We are going to simulate 30 participants presenting four different cognitive profiles, that I defined as, respectively, *verbal* aphantasics, *spatial* aphantasics, *spatial* phantasics, and *visual* phantasics. Their imagery abilities are summarised in Table 1.

To simulate these four sub-groups, we will generate multivariate normal distributions of scores on these three dimensions for each sub-group. For instance, verbal aphantasics have normally distributed visual imagery scores centred around a mean of 0 (normalized, so negative scores are possible), 0.4 for spatial imagery, and 0.7 for verbal style; Spatial aphantasics have means of 0 for visual, 0.75 spatial, and 0.3 for verbal; etc. The numbers are arbitrary, but have been chosen by trial-and-error to obtain a model that is both well-defined and not exaggerated. The 30 subjects' imagery profiles are represented in the three dimensional space of the visual-spatial-verbal dimensions in Figure 6.

Table 1: Imagery abilities of the four hypothesized cognitive profiles.

Cognitive profile	Visual imagery	Spatial imagery	Verbal style
Verbal aphantasic	–	-	++
Spatial aphantasic	–	++	-
Spatial phantasic	+	++	-
Visual phantasic	++	-	+

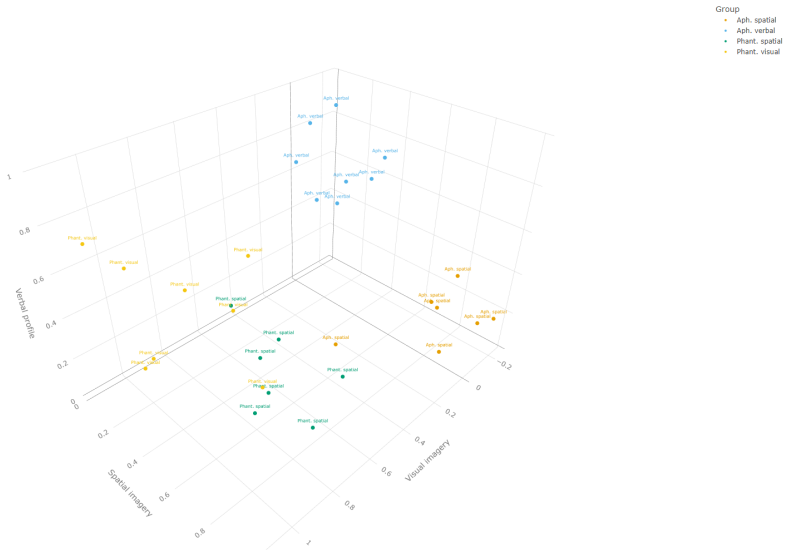


Figure 6: Imagery profiles generated for 30 subjects on the three object, spatial, and verbal dimensions.

Source: Article Notebook

3.2. Data simulation: Creating representational structures

Gardenfors (2004) invokes two scientific concepts, to wit, prototypes and Voronoi tessellations. Prototype theory builds on the observation that among the instances of a property, some are more representative than others. The most representative one is the prototype of the property. *We hypothesize that aphantasics will be more inclined to categorize items according to prototypes than phantasics.*

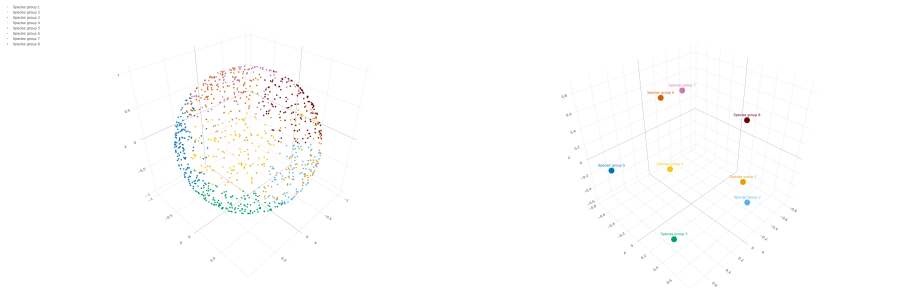
A Voronoi tessellation of a given space divides that space into a number of cells such that each cell has a center and consists of all and only those points that lie no closer to the center of any other cell than to its own center; the centers of the various cells are called the generator points of the tessellation. This principle

will underlie our data simulation, as we will build representations in a 3D space based on distances to "centroids", namely, prototypes. These representations will thus be located inside of the tessellations around these prototypes, more or less close to the centroid depending on the subject's representational structures.

