University of Zurich[UZH]

Master's thesis
presented to the Faculty of Arts
of the University of Zurich
for the degree of
Master of Science in Psychology UZH

The Role of Categorization in Short-Term Memory

Author: Alexei Fischer
Student ID Nr.: 03-925-526

Examiner: Prof. Dr. Klaus Oberauer
Supervisor: MSc. Hsuan-Yu Lyn

Department of Psychology
Cognitive Psychology Unit

Submission date: (01.01.2015)

ABSTRACT

This work explores the influence of similarity based categorization as contextual information for recollection in short-term memory. First, *cohesion*, a new index based on Nosofsky's exemplar-based random walk categorization model, is introduced as predictor for how easily a category can be formed and distinguished from other categories given a combination of elements. Second, an algorithm is introduced which identifies category exemplar clusters based on a participant's similarity judgments. Both the *cohesion* criterion and the cluster inference algorithm are tested empirically. Finally, in a modified Sternberg Task where each trial was composed of two different lists (A and B) containing 3 colors each, participants had to judge whether a probe presented with a list label (either A or B) was present in that list. Results showed that compared to a baseline of lists composed of randomly chosen colors, category-based lists constructed using the algorithm evoked better performance for recognizing "old" items and refuting intrusion probes (probes which were present but labeled with the wrong list). Additionally, a false recall effect could be evoked by presenting new colors which were judged by the algorithm to be members of the same cluster as the colors presented on the target list.

AKNOLEGMENTS

I would like to thank Prof. Dr. Klaus Oberauer and Hsuan-Yu Lin for giving me the liberty and leeway to explore and pursue a thematic of my choice. Mr. Lin's guidance, patience and willingness to review this work during his holidays so it could be finished in a timely fashion are a testimony of his extraordinary cooperativeness and were very much appreciated.

I'd further like to thank Ralf Dörig and Christian Machein for setting up the internet server on which the experiments ran; Ana Dojcinovic, Anastasios Ziogas, Gina Stiffler, Ksenia Kolesnichenko and Slavtcho Groshev for helping me test the software before the experiment went live; Prof. Dr. Adrian Schwaninger and Dr. Diana Hardmeier for providing the work environment where I could develop the programming skills necessary for the realization of this project.

CONTENTS

TABLES

FIGURES

1 INTRODUCTION

Similarity often has an important role in both categorization and memory theories. The aim of this work is to explore how similarity based rapid categorization can influence short-term memory performance.

The term *categorization* used here is different how it is used in works discussing the formation of schemata or concepts such as those by Fodor (Aydede, 1998; Fodor, 2010; Fodor & Lepore, 1996), D'Andrade (1990; 1981) or Mandler (2004, 2008) where the mechanisms of categorization are considered to be a complex developmental process which is iteratively refined as a person extends their conceptual understanding of the world around them.

In this work the term *categorization* is used as the rapid and to some extent automatic of classifying elements as members of a given group (Goldstone, 1994). The theoretical basis is provided by *exemplar based* models of categorization which have as their central tenet the assumption that elements are categorized based on their similarity to other members of a category called *exemplars*. For the operationalization, item construction and hypothesis testing, Nosofsky's (2011) generalized context model of categorization was used for its extensibility and relatively simple and malleable computational framework.

1.1 Dual Process Theories and Core Assumptions about Recollection

Most of the memory research this work is based upon, or presupposes a dual process theory of recall and recognition (for an overview, see Smith & DeCoster, 2000).

Dual process models differentiate between two (or more) modes of memory retrieval. One is referred to as *familiarity* which is mostly associated with a "feeling" of knowing an element during recognition. Familiarity is thought to be a unimodal, automatic and nearly instant process which requires no higher order processing. The other is *recollection* which commonly regarded as a laborious and (mostly) conscious

process of retrieving information from memory. While familiarity is context-free and is only associated with the item itself, recollection might also contain contextual or associative information about the item such as the episode where the item was last seen and further semantic information about the item or other items related to it in memory.

From particular interest is the notion that recollection of the context in which an item was presented can serve as a cue for recalling the item itself (Oberauer, 2008). This means that remembering the category which an item belongs to could facilitate remembering the item itself. Conversely, remembering an item might also activate information about its category, which in turn might aid recollection of other exemplars belonging to the same category.

In the following paragraphs I'll summarize the core assumptions of this study regarding the previously described categorization and recollection processes.

First, given a set of elements which share similarities across psychological dimensions (e.g. "roundness" or "edibility"), a person will tend to attribute a category to the set dependent on the shared similarities among elements. For example, when presented with the set {*dachshund*, *dingo*, *red fox*} the category *canine* will be automatically attributed to the set.

This assumption is made based primarily on studies conducted by Shepard (2001), Tversky (e.g. Gati & Tversky, 1984) and Nosofsky (1986, 1987, 1988) which show that identification and categorization is guided by similarity of elements. The exact theoretical framework for similarity based categorization will be discussed in Chapter 2.

Second, when tasked with remembering elements of memorized sets (conceptual group of elements), the category attributed to a given set during memorization serves as contextual information aiding recollection. The degree to which categorization aids recollection is directly related to how coherent the set is, and how easily a category can be attributed to it. For example, the category *edible fish* will aid the recollection of the set

{*anchovy, tuna, salmon*} more than the category *household* will aid the recollection of {*pillow, knife, television*}.

This notion is partially supported by studies on false memory and research on the effects of context information on recall which will be discussed later in this chapter.

Finally, when remembering whether an element is a member of a given set, the distinctiveness between categories attributed to the memorized sets will hinder misclassification. For example, because the category *furry animal* attributed to Set A is very different from the category *electronics* attributed to Set B, the element *cat* is unlikely to be judged as being a member of Set B. Conversely, if the categories attributed to sets A and B were *philosophers* versus *scientists*, finding Noam Chomsky's or Werner Heisenberg's set of origin would hardly be aided by the category information.

This assumption is only vaguely supported by two-list recollection studies (discussed below). While it is clear that context information aids distinction of an element's origin, it is not so clear how the *distinctiveness* between categories influences misclassification in short-term recognition tasks. One of the goals of the present work, is to determine whether this distinctiveness, which will later be formalized as *cohesion* (see Chapter 2.5), has an influence on misclassification.

1.2 Context Information and Intrusions in Two-List Recognition Tasks

Two-list paradigms, where subjects are tasked with memorizing items divided into two different lists and subsequently asked to remember items of the "*target*" or "*relevant*" list while ignoring items from the *irrelevant* list, have long been used for differentiation between *familiarity* and *recollection* within dual process theories of recall.

Weiskrantz and colleagues (Warrington & Weiskrantz, 1968; Winocur & Weiskrantz, 1976) have shown that amnesiac patients, when asked to memorize words in different lists, were more susceptible to interference from previously memorized lists.

That is, they were more prone to recognize words from previously memorized lists as being on the list currently being recalled.

Later, Jacoby (1991) proposes what he calls a 2-study-list process-dissociation paradigm for distinguishing between *familiarity* which was deemed automatic, and *recollection* which was proposed as a conscious process. On a first experiment, a mixture of words and anagrams were presented in a word-by-word basis. If the trial was a word, the participant should read it out loud, if it was an anagram, the participant should speak the solution out loud.

Subsequently, a recognition task was performed. If a word was present in the previous task, subjects were to judge them as "old" otherwise they were to be judged as "new". Jacoby differentiated between *divided* and *full attention* groups. The divided attention group was given an additional auditory task to be performed simultaneously with the "old" versus "new" recognition task. The full attention group performed the recognition task undisturbed. Differences in *familiarity only* versus *familiarity plus recollection* were calculated through the difference in performance between divided and full attention.

In a second experiment, the same task as the first experiment was repeated but this time with different words. The difference was that in the recognition task, words from the previous experiment were included in the recognition phase. This type of probe will be referred to as *intrusion*. Participants were explicitly told only to judge a word as "old" if they were present in the second experiment regardless as to whether the word was present in the first experiment or not.

The results showed a statistically significant interaction between divided attention and probability of responding "old" for intrusion probes. This result was interpreted as supporting the notion that when the conscious process of recollection is hindered, familiarity takes over causing a higher false-alarm rate for intrusion probes, confirming propositions previously postulated by Weiskrantz and colleagues.

Some years later, Gruppuso and colleagues showed that the ability to reject intrusions in Jacoby's task is heavily dependent on the perceived similarity[1] between lists (Gruppuso, Lindsay, & Kelley, 1997). On a two-list memorization task with everyday objects, perceived similarity between the two lists was manipulated by the type of encoding.

The first list was divided in to two halves. In one half of the items, participants had to judge the price of the items listed. In the other half, participants had to evaluate with which frequency they've encountered those items in their everyday lives.

For the memorization of the second list, participants were divided into two groups. One group made value judgments while the other judged the frequency with which they have encountered these objects.

Results showed that participants were more susceptible to intrusions which were encoded in the same way as they encoded the second list. That is, participants of the group which made value judgments in the second list were more prone to think items of the first list, for which they made value judgments, were present in the second list. The same was true for the frequency group regarding frequency judgment items in the first list.

An important aspect these results imply is that other forms of association takes place than the ones explicitly instructed. Although participants were told to classify items dependent solely on which list they belonged to, elements of both lists were inadvertently grouped based on other contextual factors. In this case, the type of judgment (value or frequency) performed on an item. During recognition, this contextual information heavily influences judgments about an element's classification (whether it belongs to the target or irrelevant list).

[1] The authors used the word "similarity" referring to the mode of encoding rather than similarity among elements of both lists. This is rather unusual. For the remainder of this work, whenever I refer to similarity, I'll be referring to the more common usage, i.e. the perceived similarity among elements.

Recall is thus not an "all-or-none" process. People can retrieve some aspects of a past event without retrieving other. In the study of Gruppuso and colleagues, the mode of encoding served as a more salient contextual information about a given item than temporal-episodic information of when and in which list this item was encountered.

In this study, I'll attempt examine how contextual information can be made use of in order aid short-term memory and reduce intrusion effects. The core assumption is that the easier it is to distinguish the two lists by means of contextual information, the better the performance will be distinguishing "old" and "new" items, and the less likely it will be for someone to confound items from one list with items from another.

1.3 Categorization, Implicit Context, and False Recall

One of the main tenets of the current work is that contextual information doesn't need to be imposed extraneously like in the previously discussed study by Gruppuso and colleagues. In certain constellations, contextual information can arise from the combination of items being presented. For example, a list containing the words "truck", "SUV", and "motorcycle" all belong to the category "automobile" and are likely illicit the association with said category when memorized. On the other hand, if the word "motorcycle" is memorized in conjunction with "beard" and "Rock 'n Roll", the combination of words is likely to illicit associations with "biker gangs" from the 60ies and 70ies.

Figure 1 depicts how one might create a contextual category based on the combination of color triads being presented. If color triads are constructed row-wise, the shared hue might elicit contextual classification. That is, when 3 different tones of green are presented in a list, the person might remember the color as being a member of "the green list". Alternatively, when colors are presented column-wise, contextual classification might be evoked based on luminosity and saturation, leading a person to remember a list as the "vibrant color list" or "the pastel tone" list.

Categorization guided by similarity across relevant dimensions commonly shared by elements of a group (in the above case across hue, brightness or saturation) has been thoroughly studied and formalized (Estes, 1986; Goldstone, 1994; Nosofsky, 1988; Tversky & Gati, 1978). I will discuss the intricacies of similarity based categorization in Chapter 2. For now, it is just important to note that categorization is similarity driven, context dependent, and can influence recall.

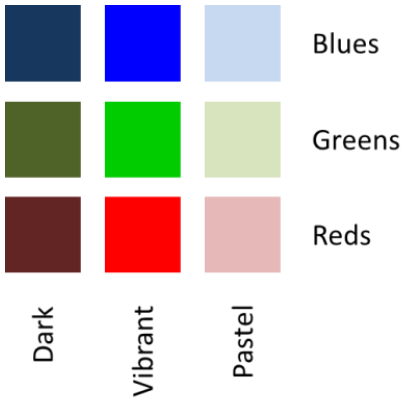


Figure 1: *Illustration of how color triads could be generated. When lists are formed row-wise, hue is utilized as classificatory criterion resulting in lists categorized as "blue", "green" and "red" lists. When lists are formed column-wise, brightness and saturation are utilized as classificatory criteria resulting in "dark" "vibrant" and "pastel tone" lists.*

Probably the best known evidence for the influence of implicit categorization on recall is provided by false memory studies utilizing the *Deese–Roediger–McDermott* (DRM) paradigm (Gallo, 2010).

DRM typically involves the memorization of a list of related words (e.g. *bed*, *rest*, *awake*, *tired*, *dream*, *snooze*, *blanket*, *snore*, *pillow*, *drowsy*) followed by either a free recall or a recognition task. During free recall tasks, participants often remembered words which were not in the original list (e.g. *sleep*). A proposed explanation for this effect is linked to the associative nature of memory. When the words in the list are learned, they activate related words. The probability that a word prototypical for the category is activated increases with every word being learned (Roediger & McDermott, 1995; Roediger, Watson, McDermott, & Gallo, 2001; Stadler, Roediger, & McDermott, 1999). Furthermore, it could be shown that focusing on category-specific attributes instead of item-specific characteristics during the learning phase (e.g. judging whether an animal has fur when learning a relevant list of felines and canines versus an irrelevant list of amphibians) not only improves recognition in the target category, but also increases the probability of falsely remembering a category member which was not present on the list (Fisher & Sloutsky, 2004).

These results further support the notion that extraneous information about implicit categories influence memory performance. In other words, categories flow into the contextual information during recollection.

1.4 The Sternberg Task and Rapid Categorization in Short Term Memory

In the previously mentioned studies memory tasks always involved a rather prolonged learning phase followed by either a recall or recognition task. Subjects had time to reflect about the items to be memorized and the encoding process was mostly deliberately manipulated to enhance the influence of contextual information by either forcing a type of encoding, focusing attention to similarity among items or eliciting active reflections about the item's category.

My interest lays in finding out whether categorization can be elicited in a rapid memorization setting and whether it will aid recollection, reduce intrusion effects and generate similar effects to those observed with the DRM paradigm.

For this purpose, a modification of a short term recognition task first introduced by Sternberg (1966) will be used to examine whether elements presented sequentially where also scanned sequentially and exhaustively in short-term memory. Oberauer (2001) then utilized the modified version of the task to study intrusion effects and later (2008) to measure a formalization of single-process against dual-process of recognition.

In the original Sternberg Task a list of items such as numbers or words are presented sequentially for a short amount of time (usually between 300ms and 400ms per item). After presenting the list, a *probe* is presented and the participant is asked to judge whether that probe was present on the list. The probe can be either a *positive*, meaning that the probe was presented on the memorization list, or a *negative*, meaning that the item as not present on the list.

Within signal detection theory (Macmillan & Creelman, 2004), which will later be used to analyze the data of the second experiment in this study, the response coding can

be divided into 4 different types (Table 1). If the probe is a positive and the response is "yes", the response is coded as a *hit*; if the response is "no", the response is coded as a *miss*. If the probe is a *negative* and the participant responds with a "yes" the response is coded as a *false alarm*, otherwise the response is coded as a *correct rejection*.

Table 1: *Response types*

	"Yes" Response	**"No" Response**
Positive	Hit	Miss
Negative	False Alarm	Correct Rejection

The modified version of the Sternberg Task (Figure 2) divides the presented items into two lists: a target list and an irrelevant one. Participants are not informed about which of the two lists is the target list beforehand.

In the recognition phase, additionally to judging whether the probe item was presented or not, participants need to decide whether the probe was also a member of the target list. If the probe was both present and a member of the target list, the participant should respond with "yes", otherwise the participant should respond with "no".

The division into two lists also adds another type of probe: the *intrusion* probe. The intrusion probe is a member of the irrelevant list which is presented as a member of the target list during recognition. The response for intrusions is coded in the same way as negatives: a "yes" response to the question as to whether the probe was present in the target list yields a *false alarm* and a "no" response yields a *correct rejection*.

Since in this study the modified Sternberg Task utilized by Oberauer is going to be extended in order to harbor yet another type of intrusion (see below), I'm going to refer to intrusions originated from the irrelevant list as *list-wise intrusions* or *L-Intrusions*.

In order to test the influence of categorization in recognition, I'll further expand the task with trials in which the two lists also represent two different categories. For

example, the first list has the category "public transportation" with members *bus, tram* and *train* while the second list has the category "house pet" containing *cat, dog* and *hamster*.

In addition to the list-wise intrusion, I'll introduce a *category-wise intrusion* or *C-Intrusion* which will be a member of the target list's category but not a member of the list itself. Going back to our previous example, let's suppose the second list with the category "house pet" was the target list. A possible C-Intrusion in this case would be another house pet not present in the original list such as a *parrot*.

The introduction of C-Intrusions allows us to test whether DRM-like false recognitions take place and whether category *cohesion* plays a role in the probability of falsely accepting a C-Intrusion as a positive.

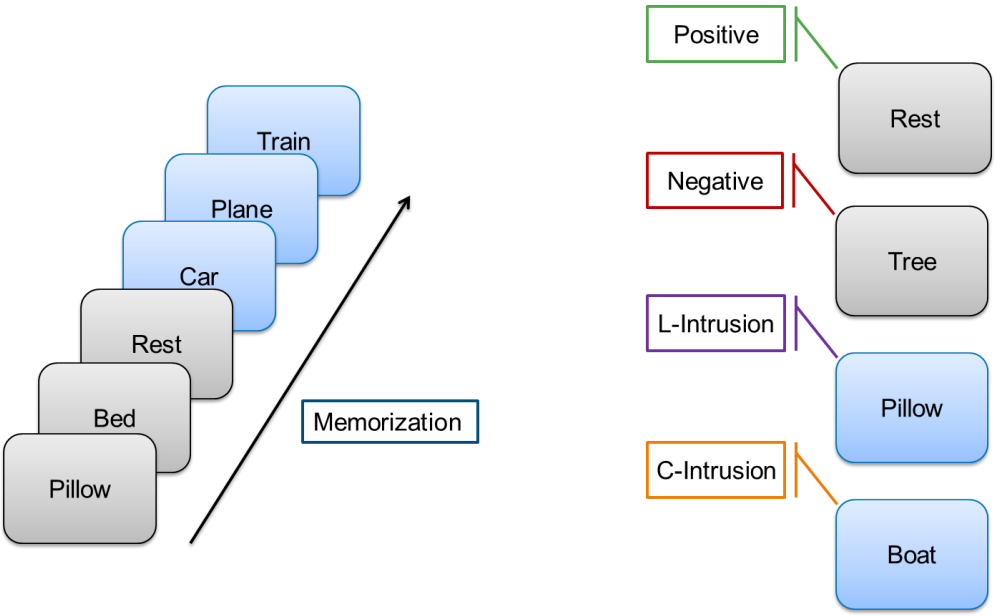


***Figure 2**: The modified Sternberg Task schematic. During memorization (left) items divided into two lists are presented one after another together with a contextual list cue (grey and blue boxes). Apart from the baseline, each list has a category defining its items. The category for the grey list is "sleep" and the category for the blue list is "means of transportation".*
In the recall phase a probe is presented (right) and the participant needs to decide whether that probe was shown during memorization in the same list as it is being shown during recognition. Positives are probes which are presented as members of the same list as they were shown during presentation ("rest" in the grey list). Negatives are probes which were neither on the list, nor a member of each category ("tree" presented as member of the grey list). L-Intrusions are items which were shown during memorization, but are presented in the wrong list during recognition ("pillow" presented as member of the blue list). C-Intrusions are probes which were not shown during memorization but are members of the target list's category ("boat" presented as member of the blue list which has the category "means of transportation")

1.5 Hypotheses

With the previously described modified Sternberg Task and the speculated influence of categorization we can make predictions regarding the expected experimental results. The task will be divided into 2 types of trials, the baseline where both lists will be composed of randomly selected items without a particular category (e.g. L1 = {*tile, dog, sky*}; L2 = {*bus, water, arm*}), and category-based based trials where the elements of each list will be selected in such manner that each list forms a category (e.g. L1 = {*pen, ink, paper*}; L2 = {*ravioli, lasagna, tagliatelle*}).

The performance on category-based trials is expected to differ from the baseline in the following fashion:

1. **Increased Recognition of Positive Probes:** The only contextual information available in the baseline trials which can be used by the recollection process will be the list label. Conversely, category-based trials will have the category as additional contextual information available for the recollection process. This information is also speculated to be more salient and meaningful than the neutral and repetitive list labeling. Therefore, the hit rate (proportion of "yes" responses on positive probes) is expected to increase.

2. **Increased Rejection of Negatives:** Category-based trials will also have information which will facilitate the rejection of negative probes. I speculate that in some cases, the information about the category alone will be sufficient for rejecting a negative. For instance, when the list {*hummingbird, eagle, seagull*} is presented, remembering that the list was composed of birds is enough information to reject *car* as a possible element of the list. Hence, the false alarm rate for negative probes is expected to decrease.

3. **Increased Rejection of L-Intrusions:** When category pairs display high *cohesion*, that is, when the categories for both lists are formed easily upon item presentation and are easy to distinguish from one another, the risk of confounding an element from one list as being a member of the other list is expected to decrease. The rejection of L-Intrusions in high cohesion settings can occur solely based on contextual information about one of the presented lists. Either the subject remembers the category of the target list and rejects the intrusion probe on the grounds of the element not being a member of the relevant list, or the subject remembers the category of the irrelevant list and knows the intrusion to be a member of the irrelevant list.

4. **Decreased Rejection of C-Intrusions:** Contextual information can work against a participant in a way similar as observed in the DRM paradigm. Because the C-Intrusion share communality with all other exemplars, participants are likely to be "lured" into believing the item was present in the target list when it was not. Thus, I expect the false alarm rate for C-Intrusions to be higher than that of negatives across all conditions.

2 Computational Framework

Constructing lists that will implicitly elicit an association to its category requires a priori knowledge about said category and its elements. False memory studies which utilize the DRM paradigm have accumulated a substantial body of techniques for generating wordlists which will evoke associations with categories or category prototypes (Gallo, 2010; Geraci & McCabe, 2006; Stadler et al., 1999). Unfortunately, the characteristics of these methods make them unfit for generating lists which could be utilized in a short-term memory task.

First, the word lists are fixed and tend to be much larger than what would seem sensible in a short-term memory task such as the one proposed in the last chapter. Second, though they do make predictions about the probability of eliciting a false recall, they lack a quantifiable measure of how easy it would be for a subject to associate the presented elements with the category they stem from. Third, it is possible to indirectly imply category consistency through the previously measured list's tendency to generate false recalls, but there is virtually no way to determine how *distinctive* from one another the two categories presented in the same trial would be without first examining each list against each other. Even after such examination, the results would be "static", meaning that there would be no guarantee that the same distinctiveness would still apply when subsets of each list are selected e.g. {*cat, tiger, lion*} out of {*panther, lynx, cat, jaguar, lion, leopard, tiger*}.

In order to examine the influence of ease of categorization in short term memory in an analytical manner, a theoretical framework allowing a priori assumptions about constructed lists must be utilized. Said framework would need to satisfy the following requirements:

i. Provide a basis for quantifying *ease of categorization* and *distinctiveness* between categories.
ii. Allow for ad-hoc stimuli construction based on formalization.
iii. Make verifiable predictions which could then be used for model validation.

For this study, a variation of Nosofsky's (1997) *generalized context model* (GCM) called *exemplar based random walk* (EBRW) model was selected to provide a computational framework for the experimental design, item construction and for validating assumptions made about ease of categorization.

EBRW assumes categorization decisions are guided by similarity between the stimulus being presented and *exemplars* of a category stored in memory. These exemplars are distributed in a multidimensional psychological space according to their characteristic along each dimension. The closer an item is to these exemplars, the more *representative* this item is to the category and the easier it is for that item to be categorized.

For stimulus construction, the CIE color space was selected due to its psychological uniformity (Wyszecki & Stiles, 2000, p. 164), meaning that the distance between two colors corresponds to their subjectively perceived dissimilarity. The goal is to take advantage or the distances between colors in the CIE color space to infer ease of categorization from spatial distribution. That is, colors which are close together in the CIE color space are easier to ascribe to one category than colors which are far apart. The detailed plan for item construction using CIE will be described in Chapter 2.6.

2.1 Underlying Computational Models

The *exemplar-based random walk* (EBRW) model proposed by Nosofsky and Palmeri (1997) is a model of speeded classification which builds upon a family of variants (Nosofsky, 1986, 1987, 1989, 1990) of a previous model called the *generalized context model* (GCM). Both GCM and EBRW predict categorization decisions by evaluating an item's similarity to exemplars of a category.

The model is based on Medin and Chaffer's (1978) *context theory of classification learning* which stipulates a multidimensional representational space where so called *exemplars* of a category are organized. Their position in this representational realm is determined by their attribute characteristics. For example: compared to a needle, an exercise ball would have a lower value along the "elongation" dimension and a high value along the "size" dimension.

The presentation of a novel item acts as a retrieval cue activating exemplars stored in memory based on how similar those items are to the probe being presented. To semantically illustrate how this similarity-based activation works, consider the following example:

Imagine an individual is familiar with a variety of animals categorized as either *cat-like* (feliformia) or *dog-like* (caniformia). These animals can be organized in a representational space along the dimensions *snout* (elongation, form, nostril shape, etc.), *paw-pastern-leg anatomy* (proportions of members, walk movement, rigidity, etc.) and *fur type* (length, smoothness, thickness, color pattern). When said individual sees a lynx for the first time, it will cause animals stored in memory to be activated depending on how similar their morphological characteristics are to those of the lynx. Animals such as house cats, lions or panthers will come to mind.