Generating "prototype" embeddings from a sphere

Source: Article Notebook

A function will be used to generate embeddings. These spherical embeddings are displayed in Figure 7. We get 8 nicely distributed clusters. We'll retrieve the centroids of each cluster, which would be the "perfect" categories of each species group (say, generated by a computational model on categorical criteria).



(a) Generated spherical distribution of 1000 observations grouped in 8 equal clusers with Gaussian Mixture Clustering.

(b) Centroids of the 8 clusters created on the sphere.

Figure 7: Initial random generations of 1000 points grouped in 8 clusters to represent the theoretical embeddings of 8 groups (i.e. groups of species here). ***Interact with the figures to see the details.***

Now we want two sets of embeddings: one where the observations are very concentrated around the centroids, which would be the **categorical model**, and one where the observations are more spread out, which would be the **visual model**.

We need to select 8 observations per cluster, which would be our animals per group. These observations will be subsets of the 1000 observations we generated.

Categorical model embeddings

The selection procedure for the **categorical model** will consist of selecting points that are rather *close to the centroids*. Thus, we will filter the observations of the large sets to keep only points for which the distance to the centroid is inferior to a given value. That is, points for which the Euclidean norm of the vector from the observation to the centroid:

$$d(centroid, observation) = \sqrt{(x_c - x_o)^2 + (y_c - y_o)^2 + (z_c - z_o)^2}$$

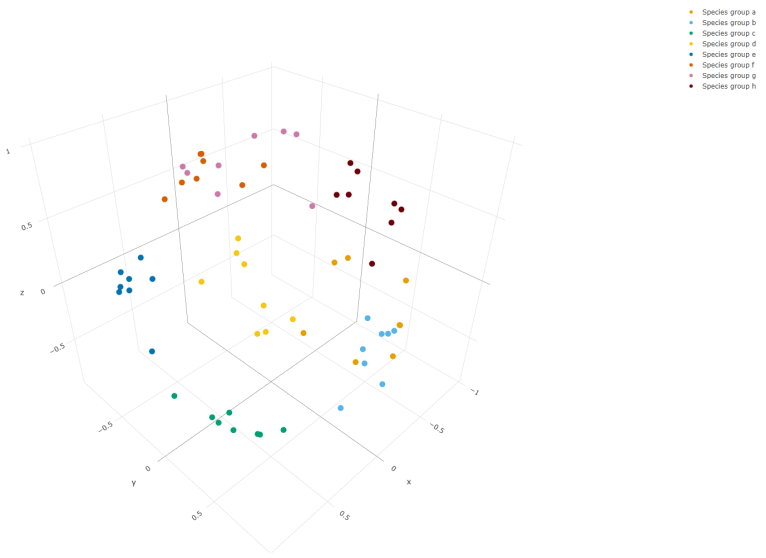


Figure 8: Selection of 64 points to represent prototypical categorical embeddings, based on the distances to each groups' centroid. These will be the bases of the verbal aphantasics' embeddings. ***Interact with the figure to see the details.***

Source: Article Notebook

Visual model embeddings

In the case of the **visual model**, we would like approximately evenly distributed embeddings, that could also dive *inside* the sphere, i.e. representing species that are visually close although diametrically opposed when it comes to taxonomy. To do this we can simulate multivariate normal distributions around the centroids.

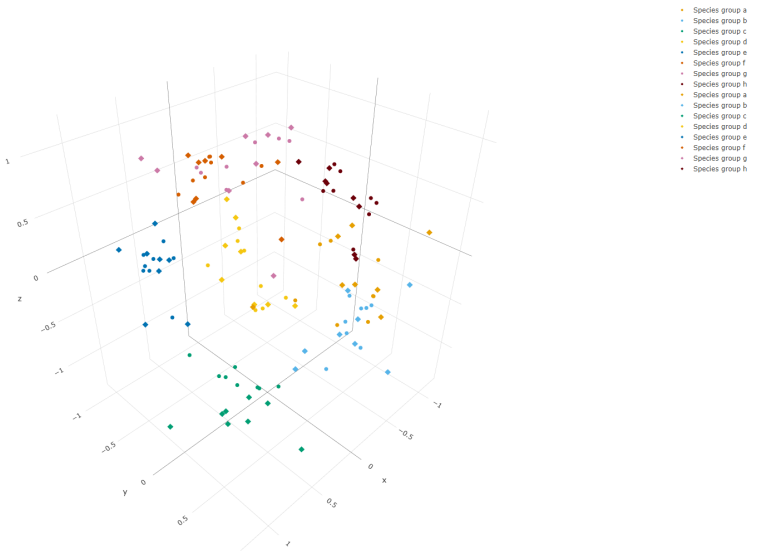


Figure 9: Selection of 64 points to represent prototypical visual embeddings, chosen randomly in multivariate distributions centered around each categorical embedding. The visual embeddings are overlaid as diamonds along with categorical ones as dots. The two distributions keep the group structure, but are pretty far apart at times. ***Interact with the figure to see the details.***

Source: Article Notebook

Intermediate embeddings

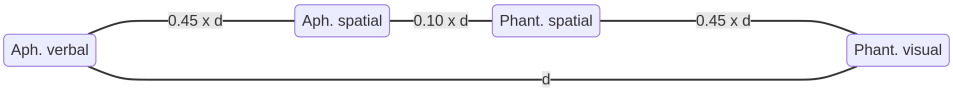


Figure 10: Model of the distances between participants' representations. Note that here d is a one-dimensional distance between the representations, but it will be computed as a three-dimensional distance in our toy-model. The verbal aphantasic profile is hypothesized to be very categorical, thus diametrically opposed to the visual phantasic profile, by a given distance d. Spatial profiles are in-between: they are close to each other (10% x d), but the spatial aphantasic profile is a bit closer to the verbal aphantasic one (45% x d), and the spatial phantasic is a bit closer to the visual phantasic one (45% x d).

Source: Article Notebook

Source: Article Notebook

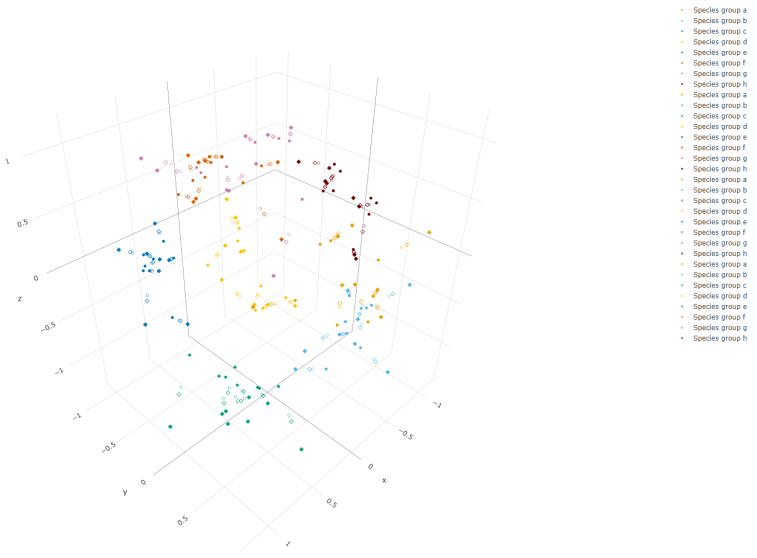


Figure 11: Space of embeddings with 128 additional points based on the euclidean distances between the visual and categorical embeddings. The empty dots are the *aphantasics-spatial* ones, and the empty diamonds are the *phantasic-spatial* ones. Some can be very close together, and sometimes further apart due to the various pairs of visual and categorical points used to create them. A network-like structure seems to appear, with empty points seemingly 'connecting' the dots and diamonds. ***Interact with the figure to see the details.***

Source: Article Notebook

The distributions created are still gathered around the centroids of each group, but they are much more widespread, each group getting close to each other and even reaching inside the sphere.

Perfect! Now we have two 3D embeddings per animal, in a categorical or a visual description of their features. Thus, we have four sets of coherent coordinates, around which we will simulate the embeddings of the 30 participants, depending on their groups.

Source: Article Notebook

Generating the subject embeddings

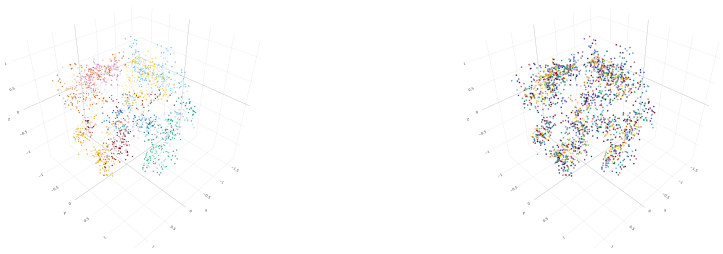
We have four "reference" sets of embeddings which represent animals either judged according to their similarity in categorical terms (namely, species), or in visual terms (namely shape or color similarities, assuming that these similarities are more evenly distributed, e.g. the crab looks like a spider, but is also pretty close to a scorpion, etc.).