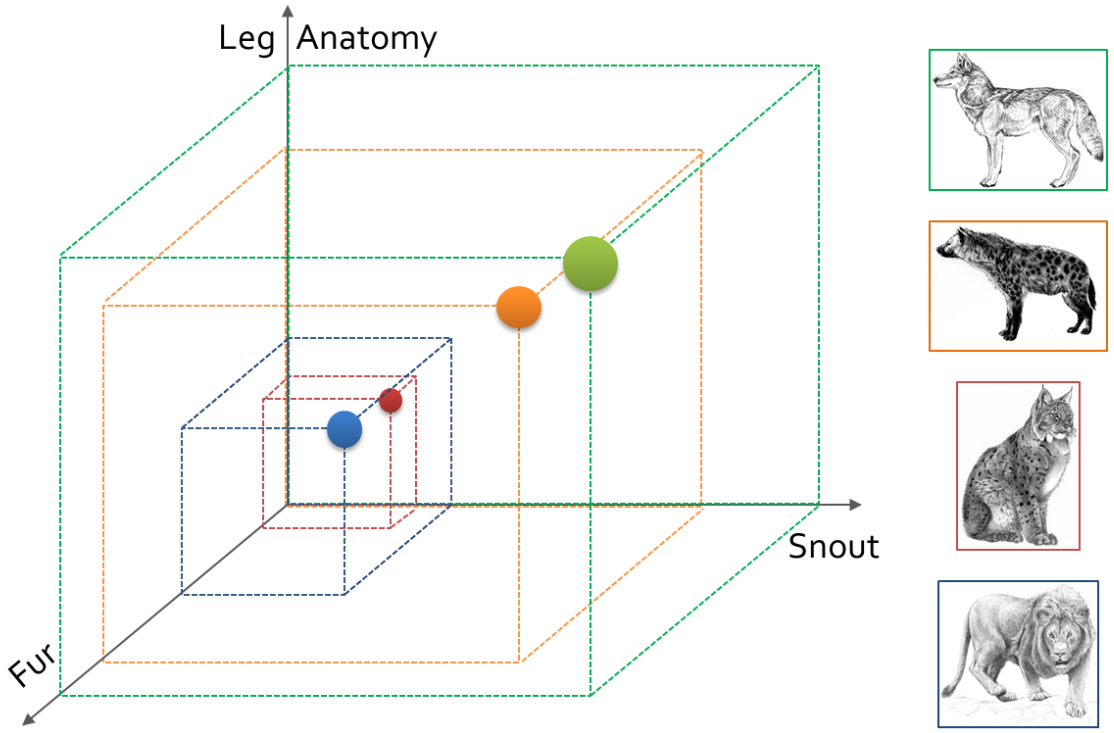


Figure 3: *Depiction of the positioning of a gray wolf (green), hyena (orange), lynx (red) and lion (blue) multidimensional psychological space. The dimensions are fur (from soft to rough), snout (from short to elongated) and leg anatomy (from cat-lyke to dog-like). The soft fur, short-broad snout, and very cat-like walk of the lynx positions it in the proximity of other felines such as the lion which shares a lot of common characteristics, except for the shorter, rougher fur and its mane. The hyena's elongated snout, longer, rougher fur and dog-like posture and legs positions it in the nearer proximity of canines such as the gray wolf.*

The lynx's broad paws, visible flexibility and furtive movements position it among most felines while keeping a considerable distance from most canines (Figure 3). Though short snouts are also present in some canines, the lynx's snout elongation and form is more representative of felines. The same can be said about its smooth fur and color patterns. The lynx's proximity to other exemplars of the category *cat-like* (such as lions, tigers, domestic cats, etc.) will increase the probability of these exemplars to be activated, which in turn will increase the probability that the individual will classify the lynx as a member of the category *cat-like*.

If this individual were to see a hyena for the first time, she would likely categorize the animal as *dog-like*, due to its elongated snout, slender paws, trotting walk and rough fur. Even after learning that hyenas are more closely related to cats then they

are to dogs, this individual would be more likely to have the exemplar *hyena* activated when they see a dog than when they see a cat due to its morphological proximity along the dimensions *snout*, *paw-pastern-leg* and *fur type*.

GCM attempts to formalize the above illustrated representational space (Figure 3), its constituent exemplars and their respective positions in order to make predictions about the probability with which a probe will be classified as member of a stipulated category.

2.2 Selection of Appropriate Distance Measure

GCM utilizes the distance between probe and exemplars in a geometric psychological space as core determinant for exemplar activation. There is research which indicates that some types of stimuli escape core assumptions and axioms (see below) associated with geometric spaces (Nosofsky, 1986; Roberson, Davidoff, Davies, & Shapiro, 2005). Still, a geometric distance measure was utilized in this study for the following reasons:

First, there are some computational constraints imposed by EBRW which require most characteristics of a geometrical space so that an *ease of categorization* can be inferred. These constraints and computational processes will be discussed later in this chapter.

Second, a significant body of research (Indow, 1988; Nosofsky, 1989; Roberson et al., 2005; Wuerger, Maloney, & Krauskopf, 1995; Wuerger et al., 1995) accumulated over the past decades suggests that geometrical distance measures, in particular the Euclidean and Manhattan distances, provide a good computational representation of psychological spaces for color stimuli.

Third, a geometrical space provides significant computational advantages for ad hoc stimulus generation. Since similarity can be deduced form spatial proximity, mean similarity between stimuli can be manipulated by varying the area size for stimulus

selection in the psychological space. A selection from a larger area should yield low similarity among its exemplars and vice-versa (see Chapter 2.6).

According to Tversky (Tversky & Gati, 1978; Tversky & Krantz, 1970), geometric (distance-based) psychological spaces should incorporate the following fundamental assumptions:

(a) **Decomposability:** The function which describes the distance between two points is expressed by their component-wise contributions.
This means that, in a 2 dimensional vector space, the distance between $P_1(x_1, y_1)$ and $P_2(x_2, y_2)$ is dependent on their values along both dimensions x and y.

(b) **Intradimensional Subtractivity:** Each component-wise contribution is the absolute values of their component-wise differences.

(c) **Interdimensional Additivity:** The distance between two points can be expressed as a function of the additive combination of the contribution of their components.

As an illustration of assumptions (a) and (b) consider the Euclidean distance between points $p(x, y)$ and $q(x, y)$ in a 2-dimensional space:

$$d(p, q) = \sqrt{(q_x - p_x)^2 + (q_y - p_y)^2}$$

Here, the *intradimensional subtractivity* is satisfied by $(q_x - p_x)$ and $\left(q_y - p_y\right)$, while *interdimensional additivity* is satisfied by the addition of those two terms.

Other important axioms (Ashby & Perrin, 1988; Tversky & Gati, 1982) regarding the relationship between distances in psychological spaces and similarity spaces are:

i. **Equal self-similarity:** exemplar-probe pairs with equal distances also have equal similarity.
ii. **Minimality:** two different exemplars are at least as dissimilar as either exemplar to itself.
iii. **Symmetry:** the distance from exemplar A to exemplar B is equal to the distance from exemplar B to exemplar A. The same is true for their similarities.
iv. **Triangle Inequality:** for every exemplar triad in the space it must be true that $d(A,B) + d(B,C) \geq d(A,C)$ where $d(p,q)$ describes the distance between points p and q.

(For a comprehensive summary see Ashby & Ennis, 2007)

The distance metric utilized by Nosofsky in many of his publications can be interpreted as a weighted form of the Minkowski distance where w_k represents the weighting of each individual dimension k

$$d_{ij} = \left[\sum_{k=1}^{K} w_k \left|x_{ik} - x_{jk}\right|^{\rho}\right]^{\frac{1}{\rho}} \qquad (2.1)$$
$$0 < w_k, \sum w_k = 1$$

which has been proven to be robust in color and image (dis)similarity paradigms (Li, Chang, & Wu, 2003) and has the advantage of rendering the model applicable in a multitude of L^p spaces (Beals, Krantz, & Tversky, 1968; Tversky & Krantz, 1970). For example, in non-continuous representational spaces ρ can either be set to 1 when distances across dimensions (diagonals) are not permitted, in which case the resulting metric would equate to that of the Manhattan distance, or let it approach infinity when

cross-dimensional (diagonal) steps are allowed, in which case the resulting metric will be analogous to the Chebyshev distance.

For *integral-dimension* stimuli such as colors and tones, which differ along dimensions that do not correspond to distinct and uniquely defined independent variables (as opposed to so called *separable-dimension* stimuli such as *length* vs. *orientation*), the L^2-norm should be applied (Santini & Jain, 1999; R. N. Shepard, 1987). Though some researchers had suggested that Manhattan distances are more appropriate when dealing with dissimilarities among colors (Wuerger et al., 1995), for stimuli generated with allegedly psychologically uniform color spaces with integral dimensions such as CIE, the Euclidean distance (L^2-norm) seems to work as a more reliant distance measure when trying to model similarity judgments (Garner, 2014; Nosofsky, 1986; R. N. Shepard, 1964). Thus, the weighted Euclidean distance $\rho = 2$ in Equation 2.1 was utilized for simulation, stimuli construction and validation (see chapters 2.6 and 7.1).

2.3 Similarity and Activation as an Exponential Function of the Distance Metric

A multitude of studies have shown that subjective similarity ratings do not scale proportionally according to their distance metric in the psychological space (Ashby & Perrin, 1988; Nosofsky, 1984, 1992; Richler & Palmeri, 2014; Santini & Jain, 1999).

Shepard (1987) has attempted a generalization rule for the relationship between distance metric and similarity by analyzing the data of numerous similarity judgment experiments. Shepard's strategy was to use nonmetric multidimensional scaling to determine the space with the lowest number of dimension which would still explain the data. He found that across multiple types of stimuli and spaces, the relationship between metric and similarity could be expressed by an exponential function with decay rates varying according to stimulus type (Figure 4).

Another important finding is that the relationship between the ease with which categories are learned and the distance between items in the psychological space also seem to follow an exponential decay function (Richler & Palmeri, 2014).

As it will be demonstrated later in this work, an exponential relationship between hypothesized ease of categorization and category "cohesion" has both been predicted using Nosofsky's model (see Chapter 4.4.1) as well as validated in the categorization experiment (Chapter 8).

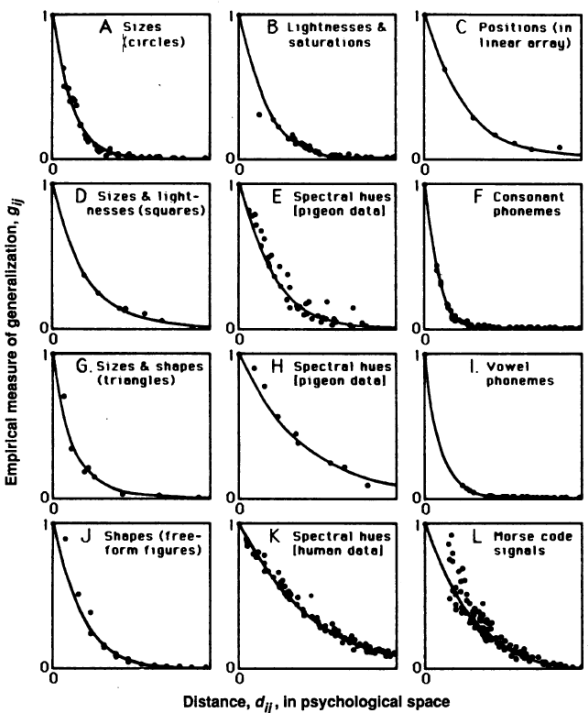


Figure 4: *Twelve gradients of generalization (similarity against distance). Though all follow an exponential decay, their decay rate is dependent on which stimuli were used (R. N. Shepard, 1987, p. 1318)*

When dealing with categorization tasks, be it the classification of a probe or learning of categories, GCM follows previously proposed, and relatively robust *similarity choice models* (Estes, 1986; Medin & Schaffer, 1978) where the probability of assigning an item to a certain category is defined by the ratio between the similarity rating for that category to the conjoint similarity with all categories (Nosofsky, 1986, 1990).

However, when exemplars are queried from memory, the model assumes an activation pattern dependent on both similarity as well as *memory strength* for each individual exemplar as formalized as

$$a_{ij} = M_j \cdot e^{-c \cdot d_{ij}} \qquad 0 \leq c \leq \infty \qquad 0 \leq M_j \leq \infty \tag{2.2}$$

where a_{ij} represents the degree to which exemplar j is activated when item i is presented, c is the decay rate of similarity with increasing inter item distance, and d_{ij} the estimated distance between items *i* and *j* in the psychological space.

The addition of a memory strength dependent parameter allows the model to account for factors such as primacy or recency effects in serial presentation or presentation frequency when within a learning paradigm (Nosofsky, 1991).

2.4 Reaction Time Estimation

To account for differences in reaction time, Nosofsky and Palmeri (1997) propose a mechanism based on Logan's (1988) *instance-based model* of automaticity and a sequential-sampling drift similar to Ratcliff's (1978) *diffusion model*.

According to Logan's instance based model, people have a "base algorithm" which is used to solve a task. As a subject learns new techniques or solutions to said task, every episodic instance of these solutions enter a race against the base algorithm when the task is presented anew. If a solution finishes the race before the base algorithm, the subject uses this solution instead of the one provided by the base algorithm. Task automatization then, reflects the transition from algorithm-based performance to memory-based performance.

Analogue to Logan's instance based model, when an item is presented it sets of a *race* among all exemplars stored in memory. The time it takes for each exemplar to finish the race is given by random variables proportional to how strongly exemplar *j* is activated by the presentation of item *i*. The density function $f(t)$ which describes the probability that an exemplar *j* will have finished the race by time *t* follows a standard probability density function of an exponential distribution

$$f(x;\lambda) = \lambda e^{-\lambda x}, x \geq 0$$
$$f(t) = a_{ij} \cdot e^{-a_{ij} \cdot t} \tag{2.3}$$

where $f(x;\lambda)$ is the probability density function for exponentially distributed random variables and $f(t)$ the adapted probability density function dependent on item for exemplar j finishing the race at time t dependent on activation a_{ij} when item i is presented.

While Logan's model gives us an account of how rapidly evidence for a particular category can be retrieved, Ratcliff's diffusion model attempts to explain rapid decisions made in binary decision tasks where no complex cognition or problem solving is required. The model assumes that decisions are made by a noisy process that accumulates information over time (Figure 5).

The process starts at a neutral state. Upon trial presentation, evidence is collected in favor of either decision A or B. This evidence accumulation guides the "drift" towards either response A or B. For either decision, there is a decision boundary. Once this boundary is crossed, the participant gives his or her response. The rate with which the process drift approaches a decision boundary is directly dependent on the strength of the evidence for that decision.

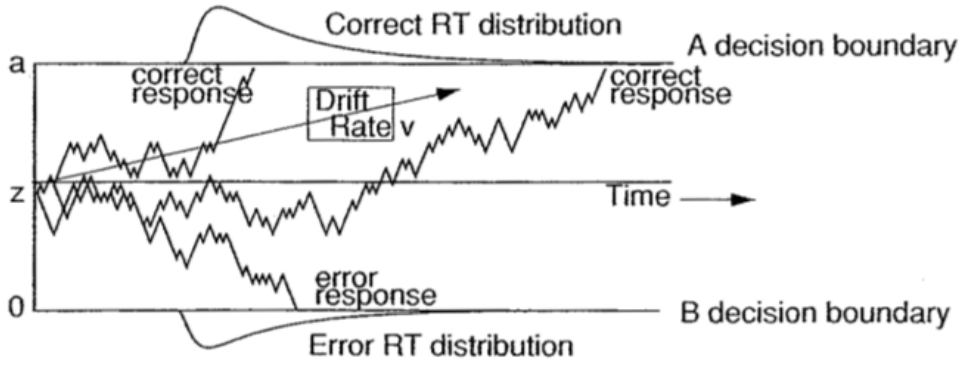


Figure 5: *Illustration of the diffusion decision model. The process starts at point Z and drifts either towards a correct response (a) or an incorrect response (0). The drift rate (v) depends on how fast evidence can be accumulated for each case. The decision boundaries A and B determine the threshold at which a decision is executed (Ratcliff & McKoon, 2008, p. 31).*

EBRW utilizes the diffusion model's core concept to explain rapid classification decisions. Nosofsky (1997) proposes a random walk guided by the "race winner" exemplars retrieved after an iteration of the *instance based model's* race is complete.

During a trial, when a probe is presented, all stored exemplars enter a race as described by the *instance based model*, where the probability an item finishes the race under a certain time is given by Equation 2.3. The exemplar that wins the race influences the random walk towards the direction of the category the "winner exemplar" belongs to. This process repeats itself until the walk reaches a response threshold.

From Equation 2.3 we can read that EBRW treats exemplar activation as the decay rate in a negative exponential distribution ($a_{ij} = \lambda_j$). This is advantageous since the probability density function for retrieving a particular exemplar j follows the distribution of the minimum of exponentially distributed random variables $\Pr(\mathrm{X_k} = \min\{\mathrm{X_1}, \ldots, \mathrm{X_n}\})$. In our case, these would be all exemplars entering the race. The probability that a particular item will be the one to finish the race first can thus be computed through the distribution of the minimum of all activations (see Nelson, 1995, Chapter 4, Example 4.21; Schinazi, 1999, Chapter 7):

$$Pr(X_k = min\{X_1, \ldots, X_n\}) = \frac{\lambda_j}{\lambda_1 + \cdots + \lambda_n} = \frac{a_{ij}}{\sum a_{ik}} \tag{2.4}$$

Given a set K the probability that the random walk will take a step towards category A can be calculated by summing the probabilities that a given exemplar from category A will finish the race first. Thus we have probabilities p_i and q_i of a step to go in direction of A or B respectively as described by:

$$p_i = \frac{\sum_{j \in A} a_{ij}}{\sum_K \sum_{k \in K} a_{ik}} \tag{2.5}$$

$$q_i = \frac{\sum_{h \in B} a_{ih}}{\sum_K \sum_{k \in K} a_{ik}} \tag{2.6}$$

With probability estimates for each step, reaction time predictions can be made by estimating at which point the random walk will cross either threshold A or B. This

estimate is determined jointly by the total number of steps made during the random walk and by the speed with which each individual step is made.

An estimation of steps necessary to reach threshold A or B can be attained by modeling the *gambler's ruin problem*[2] in a one-dimensional random walk. The number of steps necessary to reach a threshold corresponds to the game duration and is given by E(N|i) in Equation 2.7. The probability of a category A response is given by P(A|i) in Equation 2.8 and the probability of a category B response is given by P(B|i) in Equation 2.9.

$$E(N|i) = \frac{B}{q_i - p_i} - \frac{A+B}{q_i - p_i}\left[\frac{1-\left(\frac{q_i}{p_i}\right)^B}{1-\left(\frac{q_i}{p_i}\right)^{A+B}}\right] \tag{2.7}$$

$$P(A|i) = \frac{1-\left(\frac{q_i}{p_i}\right)^B}{1-\left(\frac{q_i}{p_i}\right)^{A+B}}, \quad if\ p_i \neq q_i$$
$$P(A|i) = \frac{B}{A+B}, \qquad if\ p_i = q_i \tag{2.8}$$

$$P(B|i) = \frac{\left(\frac{q_i}{p_i}\right)^B - \left(\frac{q_i}{p_i}\right)^{A+B}}{1-\left(\frac{q_i}{p_i}\right)^{A+B}}, \quad if\ p_i \neq q_i$$
$$P(B|i) = \frac{A}{A+B}, \qquad if\ p_i = q_i \tag{2.9}$$

Where A and B represent the threshold with which a decision is made, and p_i and q_i represent the probability that an exemplar belonging to category A (or B) finishes the walk first as formalized in equations 2.5 and 2.6.

[2] In probability theory, the gambler's ruin problem describes a wealth process where a gambler with an *x* amount of chips plays a game repeatedly until his wealth reaches a prescribed level (in our case threshold A) or until his wealth drops to zero (in our case threshold -B). For a complete description and solution of the gambler's ruin problem in an one-dimensional random walk see Lawler & Limic (2010), Chapter 5.1 or Feller (1968), Chapter XIV.3.

For reaction time computation, the number of steps is multiplied with the expected step duration. The expected number of steps is given by Equation 2.7. For the estimated time each step takes ($E(t_s|i)$ in Equation 2.11), Nosofsky proposes adding a constant (α) which is treated as an individually fitted free parameter and represents the time needed for *category-label* extraction, to the estimated finishing time for the *exemplar race* which guides the random walk step direction (Equation 2.10). The expected finishing time for the exemplar race winner can be computed through the expected minimum value of T_n exponentially distributed random variables:

$$E(\min(T_1, \dots, T_n)) = \frac{1}{\lambda_1 + \cdots + \lambda_n} = \frac{1}{\sum_{j=1}^{n} a_{ij}} \tag{2.10}$$

$$E(t_s|i) = \alpha + \frac{1}{\sum_{j=1}^{n} a_{ij}} \tag{2.11}$$

$E(t_s|i)$ is listed here solely for the sake of completion. Since no direct reaction time predictions or individual model fitting were conducted in this study, and since the step count and expected minimum of T_n (Equation 2.10) suffice for validating the assumptions made about ease of categorization (Chapter 4.4), I'll refrain from utilizing the walk duration estimate (Equation 2.11).

2.5 Ease of Categorization

With the afore-discussed framework in place, ad hoc definitions of *category* and *ease of categorization* can be proposed.

There is a considerable body of evidence to support the notion that categorization of visual stimuli is driven at least in part by similarity (Cavanagh, 2011; Palmeri & Gauthier, 2004; Richler & Palmeri, 2014). In regards to color in general, and uniform color spaces such as the CIE in particular (see Chapter 2.6), there is reason to believe that categories exists within spatial regions delimited by their position within the color space (Derefeldt, Swartling, Berggrund, & Bodrogi, 2004; Xiao, Kavanau,

Bertin, & Kaplan, 2011). Moreover, color categories seem not to be limited to hue-based *basic colors* (Boynton & Olson, 1987; Sturges & Whitfield, 1995), but expand across all dimensions in uniform color spaces (Lindsey & Brown, 2009). In other words, color categories are not limited to "reds", "greens" or "yellows" but expand across hue, as for example "warm" and "cold" colors or pastel tones.

In accordance with these findings the stipulation is made that confined regional areas within the color space will tend to be associated with a category. The *cohesion* of these categories is related to the size of the selected areas. Small areas will produce highly coherent category groups, since the dissimilarity among its constituent colors is expected to be relatively small. On the other hand, large areas will produce somewhat incoherent category groups, since they are likely to encompass a broader variety of colors.

The cohesion of these categories should directly influence the *ease of categorization*. That is, how easily someone can identify a member of that group as belonging to the same category. Because EBRW stipulates that exemplar similarity is related to walk speed (high activation exemplars finish the exemplar race faster), this ease of categorization should produce observable differences in reaction time during a categorization task.

From Equation 2.2 we can deduce that as the distance d_{ij} between exemplar j and probe i approaches null, the exemplar activation a_{ij} will approach the value of its memory strength M_{ij}. At the same time, as the distance between exemplar and probe increases, the memory strength for exemplar j plays a less and less important role, until the entire term approaches zero:

$$\lim_{d_{ij}\to 0}\left(a_{ij}\right) = M_j$$
$$\lim_{d_{ij}\to \infty}\left(a_{ij}\right) = 0 \qquad (2.12)$$

From Equations 2.5, 2.6 and 2.7 we can predict the number of expected steps in the random walk depending on how categorical regions are selected.

This study uses the modified Sternberg task with two lists of the same length, each constituting a category (apart from the baseline condition). Because a forced-choice response paradigm is implemented, we can assume that p_i in 2.5 and q_i in 2.6 add up to 1 since there are no further elements of set K which are not elements of either category A or B. Therefore we can say that $q_i = 1 - p_i$.

In the modified Sternberg task, when probes are balanced in regards to their target list (meaning that the probe is presented as being a member of either the first or the second list), we can expect the random walk category thresholds A and B to be equal on average, meaning that $\widehat{A} = \widehat{B}$.

Applying these two assumptions to Equation 2.7 will give us

$$E_2(N|i) = \frac{\widehat{A}}{1-2p_i} - \frac{2\widehat{A}}{1-2p_i}\left[1 + \left(\frac{1-p_i}{p_i}\right)^{\widehat{A}}\right]^{-1} \tag{2.13}$$

As p_i in Equation 2.13 approaches 0 or 1, the walk approaches its minimum estimated number of steps, namely $E_2(N|i) = \widehat{A}$. As p_i reaches 0.5, the gambler's ruin problem solves to $\widehat{A}^2$ (Feller, 1968), which corresponds to the maximum for the expected value for the number of steps until a boundary is reached.

From Equation 2.2 we learn that we can maximize activation by minimizing the distance between exemplar and probe as depicted in Equation 2.12. In a two category, forced choice paradigm where A denotes the target category and B the irrelevant category, the denominator in Equation 2.5 resolves to the sum of the activation of exemplars in both category A and B. We can maximize p_i by maximizing the distance between the probe and all exemplars of category B because as $\sum a_{ib},\ b \in$ B approaches 0, p_i approaches 1.

The second term in the time estimate for Nosofsky and Palmeri's random walk model is the expected step time $E(t_s|i)$ (Equation 2.11). Since α is a constant and assumed to converge to a mean across trials (Cohen & Nosofsky, 2003; Nosofsky & Palmeri, 1997), we can achieve a minimum by maximizing the activation sum in the denominator. This means decreasing the distances between exemplars and probe in the target category. One way to achieve this is by reducing the region's n-volume where category exemplars are selected from.

In short, Equation 2.7 can be minimized by maximizing the distance between the target and irrelevant categories. Equation 2.11 can be minimized by reducing the size of the target category in the psychological space. When a category pair is distant from one another and their n-volume in the psychological space is small, they display *high cohesion*. When the n-volume of both categories is large in proportion their distance to one another, they have *low cohesion*. We can thus say that the cohesion measure $\mathcal{C}_{T,I}$ between a target category T and an irrelevant category I is defined by the proportion of the distance metric $D_{T,I}$ between the category spaces and the n-volume ${}^{n}V_{T}$ of the target and irrelevant categories:

$$\mathcal{C}_{T,I} \sim \frac{D_{\mathrm{T,I}}}{{}^{n}V_{T}} \tag{2.14}$$

Because cohesion is defined entirely with metric properties of the category space and because EBRW utilizes that same metric to predict speed of categorization in terms of activation speed and competing evidence among categories, $\mathcal{C}_{T,I}$ should be measurable in reaction time differences, higher cohesion values producing shorter reaction times, and directly reflect *ease of categorization*.

2.6 The CIE Color Space as Stimulus Pool

CIE stands for *Commission Internationale de l'Eclairage* or International Commission on Illumination. The CIE color space was defined as a quantitative link between the wave length in the visible color spectrum and the physiologically perceived colors in human color vision. It is based on *color sensations* caused by the excitation of cone cells.

The excitation of the three types of cone cells (Short, Medium and Long) are mapped onto the coordinates XYZ distributing the colors (wavelengths) along the Euclidean space depending on which color sensation they generate in normal lightning conditions.

Because CIE XYZ is based solely on spectral distribution and cone activation, some ranges within the CIE might be perceived as the same color. To counteract this, a perceptually uniform transformation across lightness *L* and two color components *a* and *b* was created based on the Munsell color space (Hunter, 1987). The CIELAB or *Lab* color space has the advantage of being organized in a Euclidean dissimilarity space, meaning that the perceived similarity between two colors bares the same distance throughout the entire space.

Multiple studies (e.g. Franklin, Pilling, & Davies, 2005; Xiao et al., 2011; Yoshioka, Dow, & Vautin, 1996) have utilized CIE in color categorization and memory experiment with good results. However, I was unable to find a stimulus generation paradigm which would fit the here stipulated *cohesion* criterion (Equation 2.14).

Fortunately, the CIE color space is defined in an Euclidean space. As discussed in Chapter 2.2, the chosen similarity measure for the computational model can be regarded as a weighted form of the Euclidean distance. And though it is advantageous that the spatial organization of the CIE can be translated directly onto GCM's metric, it's hard to heuristically predict what the weight for each dimension should be. For example, it could be that differences within the hue dimensional components a* and b*

of CIELAB carry much more weight in regards to similarity than the luminosity dimension L.

Still, this shouldn't hinder us from stipulating categorical groups within the CIE utilizing an Euclidean measure of equal weights (i.e. $\vec{w} = [0.\overline{33},\ 0.\overline{33},\ 0.\overline{33}]$). As long as appropriate methods of cohesion manipulation are applied, these weighting differences should not interfere with our hypothesis testing. Small category spaces which are far apart from each other would still generate trials with higher cohesion than larger category spaces which are closer together.

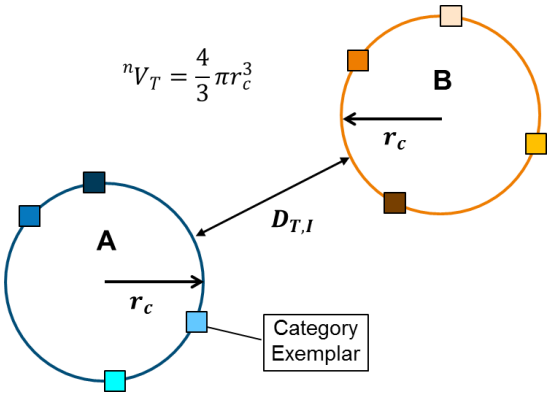


***Figure 6**: Spatial illustration of planned element sampling in the CIE. Two spatial regions of radius r_c and $D_{T,I}$ apart from each other are defined in the color space. Category exemplars within the spatial definition are taken as list elements for the modified Sternberg Task.*

Discrepancies caused by dimensional weighting could also be counteracted by varying the relational position of the categories. For some trials, category space A would be 0.3 units away from category space B across the a* dimension, for other trials it would be 0.3 units away from B across the b* dimension, thus counterbalancing the error originated by false weighting.

With all that in mind, the trials for the modified Sternberg Task could be constructed with the following method:

First, a category space radius r_c is defined and set to a short distance for high cohesion lists, or to a larger distance for low cohesion lists respectively. Then, two category centers are selected at random within the color space. The distance between these points is defined by $D_{\mathrm{T,I}}$ and the category volume can be calculated as $^nV_T = \frac{4}{3}\pi r_c^3$ (Figure 6).

For both categories, pick 4 random colors which are positioned within radius r_c from the category's center. The lists in the Sternberg Task are populated with 3 of the

randomly picked items of each category. The fourth randomly chosen item is used as a C-Intrusion.

With the values for both $D_{\mathrm{T,I}}$ and ${}^{n}V_{T}$ the cohesion metric can be inferred a priori (see Equation 2.14), and for baseline trials, colors for both lists will be randomly selected across the entire color space.

There are still some concerns regarding the utilization of the CIE as stimulus pool. Though there are minimal noticeable difference measures for the color space (McDermott & Webster, 2012; Wyszecki & Stiles, 2000), these differences are perceived when colors are presented simultaneously. In the Sternberg Task, however, subjects will need to be able to differentiate between colors from memory. Due to *change blindness* where differences which are easy to notice become almost unnoticeable when the stimuli presented in a sequential manner are interrupted by an inter-stimulus interval (Simons & Rensink, 2005), the just noticeable interval might not be sufficient to ensure a participant can distinguish between two colors in memory. Therefore, a minimal distance between two selected stimuli in the CIE color space with which participants are able to distinguish said stimuli from memory alone, must first be defined.

Though psychologically uniform spaces seem to conserve the dissimilarity-to-distance relation well for localized regions, failure to maintain this uniformity across larger distances have also been reported (Seaborn, Hepplewhite, & Stonham, 2005). This means that similarity measures might only be allowed within small distances. If this is true, a reliable measure for cohesion might be undermined.

Finally, it might be the case that an universal Euclidean arrangement for the CIE based on perceived color similarity is not at all possible. Either due to the metric itself (Kuehni, 2001; Wuerger et al., 1995), due to the fact that perceived similarity among colors is heavily influenced by other developmental factors such as language and culture (Roberson, Davies, & Davidoff, 2000; Roberson & Hanley, 2007), or even due to

physiological reasons not related to color blindness (Bimler, Kirkland, & Jameson, 2004; Feitosa-Santana et al., 2006).

All this points to the necessity of a pilot study to test whether the CIE can serve as an adequate system for allowing the inference of implicit categories for polled colors. The experiment design to test CIE's metrics as well as its results will be discussed in Chapter 3.

2.7 Summary

In this chapter I've discussed the theoretical framework whereupon the study will be constructed. The CIE color space will serve as stimulus pool. Sampling items within a small region in the color space is expected to yield a list of exemplars which are highly similar to each other. GCM tells us that this similarity will increase the probability that related exemplars are activated when a probe from within that area is presented. EBRW tells us that a participant should be able to classify the probe as a member of the same category as the sampled exemplars with greater ease when the distance between the target category and the irrelevant category is large in relation to the target category volume.

Here I've also discussed the computational constraints which must be satisfied by the stimuli. Activation in the generalized context is directly dependent on similarity measures. These measures must conform, at least in part, to the assumptions and axioms discussed in the beginning of this chapter (see Section 2.2) otherwise our assumptions about ease of categorization are invalid.

Further concerns were raised regarding the CIE metrics, namely, that the relation between similarity and distance metric might not be uniform with increasing distances; that a minimal distance between stimuli which can be sample must be established; and that the CIE might not be universal, meaning that some people perceive similarity among colors differently than others.

3 CIE Metric Evaluation

As discussed in Chapter 2.6, the CIE color space needs to be tested for its appropriateness. For this purpose, a pilot study involving similarity judgments among equidistant colors within CIE was conducted. The goal of the pilot study is to (a) determine the minimal distance with which colors can be sampled so that a participant can still distinguish between them when presented within a change blindness paradigm (i.e. with an inter stimulus interval between the presentation of the first and second colors). And (b) the Euclidean distance metric still relates to similarity even across larger distances.

3.1 Experiment Design

For the metric evaluation, the CIE-XYZ color space was divided into a grid of isometric nodes along its dimensions (Figure 7). The interval between nodes was set to 2.5% of the total length of the color space, creating a $40 \times 40 \times 40$ node matrix.

From this grid, a selection of nodes, referred to as *probes*, had their similarity measured against other nodes, referred to as *neighbors*, across a single dimension (e.g. compared with other nodes with the same X and Y values but different Z values).

The single dimension comparison was repeated across all 3 dimensions with equidistant neighboring nodes. For example, the node at $n_A = [.5, .5, .5]$ is compared against its *second degree* neighbors $n_{A-x} = [.45, .5, .5]$ and $n_{A+x} = [.55, .5, .5]$ (inter-node distance $= 2 \times 0.025$) along the X-dimension, against $n_{A-y} = [.5, .45, .5]$ and $n_{A+y} = [.5, .55, .]$ along the Y-dimension, and finally against $n_{A-z} = [.5, .5, .45]$ and $n_{A+z} = [.5, .5, .55]$ along the Z-dimension.

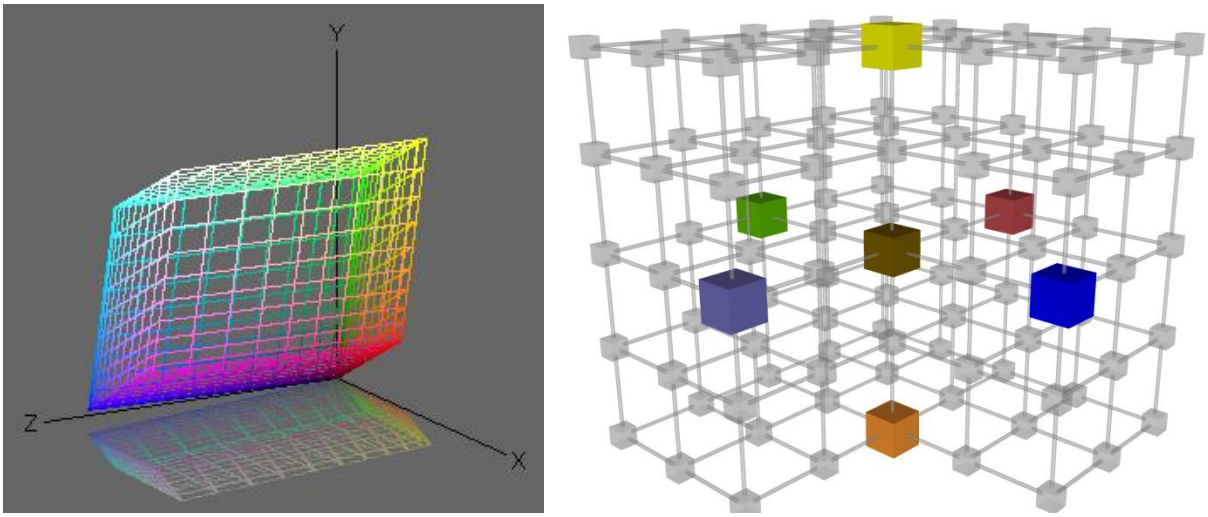
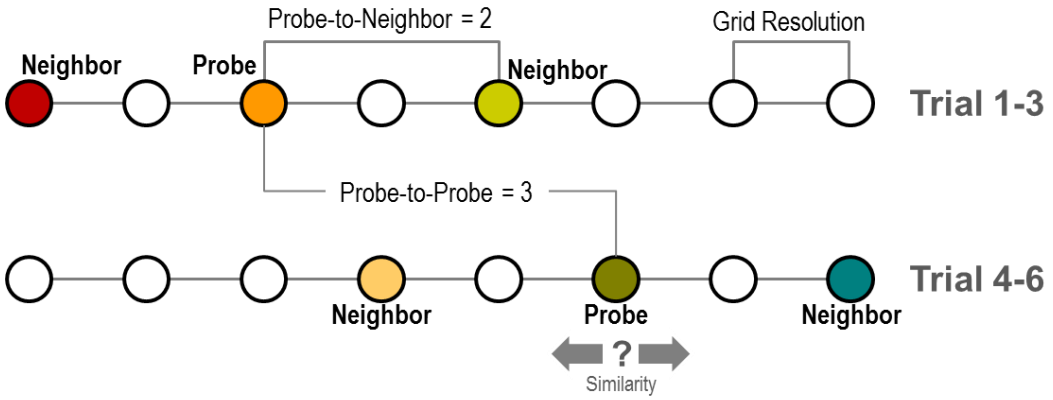


***Figure 7**: The CIE color space subdivided into a grid (left). Each vertex represents a node. For each probe sampled from the grid, a similarity judgment is made against neighboring nodes along each dimension of the color space (right). The probe is located in the center (brown) and the neighbors it's being compared against are equidistantly situated along each axis (yellow, green, violet, blue, red and orange)*

The trials were constructed according to two measures, a *probe-to-neighbor* distance (d_{PN}) which stands for the distance between the node being evaluated and its fixed neighboring nodes along a single dimension, and the *probe-to-probe* (d_{PP}) distance which describes the distance between the probes across trials. Both d_{PN} and d_{PP} were defined as multiples of the grid resolution (Figure 8). First, a probe and its neighbors are selected. For each probe, two neighbors located at distance d_{PN} from the probe across a single dimension are selected for a trial. This process is repeated for the two other dimensions constituting a total of 3 trials with 3 colors each. Then, the next probe located at distance d_{PP} form the previous probe is selected and the process is repeated until the entire grid has been covered.

A)



B)

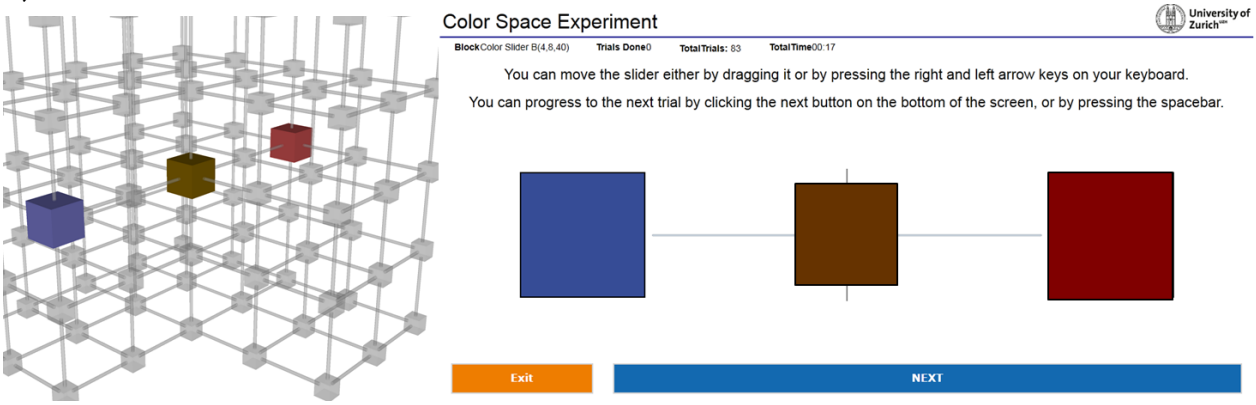


Figure 8: *A) Trial construction method. Each circle represents a node along one dimension of the grid. The distance between the nodes constitutes the grid resolution. First, a probe is selected. Then, along a single dimension, two neighbors are selected which are p_{PN} nodes apart from the probe (Probe-to-Neighbor distance). A probe and its two neighbors constitute a trial. The process is repeated for the other 2 dimensions yielding trials 1 to 3. The algorithm then selects the next node which is d_{PP} nodes apart from the original probe (Probe-to-Probe distance).*
B) The probe and its two neighbors are selected as a trial. Similarity judgment is effectuated by sliding the probe (color in the center) left or right according to how similar to its neighbors the participant judges them to be.

The experiment trials were divided into 4 blocks with varying d_{PN} and d_{PP} values (Table 2) ensuring that similarity comparisons are made across different distances and interlacing comparison points.

Table 2: *Node selection directives for all 4 experimental blocks*

Block	Probe-to-Neighbor (d_{PN})	Probe-to-Probe (d_{PP})
A	12 Nodes	6 Nodes
B	8 Nodes	4 Nodes
C	3 Nodes	6 Nodes
D	6 Nodes	3 Nodes

The pilot test was conducted with 15 participants (11 male, 4 female). To test their color vision, the Ishihara Test for color blindness (Ishihara, 1981) was administered.

The trials evaluated a total of 412 nodes. On each trial the probe color was plotted on the center of the screen (Figure 8 B). Left and right of it were the two neighboring colors. Participants were asked to slide the color in the center (i.e. the probe) according to how similar they perceived it to be to the color to the left or the color in the right.

The distribution of the new positioning of the probe color by participants would provide some insights about CIE's psychological uniformity and at the same time test whether the smallest interval of 3 nodes, i.e. 0.075 is likely to be sufficient to make two colors distinguishable from each other.

3.2 Results and Discussion

The positioning distribution for all 412 nodes examined was evaluated by combining the data of all three dimensions. The entire analysis and data is available in the additional materials. Of particular interest for the current work are following findings:

First, only 31 of the 412 nodes cold be confidently positioned within acceptable range. For other nodes, the distributions were either not normal (i.e. did not pass the

Shapiro-Wilk test of normality), or the standard deviation of the node's positioning overlapped that of another node. That is, along one dimension, it was no longer possible to determine whether a node A came before or after a neighboring node B.

Second, with increasing d_{PN} the confidence interval for a node's subjective position in the color space also increases (Figure 9). This finding concurs with other findings in the literature (e.g. Seaborn et al., 2005) which argue that psychological uniformity of the CIE is only present within short distances.

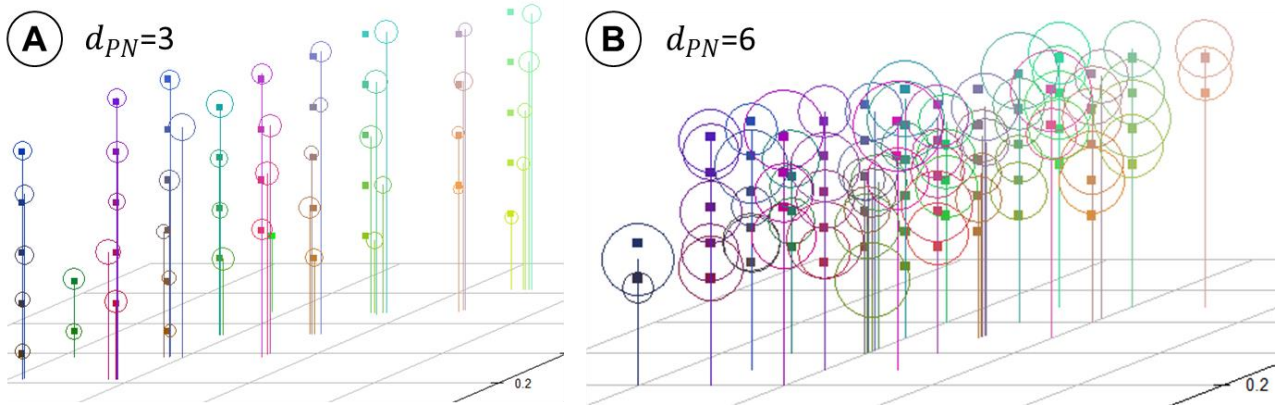


***Figure 9**: Positioning and variance of selected probes under different d_{PN} values. Each data point is plotted using its CIE color. The squares represent each probe's original position on the CIE color space. The circles represent the new positioning according to the participants' similarity judgments. The circle's center stands on the mean of the offset judgments over all participants and dimensions. The circle's radius is the standard deviation of the offset positioning. For small probe-to-neighbor distances (A) both positioning offset and deviation are small, when the probe-to-neighbor distance increases (B) uncertainty about a nodes position also increases.*

Third, the Hartigan's Dip Test (Hartigan & Hartigan, 1985) was performed on all nodes and revealed bimodality for at least 24 of them ($p < 0.05$). This reflects the previously discussed lack of universality regarding color similarity judgments. In other words, for some people, along a single dimension, a given color A was more similar to another color A+ with a higher value in that dimension, than a color A- with a lower value; for others, the same color A was more similar to the color A- as the color A+ (Figure 10).

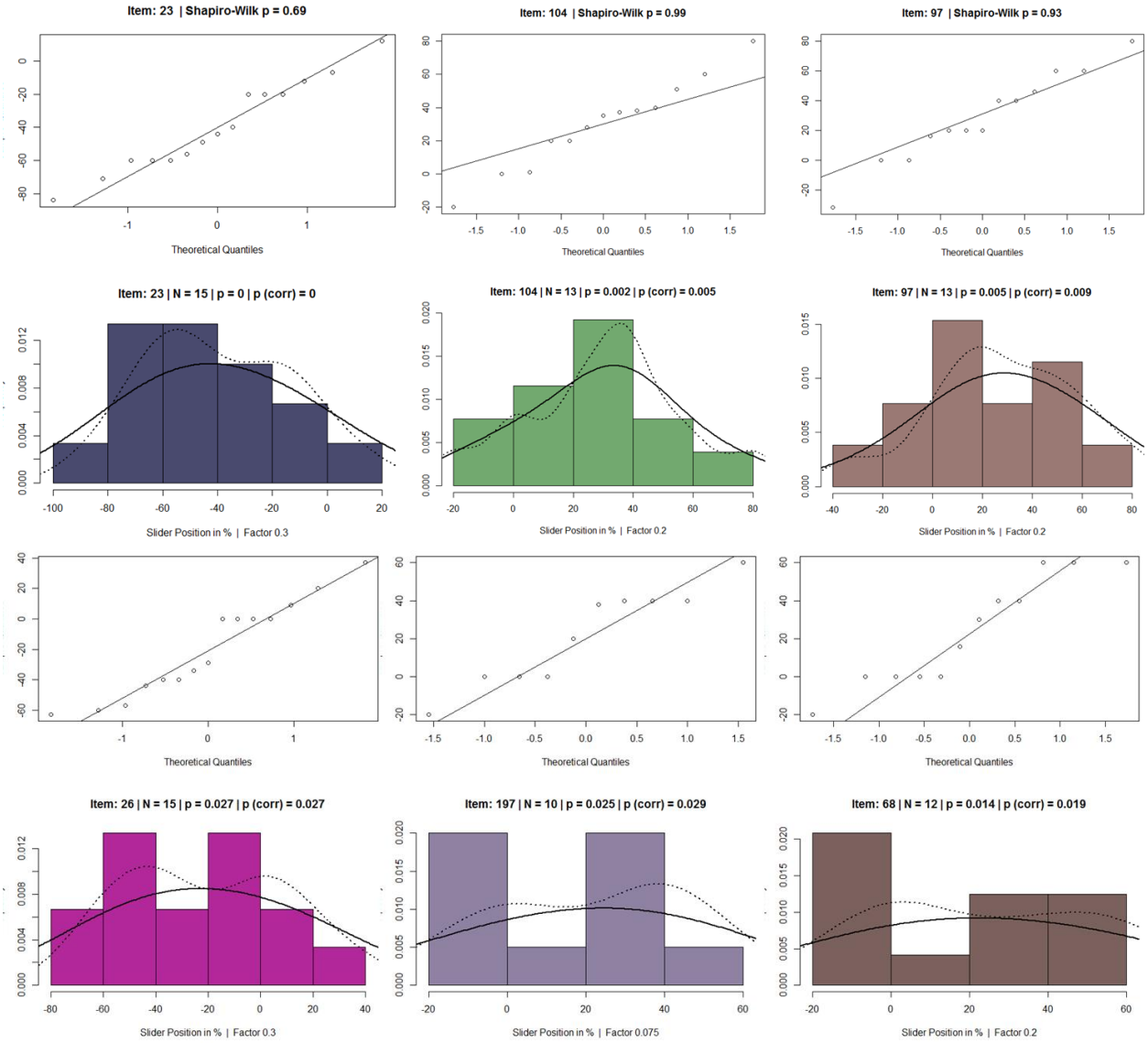


***Figure 10**: Distribution and modality of probe positions. **Top row**: colors for which the distribution was unimodal and dimensional offset was one-directional, meaning that the similarity judgments of most participants were towards the same direction along a common dimension and neighbor. **Bottom row**: examples of probes for which the distribution was bimodal while having each mode point towards opposite offset directions. This indicates two distinct groups of people. Along a single dimension, one group judges the color to be more similar to neighbors with lower values, the other judges the color to be more similar to neighbors with higher values.*

Fourth[3], some participants contradicted their own similarity ratings when d_{PN} was increased or when probe and neighboring nodes interlaced with comparable trials,

[3] Directionality incongruence is not included in the analysis script for R, instead it can be queried case-wise directly from the SQL database using the `GetIncogurentDirections.sql` query available in the additional materials.

meaning that their judgment of a probe's position reversed directions when compared with different probes along one and the same dimension (Figure 11). This inconsistency in directionality indicates that, for some participants, the CIE colors were not organized in the "correct" order along one dimension. For example, when progressing along one dimension, instead of organizing tones as $blues \rightarrow reds \rightarrow yellows \rightarrow greens$, a participant's representational space organizes them as $reds \rightarrow blues \rightarrow greens \rightarrow yellows$.