To generate the embeddings of each subject in each condition, we will start from these reference embeddings and generate random noise around *each item*,

i.e. for all 64 animals. For 100 subjects, we would thus generate 100 noisy points around each animal, each point corresponding to a given subject.

The visual and verbal groups will be generated with slightly more intra-group variance, so as to try to make the spatial groups as coherent as possible (and avoid blurring everything and making the groups disappear in noise).

Although the groups and species in Figure 12a look fairly obvious when we colour the embeddings using the knowledge about how we built them, the algorithm will only be fed with the data for each subject, without any labelling or additional information. Thus, Figure 12b is what the algorithm will actually "see" (and what it will try to decrypt). Said otherwise, its objective will be to find all the correct colours and shapes in Figure 12a using only 30 sub-datasets (one for each subject) that are illustrated in Figure 12b. Admittedly, that looks a lot more complicated.



(a) Distribution of the embeddings of the 30 sub-jects,***colored by the species groups*** they represent.The symbols represent the four imagery groups (Aph. verbal, spatial, etc.)

(b) Distribution of the embeddings of the 30 subjects, ***colored by subject***.This is the only information the unsupervised algorithm will have to work with.

Figure 12: Final distribution of the 64 embeddings of all the 30 subjects, amounting to 1920 points total. ***Interact with the figures to see the details.***

Source: Article Notebook

Source: Article Notebook

3.3. Data analysis: Aligning representational structures

For all this section, we need to adapt a simple version of the explanations from Kawakita et al. (2023) and Sasaki et al. (2023) and avoid any technical aspects in the main manuscript.

From there on, most of the code follows the instructions from the open-source scientific toolbox by Sasaki et al. (2023). I added a few explanations on the purpose of each step, without diving into unnecessary details.

Step 1: Importing the embeddings in the Python instances
Step 2: Setting the parameters for the optimization of GWOT
Step 3: Gromov-Wasserstein Optimal Transport (GWOT) between Representations
Step 4: Evaluation and Visualization
Clustering the subjects by alignment accuracy.

First, we evaluate the accuracy per subject and group the subjects based on the alignment accuracy via hierarchical clustering. This procedure is represented in Figure 13. Second, we evaluate the accuracy of the alignment between these clusters.

Source: Article Notebook

Source: Article Notebook

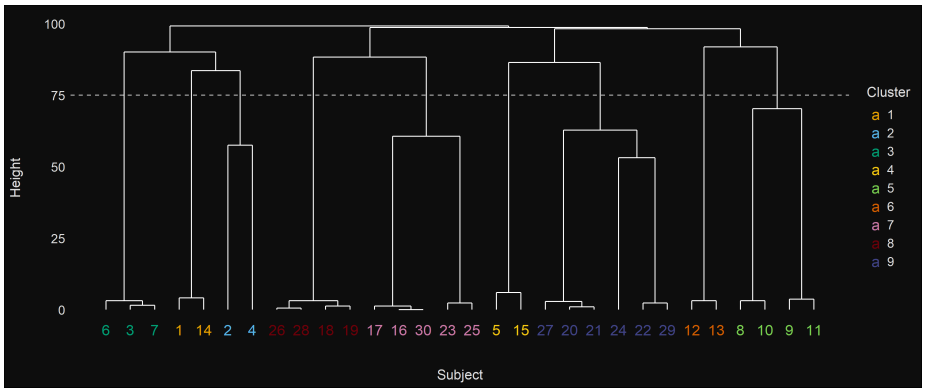


Figure 13: Hierachical clustering of the 30 subjects based on their representational alignment.