Though it is possible that a L* u* v* transformation of the CIE (see Mahy, Van Eycken, & Oosterlinck, 1994) would eliminate this effect, since the transformation "bends" the XYZ dimensions which might cause the colors to be positioned in the correct order again. But because the results also show bimodality (two groups of people contradicting each other on similarity judgments), it is unlikely that any transformation will eliminate this problem for all subjects and achieve a universally uniform color space.

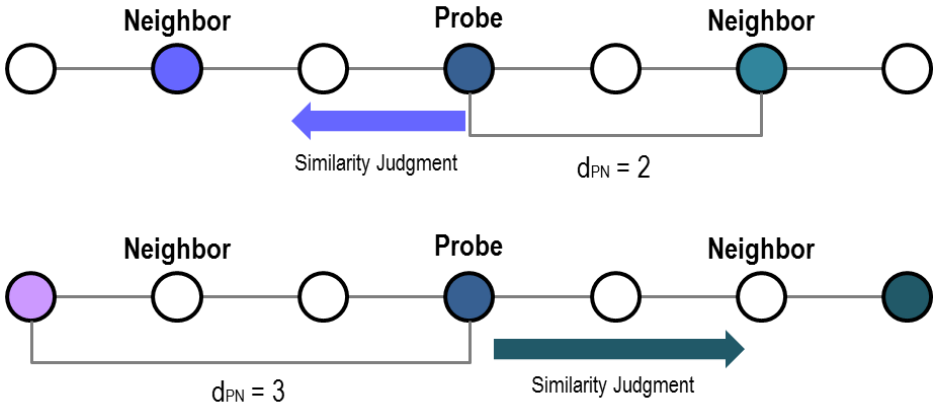


Figure 11: *Incongruent similarity judgments with differing d_{PN}. When the probe-to-neighbor distance is 2 nodes, the participant judges the color to be more similar to the color with lower value along the dimension being examined. When probe-to-neighbor distance is 3 nodes, the participant judges the color to be more similar to the color with higher value along the dimension being examined. This indicates that a probe's positioning along the same dimension changes depending on which neighbor it's being compared to, contradicting the order with which the colors are arranged in the CIE.*

With these findings, utilizing the CIE for stimulus generation under the assumption that (a) proximal colors will correspond to a category and (b) category radius and distance between categories will be directly related to cohesion seems rather

risky. Not only are there strong indicators that the assumption of universality – i.e. that the distance metric corresponds to one and the same dissimilarity metric for all individuals – cannot be trusted, but violations of the assumptions discussed in Chapter 2.2 are also likely to occur.

Since for some probes it has been shown that their position after a similarity judgment deviates considerably from its position as postulated by the CIE (see Figure 9), the principle of *equal self-similarity* (exemplar-probe pairs with equal distances also have equal similarity) is very likely to be violated if the CIE Euclidean distance between colors is taken to reflect distance between colors in the psychological space.

Because of bimodality on the similarity distribution of some nodes, and because of the cases where incongruent directionality was encountered, the principles of *intradimensional subtractivity* and *triangle inequality* might be violated as well (see Chapter 2.2). That is, the component-wise addition within one dimension might yield different distances depending on which node combination is selected.

For these reasons I've decided against using the CIE and jeopardizing the rather laborious experiment battery which would follow. Instead, I would develop a system for acquiring distances between colors at an individual bases. This means that every participant would have a "color-space" tailored to their psychological representation, and thus providing stronger support for a priori assumptions about similarity during stimulus construction.

In the following chapters I will discuss the method developed for achieving this.

4 A Data Driven Method for Constructing Category-Based Stimuli

Various studies on categorization have used have used multidimensional scaling methods in an attempt to reconstruct individual representational spaces of study participants (Bimler & Kirkland, 2009; Lee, 2001; Nosofsky, 1991; Tokunaga & Logvinenko, 2010), there are however two major issues these methods suffer from which represent a significant barrier for the current study.

First, multidimensional scaling (MDS) makes assumptions about the number of dimensions of a metric space. Since the addition of dimensions reduces the *stress* of an MDS (difference between measured and reconstructed distances), we could erroneously assume the existence of dimensions which are not really there.

By abandoning the CIE as stimulus pool, we also abandoned certitude about the number of dimensions. Without this certitude errors in the assumption of the number of dimensions flow into the distance metric (see Chapter 2.2). This could then lead to false assumptions about cohesion of category groups.

It is thus necessary to establish a method of estimating a reasonable number of dimensions, or avoid the dimensionality assumption altogether.

Second, conventional methods are very time consuming and risk drop outs. In many cases, studies which utilize methods for individualized multidimensional scaling such as INDSCAL (Gazda & Mobley, 1981) have two measurement appointments. One for gathering the data for the reconstruction of individual representational spaces, the other scheduled after the MDS analysis where the actual experimental data is collected.

In such designs there is always a risk that a portion of the participants will not show up for the follow up experiment. Due to the time constraint within which this project needed to be completed and the limited number of participants at my disposal, a treatment-analysis-treatment design would prove itself unfeasible.

Third, conventional stimulus triad methods such as the *odd-one-out* (e.g. Bimler & Kirkland, 2009; Bimler, Kirkland, & Jacobs, 2000), where tree elements are presented

and the participant must choose the one which is less similar to the rest, might be economical in terms of the time needed to complete a session, but they tend to produce rather high residuals which lead to less accurate spacing (`[Test MDS.R]` in the additional materials provides a simulation). This is because on each trial, similarity judgments can only be measured in a binary basis (similar vs dissimilar). These judgments are then aggregated to form an ordinal similarity scaling among elements. For many studies this ordinal scale might suffice, but the current study requires a metric organization of the colors utilized as stimuli. A further disadvantage is the number of comparisons this method requires, namely, all possible 3-combinations of n elements. It might be feasible with a stimulus pool as small as 13 elements, but with a larger stimulus pool of, for example, 50 elements this method would require ${}_{50}C_3 = 19600$ trials in total to get comparisons among all of them.

Some researchers (e.g. Nosofsky, 1992) circumvent both problems by collecting similarity ratings between 2 items using a rank scale (e.g. from "very similar" to "very different"). This eliminates the high inaccuracy of the odd-one-out method and greatly reduces the number of trials (i.e. $N_{trials} = N_{elements} \cdot [N_{elements} - 1]$), but still requires a considerable amount of trials to get all comparisons with a larger stimulus pool. Furthermore, similarity judgments between two elements might fall differently with the presence of other elements than when they are presented alone (Goldstone, 1994, 1995). For example, one might judge *cadet blue* (■) less similar to *avocado green* (■) when the two are judged by themselves than when they are presented in conjunction with other colors for which similarity judgments are also required, since *cadet blue* is much more similar to *avocado green* than both *cadet blue* and *avocado green* are to *red* or *yellow*.

4.1 The Color Panel Arrangement Method

The ideal similarity judgment task would let the participant judge all color stimuli at once, in a metric scale, in such way that their contextual relation to other color groups is also reflected.

Something like this could be accomplished by letting participants spatially arrange a group of elements in a two dimensional plane (Figure 12). The elements should be arranged in such way that similar elements are positioned close together, while dissimilar elements are kept far apart from each other. This way, participants can make similarity judgments while maintaining the context in which they are making said judgments always visible. The visibility of the similarity ensemble should work best with visual stimuli which require little to no conscious processing such as colors. That is, the participant can process at a glance whether or not their composition makes sense as a whole.

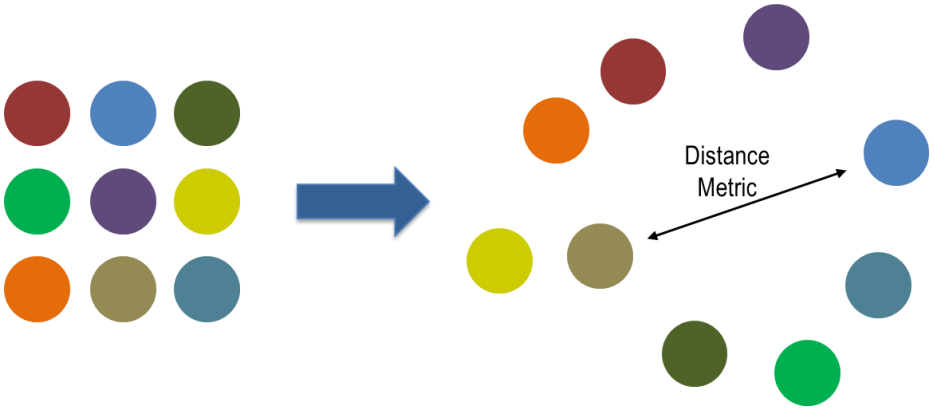


***Figure 12**: 9 Colors being positioned on a plane according to how one subjectively perceives their similarity among each other to be. The distance between the colored circles can be used as an indirect measure of dissimilarity.*

A further advantage of the method here proposed, is that the similarity judgments are (near) metric. The distance between elements can be read almost directly as a dissimilarity measure, without having to infer it from data aggregated over multiple trials as it is the case with *odd-one-out* tasks, and without a context detached ranking scale as it is the case with pair-wise similarity tasks.

There are however, some caveats regarding this method. First, an overflow of stimuli should be avoided. Positioning 9 colors among each other according to their similarity is easy; positioning 50 of them is rather difficult.

Due to the limitation to 2 dimensions, the composition might never really reflect actual similarity representation. The reason is that the further away one goes from the center of the plane the more likely it is to encounter dissimilarities which are not actually there. This is a phenomenon implicit in all reductions of dimensionality and the world map serves as a good example.

It is impossible to represent the world map in 2 dimensions without compromising the accurate representation of the area or proximity of some geographical regions. Greenland is represented as being much larger than it actually is due to stretching out the Earth's curvature near the North Pole into a plane centered at the equator. Another example is the relation of the US state of Alaska to East Russia. They are plotted on opposite sides of the map even though they neighbor each other.

4.2 The Set Cover Problem and Its Solution

The previously illustrated problem of stimulus quantity and reduced dimensionality can be circumvented by letting participants organize different combinations of stimuli and then combining the observed distance measures. That is, when 50 colors need to be compared against each other, instead of tasking participants with organizing all 50 at once, the colors could be distributed among multiple panels of 10 colors each. The participant could then organize these colors without being overwhelmed or compromising on distance measures for colors on the periphery of the plane.

The problem then lays in determining what color combination should be chosen per panel and how many panels are necessary until we have gathered distance metrics of every possible color pair. This task can be formalized as a set cover problem:

> Given an universe $U = \{e_1, e_2, e_3, \dots, e_n\}$ composed of n elements, find the smallest set of sets $Q = \{S_1, S_2, S_3, \dots, S_q\}$ containing q subsets composed of a maximum of k elements of U, so that for every possible element pair $\{e_x, e_y\}$ in U there is a subset S_z which contains both elements and is itself an element of Q:

$$\forall x \forall y \left((\{e_x, e_y\} \in U) \rightarrow \exists z \left((\{e_x, e_y\} \in S_z) \wedge (S_z \in Q) \right) \right) \tag{4.1}$$

For example, for an universe with 5 elements $U = \{1,2,3,4,5\}$ the minimal solution which satisfies Condition 4.1 for a maximal subset size k of 3, would be $Q = \{S_1\{1,2,4\},\ S_2\{2,3,5\},\ S_3\{3,4,5\},\ S_4\{1,3,5\}\}$ which is irreducible and contains all possible $\{e_x, e_y\}$ pairs within a subset S_i at least once (Figure 13).

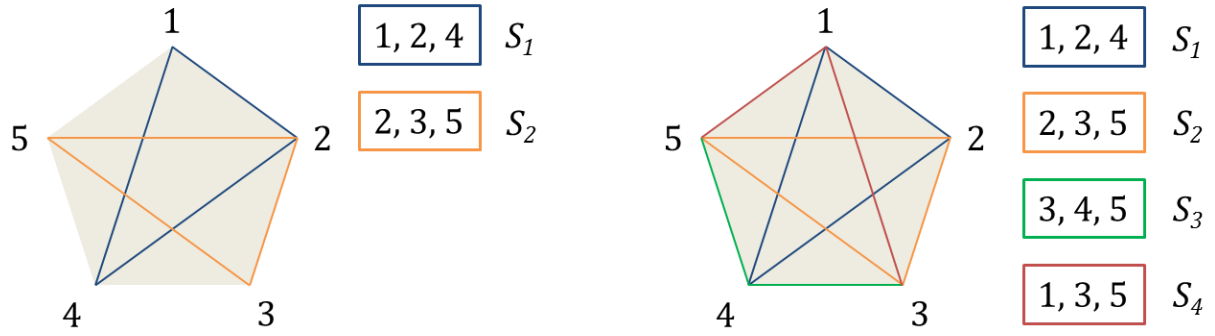


Figure 13: *Graphical Solution for 4.1 with* $n = 5$ *and* $k = 3$. *The vertices of the pentagon represent the elements of the universe U, each line represents an element-pair contained in a subset of Q, and each color represents a different subset* S_z *in Q. First (left), we pick any 3 elements of U to make* S_1*(blue), then we select the 3 elements for the next subset in such way that* S_2*contains no element pair which is already contained in Q, in other words, without drawing over existing lines (orange). The next two subsets (red, green) cannot be drawn without repeating already compared pairs.* S_3 *(green) is composed of elements 3, 4 and 5 the pair {3,5} was already contained in* S_2*; pairs {3,4} and {4,5} are new. Finally, the last set (red) is drawn to include the remaining missing pairs {1,3} and {1,5}. The solution produces Q with 4 different sets. The pentagon on the right provides proof that all elements have been grouped with every other element at least once because all vertices of the pentagon have been connected.*

The algorithmic solution presented next is based on the conjecture that a minimum for q which satisfies Condition 4.1 can be found by dividing U into subgroups $\{G_1, G_2, \dots, G_m\} \subset U$ of sizes $\lceil k/2 \rceil$ and $\lfloor k/2 \rfloor$, and having Q correspond to all

combinations among elements of G. The justification for this conjecture is implicit in the algorithmic solution explained below.

First, let us divide U into distinct groups ($\{G_1, G_2, \dots, G_m\} \subset U$) which never share the same element and can only contain a maximum of $k/2$ elements (Figure 14). Because the size of each group never larger than $k/2$, whenever we combine two groups (G_i, G_j) we create a set $S^*_{i,j}$ which contains all elements from both groups and is itself never larger than k.

If we were to create a set of sets Q^* which contained all possible $S^*_{i,j}$ without repetition and then examine the properties of Q^*, we would see that every element of U occurs in conjunction with every other element at least once within some subset in Q^*. This is the case because for any element pair $\{e_i, e_j\}$, either (a) e_i was in the same group ($G \subset U$) as e_j, in which case they are bound to be in some subset of Q^* because elements Q^* are composed of U-groups; or (b) e_i was in a different group than e_j, in which case both elements would end up together in some $S^*_{i,j}$ because $S^*_{i,j}$ is defined as a combination of two groups of U, and Q^* contains all possible $\{G_i, G_j\}$ combinations. We can thus say that Q^* satisfies Condition 4.1.

Input: $U = \{e_1, e_2, \dots, e_8\}$, $k = 4$.
Output: $Q^* = \{S_1, S_2, \dots, S_6\}$.
Algorithm:
Divide U into groups $\{G_1, G_2, \dots, G_{\frac{2n}{k}}\}$ of size $\frac{k}{2}$;
Define set of sets Q^*;
For each 2-combination $\{i, j\}$ of $\left\{1,2, \dots, \frac{2n}{k}\right\}$
{
  Select all elements from $G_i \cup G_j$ into $S^*_{i,j}$
  Add $S^*_{i,j}$ to Q^*
}
Return Q^*;

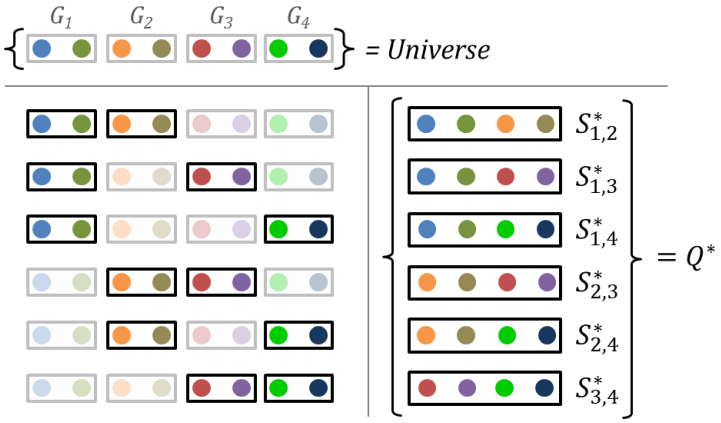


***Figure 14**: Algorithmic solution for the set cover problem with $n = 8$ and $k = 4$. 8 colors must be divided into subsets of 4 colors in such way that every color is combined with every other color at least once (Condition 4.1). This must be done with the least amount of subsets possible.*
First, the colors are divided into distinct groups (grey rectangles labeled G_1 to G_4; top row) half the size of the subsets found in Q^*. The combination of all groups (left: black rectangles) will ensure that all colors of one group is compared with all colors of all other groups. Thus, the resulting subsets $S^*_{i,j}$ (right) correspond to the minimal number of q subsets.

Furthermore, we would see that Q^* is minimal, because it is impossible to remove a $S^*_{i,j}$ from Q^* without eliminating an unique occurrence of some pair $\{e_i, e_j\}$. In other words, we cannot reduce the size of Q^* without violating Condition 4.1.

In order to minimize the number of possible $S^*_{i,j}$ one must maximize the size of the groups which divide U (smaller group size means more groups, which in turn means more possible combinations), and because $k/2$ is the maximum group size which allows us to create $S^*_{i,j}$s that are smaller than k, Q^* must contain the smallest amount of subsets possible without violating Condition 4.1.

With this conjecture we can postulate maximal and minimal expected values for q. Since all subsets in Q are composed by all two group combinations of U, we can use the binomial coefficient ${}_N C_K$ to calculate the expected values. N is the total number of groups in U given the allowed maximum size k of a subset $S^*_{i,j}$, and K is set to 2 because groups of U are combined pair-wise to form $S^*_{i,j}$.

The largest expected value of an optimal solution for Q on an universe with n elements using subsets with a maximum size of k is given by:

$$q_{max} = \frac{\left\lceil 2n/k \right\rceil !}{2\left(\left\lceil 2n/k \right\rceil - 2\right)!} \tag{4.2}$$

and the smallest expected value is given by:

$$q_{min} = \frac{\left\lfloor 2n/k \right\rfloor !}{2\left(\left\lfloor 2n/k \right\rfloor - 2\right)!} \tag{4.3}$$

Going back to the color panel method, the universe U represents all the color stimuli for which we need similarity judgments, the subsets $S^*_{i,j}$ represent the 2D planes, or panels, where the participants can arrange the colors (Figure 12), and k determines the number of colors present in every panel. Q tells us how to compose our panels so that we get distance measures among all colors.

As the equations above reveal, the color panel arrangement method is far more efficient than conventional methods for collecting similarity data. While pair-wise similarity ratings and the odd-one-out method would need 435 and 4060 trials respectively in order to gather the necessary similarity ratings among 30 elements, the panel arrangement method only needs $q = 15$ panels with $k = 10$ colors each (Table 3).

Table 3: *Solution for* $U = \{1,2,3 \ldots,30\}$ *with* $k = 10$

Panels	Element Indexes									
S_1	1	2	3	4	5	6	7	8	9	10
S_2	1	2	3	4	5	11	12	13	14	15
S_3	1	2	3	4	5	16	17	18	19	20
S_4	1	2	3	4	5	21	22	23	24	25
S_5	1	2	3	4	5	26	27	28	29	30
S_6	6	7	8	9	10	11	12	13	14	15
S_7	6	7	8	9	10	16	17	18	19	20
S_8	6	7	8	9	10	21	22	23	24	25
S_9	6	7	8	9	10	26	27	28	29	30
S_{10}	11	12	13	14	15	16	17	18	19	20
S_{11}	11	12	13	14	15	21	22	23	24	25
S_{12}	11	12	13	14	15	26	27	28	29	30
S_{13}	16	17	18	19	20	21	22	23	24	25
S_{14}	16	17	18	19	20	26	27	28	29	30
S_{15}	21	22	23	24	25	26	27	28	29	30

4.3 Using Partition Around Medoids To Acquire Color Category Groups

The color arrangement method previously described is expected to yield fairly accurate similarity ratings. We now need a method for identifying possible color categories in the representational space. A viable solution is to search for clusters of elements within the gathered similarity data. However, most clustering algorithms such as the *k-means* (Jain & Dubes, 1988) require spatial data to perform the analysis. This means we would have to employ some form of multidimensional scaling which in turn would require us to make assumptions about the number of dimensions in the representational space.

As mentioned in the introduction of this chapter, assumptions about dimensionality in an MDS might distort the "real" representational space by forcing element positions in such way that the overall stress is reduced (difference between measured distances and distances calculated after ascribing a position to an item in an n-dimensional space). There are linkage-based methods of hierarchical clustering which are not dependent on any assumption about dimensionality (Defays, 1977), but the main problem with such methods is exactly their lack of spatial "metricity".

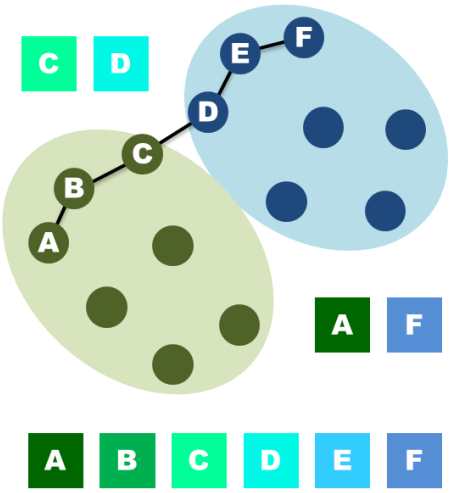


***Figure 15**: Illustration of the chaining phenomenon in linkage-based clustering algorithms. A chain is formed across categories (green and blue) linking the nearest neighbor from element A to element F. Below is an illustration of a color combination from different tones of green and blue which could lead to such chaining. Notice that color A is very different from color F and would otherwise not be classified as belonging to the same category.*

Linkage-based methods use the ranks between distance measures going from one element to another rather than analyzing "regions" of elements which would be more in line with the assumptions made about ease of categorization. This problem is most pronounced in the so called *chaining phenomenon* (Blashfield, 1976).

In single linkage, an element is added to a cluster when its dissimilarity to the cluster element it's being compared against is very small. The algorithm then iteratively compares the newly linked element and its neighbors for further links. The problem with this procedure is that it leads to one element linking to the next one in an elongated *chain* of elements. In the end, the first element of that chain is very dissimilar to the last and should actually not be grouped in the same cluster (Figure 15).

More modern variants of linkage-based clustering seem to be less susceptible to the chaining phenomenon by using methods such as *must-link* and *cannot-link* directives (Jain, 2010) or by applying hierarchical *farthest neighbor clustering* (Everitt, Landau, Leese, & Stahl, 2011).

Still, the chosen computational framework together with the measure for ease of categorization stipulated in this work (Chapter 2.5) are likely to work best with algorithms such as the *k-means* which generate clusters derived from an estimated "category center".

Just like the previously planned stimulus construction using the CIE (Chapter 2.6), deriving clusters from the center of a category is likely to yield items which provide a better estimate for cohesion since all members will be selected based on their distance to the "most prototypical" category representative.

The method for finding color clusters in the date from the similarity measures should thus:

(a) Not require spatial organization of the elements, since we want to avoid multidimensional scaling.
(b) Build ellipsoid clusters departing from an estimated category center in order to avoid high dissimilarity among elements on the periphery, and to prevent the inclusion of category external elements.
(c) Be able to compute clusters from an arbitrary dissimilarity matrix.

Given the above constraints and considerations, the Partition Around Medoids (PAM) method (Kaufman & Rousseeuw, 1990) seem the most appropriate technique to find clusters in the data produced by color similarity judgments.

PAM is able to construct clusters based solely on a dissimilarity matrix by finding medoids, elements of a dataset whose average dissimilarity among other members of a given neighborhood is minimal.

Yet another advantage of PAM is that it finds the best configuration by minimizing the sum of the distance $dis(e_i, \mathfrak{m}_h)$ of every data point e_i to its closest medoid $\mathfrak{m}_h$; described as the cost $\mathbb{C}_P$ of a configuration with k clusters:

$$\mathbb{C}_P = \sum_{h=1}^{k} \sum_{e_i \in C_h} dis(e_i, \mathfrak{m}_h) \tag{4.4}$$

This falls in line with GCM, which states that the probability of classifying an element as a member of a category (Equation 2.5) is inversely related to that element's distance to other exemplars of said category (Equation 2.2). In other words, a configuration which ensures the lowest cost will also ensure that elements within a cluster will generate higher exemplar activation within the cluster than outside of it.

4.4 The Dunn Index as Ease of Categorization

In chapter 2.5 a metric for ease of categorization was proposed (Equation 2.14) dependent on the relation between the n-volume of the target category and the distance between categories. Since the CIE color space has been abandoned as a means of item generation, and because we want to avoid making unnecessary assumptions about the number of dimensions in a participant's specific representational space, a new metric for ease needs to be chosen which does not depend on spatial coordinates.

Ideally, one would have to infer the volume metric for the target cluster form the cluster's silhouette in combination of distances within a category, versus the combination of distances between elements of different categories. Unfortunately, such approach would prove itself too computationally expensive [4], likely making it impossible to generate user-specific trials on the fly.

[4] The algorithm for trial generation in the modified Sternberg Task needs to compute every permutation of every 2-cluster combination (see Chapter 7.1).

As an alternative, a special case of a metric for evaluating clustering algorithms introduced by J. C. Dunn (1973) called the Dunn Index (*DI*) was used. The index is used by a multiplicity of clustering algorithms in order to identify compact, well separated clusters while minimizing inner cluster variance (Maulik & Bandyopadhyay, 2002; Pakhira, Bandyopadhyay, & Maulik, 2004). *DI* is defined as the minimization of the ratio of function $\delta(C_i, C_j)$ which calculates the inter cluster distance metric between clusters i and j, to the maximization of function $\Delta(C_k)$ which calculates the inner diameter metric for the largest cluster k of configuration m.

$$DI_m = \min_{1 \le i \le m} \left\{ \min_{1 \le j \le m, j \ne i} \left\{ \frac{\delta(C_i, C_j)}{\max_{1 \le k \le m} \Delta(C_k)} \right\} \right\} \tag{4.5}$$

Two aspects were key when choosing Dunn's index as a proxy metric for *ease of categorization*. First, it's only dependent on distance measures, not actual element spatial coordinates. Second, its formulation (Equation 4.5) is conceptually similar to the previously proposed measure for ease of categorization (Equation 2.14).

DI doesn't directly impose a specific method for calculating the inter-cluster distance $\delta(C_i, C_j)$ nor the inner diameter of clusters $\Delta(C_k)$. Though the partition around medoids implementation utilized in this study does provide a cluster silhouette metric (Maechler et al., 2014), its calculation for 4 element clusters would be rather wasteful. Another common method (Maulik & Bandyopadhyay, 2002) is to search for the maximal dissimilarity among elements within a cluster, and utilize that distance as an inner diameter measure. For the inter-cluster distance $\delta(C_i, C_j)$ the shortest distance between the element pair $d(c_i, c_j)$ where $c_i \in C_i$ and $c_j \in C_j$ is taken.

Reaction times are expected to decrease with both decreasing distances between probe and exemplars of the target category and increasing inter-cluster distance.

From Equations 2.2 and 2.3 we learn that the relation between distance and exemplar activation, as well as the relationship between activation and the expected time for an item to finish the race follow an exponential decay.

From Equation 2.10 and 2.5 in conjunction with 2.7 we take that the expected reaction time is inversely proportional to the sum of activation. Gardner and colleagues (1959) demonstrate that the sum of exponential decay functions with varying decay rates expressed as:

$$s(t) = \sum_{i=1}^{n} N_i \cdot e^{-\lambda_i \cdot t} \tag{4.6}$$

solves to an exponential decay function for $(N_i \wedge \lambda_i) > 0$.

This means we can expect reaction times to follow an exponential decay (see also: Chapters 2.3 to 2.5), the formalization of which is stipulated here as:

$$\begin{aligned} E(RT|DI_{\mathrm{AB}}) &= \Delta_{RT} \cdot e^{-\lambda \cdot DI_{\mathrm{AB}}} + RT_{min} \\ \Delta_{RT} &= RT_0 - RT_{min} \end{aligned} \tag{4.7}$$

where $E(RT|DI_{\mathrm{AB}})$ is the estimated reaction time for a categorization task involving two categories (A and B) given a DI value. RT_{min} represents the shortest possible reaction time, and RT_0 represents the reaction time expected when DI is zero, which in turn coincides with $\hat{A}^2$, the maximal estimated number of steps until a decision boundary is reached (Chapter 2.5).

The decay rate λ is translated to accommodate minimum and maximum reaction times. The vertical translation is given by RT_{min} and implies that as DI increases, the reaction time asymptotically approaches the fastest possible response time.

By solving the exponential decay function $f(DI_{AB} + DI_{\emptyset})$ for $DI_{\emptyset}$, the smallest Dunn Index value for which reaction times predictions can be made, we get the horizontal shift:

$$DI_{\emptyset} = \frac{\ln(\Delta_{RT})}{\lambda} \tag{4.8}$$

I'd like to remark that this work lacks the mathematical derivation for the expected step count in the *gambler's ruin* problem (see Chapter 2.4) as a function of *DI* and decision boundary $\hat{A}$. Thus the derivation of expected reaction time given a Dunn Index value is heuristic, not formal. That's why the validity of Equation 4.7 was tested through both simulation (Chapter 4.4.1), and experimental data (Chapter 8).

Another characteristic inherent to the computational model when *DI* is taken as measure of category cohesion is that for small *DI* values, the error variance for RT estimates is expected to grow. This is due to the fact that the closer together the two categories are, and the larger their volumes, the stronger the influence of the probe's position relative to exemplars is (**Figure 16**).

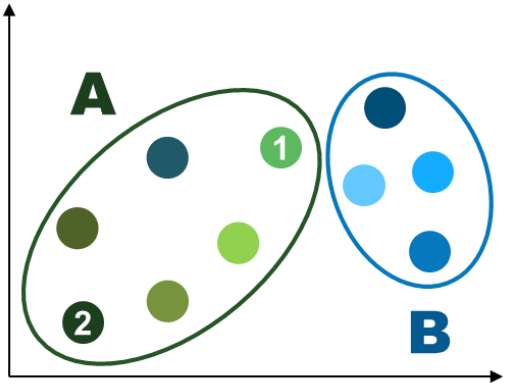


Figure 16: *Higher error variance for RT predictions for low values of the DI. A low DI configuration with two neighboring categories A and B. Element 1 of category A is nearer to exemplars of category B then many other elements of its own category. This is not the case for element 2. Due to the higher relative activation of the irrelevant category B, RT is expected to be higher for element 1 than it is for element 2.*

When two categories, A and B, are close together and an element bordering the irrelevant category is selected as probe (element 1 in **Figure 16**), we expect higher reaction times because the activation of the exemplars in the irrelevant category in relation to the activation of the exemplars in the relevant category will also be high. Conversely, if we select an element which is distant from the irrelevant category (element 2 in **Figure 16**) we would expect much shorter reaction times.

This variability in reaction times is not expected when categories are far apart from each other and should nearly disappear because the relative distances between target-category exemplars and irrelevant-category exemplars should remain more or less the same regardless of which element we choose as probe.

4.4.1 Simulation of the Correspondence of the Dunn Index to Ease of Categorization

Because there is no direct mathematical relation between either $\delta(C_i, C_j)$ and the denominator in Equation 2.11 or $\Delta(C_k)$ and the numerator in Equation 2.5 (the optimization of which was used to derive Equation 2.14; see Chapter 2.5) a simulation was conducted in order to assure that color groups with high *DI* would indeed correspond to lower reaction times according to EBRW, and could thus be considered as *high cohesion* lists.

The simulation was run using R (version 3.0.3), the library clValid (Brock, Pihur, Datta, & Datta, 2011) and the implementation of EBRW written for this study (the code for the simulation as well as the R implementation of EBRW are available in the additional materials).

The simulation randomly constructed cluster pairs with 41 varying element *dispersions*. The elements were distributed normally with a standard deviation equal to a fixed set inter-cluster distance times a *dispersion factor*. From each cluster, the element closest to the neighboring cluster was selected as probe. This selection method was expected to give the most conservative measure of ease

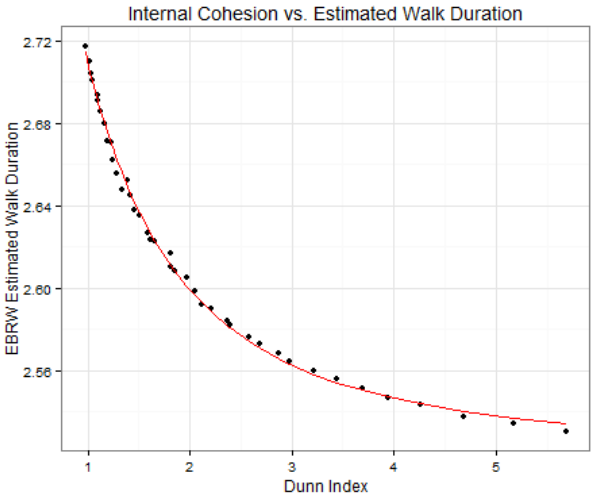


***Figure 17**: Reaction times estimated by EBRW versus Dunn Indexes of simulated categorization tasks. The data shows an exponential decay of reaction time with increasing Dunn Indexes. The red line represents the nonlinear regression fit of Equation 4.7.*

of categorization because it maximizes the activation of the irrelevant category. The distance measure was Euclidean with equal weights for all dimensions.

For each cluster pair, the Dunn Index was calculated. For the expected reaction times, a categorization task was simulated using the R implementation of EBRW.

The cluster construction and categorization task simulation was repeated 100 times under the same configuration (element distribution mean and standard deviation).

The mean reaction time and *DI* of these 100 simulations were calculated and added as a data point. The entire process was repeated 1'000 times for each dispersion value, yielding reaction time estimations for increasing *DI* values (Figure 17).

The data from the simulation was used to fit Equation 4.7 using a nonlinear least squares regression (Bates & Watts, 2007) using the R standard statistics package (Fox, 2002).

Convergence was achieved after 5 iterations with tolerance equal to 6.82E-6. All free parameters estimations (Δ_{RT}, RT_{min} and λ) were statistically significant with $p < 0.001$.

The simulation suggests that the model for expected reaction time described in Equation 4.7 provides a good estimate. This means that we should be able to generate trials with varying degrees of ease of categorization by manipulating the Dunn Index through element selection.

4.5 Stimulus Generation and Construct Validation

With the methods presented in this chapter, it should be possible to generate trials for the Sternberg Task just as previously intended with the CIE color space (Chapter 2.6). The detailed procedure for stimulus construction will be discussed with their respective experiments in the next chapters. For now, I'll just present a brief overview of the construction plan.

To serve as stimulus pool, a collection of colors is selected where, from memory alone, every color is distinguishable from every other color. That is, no change blindness should occur when two different colors are presented sequentially separated by an inter-stimulus time interval. These colors are distributed into subsets according to the cover set solving algorithm proposed in Chapter 4.2. These subsets constitute the individual panels in a color arrangement task.

With the similarity measures gathered in the color arrangement task, a dissimilarity matrix is constructed for every participant individually, and then ran through a PAM algorithm in order discover how these colors are grouped together. The color clustering information is crucial because (a) when creating *baseline* items for the Sternberg Task, populating a list with colors of the same category should be avoided; and (b) when constructing items for *category-based* trials, we want to have some information about which group of colors is most likely to form a category according to an individual participant's subjective perception.

Finally, we can construct lists for the Sternberg Task and calculate their *cohesion* utilizing the Dunn Index. By selecting lists with differing *DI*s, the formalized ease of categorization (Equation 4.7) can be tested empirically through a categorization task. This should shed some light as to whether the operationalization and the methods described in this chapter actually work. That is, whether the response patterns produced by the participants do produce the same exponential decay as expected by the model.

5 Study Design and Methods

With a computational framework, proper operationalization, and a strategy to produce the required stimuli, a 5 part experiment battery was designed to study the influence of categorization in short term memory. The battery was constituted of:

The Ishihara Colorblindness Test (Ishihara, 1981): Used to ensure no participant was color blind.

Color Arrangement Panels (Experiment 1): Designed to gather similarity judgment of each participant individually which would then be used to construct the items for experiments 2, 3 and 4.

Modified Sternberg Task (Experiment 2): A new task similar to the one described in Chapter 1.4, where participants need to distinguish whether a probe color presented with a list label was present in the correct list. Trials were constructed algorithmically utilizing a participant's color similarity judgments after they finished Experiment 1.

Categorization Task (Experiment 3): A task where the participants were asked to classify C-Intrusions form Experiment 2. Used to test whether the algorithm used to construct items for Experiment 2 actually produced groups constituent of the same category, and whether the Dunn Index provides a good measure for ease of categorization.

Change Blindness (Experiment 4): Conducted with selected elements from high cohesion trials of Experiment 2, in order to ensure colors which had been judged as highly similar in Experiment 1 could still be distinguished in memory.

5.1 Materials and Participants

In order to fit the execution of the experiment battery in a feasible timeframe, the choice was made to host the study online maximizing the number of potential participants and avoiding other constraints inherent to lab management.

For this purpose, a custom web application was build using ASP.NET to host the experiments. The Microsoft Silverlight browser plugin was used to create a client-side application which would handle trial presentation and user input. The choice for using Silverlight over JavaScript was met mainly due to extensibility (the experiment battery was constituted of trials which varied greatly in their presentation) and higher accuracy when measuring reaction times.

Experimental data was stored using a relational database in a MS-SQL 2008 R2 server. For statistical computation and stimulus construction, the R engine was used and integrated in the application using the R.NET framework.

The complete source code for the web application, Silverlight client and R-Engine as well as code- and architectural documentation are available in the additional materials.

In order to participate, subjects had to register with a username and password of their choice. They were required to enter their age and sex when registering. Participants had also the option of leaving their full name and email address in order to receive their experimental data and a report containing their personal results.

After logging in, participants were redirected to a welcome screen containing the list of experiments, which could be started depending on that participant's progression through the battery. Experiments were "unlocked" sequentially depending which experiments the participant had completed up until that point.

In the beginning, all experiments are locked except for the Ishihara Test for Colorblindness. Once completed, it unlocked Experiment 1, which in turn unlocks the remaining 3 experiments. Participants were given the choice to complete different

experiments at different times. They could log out after completing an experiment and then login with their username and password at a different date to proceed with the remaining experiments.

The experiments ran between the 25th of October and 10th of November 2014. A total of 36 participants (18 male, 18 female) between the ages of 18 and 38 years ($M = 27y$) completed the entire experiment battery within that period.

Participants were not compensated but received a personalized report containing information about their performance and reconstructed color-space. Psychology undergraduates who participated (9 in total) received 1.5 hours of participatory accreditation.

A considerable drop out occurred after the first block of the modified Sternberg task (Experiment 2). From the 73 participants who completed the first experiment, 32 quit after the first block of the modified Sternberg Task. 3 other participants were removed from the final analysis: one because of faulty reaction time data, two other due to a period of inactivity registered during the Sternberg Task (Experiment 2).

Three further candidates reported not being able to run the software in their computer.

6 Experiment 1 – Color Arrangement

The first experiment used the color arrangement method proposed in Chapter 4 to gather similarity judgments from the participants. These similarity judgments were then combined into a similarity matrix which would serve as basis for clustering and list construction for the Sternberg Task.

The stimulus pool was constituted of handpicked colors. No preliminary study was conducted to test for change blindness. Instead, colors which were similar to each other were put in a timed Microsoft PowerPoint presentation where the first color was presented for 1 second, followed by a blank screen also presented for 1 second, which was then followed by either the same or a very similar color. Distinction judgments as to whether the second color was the same as the first were made by me and two other judges.

Whenever a color pair could not be distinguished, one of the colors was eliminated from the pool. This process was repeated until all colors could be distinguished by all 3 judges. The final pool was constituted of 48 different colors (a list containing the hex-triplet for all 48 colors can be found in the additional materials under `[ColorDefinitions.csv]`). These 48 colors were distributed among 15 panels with 16 colors each using the set cover algorithm discussed in Chapter 4.2.

In order to prevent context bias (the influence of a specific configuration on similarity judgment; see Chapter 4), the color to panel assignment was randomized on a participant basis. That is, each panel was unique meaning that no two participants arranged the same panel.

Every panel was constituted of colored circles presented in grid formation against a white background (Figure 18). Participants were instructed to use the mouse to drag and drop colors into a configuration which would reflect how similar they perceived the colors to be. Similar colors should be positioned close together, while dissimilar colors should be kept farther apart.

Once a participant was happy with their composition, they could click the "next" button to progress to the next panel. When all panels were arranged, the participant was informed that the follow-up experiments were being calculated and would be ready within the next 3 minutes.

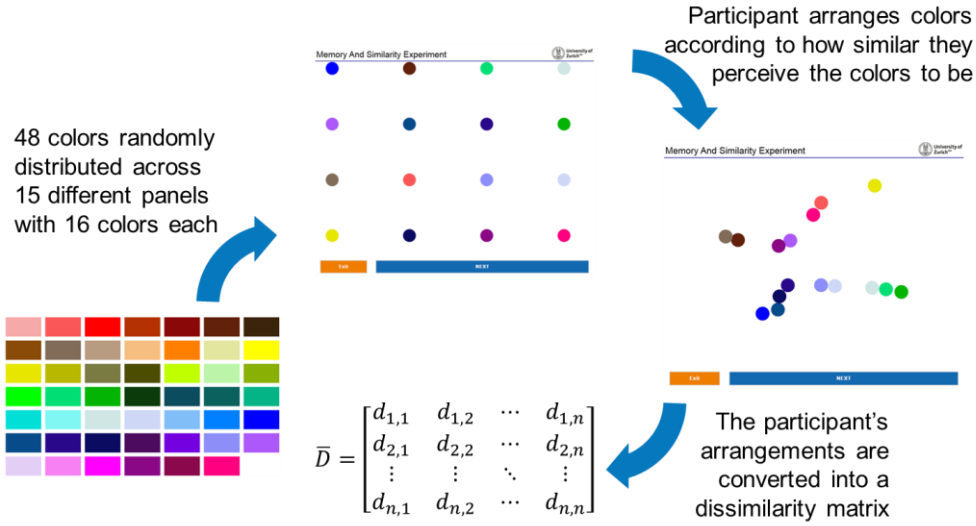


***Figure 18**: Schematic representation of the color arrangement experiment. 48 colors are distributed among 15 panels with 16 colors each. The color assignment is different for every participant to avoid context bias. Participants arrange colors depending on how similar they perceive the colors to be. The relational distances between the colors in all panels was converted into a dissimilarity matrix.*

6.1 Analysis and Clustering

The distances were calculated as a relative of the *total utilized area length*, which was defined as being the area of a rectangle with bounds defined by the smallest and largest x and y coordinates among all colors in a given panel. There are two main reasons for calculating distance measures this way:

First, because the experiment was conducted online, it was not limited to a particular monitor resolution. Participants with lower resolution monitors would produce shorter distances (measured in pixels) than participants with higher resolution monitors. In order to normalize distances, the number of pixels between colors relative to the number of available pixels in the x and y coordinates was used.

Second, the software gave the users the liberty to use almost the entire screen for moving and organizing colors. In many cases, especially with large monitors, only a

portion of the available area was used. If the total available area was used for normalization, participants with large monitors would end up producing too small dissimilarities in cases where they only used a portion of the available space, and too large dissimilarities in cases where they used the entirety of the space available to them.

Once the dissimilarity data was gathered and normalized, it was combined into a symmetrical 48×48 dissimilarity matrix by averaging duplicate distance measures. The dissimilarity matrix was then used as input for the cluster analysis conducted using the PAM algorithm implementation available in the clValid library (Brock et al., 2011).

clValid's implementation of PAM follows the implementation of most partition-based clustering algorithms (e.g. k-means) in that it is not concerned in estimating the optimal number of clusters which can most coherently describe the data.

There are multiple approaches to determining the appropriate number of clusters in a data set. My main concern was that applying a method which is solely dedicated to optimizing partitioning and cluster distinctiveness, for example by maximizing cluster validity indices (Maulik & Bandyopadhyay, 2002), might produce clusters which cannot be used for item construction because either (a) they are composed of too few items (in order to generate a list for the Sternberg Task, at least 4 items are needed: 3 for the list and 1 for the C-Intrusion), or (b) they might produce too few clusters, complicating the generation of baseline items or incapacitating a proper variation of list *DI* measures.

For this reason, a pseudo-hierarchical method for determining the number of medoids was implemented. First the number of medoids was stipulated at $k = \lfloor n/e_{min} \rfloor = 12$. That is, the highest integer which is smaller than the number of elements in the universe U (48 colors), divided by the minimum number of elements which should be present in every cluster. Any number larger than that, and U would necessarily have at least one cluster with less than e_{min} elements.

The algorithm than uses clValid's PAM to cluster the data into k clusters and the number of element in each cluster is examined. If the number of clusters containing less

than 4 items was larger than a threshold[5], the entire process was repeated with $k = k - 1$ until the number of clusters which contained less than e_{min} elements did not exceed the threshold.

The resulting clusters (Figure 19) were then used as basis for list construction in the Sternberg Task (the list construction method will be discussed in detail in Chapter 7).

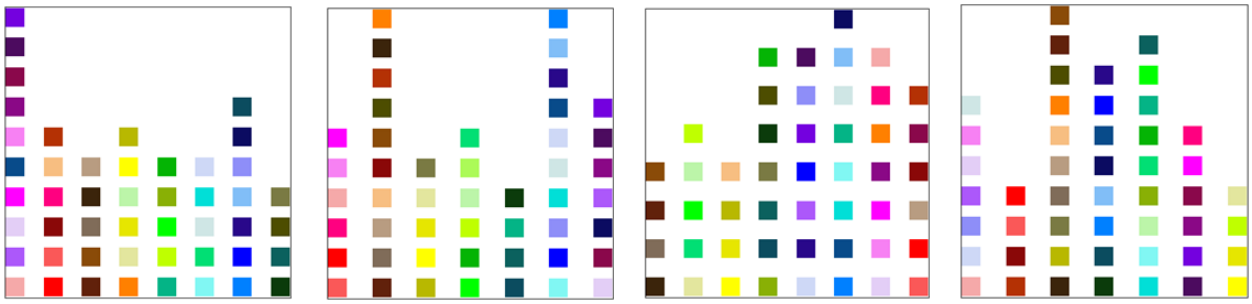


***Figure 19**: Example of color clusters of 4 different participants found using PAM. Each column represents a cluster and each square a different element.*

The distance data was also used to produce color space plots for the personalized results sent to every participant (Figure 20). For the spatial plots, Kruskal's non-metric multidimensional scaling (Cox & Cox, 2000) was used. The number of dimensions was determined by analyzing the gain in stress reduction when dimensions were added to the system.

The first MDS fit was executed with 3 dimensions. Then the total stress for that fit was calculated. The process was then repeated with an additional dimension. The stress from the previous MDS fit (3 dimensions) was compared against the stress of the MDS fit with the additional dimension (4 dimensions):

[5] The threshold represents the number of clusters which are allowed to have less than e_{min}items. These clusters would then not be used for constructing trials for the Sternberg Task. In the productive version of the software the threshold was set to 1, meaning that some cases could exclude a maximum of 3 colors was not included in the Sternberg trials. No such cases were encountered in the experimental data.

$$Stress\ Loss = \frac{[Previous\ Stress] - [Current\ Stress]}{[Previous\ Stress]}$$

If the stress loss was higher than 0.02 (2% stress reduction), another dimension was added and the process was repeated, otherwise the *previous* fit was taken.

Interestingly, the color spaces of all participants constructed with this method yielded between 4 and 6 dimensions. This could indicate that, as already suggested by other researchers (Kuehni, 2001), when it comes to psychologically uniform color spaces, 3 dimensions might not be enough.

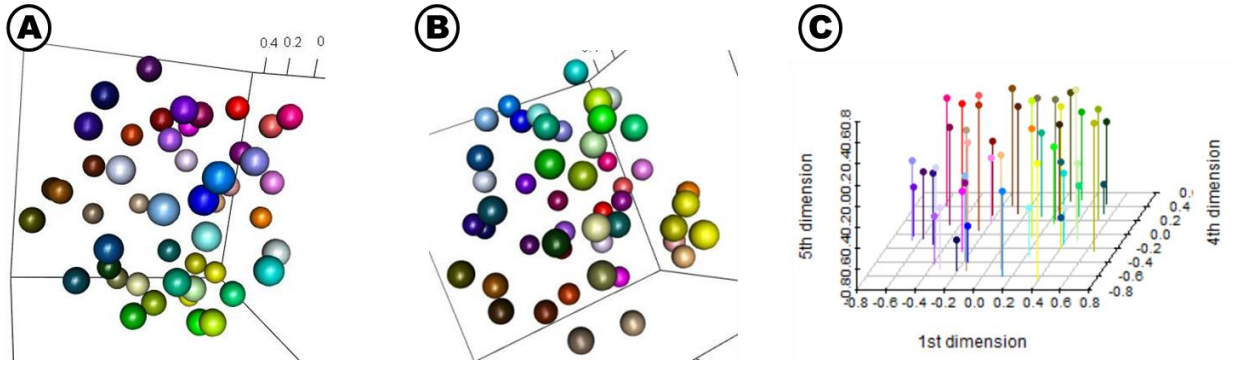


***Figure 20**: Participant's color spaces reconstructed using Kruskal's non-metric MDS method. A and B show the color space in from different angles when MDS is coerced into finding a 3 dimensional solution. C shows the 5-dimensional non-coerced spaced plotted against its 1st, 4th and 5th dimensions*

Because the data is based on subjective similarity ratings of a particular individual, a prima facie analysis of how well the clusters and color spaces match a person's representational realm and classification is rather difficult. But we can safely assume that if the Sternberg Task produces the hypothesized results, and if the categorization task matches what the model predicts, the method for examining similarity judgments and finding categorical groups has indeed worked.