Source: Article Notebook

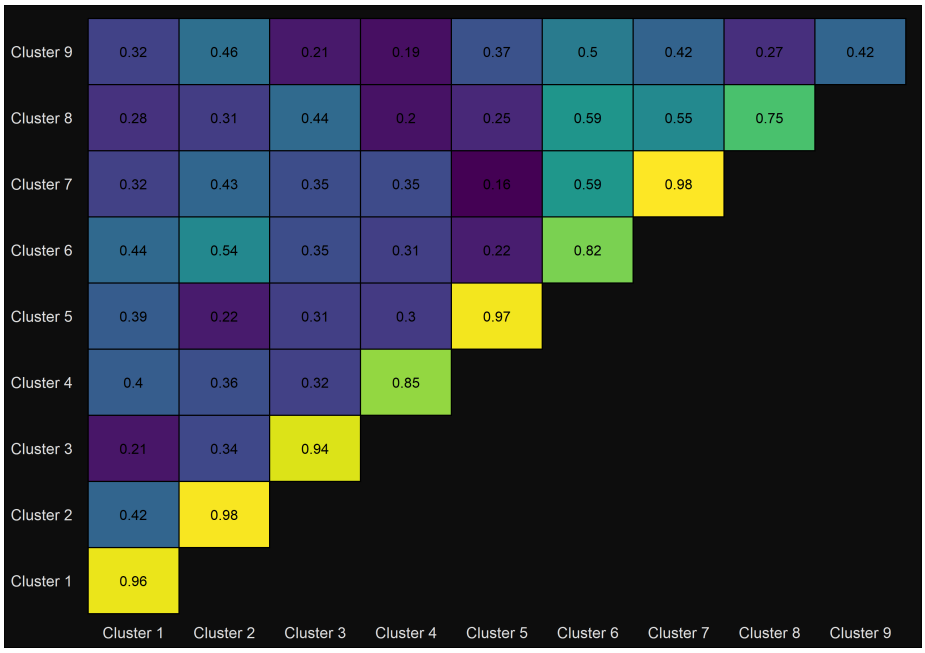


Figure 14: Accuracy of the alignments between the subject's embeddings in each cluster. An alignment of a cluster with itself (e.g. Cluster 7 - Cluster 7) is the evaluation of the alignment of the subjects ***inside*** the cluster.

Source: Article Notebook

Evaluating the clusters in light of our theoretical OSV model.

Let's see the composition of the clusters in light of our initial O-S-V model. The cognitive profiles of the subjects are represented in Figure 15, and the distribution of the cognitive profiles in the clusters is represented in Figure 16.

Source: Article Notebook

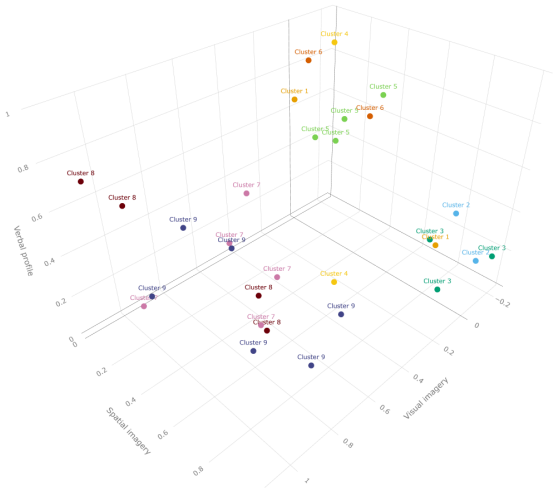


Figure 15: Imagery profiles of the nine identified clusters on the three object, spatial, and verbal dimensions. ***Interact with the figure to see the details.***

Source: Article Notebook

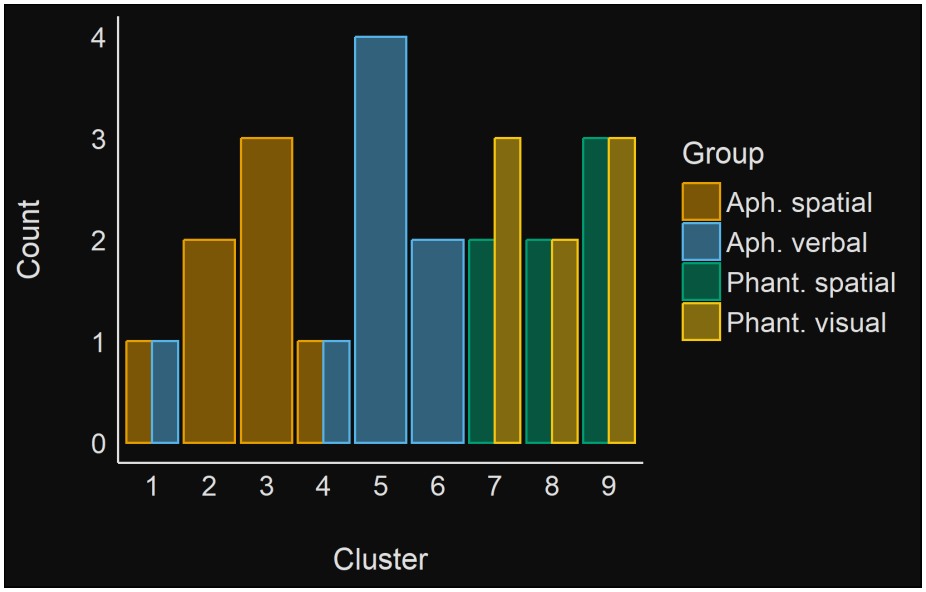


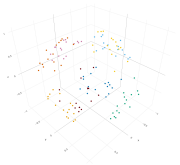
Figure 16: Repartion of our intial O-S-V groups in the clusters created by the unsupervised alignment.

Source: Article Notebook

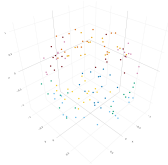
Now let's visualize the embeddings of the subjects in each cluster to get a visual idea of their representational structures and the intra-cluster alignment between the subjects.

Let's try referencing Figure 17a, or Figure 17i, the last one.

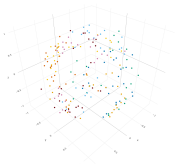
Figure 17: Psychological spaces (embeddings) of the 30 subjects, aligned and clustered with other subjects having the most similar representations. The eight colors represent the initial eight groups of species that each subject had to 'represent' with imagery (legends for these colors have been taken out for display clarity purposes). ***Interact with the figures to see the details.***



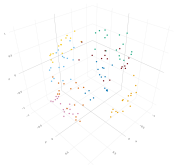
(a) **Cluster 1** embeddings (Aphantasic cluster).Within-cluster aligment accuracy = 95.83%.



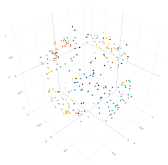
(b) **Cluster 2** embeddings (Spatial aphantasic cluster).Within-cluster aligment accuracy = 97.66%.



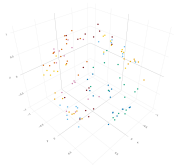
(c) **Cluster 3** embeddings (Spatial aphantasic cluster).Within-cluster aligment accuracy = 94.01%.



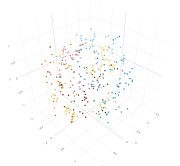
(d) **Cluster 4** embeddings (Aphantasic cluster).Within-cluster aligment accuracy = 85.03%.



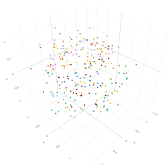
(e) **Cluster 5** embeddings (Verbal aphantasic cluster).Within-cluster aligment accuracy = 96.88%.



(f) **Cluster 6** embeddings (Verbal aphantasic cluster).Within-cluster aligment accuracy = 81.80%.



(g) **Cluster 7** embeddings (Phantasic cluster).Within-cluster aligment accuracy = 98.23%.



(h) **Cluster 8** embeddings (Phantasic cluster).Within-cluster aligment accuracy = 74.64%.

(i) **Cluster 9** embeddings (Phantasic cluster).Within-cluster aligment accuracy = 42.45%.

3.4. Summary of the simulation analysis

- We generated subject data on subjective imagery based on the Object-Spatial-Verbal model of cognitive styles and representations

- We translated this model into a model of the distances between participants' mental representations
- We generated random data based on this theoretical model, on the representations of 64 items with categorical and visual properties by 30 participants
- We used an unsupervised alignment algorithm to judge the similarity between the representations of the subjects without any knowledge of their initial groups and relations
- The algorithm aligned with high precision 9 clusters of participants, which were coherent with the initial model we created, with several differences and unexpected alignments due to the randomization.
- The 9 clusters revealed a distinction between verbal aphantasics, spatial aphantasics, and phantasics in general. This interesting result shows that even though we tried to model spatial aphants and phantasics closer to each other, they all ended up separated based on visual imagery. This unexpected outcome, that went besides our initial intentions, shows that such an unsupervised method could reveal coherent patterns of representations that we did not expect, even with a relevant psychometric model.

This simulation motivates the idea that, should the imagery of participants be accurately fitted by our OSV model (or any other model to be tested), this paradigm and analytic method would be able to align the representations of participants with the same subjective imagery abilities.

I insist on a key finding : I did not use the data of the OSV model presented in Section 3.1 to generate the subject embeddings. The only hypothesis that guided how I simulated the subjects' embeddings was the distance model I envisioned, which is represented in Figure 10. Thus, **the algorithm managed to reverse-engineer my logic**, to find the subjects groups I simulated with this logic, and it so happens that these groups' matched the cognitive profiles groups I built in the beginning.

In other words, the algorithm managed to find the common pattern - which was the groups pattern - between two models built differently, ***a common pattern that existed originally only in my head.***

I think this "mind-reading"[2] further argues for the potential of this method to reveal hidden patterns of inter-individual differences in subjective experience. These patterns could help build models of subjective mental imagery, one of the most challenging tasks in cognitive psychology to date.

[2]Or, less prettily put, "this unsupervised extraction of hidden representational features".