7 Experiment 2 – Modified Sternberg Task

The modified Sternberg Task constitutes the main experiment with which the hypotheses formulated in Chapter 1.5 would be tested. The main design of the task was kept similar to the one described in Chapter 1.4: During a memorization phase colors divided into two lists (A and B) were presented sequentially in the middle of the screen. The first 3 colors were labeled with A and the 3 subsequent ones were labeled with B. After the memorization phase a probe color was presented in the middle of the screen with a list label (either A or B) participants were asked whether that color was present and with the correct label.

Lists were created with 3 conditions: a *baseline* condition where colors were picked at random from varying clusters and 2 *category-based* where each list represented a different category. The *category-based* trials where divided into *low-* and *high-cohesion* conditions. The low-cohesion condition contained list pairs with particularly low *DI* values, and vice versa.

The participants' detection performance and reaction times were analyzed in each condition separately. The results confirmed all hypotheses except for hypothesis number 2 which stated rejection of negatives would improve with category-based trials.

7.1 Experiment Design and Item construction

The main focus of operationalization was the degree of *cohesion* and the *probe types*. As discussed in Chapters 2.5 and 4.4, trial category cohesion as measured by the adaptation of the Dunn Index would serve as an indirect measure for ease of categorization. This ease of categorization would give us some insight into how recognition can be aided by categorical context information by observing differences in reaction time and accuracy for the different type of probes (Table 4; see also: Chapter 1.4).

Table 4: *Probe Types: Short Overview*

Probe Type	Description
Positive	A probe color present in the list it was labeled with
Negative	A probe color not present in either lists
L-Intrusion	A probe color which was present in the irrelevant list, i.e. labeled incorrectly.
C-Intrusion	A probe color not present in either list, but taken from the same category as the target list (the list it is labeled with)

In category driven trials, participants are expected to perform better with Positives and Negatives because the category serves as contextual information which aids recollection.

Another expected effect is that C-Intrusions will be harder to reject than Negatives because they belong to the same category as the presented list. If contextual category information is being used for recollection, this information could fool participants into thinking they saw the C-Intrusion during memorization.

Predictions regarding cohesion are that high cohesion items will generate highly distinct lists while low cohesion items will tend to generate more homogenous lists (Figure 21). When lists are homogenous it becomes harder to distinguish between them. The performance on L-Intrusions should then drop when compared with high cohesion lists.

With this in mind, item cohesion was not selected as a continuum of Dunn Indexes, but rather divided into high and low cohesion conditions for which the data would be analyzed separately.

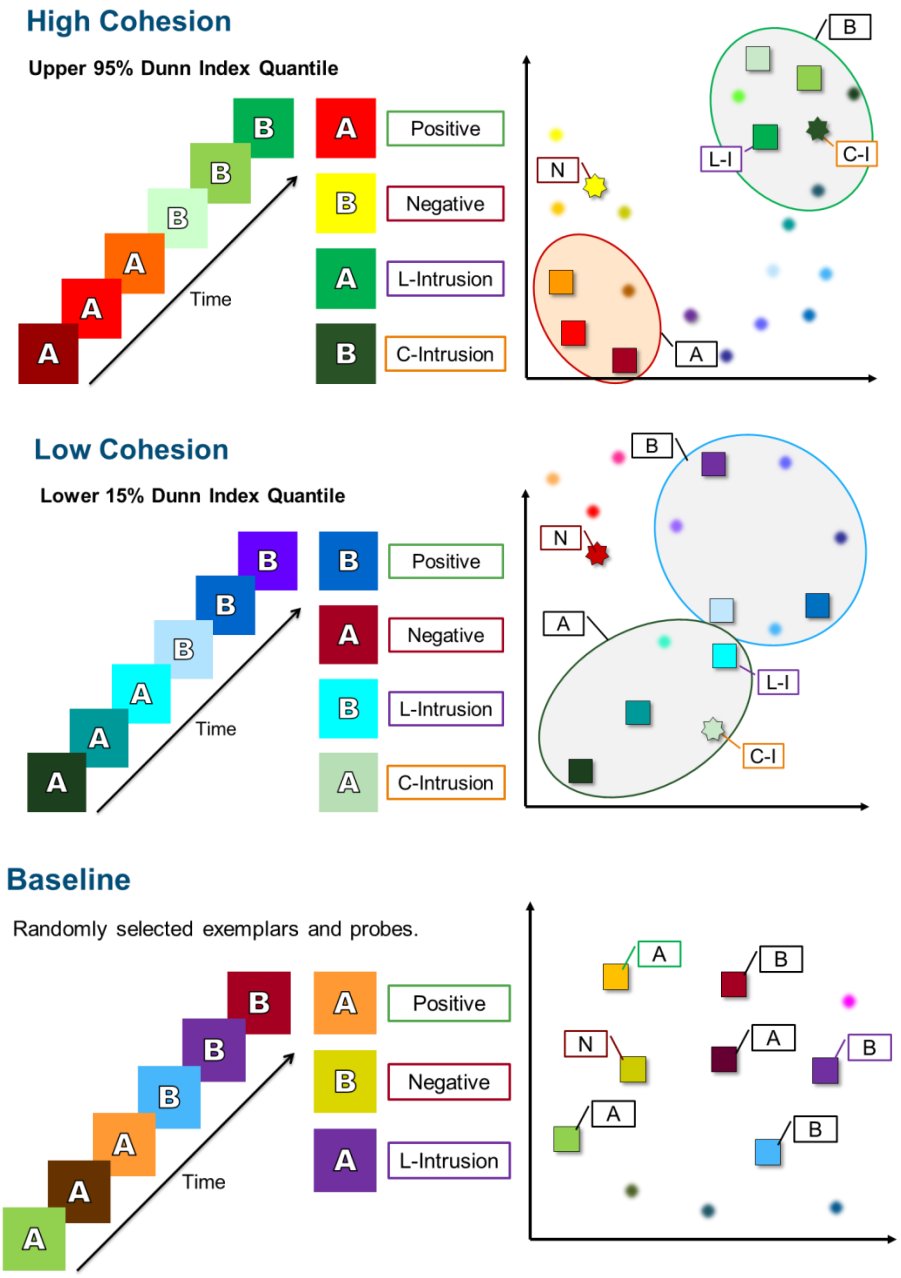


***Figure 21**: Schematic representation of trial composition. To the left are schematic representations of Sternberg two-list trials with the respective probe types under the 3 different cohesion conditions (high, low and none). To the right are graphs illustrating a participant's representational space. The spheres represent a category assigned to a specific list (A or B); the squares represent items included in the list; stars are probes which were not included in either list: Negatives when outside the category, C-Intrusions when inside the category. The labels are: list (A and B); Negatives (N); list-wise intrusions (L-I); category-wise intrusion (C-I).*

High cohesion trials (top) are composed of highly distinct lists. The items selected for list composition originate from spatially compact clusters which are far apart from each other. Low cohesion trials (center) are composed of less distinct lists. The items selected for list composition originate from larger clusters which are near each other. For this reason, the color composition tends to be homogenous across lists. Baseline trials (bottom) are composed of randomly selected colors assigned to either list. Because item generation for the baseline was not category-driven, these trials did not have category-wise intrusions.

The experiment constituted of a total of 320 trials divided into 4 blocks of 80 trials each. Items were constituted of 6 colors divided into two lists (A and B) and a probe. Items were constructed according to the *target list* (the relevant list to be remembered during recall), their *cohesion value*, the probe type and the probe's serial position during presentation (Table 5).

A total of 10 trials per condition configuration (Target List x Cohesion x Probe Type) was constructed. Since there were 3 types of negatives to counterbalance (Negative, L-Intrusion, C-Intrusion), 30 Positive trials (instead of 10) were assigned to all configurations. The serial position of Positives and L-Intrusions was assigned randomly because the counterbalancing of serial position across conditions would require over 1000 trials in total.

Table 5: *Experimental conditions of the modified Sternberg task*

Parameter	Value
Target List	A, B
Cohesion	High, Low, None
Probe Type	Positive, Negative, L-Intrusion, C-Intrusion
Serial Position	1, 2, 3, 4, 5, 6

Experiment 1 provides us with the dissimilarity data necessary to generate trials which are tailored to an individual participant. As soon as a participant finished Experiment 1, the data was sent to the server which then used a trial generation algorithm to prepare the Sternberg Task for that particular participant.

The algorithm utilized the distance matrix and cluster analysis (Chapter 6.1) to construct lists sampling from the color pool. The process was divided into two distinct types of trials: category-driven and baseline.

The baseline items were constructed by sampling random colors in alternating clusters. For example, when the pseudo-hierarchical implementation of the PAM algorithm yielded 8 clusters for a particular participant, the item construction algorithm would choose 7 clusters randomly, say [2,5,8,3,9,4,7]. It would then proceed to sample randomly from clusters 2, 5 and 8 for elements of list A, from cluster 3, 9 and 4 for elements of list B, and from cluster 7 for a potential negative probe. The same process was repeated for each trial.

The target list (the list the probe would be labeled with), was assigned as A for one half of the trials and as B for the other. The probe type was also distributed among trials according to the specified number of items per condition and counterbalancing positive and negative response items: 40 Positives, 20 Intrusions and 20 negatives.

For the serial position, a number 1 to 3 was assigned randomly to every trial. The actual serial position, color and label presentation was executed by the logic in the client-side software. For example, a Positive with B as its target list and 3 as its serial position number would be constructed by picking the third color of list B (6^{th} color in the serial presentation). Conversely, an L-Intrusion with **target list A** and 2 as its serial position number would be constructed by taking the second element of **list B** (5^{th} element in the serial presentation). The reason the color is sampled from list B and not A is that the *target list* parameter defines the list the *probe* will be labeled with, regardless of what kind of probe it is. Because it is an L-Intrusion we are dealing with, the color would be taken from the *irrelevant* list, i.e. list B.

The category driven items had a somewhat more complex construction method. First, every possible 2-combination among all clusters was retrieved. Then, for every resulting cluster pair, every possible 4-permutation of cluster items was constructed. That is, all permutations with 4 elements from the first cluster (3 elements for the list and one more for the C-Intrusion) with all possible 4 element permutations of the

second cluster. The negative probes were selected from a different cluster than the ones used for list construction.

For every combination of permutations, the Dunn Index was calculated. From this list of permutation combinations with respective Dunn Indexes, elements were sampled belonging to the top and bottom 15% quantiles. The top quantile would become the *high cohesion* trials and the bottom quantile the *low cohesion* items.

Both data matrices containing the list elements, C-Intrusions and Dunn Indexes were sorted by their calculated Dunn Indexes. The high cohesion matrix descending and the low cohesion matrix ascending.

The first row (with the highest/lowest *DI*) was as picked for inclusion in the *trials set* (the set which would eventually be used in the Sternberg Task). The algorithm than scanned all subsequent rows of the matrix, selecting rows in which no color was already present in the trials set. Whenever such row was found, it was included in the trial set. The colors of this row then counted as colors to be avoided as well. Once the algorithm reached the bottom of the matrix and there were no rows with unique colors left, the algorithm increased its tolerance (e.g. the repetition of one color was allowed). This process was repeated until the number of trials required for high and low cohesion conditions was reached (120 trials each).

Once all the trials had been selected based on their Dunn Index while avoiding color repetition, the conditions were assigned in the same way that they were assigned to the baseline trials. The only difference being that category based trials also included C-Intrusions (60 Positives, 20 Negatives, 20 L-Intrusions, 20 C-Intrusions).

The trials from all 3 lists (baseline, high cohesion and low cohesion; Figure 22) were combined into one matrix and the Dunn Index was calculated again, but this time it would only include the colors on both lists and the probe. This value would then become the *DI* (labelled "TargetDunn" in the experimental data) which predicts ease of categorization in Experiment 3.

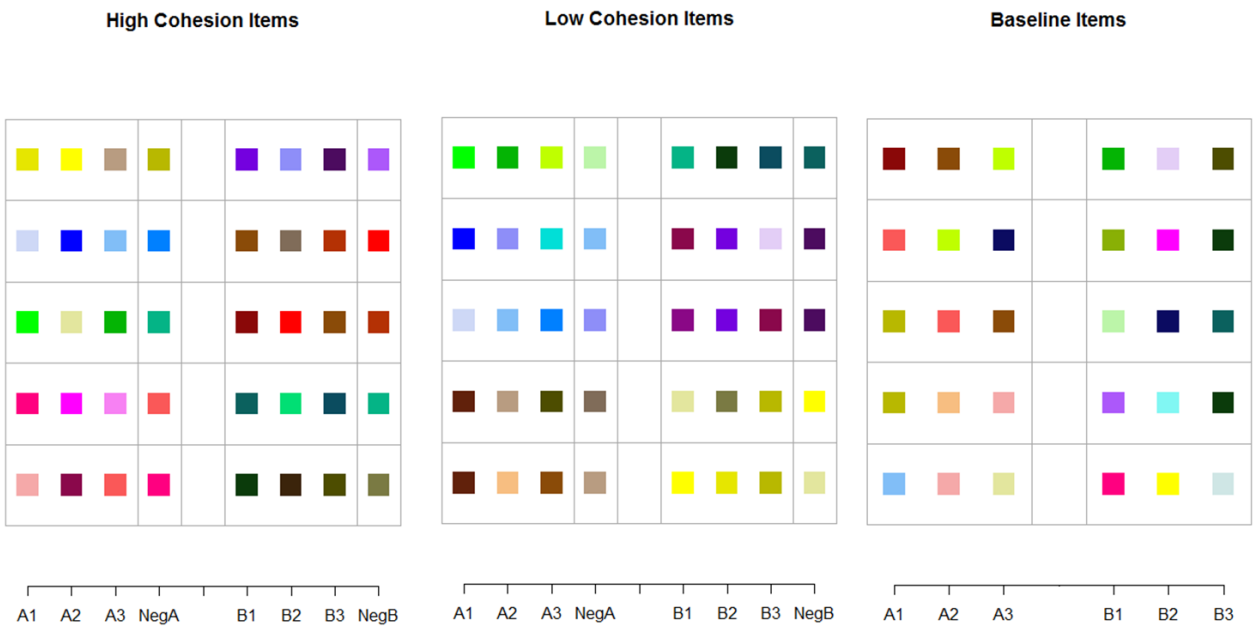


***Figure 22**: Examples of trials constructed by the algorithm. Each plot contains 5 trials constructed by the algorithm sampled at random from participant data. The first plot (left) contains trials from the high cohesion group, the second (center) from the low cohesion group, and the third (right) from the baseline. Colors are grouped together according to the list and order they were presented in. A1, A2 and A3 in the first list; B1, B2, B3 in the second list. Included in the high cohesion plots are the selected C-Intrusion colors for each list (NegA for list A, NegB for list B). Note how the high cohesion group form lists which are highly distinct from each other while lists of the low cohesion group are more similar to one another.*

7.2 Method

At the beginning of the experiment instructions, participants were urged to close all other applications in their computers and to conduct the experiment in an environment where they could concentrate free from distractions. Participants were then instructed they were taking part in a short term memory task were 6 colors would flash on the center of the screen sequentially. Each color would be presented with a label. The first 3 colors would be labeled "A" and the following 3 would be labeled "B". After all six colors had been presented, a final labeled color would appear in the center of the screen followed by the question "Was this item present on list [X]?" (X changed dynamically and stood for A or B depending on the how the probe color was labeled). The participant could then press the right arrow key on their keyboards to answer with "yes" or the left arrow key to respond with "no". The arrows with their respective answers were plotted on the screen when the probe was presented.

Participants were explicitly informed that in some cases a color which was present before would appear with the wrong label (L-Intrusions). When this was the case, participants should respond with "no".

After this introduction, participants tried out 4 examples and received feedback for their answer. The feedback contained all labeled colors presented during the memorization phase, the probe, the correct answer and the answer given by the participant.

After the try-out trials were finished, the participant was informed that the real experiment would start and that feedback about their answers would no longer be provided.

The experiment was divided in 4 blocks with pauses in between them. Once a block was complete, the experiment was interrupted with a pause screen where participants were encouraged to take a break and look away from the monitor for a few moments. Participants could continue the experiment at their own discretion.

The trials were divided into a memorization phase and a recollection phase. During the memorization phase 6 colors were flashed sequentially on the screen. First, a fixation cross appeared on the middle of the screen for 1.5 seconds. Then each color was presented on the middle of the screen against a white background for 400ms each. The inter-stimulus interval was 100ms, during which the screen was blank.

Once all 6 labeled colors had been presented, the screen went blank for another 500ms and then the probe was presented. Underneath the probe, the question "Was this item present on list X?" (X being A or B depending on the target list) and two arrows labeled "yes" and "no" (representing the keyboard right and left arrow keys respectively) were presented. Participants had 5 seconds to respond. The experiment progressed to the next trial as soon as the participant gave an answer or after the 5 second time limit had expired.

After all 4 blocks were completed, participants were informed the experiment was over and that they could proceed with the remaining experiments.

7.3 Results and Discussion

The results were examined individually according to the predictions made on Chapter 1.5. I'll first provide a comparative overview (Table 6) of the predictions and experimental results. Later, I will go into the detailed analysis of every probe type and cohesion condition.

Performance on positive probes would increase since the additional contextual information derived from the category is expected to work as contextual information which aids recollection. This effect is expected to be observed in both high and low cohesion conditions. However, because low cohesion trials will produce categorical ranges which are more disperse, this effect is expected to be more pronounced on high, than on low cohesion trials.

Because negative probes will stem from outside the categories utilized for list construction, rejection of negative probes is expected to be somewhat improved since it will not match the remembered information about either category. The cohesion dependent effects were somewhat harder to predict. On one hand, high cohesion trials have more concise and coherent category spaces than low cohesion trials. This is likely to facilitate exclusion of an item exterior to the target category. For example, it is easier to exclude a *brown* from a category composed solely from vibrant tones of red, than it is from a broader categorical space which also includes darker tones of orange. On the other hand, low cohesion trials tend to be constructed with broader neighboring category spaces. This produces homogenous lists which could in theory be interpreted as one single category, thus reducing the amount of information necessary to reject a probe external to both categories.

The ability to reject list-wise intrusions (L-Intrusions) is also expected to benefit from categorization. High cohesion trials are expected to form two very distinct and very easily recognizable categories. This high distinction between lists will prevent the participant from confounding an item from one list as being an item from another list.

Conversely, in low cohesion trials the categories might "fuse together" to form a single "super-class" category. When this is the case, categorization will either no longer help in holding lists apart, or even hinder one's ability to distinguish between lists.

In regards to category-wise intrusions (C-Intrusions), an effect similar to the elicitation of false memories in the DRM paradigm is expected. That is, when participants rely too heavily on information about the category to remember elements, they might be lured into thinking an element which was not included in the target list was present because it fits the category perfectly. This effect is expected to be more pronounced on high cohesion trials because in such cases the categorical space is more concise, thus strengthening the criterion for inclusion.

Table 6: *Predictions and results regarding performance on different probe types compared to the baseline.*

Cohesion		Positives	Negatives	L-Intrusions	C-Intrusions
Predictions	High	++	+	++	--
	Low	++	+	0 / -	- / 0
Results	High	++	0	++	--
	Low	+	0	0	--

The results confirmed most of the predictions made in the previous chapters with some exceptions. What I found surprising was the fact that detection performance was improved on positive, but not negative probes. Intuitively, category information should both facilitate acceptance of a positive by means of category inclusion, as well as

aid rejection of a negative by recognizing the probe cannot be part of the target category.

A second observation which did not exactly match the predictions was that C-Intrusions were equally harder to detect than negatives and L-Intrusions on both high and low cohesion conditions. This means that the target category distance to the irrelevant category plays little to no role in the strength with which category information will elicit false recognition.

Still, the fact that false memory was elicited at all provides some support for the construct validity of the previously discussed models and algorithms, meaning that the software could indeed detect categorical groups and successfully produce lists which would illicit associations with the list's category of origin.

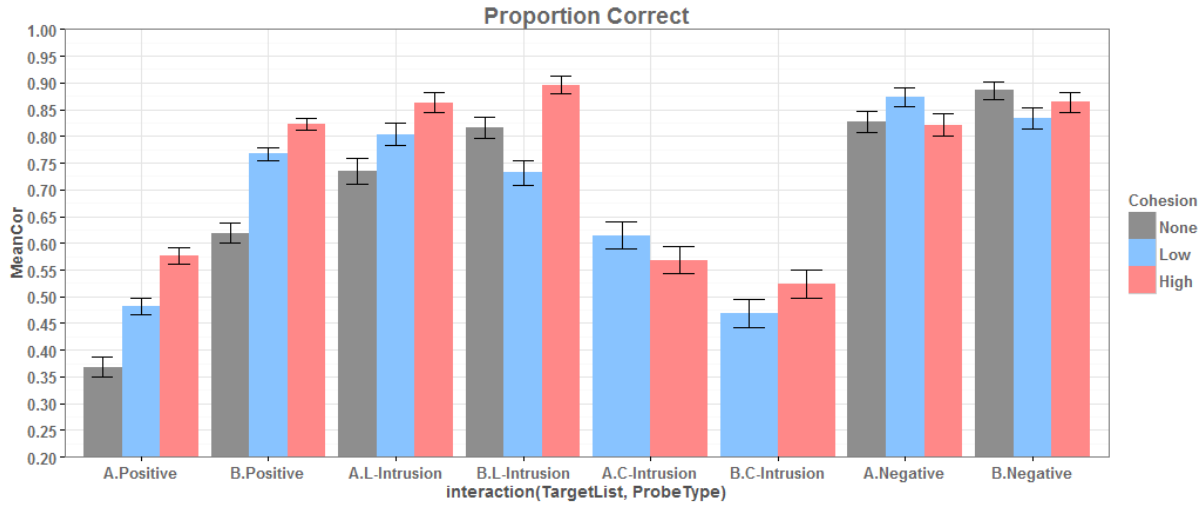


***Figure 23**: Proportion of correct responses per target list (A and B), probe type (Positive, Negative, L-Intrusoin and C-Intrusion) and cohesion (high, low and none) over all participants.*

7.3.1 Performance on Positive and Negative Probes

To examine the performance on positive probes (Figure 24), a binomial logit regression was conducted with cohesion and serial positions as predictor of the proportion of correct responses. Performances increased in both low ($b_1 = 0.6,\ Z = 8.17,\ p < 0.001$) and high cohesion ($b_2 = 0.97,\ Z = 12.9,\ p < 0.001$) conditions. A serial

position effect was also detected ($b_3 = 0.42,\ Z = 23.15,\ p < 0.001$) indicating that items recognition performance increased on later items. No primacy effect was detected.

Performance on baseline positives was generally low ($M = 0.49,\ SD = 0.22$), particularly for the first 3 items on the list ($M = 0.36,\ SD = 0.05$) where it mostly remained below chance level, rarely surpassing it.

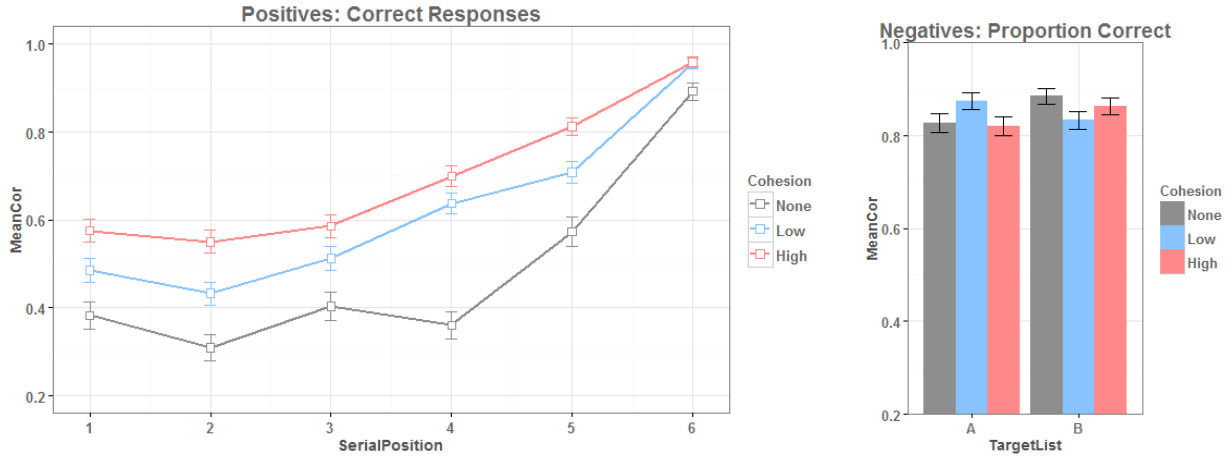


Figure 24*: (left) Proportion of correct responses on positive probes by serial position and cohesion. The baseline condition (Cohesion: None) is plotted in grey, high cohesion trials in red, and low cohesion trials in blue. The bars represent the standard error. (right) Proportion of correct responses on negative probes by target list and cohesion.*

A logit binomial regression with cohesion as predictor for performance on Negatives did not yield any significant results. When the target list was added to the model, a statistically significant interaction difference between target list and low cohesion could be observed ($b = -0.76,\ Z = -2.57,\ p < 0.05$). None of the case-wise comparisons of means yielded statistically significant differences.