4. Feasibility

4.1. Stimuli and study design

The simulation study presented here focused on aligning the representations ***between*** various participants, but a real study should go further and also analyse the similarities of representations ***within*** participants, for instance with a perception and an imagery condition. This was the basis of the study of Shepard and Chipman (1970), and remains a good starting point to design our own.

4.2. Online study materials

4.3. Analytic methods and collaborations

I have proved (*mostly to myself*) that I was capable of implementing unsupervised GWOT alignment analysis in Python and R using the open-source toolbox provided by Sasaki et al. (2023), firstly to demonstrate that this key feasibility aspect was not out of reach (and that I was prepared to handle it). This toolbox is very recent (the associated article was posted on bioRχiv last September) but is based on long-standing theory on similarity and cutting-edge topology analysis research. Consequently, I may not be confident enough on my expertise in these fields to state with confidence that my analyses of this project's data would be solid, which is a very important aspect for me. I need to be convinced by analytical choices, which are often taken for granted, to believe in their results and implications. I'm convinced of the relevance of what I've done here, but in the context a real application, even more expertise (and other analytical points of view) would be most welcome.

Therefore, I think that this project could be the opportunity to collaborate with several teams working in these fields that have great data analysis expertise.

- Starting of course with **Ladislas Nalborczyk** (who gave me this idea), who works on synesthesia and inner speech aphantasia at the *Paris Brain Institute* with **Laurent Cohen** and **Stanislas Dehaene**.
- **Nikolaus Kriegeskorte**, one of the creators of the famous Representational Similarity Analysis (RSA, another *supervised* alignment method, see Kriegeskorte et al., 2008) and his colleagues could be precious collaborators for alignment analyses and study materials.
- The Japanese team of **Masafumi Oizumi** behind the GWOT toolbox is of course also very knowledgeable on the subject, the method, and its technical implementation.
- They are collaborating with the Australia-based team of **Naotsugu Tsuchiya**, with whom they recently published several very interesting articles on similarity as a concept and method for perception research (e.g. Tsuchiya et al., 2022; Kawakita et al., 2023; Kawakita et al.; Zeleznikow-Johnston et al., 2023).

- Visiting Tsuchiya's webpage, I also found an amazing chain of connections that lead us to his team. Tsuchiya also works on sleep and dreams and has collaborated several times with **Thomas Andrillon**, who works at the *Paris Brain Institute*, thus close to Dehaene, Cohen, Bartolomeo, and Nalborczyk, and is very technically knowledgeable. Even more interestingly, Andrillon is a co-author of one of Alexei Dawes' most famous papers on aphantasia, probably because they surveyed aphantasics about dreams (looking at the author contributions, he apparently took part in the study concept, data analysis, and critical revisions). Further, Tsuchiya and Andrillon are co-directors of **Nicolas Decat**, whom I met at the *Immersion and Synesthesia Conference* where he gave another talk. So Tsuchiya and Andrillon might even have indirectly heard of our work! (Provided that my talk was noticeable enough for Nicolas - or anyone else - to tell them about it…)

Conclusion

Using the unsupervised alignment method that we exposed and tested in this report, Kawakita have shown that relational properties of color representations were universally shared by color-neurotypical individuals, but structurally different from color-atypical individuals. Yet intriguingly, their results also support the hypothesis that color-atypical individuals have a different structure of their color representations, rather than simply failing to experience certain colors. This observation on color-atypical individuals, which emerges primarily from the novel consideration of color representation in a psychological space, foreshadows of the potential of this technique to demystify aphantasia.Such perspectives open up unexpected avenues of research to address the impossibility of comparing subjective experiences using psychophysical science.

I tried to show in this project report that this type of (very simple) paradigm focusing on similarities between participants' subjective representations, combined with a state-of-the-art unsupervised alignment method that I was able to implement using an open-source Python scientific toolbox, can be extremely promising for objectifying the difference (or lack of difference) between people's representational formats. This objectification is intrinsically tied to the idea of a link between similarities and representations, but we have good evidence to support this hypothesis. So, provided we create a good study design, this project would enable us to make robust inferences about the contents of analog representations (visual, spatial, auditory-verbal) of aphantasics and phantasics, independent of any subjective assumptions or relationships.

References

Bainbridge, W.A., Pounder, Z., Eardley, A.F., Baker, C.I., 2021. Quantifying aphantasia through drawing: Those without visual imagery show deficits in object but not spatial memory. Cortex

135, 159–172. URL: https://www.sciencedirect.com/science/article/pii/S0010945220304317, doi:10.1016/j.cortex.2020.11.014.