Regarding reaction times (Figure 25), a linear model of cohesion and serial position as predictors of reaction time for positive probes revealed that participants responded faster for high cohesion items ($b = -76,\ t = -3.68,\ p < 0.001$) and only marginally so for trials in the low cohesion condition ($b = -38.4,\ t = -1.86,\ p = 0.062$) . A serial position effect could also be observed ($b = -84.2,\ t = -17.8,\ p < 0.001$) .

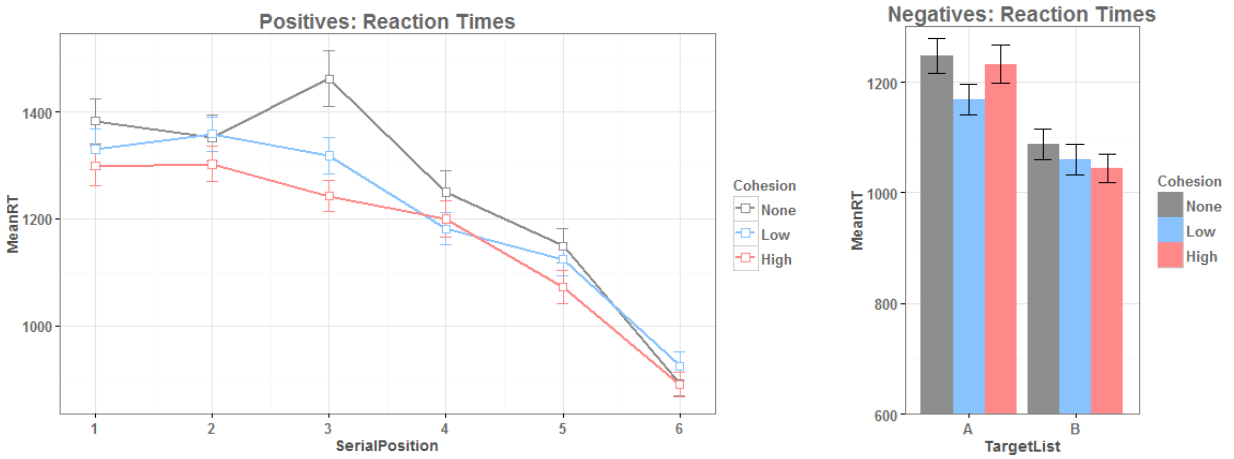


***Figure 25**: Reaction times on positive probes by serial position and cohesion. The baseline condition (Cohesion: None) is plotted in grey, high cohesion trials in red, and low cohesion trials in blue. The bars represent the standard error.*

A possible explanation for the lack of difference in performance on Negatives might be due to a shift in response bias. Participants were more "liberal", i.e. produced more "yes" responses, on category-based trials than they were on baseline trials. An analysis of variance showed a criterion position shift (Macmillan & Creelman, 2004, p. 29) for category-based trials when compared to the baseline ($F(2, 213) = 12.6$, $p < 0.001$). A linear model with cohesion and target list as predictors for criterion position showed further that both low ($b = -0.2$, $t = -2.38$, $p < 0.05$) and high cohesion ($b = -0.45$, $t = -5.33$, $p < 0.001$) conditions produced a shift in bias ($R^2 = 0.21$, $F(3,212) = 19.1$, $p < 0.001$).

A possible explanation for this difference in response bias might be linked to the relatively high difficulty of baseline trials. As already discussed, performance for baseline trials remained below chance for the first 3 serial positions. Because category-based trials were so much easier, participants might have associated the higher difficulty in recalling an item with it not being present on the list, thus negating the presence of most items and generating the observed conservative bias.

To get a measure for performance less susceptible to the influence of bias, performance on positive and negatives was calculated as a signal detection paradigm

(Macmillan & Creelman, 2004). The sensitivity parameter d' was calculated using the performance on positive probes as pHit and the performance on negative probes as $1 - \text{p}FA$ (see Table 1 for an overview of response types).

A linear regression with cohesion as predictor for sensitivity (d') was conducted for target lists A and B separately. For target list A, the linear model revealed an increased detection performance for both low ($b = 0.59,\ t = 3.1,\ p < 0.01$) and high cohesion ($b = 0.43,\ t = 2.3,\ p < 0.05$) conditions as depicted in Figure 26. For target list B, the linear model revealed gains in detection for high ($b = 0.54,\ t = 2.6,\ p < 0.05$) but not for low cohesion trials ($b = 0.17,\ t = 0.8,\ p = 0.42$).

Overall, the prediction that participants would benefit from contextual category information on positive probes was confirmed. The prediction regarding an increased performance on Negatives could not be confirmed, albeit the comparable rate of rejection of Negatives in the baseline versus category-based trials could be due to a shift in response bias. That is, participants were more conservative on baseline trials and responded in a more liberal fashion on category-based trials.

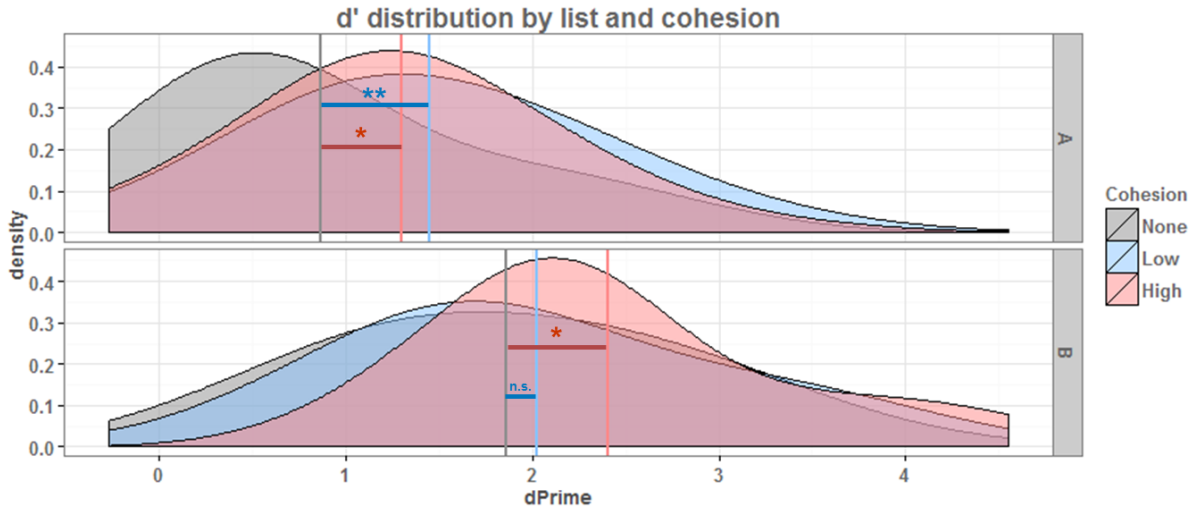


Figure 26: *Density plots of d'distribution by target list and cohesion. Distributions for A as the target list are plotted on the upper row, distributions for B as the target list on the bottom row. Participants performed better with high cohesion trials on both lists. The detection performance was only significantly better for negatives labeled with "A".*

7.3.2 Performance on L-Intrusions

A binomial logit regression was performed with cohesion and serial position as predictors of correct responses on L-Intrusions. The regression showed that high ($b = 0.70,\ Z = 4.9,\ p < 0.001$) but not low cohesion ($b = -0.06,\ Z = -0.51, p = 0.61$) had an effect on a participant's ability to reject list-wise intrusions.

Expanding the model with the target list as predictor showed that for L-Intrusions originating from list A (target list is B), the performance was worse than the baseline ($b = -0.37,\ Z = -2.1,\ p < 0.05$) as visible in the data points of the first 3 serial positions in Figure 27. This finding was confirmed by a paired, one tailed t-test ($t(35) = 1.93,\ p < 0.05,\ d = 0.07$) on low cohesion trials between list A ($M = 0.8,\ SD = 0.15$) and list B ($M = 0.73,\ SD = 0.15$).

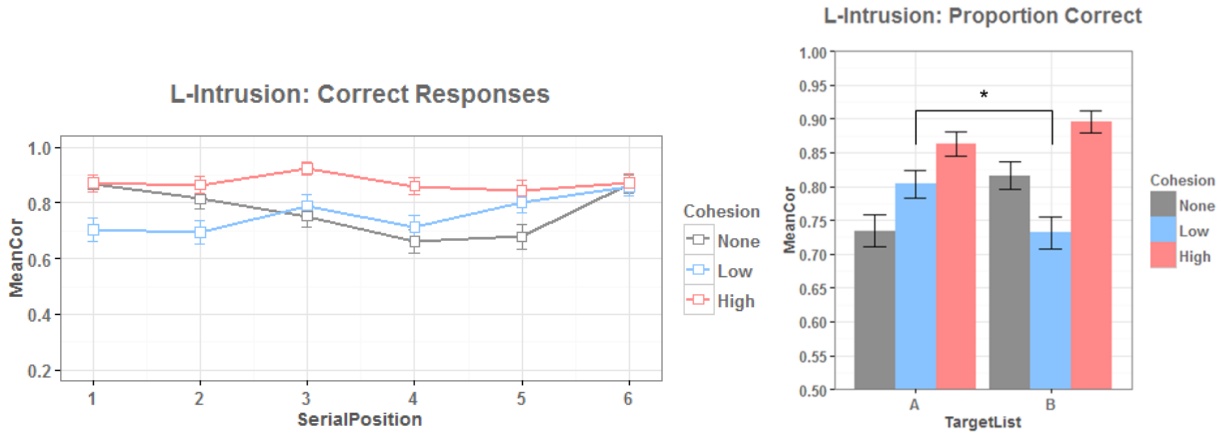


***Figure 27**: (Left) Performance on L-Intrusions by serial position. The baseline condition (Cohesion: None) is plotted in grey, high cohesion trials in red, and low cohesion trials in blue. The bars represent the standard error. (Right) Aggregated proportion of correct responses by target list. Performance for L-Intrusions labeled with "B" (color taken from the A list) is lower than for L-Intrusions labeled with "A" (color taken from the B list).*

This drop in performance was not expected (Figure 27, right). A possible explanation involves the homogeny among lists of low cohesion trials. In Figure 24 we can see that the performance for positive probes of list A (serial positions 1-3) is much lower than the ones from list B (serial positions 4-6). At the same time, there is no performance loss for negatives labeled as elements from list A. The homogeny among the two lists of low cohesion trials (see Figure 22) might be rendering the elements of

that list not salient enough to be held apart from each other. They just "blend together" when recollecting the beginning of the serial presentation.

Thus, when an element from the first list is presented as probe but labeled as an element of list B, a participant might not remember if that element came before or after serial position 3. At the same time, one and the same category might be serving as contextual cue for both lists as argued in the beginning of this chapter. This contextual cue is telling the participant to accept the element in a similar mechanism as hypothesized for C-Intrusions. On the absence of other evidence against such decision, i.e. on the absence of the memory for the element itself, the participant will be more prone to accept the element as a positive.

Conversely, elements from list B are "fresher" in memory and can still be held apart from the rest of the list. So, when they are presented as if they were elements from list A, the participant still has that additional element-based evidence they need to reject it.

7.3.3 Performance on C-Intrusions

Category-wise intrusions were analyzed differently than Negatives and L-Intrusions. Because they were postulated as a form of "false memory" elicited by the association of the elements of a list with a category, they were measured against negative probes which were also not present in either list, but at the same time had no association with the hypothesized category.

A binomial logit regression was conducted on a subset of the data containing only Negative and C-Intrusion trials. The model used the probe type as sole predictor of correct response. A statistically significant decay in rejection performance could be observed for C-Intrusions when compared with Negatives ($b = -1.5,\ Z = -19.22,\ p < 0.001$).

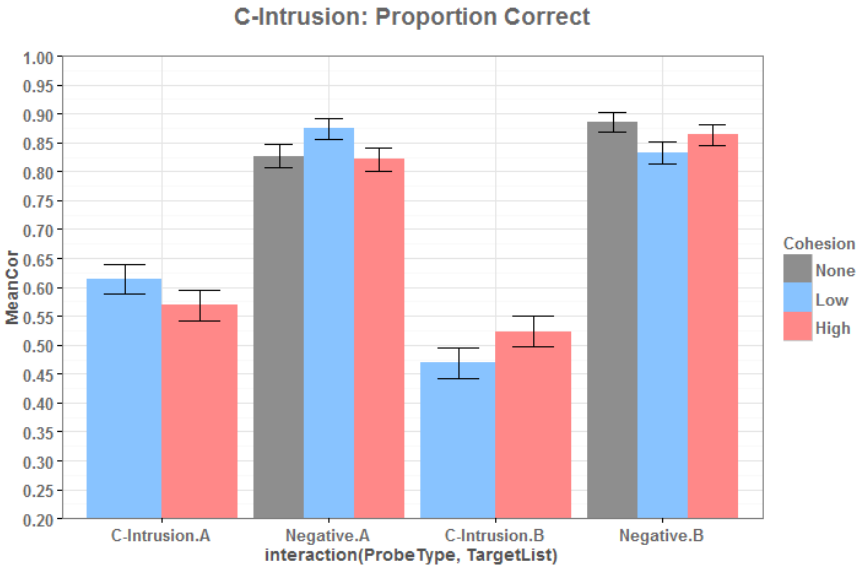


***Figure 28**: Proportion of correct responses for C-Intrusions compared to negative probes by cohesion and target list. Participants have consistently shown better performance for negatives than for C-Intrusions.*

Just like with the performance on positive probes, the performance on C-Intrusions was analyzed in a signal detection paradigm to avoid confounding a shift in bias with an increase or decrease in performance.

However, the operationalization was somewhat different. Performance on C-Intrusions was interpreted as constituting correct rejections (or $1 - \mathrm{p}FA$) and the performance on Positives was interpreted as hit rate ($\mathrm{p}Hit$). With Positives constituting signal and C-Intrusions constituting noise, the sensitivity measure (d') for C-Intrusions was calculated.

The sensitivity measure calculated with negative probes as noise (instead of C-Intrusions) was then used as baseline against which the detection performance of C-Intrusions would be compared (Figure 29).

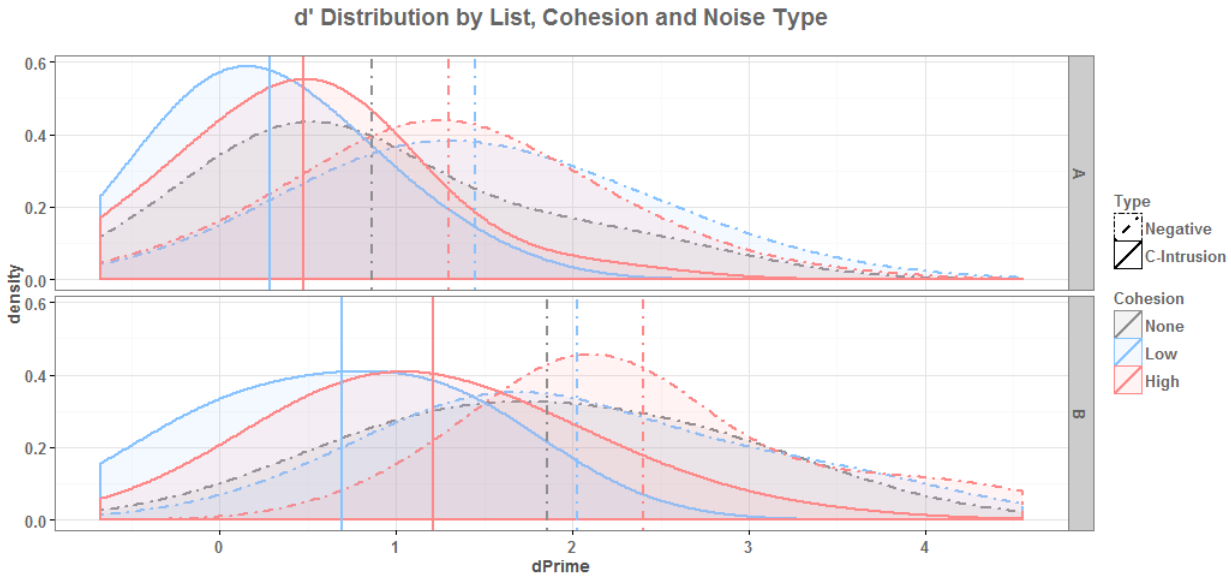


Figure 29: *Density plot of d′ distributions by target list, level of cohesion and noise type (the type of probe for which performance was interpreted as correct rejection). The upper plot contains the densities of d′ for list A, the lower plot those of list B. The vertical bars mark the mean for d′ on each condition. Sensitivity measures done with C-Intrusions are represented by full lines; sensitivity measures constructed with Negatives are represented with dash-dotted lines.*
Detection performance (d′) calculated with C-Intrusions was consistently lower than when calculated with Negatives. This was the case across all cohesion conditions and for both lists. All differences are statistically significant.

Paired, one-tailed t-tests were conducted to test the differences in mean *d′* between C-Intrusions and Negatives for all cohesion conditions and for both target lists (Table 7). All differences were statistically significant.

Table 7: *Differences in means of d′ calculated with C-Intrusions versus Negatives*

C-Intrusions		**Negative Probes**		
Target List	**Cohesion**	Low	High	None
A	Low	1.16***		0.57***
	High		0.82***	0.39*
B	Low	1.34***		1.16***
	High		1.2***	0.65***

Signif. Codes: '***' p < 0.001 | '**' p<0.01 | '*' p<0.05 All statistics were calculated using paired, one-tailed t-tests; df=35

C-Intrusions are a key component in this study. The stark contrast in performance between Negatives and C-Intrusions provides strong support for the assumption that participants are indeed associating the lists to super-ordinate classes or

categories (i.e. “bright tones” or “reds”) and using that information to aid recall. The reliance on this contextual information is so strong, that false memories can be elicited when other elements from within the category are used as probes. The fact that false recognition took place with the selected elements also strongly supports construct validity. That is, since the elements chosen by the algorithm produced an effect typical for the influence of categories on recall, we can assumed that both the method for acquiring dissimilarity measures as well as the algorithm for discovering category groups work.

8 Experiment 3 – Categorization Task

In Chapter 2.5 a measure for ease of categorization was proposed (Equation 2.14) building upon Nosofsky's EBRW model and the assumption that concise and isolated category spaces would lead to faster and easier classification of its constituent elements. In Chapter 4.4 the Dunn Index was suggested as an indirect measure for the proposed ease of categorization along with a validation model also based on EBRW's predictions regarding reaction times (Equation 4.7).

To test whether (a) the algorithm described in previous chapters does indeed construct lists which translate to a category group; (b) that the chosen framework does model categorization choices; and (c) that the *DI* can predict ease of categorization, Experiment 3 would use the category-based trials from Experiment 2 in a categorization task.

8.1 Method

The experiment was modeled in a very similar way to the simulation conducted in Chapter 4.4.1. The task consisted of classifying a single element to either of two groups depending on where the participant perceived the item to "belong" to.

In order not to strain the participants which had already spent over 1 hour in the previous Experiments, the categorization experiment consisted of only 80 trials with varying *DI*s sampled from Experiment 2.

On each trial, all elements from list A and list B were plotted on the left and right side of the screen respectively. The element selected by the algorithm for a potential C-Intrusion (see Chapter 7.1) was then plotted in the center of the screen. Participants were instructed to use the left and right arrow keys to choose which group the color in the center belonged to (Figure 30).

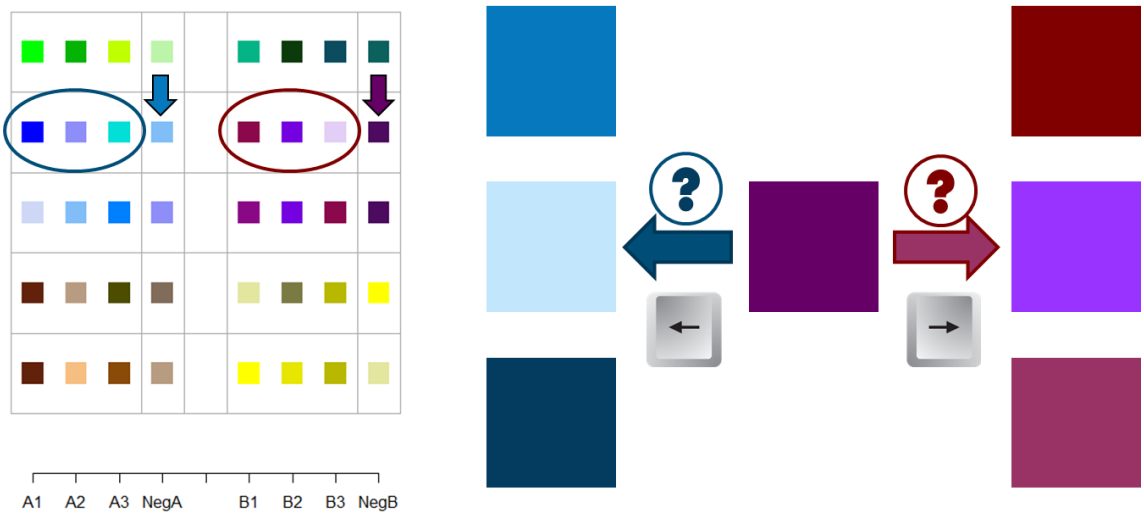


Figure 30: *Categorization experiment schematics. Category driven trials from Experiment 2 are sampled for item construction. The elements which constituted list A and B in experiment 2 are plotted in the left and right side of the screen. The element selected by the algorithm for a potential C-intrusion (NegA or NegB) was plotted in the center of the screen. Participants made a classification choice by pressing the left or right arrow key according to which group of color they perceived the color in the center to belong to.*

At the beginning of the experiment, participants were presented with introductory slides explaining the task. They would see two columns of color and one color in the center of the screen. They should then press either the left or right arrow keys depending on which group of colors they thought the color in the center belonged to. Participants were also informed that there was no "right" or "wrong" answer and when in doubt, they should just follow their gut feeling. They were further instructed to be as fast and accurate as possible. After this introduction two sample trials were conducted by the completion of which participants were then informed that the experiment was about to begin.

At the beginning of each trial, a fixation point was shown in the middle of the screen for a period of 1.5 seconds. Right after, both color groups and the probe were presented at once and participants could respond by pressing the left or right arrow keys. The next trial began as soon as the participant entered a response.

8.2 Results and Discussion

Overall, the model did well in predicting which category choices participants would make. The participants' choices coincided with model prediction 91% of the time.

A binomial logit regression was conducted with the trials' *DI* as predictor for the binary coded correct responses. A correct response meant that the participant classified the probe element as a member of the same list as the model judged it to be during the trial construction for Experiment 2. The regression found that the probability of a correct response increases with an increasing *DI* value ($b = 2.11,\ Z = 6.51,\ p < 0.001$).

In order to confirm the simulation conducted in Chapter 4.4.1, both performance and reaction times were aggregated by *DI*. Because the dissimilarity measures and subsequent clustering were based on individual data, *DI* could not be manipulated a priori. Meaning that that they could not be set to a specific value, but had to be taken from the algorithmically constructed trials.

For this reason, the *DI* values of all trials were grouped into 15 quantiles of equal probability. Within each quantile the mean, standard deviation and standard error for reaction times and correct responses were calculated (Figure 31).