Decock, L., Douven, I., 2011. Similarity after goodman. Review of Philosophy and Psychology 2, 61–75. URL: https://doi.org/10.1007/s13164-010-0035-y, doi:10.1007/s13164-010-0035-y.

Fechner, G.T., 1860. Elemente der Psychophysik. Breitkopf u. Härtel. Google-Books-ID: l5ui0NKWXyAC.

Gardenfors, P., 2004. Conceptual spaces as a framework for knowledge representation .

Goodman, N., 1972. Seven Strictures on Similarity. Bobs-Merril.

Jozwik, K.M., Kriegeskorte, N., Mur, M., 2016. Visual features as stepping stones toward semantics: Explaining object similarity in it and perception with non-negative least squares. Neuropsychologia 83, 201–226. URL: https://www.sciencedirect.com/science/article/pii/S0028393215301998, doi:10.1016/j.neuropsychologia.2015.10.023.

Jozwik, K.M., Kriegeskorte, N., Storrs, K.R., Mur, M., 2017. Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. Frontiers in Psychology 8. URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2017.01726.

Kawakita, G., Zeleznikow-Johnston, A., Takeda, K., Tsuchiya, N., Oizumi, M., 2023. Is my" red" your" red"?: Unsupervised alignment of qualia structures via optimal transport Publisher: PsyArXiv.

Kawakita, G., Zeleznikow-Johnston, A., Tsuchiya, N., Oizumi, M., . Comparing color similarity structures between humans and llms via unsupervised alignment doi:10.48550/arXiv.2308.04381.

Kriegeskorte, N., Mur, M., 2012. Inverse mds: Inferring dissimilarity structure from multiple item arrangements. Frontiers in Psychology 3. URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2012.00245.

Kriegeskorte, N., Mur, M., Bandettini, P., 2008. Representational similarity analysis - connecting the branches of systems neuroscience. Frontiers in Systems Neuroscience 2. URL: https://www.frontiersin.org/articles/10.3389/neuro.06.004.2008.

Mach, E., 1890. The analysis of the sensations. antimetaphysical. The Monist 1, 48–68. URL: https://www.jstor.org/stable/27896829. publisher: Oxford University Press.

Majewska, O., McCarthy, D., van den Bosch, J., Kriegeskorte, N., Vulic, I., Korhonen, A., 2020. Spatial multi-arrangement for clustering and multi-way similarity dataset construction. European Language Resources Association. URL: https://www.repository.cam.ac.uk/handle/1810/306834. iSSN: 2522-2686.

Marr, D., Nishihara, H.K., Brenner, S., 1997. Representation and recognition of the spatial organization of three-dimensional shapes. Proceedings of the Royal Society of London. Series B. Biological Sciences 200, 269–294. URL: https://royalsocietypublishing.org/doi/abs/10.1098/rspb.1978.0020, doi:10.1098/rspb.1978.0020. publisher: Royal Society.

Mur, M., Meys, M., Bodurka, J., Goebel, R., Bandettini, P., Kriegeskorte, N., 2013. Human object-similarity judgments reflect and transcend the primate-it object representation. Frontiers in Psychology 4. URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2013.00128.

Roads, B.D., Love, B.C., 2024. Modeling similarity and psychological space. Annual Review of Psychology 75, 215–240. URL: https://www.annualreviews.org/doi/10.1146/annurev-psych-040323-115131, doi:10.1146/annurev-psych-040323-115131.

Sasaki, M., Takeda, K., Abe, K., Oizumi, M., 2023. Toolbox for gromov-wasserstein optimal transport: Application to unsupervised alignment in neuroscience URL: https://www.biorxiv.org/content/10.1101/2023.09.15.558038v1, doi:10.1101/2023.09.15.558038.

Shepard, R.N., Chipman, S., 1970. Second-order isomorphism of internal representations: Shapes of states. Cognitive Psychology 1, 1–17. URL: https://www.sciencedirect.com/science/article/pii/0010028570900022, doi:10.1016/0010-0285(70)90002-2.

Tsuchiya, N., Phillips, S., Saigo, H., 2022. Enriched category as a model of qualia structure based on similarity judgements. Consciousness and Cognition 101, 103319. URL: https://www.sciencedirect.com/science/article/pii/S1053810022000514, doi:10.1016/j.concog.2022.103319.

Zeleznikow-Johnston, A., Aizawa, Y., Yamada, M., Tsuchiya, N., 2023. Are color experiences the same across the visual field? Journal of Cognitive Neuroscience 35, 509–542. URL: https://doi.org/10.1162/jocn_a_01962, doi:10.1162/jocn_a_01962.