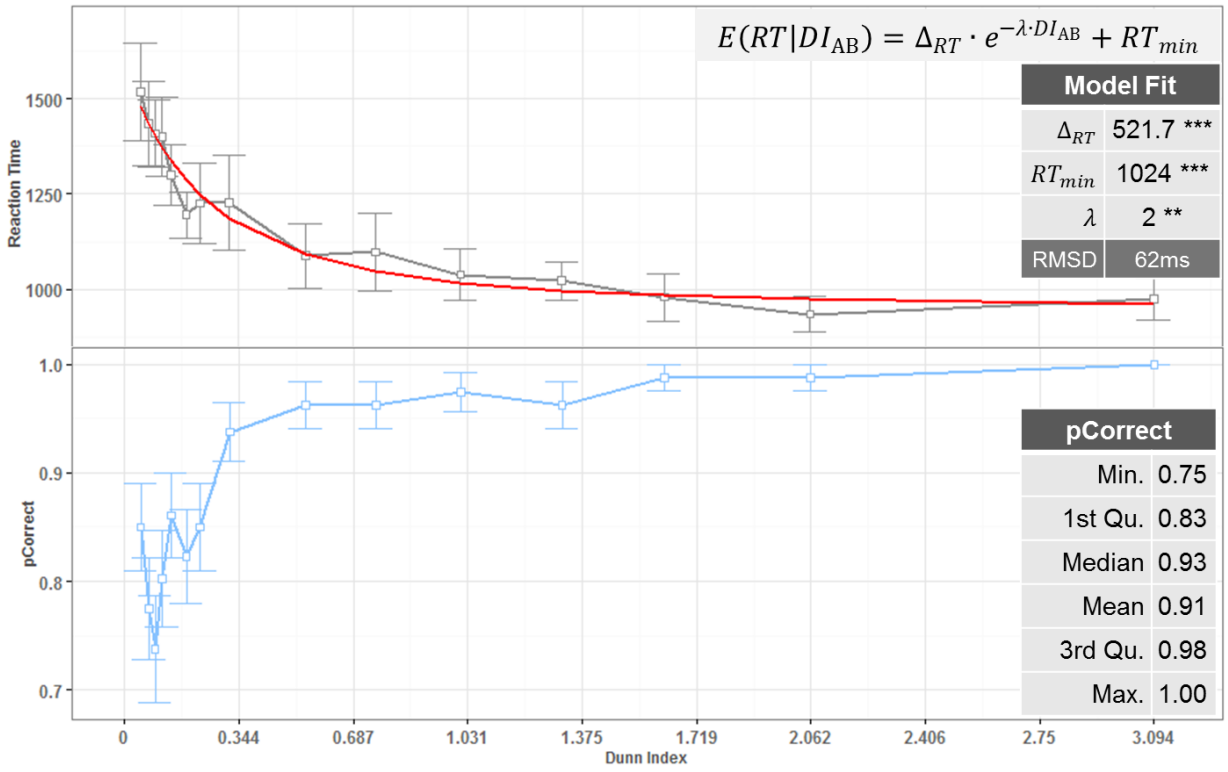


Figure 31: *Ease of categorization model fit. The aggregated reaction times (top) and proportion of correct responses (bottom) constructed by aggregating the performance data of all users within a DI 1/15-quantile. The bars represent the standard error. The fitted model is plotted in red. Reaction times follow an exponential decay with increasing Dunn Index values. The root mean square deviation for the fitted model is 62 milliseconds. All free parameters are statistically significant below p=0.01.*

The exponential decay model in Equation 4.7 was fit to the reaction time data using a nonlinear least squares regression (Bates & Watts, 2007). The fit root mean square deviation was 61.8ms and the residual standard error 69.1ms on 12 degrees of freedom. All three free parameters could be estimated with $p \leq 0.01$ (Table 8):

Table 8: *Nonlinear least squares estimation of Equation 4.7 on the experimental data*

	Estimate	Std. Error	t value	Pr(>|t|)	
$\boldsymbol{\Delta_{RT}}$	521.72	53.76	9.70	0.00	***
$\boldsymbol{\lambda}$	2.04	0.68	2.98	0.01	**
$\boldsymbol{RT_{min}}$	1024.21	44.53	23.00	0.00	***

As the fit indicates, reaction times for classification choices do indeed follow an exponential decay with increasing *DI* values in a two-category paradigm. This serves as evidence in support of both the computational model and the algorithmic implementation of the method of category group discovery in a representational discussed in Chapter 4.

Overall, the modified Sternberg task has proven itself to be a reliable way to study whether implicit association with a category during memorization aids recall.

With the exception of the hypothesized increase in rejections of Negatives, all expected effects and hypotheses were confirmed by the experiment.

9 Experiment 4 – Change Blindness

To make sure the results observed in Experiment 2 were not due to participants being unable to discern between colors, a change blindness experiment was performed using colors which participants had judged as being highly similar and were judged by the algorithm to be good candidates for C-Intrusions. That is, for every participant, in every high cohesion trial, color combinations between C-Intrusion candidates and list elements were scanned for unique color-to-color combinations of high similarity. These combinations were then transformed into change blindness trials where the probe color is different than the presentation color, i.e. Negatives.

To generate Positives, the algorithm selected random colors from the 48 color pool which did not appear in the negative trials set, and then created trials were the randomly chosen color was used as both probe and presentation color. The number of positive trials was matched with the number of previously constructed negative trials.

Because similarity judgments were specific for a particular individual (Experiment 1) and the trials in Experiment 2 differed from participant to participant, the number of trials in Experiment 4 depended on the number of available combinations of highly similar colors. That is, the number of trials (mostly around 100) was not the same for every participant.

9.1 Method

Participants cold start the experiment from the welcome screen. Introductory slides were then presented explaining the task to participants. Just like in the previous experiments, participants solved 2 examples and then proceeded to the main experiment.

The experiment consisted of two colors, one being the presentation color, the other the probe, presented sequentially in the center of the screen. First, a fixation cross was shown in the center of the screen for 1.5 seconds. Then the presentation color

flashed for 400ms. The screen went blank for another 2.5 seconds, matching the presentation length of the remaining 5 list elements in the modified Sternberg Task. Then the probe color appeared. The task was to judge whether this second color was the same as the originally presented color.

Participants could respond by pressing the arrow keys. Right for "yes" and left for "no".

9.2 Results and Discussion

The overall performance was quite high and there were no great differences between performance on positives ($M = 0.92$, $SD = 0.26$) and negatives ($M = 0.91$, $SD = 0.28$) (Figure 32).

Figure 32: *Boxplot of the mean of correct responses on change blindness trials by probe type. When the probe was the same as the presented color the target response was labeled "same" otherwise the target response is "different".*

A paired, two-tailed t-test failed to reveal a statistically significant difference between correct rejections and hits ($t(25) = 0.4,\ d = 0.008,\ p = 0.69$).

This suggests that misses (i.e. the failure to identify changes in color) were mostly due to either bias or other factors such as reaction error (pressing "no" when the intended response was "yes").

There were however some problematic color pairs involving vibrant colors such as blue (`#0000FF`) or magenta (`#FF00FF`) and some of their less vibrant counterparts.

Table 9 contains the examples of vibrant colors which sometimes could not be distinguished from other colors of comparable hue and luminosity. Correct responses were averaged across participants. Also listed is the number of occurrences over all participant data, that is, how many times this color pair appeared in the experiment when the data of all participants were combined into one dataset.

Table 9: *Performance on vibrant colors when presented together with a less vibrant tone with comparable luminosity and hue.*

Color 1	Color 2	Mean Correct	Sdt. Error	Occurrences
#00FF00	**#04B404**	0.6	0.25	5
#FF00FF	**#FF0080**	0.6	0.25	5
#0000FF	**#0080FF**	0.6	0.15	11

These colors should, in theory, be very easy to dsitinguish from one another. A possible explanation for the low performance is monitor calibration or browser color rendering. It could be that in some computers the vibrant color was not rendered as vividly as it should have been, changing either luminosity or saturation.

It is also important to note that lower performance was not exclusive to negative trials (see Figure 32). For example, the performance for the color #E3CEF6 on positive trials was 0.73 ($N = 11,\ SD = 0.46,\ SE = 0.1$).

In total, only 8% of all color pairs used in negative probes were detected less than 70% of the time. Around half of these color pairs were single instances, meaning that the color pair occurred only once in the entire dataset.

For the reasons stated above, I believe it is very unlikely that the results seed in Experiments 2 and 3 can be explained by mere incapacity to distinguish between colors in memory.

10 Conclusion

In this work I've proposed that rapid categorization, which is defined ad hoc as the similarity guided process of quickly associating a set of elements with a category, can aid short-term memory by providing contextual information which in turn influences recollection. This contextual information not only aids recollection, but also prevents cross list intrusions in a two-list memorization setting.

The modified Sternberg Task utilized in this study showed conclusively that his is the case. Recognition performance increased when categories where available as context information even when controlled for response bias in a signal detection paradigm.

Furthermore, the distinctiveness between categories showed to improve participants' performance on list-wise intrusions, further reinforcing the notion that contextual category information is utilized for evidence gathering.

A false recollection effect similar to the one elicited in the DRM paradigm could be reproduced with sets as small as 3 elements. Participants were lured into reporting having seen an item not present in the set during memorization just because it was a member of the set's category. This of course poses the question as to whether the contextual category information is really helping participants remember a specific item, or whether participants are taking a "shortcut" by remembering only the category and then judging whether or not the probe is a member of the category during the recognition phase.

The truth is probably a mixture of both. Performance for Positives and Negatives was clearly better than for C-Intrusion suggesting that participants do indeed remember specific elements and don't rely solely on information about the category. Still, the lower performance on L-Intrusions in low cohesion trials, where elements of both lists formed a homogenous group, shows that participants had greater difficulty keeping both lists apart compared to baseline trials. What is likely to be happening is that in

some settings category-driven evidence is "stronger" than element-driven evidence, causing the decision as to whether an element is "old" or "new" be guided mainly by the information about a category.

All of this is of course mere speculation. The results produced in this work serve only to validate the notion that rapid categorization does occur and that the resulting contextual information aids recognition. The degree to which categorical contexts versus item specific recollection influence evidence gathering, as well as the actual mechanics behind the recollection process cannot be answered here and would be a matter of further research.

Another question which remains open is whether the rejection of Negatives is indeed not influenced by the association of a group of elements with a category, or whether the unchanged performance on Negatives is mainly due to a shift in response bias. It could be the case that a more conservative bias was elicited by the contrast in difficulty between category-based and baseline trials. Since list items from category-based trials were easier to remember, participants might have been more prone to reject items from baseline lists because they could not recall them. This interpretation is partially supported by the analysis which showed that performance on Positives for the first list of baseline trials was consistently under the value expected by chance alone. Furthermore, a shift of the response criterion's position was observed between baseline and category-based trials.

Still, the question as to whether this shift in bias is solely responsible for the absence of an effect in performance remains open. To answer this question, the influence of categorization on bias would have to be studied in isolation.

Apart from the inquiry on the role of categorization in short-term memory, this work also proposes a method for surveying subjective similarity ratings which is more efficient than the conventionally used ones. The method includes a process for identifying categories within a representational space through cluster analysis, and a

measure of cohesion which can successfully predict how easily elements can be classified.

Instead of having to learn categories exclusively for the purpose of an experiment, categories and concepts already present in a subject's representational space can be used for stimulus construction. This can be done on the fly and without any assumptions about the dimensionality of the representational space.

A further advantage of the proposed method is that elements can be judged against one another without losing the potential influence of the context in which this judgment is taking place. For example, because multiple colors are present on the screen at once, similarity judgments among different tones of blue will not fall artificially low due to the absence of a color of another hue. Aquamarine (■) might be judged dissimilar to baby blue (■) when the two are compared in isolation, but they might seem a lot more similar when tones of red and yellow are included in the judgment process.

The method has been proven to work in the scope of this study and it could be extended to work with other forms of stimuli, different distance measures, or even other clustering strategies. The color stimuli could easily be replaced by words or images and the distance measure could be adapted accordingly.

All in all the current study has successfully demonstrated that categorization does indeed influence recognition in short-term memory and that this influence is directly related to how easily its elements can be classified as members of said category.

11 Equations

Minkovsky distance

$$d_{ij} = \left[\sum_{k=1}^{K} w_k \left|x_{ik} - x_{jk}\right|^{\rho}\right]^{\frac{1}{\rho}} \qquad (2.1)$$
$$0 < w_k, \sum w_k = 1$$

d_{ij}: Distance between elements *i* and *j* in the n-dimensional representational space.

K: The number of dimensions in the multidimensional space

x_{ik}: Position of element *i* in the dimension *k*.

w_k: Weight given to dimension k. Set to $w_{1\ldots k} = 1/K$ for all simulations conducted in this work.

ρ: Distance metric for L^p space. Set a-priori depending on stipulated characteristics of the representational space

Exemplar Activation

$$a_{ij} = M_j \cdot e^{-c \cdot d_{ij}} \qquad (2.2)$$
$$0 \leq c \leq \infty$$
$$0 \leq M_j \leq \infty$$

a_{ij}: The degree to which exemplar *j* is activated when item *i* is presented.

M_j: Represents the strength with which item j is stored in memory.

c: Decay rate of similarity with increasing inter item distance.

d_{ij}: Estimated distance between items *i* and *j* in the psychological space

Exemplar Race Time Distribution

$$f(x;\lambda) = \lambda e^{-\lambda x}, x \geq 0 \qquad (2.3)$$
$$f(t) = a_{ij} \cdot e^{-a_{ij} \cdot t}$$

$f(x;\lambda)$: Probability density function for exponentially distributed random variables.

$f(t)$: Probability density function for exemplar *j* finishing the race at time *t* when item *i* is presented.

a_{ij}: The degree to which exemplar *j* is activated when item *i* is presented.

t: Time value.

Probability of an Exemplar a_{ij} Finishing the Race First

$$Pr(X_k = min\{X_1, \dots, X_n\}) = \frac{\lambda_j}{\lambda_1 + \cdots + \lambda_n} = \frac{a_{ij}}{\sum a_{ik}} \tag{2.4}$$

$X_1, \dots, X_n$ Independent exponentially distributed random variables with decay $\lambda_1, \dots, \lambda_n$

λ_j: Decay rate for exponentially distributed random variable *j*

a_{ij}: Activation of exemplar *j* when item probe *i* is presented

Probability of Step in Direction of A or B Category Thresholds

$$p_i = \frac{\sum_{j \in A} a_{ij}}{\sum_K \sum_{k \in K} a_{ik}} \tag{2.5}$$

$$q_i = \frac{\sum_{h \in B} a_{ih}}{\sum_K \sum_{k \in K} a_{ik}} \tag{2.6}$$

p_i: EBRW: Probability of taking a step towards category threshold A during the random walk.

q_i: EBRW: Probability of taking a step towards category threshold B during the random walk.

a_{ij}: The activation of exemplar *j* upon presentation of item *i*.

A: Set of exemplars which belong to category A.

B: Set of exemplars which belong to category B.

K: Complete set of exemplars which have entered the race.

Random Walk Probabilistic Predictions

$$E(N|i) = \frac{B}{q_i - p_i} - \frac{A+B}{q_i - p_i}\left[\frac{1-\left(\frac{q_i}{p_i}\right)^B}{1-\left(\frac{q_i}{p_i}\right)^{A+B}}\right] \tag{2.7}$$

$$P(A|i) = \frac{1-\left(\frac{q_i}{p_i}\right)^B}{1-\left(\frac{q_i}{p_i}\right)^{A+B}}, \quad if\ p_i \neq q_i$$
$$P(A|i) = \frac{B}{A+B}, \qquad if\ p_i = q_i \tag{2.8}$$

$$P(B|i) = \frac{\left(\frac{q_i}{p_i}\right)^B - \left(\frac{q_i}{p_i}\right)^{A+B}}{1-\left(\frac{q_i}{p_i}\right)^{A+B}}, \quad if\ p_i \neq q_i$$
$$P(B|i) = \frac{A}{A+B}, \qquad if\ p_i = q_i \tag{2.9}$$

$\mathrm{E}(N|i)$: Expected number of steps in the random walk for trial i.

$P(\mathrm{A}|\mathrm{i})$: Probability if classifying item i to category A

$P(\mathrm{B}|\mathrm{i})$: Probability if classifying item i to category B

A, B: Classification thresholds. When the walk crosses one of these values, a response is in favor of the corresponding category.

p_i: The probability that an exemplar belonging to category A finishes the walk first. See Equation 2.5

q_i: The probability that an item belonging to category B finishes the walk first. See Equation 2.5.

Estimated Total Time for Random Walk

$$E(\min(T_1, \dots, T_n)) = \frac{1}{\lambda_1 + \cdots + \lambda_n} = \frac{1}{\sum_{j=1}^{n} a_{ij}} \tag{2.10}$$

$$E(t_s|i) = \alpha + \frac{1}{\sum_{j=1}^{n} a_{ij}} \tag{2.11}$$

$E(\min(T_1, \dots, T_n))$: Expected minimum value of T_n exponentially distributed random variables. Corresponds to the estimated finishing time for the *exemplar race* winner.

$E(t_s|i)$: Expected step time when probe *i* is presented.

α: Extraction time of category label

λ_j: The decay rate of exponentially distributed random variable T_j.

a_{ij}: The activation of exemplar *j* when probe *i* is presented.

Expected Activation Value for d_{ij} Extremes

$$\lim_{d_{ij} \to 0} (a_{ij}) = M_j$$
$$\lim_{d_{ij} \to \infty} (a_{ij}) = 0 \tag{2.12}$$

a_{ij}: The degree to which exemplar *j* is activated when item *i* is presented.

d_{ij}: Distance between exemplar *j* and probe *i*

M_j: The memory strength for exemplar *j*

Random Walk for Two Category Forced Choice Paradigm

$$E_2(N|i) = \frac{\widehat{A}}{1 - 2p_i} - \frac{2\widehat{A}}{1 - 2p_i}\left[1 + \left(\frac{1 - p_i}{p_i}\right)^{\widehat{A}}\right]^{-1} \tag{2.13}$$

$E_2(N|i)$: Estimated number of steps in the random

$\widehat{A}$: Estimated category threshold over multiple trials for balanced A and B trials

p_i: Probability that an exemplar belonging to category A will finish the race first

Category Cohesion Metric

$$\mathcal{C}_{T,I} \sim \frac{D_{\mathrm{T,I}}}{{}^{n}V_{T}} \tag{2.14}$$

$\mathcal{C}_{T,I}$: Cohesion metric for category pairs

${}^{n}V_{T}$: The n-volume of the target category

$D_{\mathrm{T,I}}$: Distance metric between target (T) and irrelevant (I) categories, taking into consideration category size.

Formalized Condition for the set cover problem

$$\forall x \forall y \left(\left(\{e_x, e_y\} \in U \right) \rightarrow \exists z \left(\left(\{e_x, e_y\} \in S_z \right) \wedge \left(S_z \in Q \right) \right) \right) \tag{4.1}$$

e_x, e_y: Elements of U

U: Universe containing all elements.

S_z: Subset of U of size k

Q: Set of sets containing all S_z

z: Subset index.

x, y: Element indexes.

Expected minimum and maximum number of subsets in Q which solves the set cover problem

$$q_{max} = \frac{\left\lceil {}^{2n}/_{k} \right\rceil !}{k! \left(\left\lceil {}^{2n}/_{k} \right\rceil - k \right) !} \tag{4.2}$$

$$q_{min} = \frac{\left\lfloor {}^{2n}/_{k} \right\rfloor !}{k! \left(\left\lfloor {}^{2n}/_{k} \right\rfloor - k \right) !} \tag{4.3}$$

q_{max}: The largest expected value of an optimal solution for Q which satisfies the condition formalized in 4.1.

q_{min}: The smallest expected value of an optimal solution for Q which satisfies the condition formalized in 4.1.

n: The number of elements in the universe U.

k: The maximal size for all sets in Q.

Partition Configuration Cost

$$\mathbb{C}_P = \sum_{h=1}^{k} \sum_{e_i \in C_h} dis(e_i, \mathfrak{m}_h) \tag{4.4}$$

$\mathbb{C}_P$: Cost of a partition *P*

$dis(e_i, \mathfrak{m}_h)$: Function which calculates the distance between an element e_i of a cluster C_h and the cluster's medoid $\mathfrak{m}_h$

k: Total number of clusters

Dunn Index

$$DI_m = \min_{1 \le i \le m} \left\{ \min_{1 \le j \le m, j \ne i} \left\{ \frac{\delta(C_i, C_j)}{\max_{1 \le k \le m} \Delta(C_k)} \right\} \right\} \tag{4.5}$$

DI_m: Dunn Index for cluster set of size *m*

$\delta(C_i, C_j)$: The function that calculates the inter cluster distance metric between clusters *i* and *j*.

$\Delta(C_k)$: The function that calculates the inner diameter metric for cluster *k*

Expected Reaction Time with Increasing Dunn Index Values

$$E(RT|DI_{\text{AB}}) = RT_{max} \cdot \exp(-\lambda \cdot DI_{\text{AB}}) + RT_{min} \tag{4.7}$$

$E(RT|DI_{\text{AB}})$: Expected reaction time given a specific Dunn Index

DI_{AB}: The Dunn Index for a two category superset.

RT_{max}: Reaction time expected when the Dunn Index approaches 0.

RT_{min}: Shortest possible reaction time.

λ: Decay rate

Smallest Dunn Index Value with Predictive Power

$$DI_{\emptyset} = \frac{\ln(RT_0)}{\lambda} \tag{4.8}$$

$DI_{\emptyset}$: Smallest Dunn Index value for which reaction times predictions can be made.

RT_0: Theoretically expected reaction time when the Dunn Index approaches 0.

λ: Decay rate.

12 Bibliography

Ashby, F. G., & Ennis, D. (2007). Similarity measures. *Scholarpedia*, *2*(12), 4116. doi:10.4249/scholarpedia.4116

Ashby, F. G., & Perrin, N. A. (1988). Toward a unified theory of similarity and recognition. *Psychological Review*, *95*(1), 124–150. doi:10.1037/0033-295X.95.1.124

Aydede, M. (1998). Fodor on concepts and Frege Puzzles. *Pacific Philosophical Quarterly*, *79*(4), 289–294. doi:10.1111/1468-0114.00063

Bates, D. M., & Watts, D. G. (2007). *Nonlinear Regression Analysis and Its Applications*. Wiley.

Beals, R., Krantz, D. H., & Tversky, A. (1968). Foundations of Multidimensional Scaling. *Psychological Review*, *75*(2), 127–142. doi:10.1037/h0025470

Bimler, D., & Kirkland, J. (2009). Colour-space distortion in women who are heterozygous for colour deficiency. *Vision Research*, *49*(5), 536–543. doi:10.1016/j.visres.2008.12.015

Bimler, D., Kirkland, J., & Jacobs, R. (2000). Colour-vision tests considered as a special case of multidimensional scaling. *Color Research & Application*, *25*(3), 160–169. doi:10.1002/(SICI)1520-6378(200006)25:3<160::AID-COL4>3.0.CO;2-N

Bimler, D., Kirkland, J., & Jameson, K. A. (2004). Quantifying variations in personal color spaces: Are there sex differences in color vision? *Color Research & Application*, *29*(2), 128–134. doi:10.1002/col.10232

Blashfield, R. K. (1976). Mixture model tests of cluster analysis: Accuracy of four agglomerative hierarchical methods. *Psychological Bulletin*, *83*(3), 377–388. doi:10.1037/0033-2909.83.3.377

Boynton, R. M., & Olson, C. X. (1987). Locating basic colors in the OSA space. *Color Research & Application*, *12*(2), 94–105. doi:10.1002/col.5080120209

Brock, G., Pihur, V., Datta, S., & Datta, S. (2011). clValid, an R package for cluster validation. *Journal of Statistical Software (Brock et Al., March 2008)*, *25*(4), 1–22.

Cavanagh, P. (2011). Visual cognition. *Vision Research*, *51*(13), 1538–1551. doi:10.1016/j.visres.2011.01.015

Cohen, A. L., & Nosofsky, R. M. (2003). An extension of the exemplar-based random-walk model to separable-dimension stimuli. *Journal of Mathematical Psychology*, *47*(2), 150–165.

Cox, T. F., & Cox, M. A. A. (2000). *Multidimensional Scaling, Second Edition*. CRC Press.

D'Andrade, R. (1990). Some propositions about the relations between culture and human cognition. In J. W. Stigler, R. A. Shweder, & G. Herdt (Eds.), *Cultural psychology: Essays on comparative human development* (pp. 65–129). New York, NY, US: Cambridge University Press.

D'Andrade, R. G. (1981). The cultural part of cognition. *Cognitive Science*, *5*(3), 179–195. doi:10.1016/S0364-0213(81)80012-2

Defays, D. (1977). An efficient algorithm for a complete link method. *The Computer Journal*, *20*(4), 364–366. doi:10.1093/comjnl/20.4.364

Derefeldt, G., Swartling, T., Berggrund, U., & Bodrogi, P. (2004). Cognitive color. *Color Research & Application*, *29*(1), 7–19. doi:10.1002/col.10209

Dunn, J. C. (1973). A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *Journal of Cybernetics*, *3*(3), 32–57. doi:10.1080/01969727308546046

Estes, W. K. (1986). Array models for category learning. *Cognitive Psychology*, *18*(4), 500–549. doi:10.1016/0010-0285(86)90008-3

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). Hierarchical Clustering. In *Cluster Analysis* (pp. 71–110). John Wiley & Sons, Ltd. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/9780470977811.ch4/summary

Feitosa-Santana, C., Oiwa, N. N., Paramei, G. V., Bimler, D., Costa, M. F., Lago, M., … Ventura, D. F. (2006). Color space distortions in patients with type 2 diabetes mellitus. *Visual Neuroscience*, *23*(3-4), 663–668. doi:10.1017/S0952523806233546

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications, Vol. 1, 3rd Edition* (3rd edition.). New York: Wiley.

Fisher, A. V., & Sloutsky, V. M. (2004). Categorization and memory: Representation of category information increases memory intrusions. In *Proceedings of the XXVI annual conference of the Cognitive Science Society* (pp. 387–391). Retrieved from http://www.cogsci.northwestern.edu/cogsci2004./papers/paper409.pdf

Fodor, J. (2010). *LOT 2: The language of thought revisited*. OUP Oxford.

Fodor, J., & Lepore, E. (1996). The red herring and the pet fish: why concepts still can't be prototypes. *Cognition*, *58*(2), 253–270. doi:10.1016/0010-0277(95)00694-X

Fox, J. (2002). *Nonlinear regression and nonlinear least squares*. January. Retrieved from http://fs6.fudan.edu.cn/mirror/CRAN/doc/contrib/Fox-Companion/appendix-nonlinear-regression.pdf

Franklin, A., Pilling, M., & Davies, I. (2005). The nature of infant color categorization: Evidence from eye movements on a target detection task. *Journal of Experimental Child Psychology*, *91*(3), 227–248. doi:10.1016/j.jecp.2005.03.003

Gallo, D. A. (2010). False memories and fantastic beliefs: 15 years of the DRM illusion. *Memory & Cognition*, *38*(7), 833–848. doi:10.3758/MC.38.7.833

Gardner, D. G., Gardner, J. C., Laush, G., & Meinke, W. W. (1959). Method for the Analysis of Multicomponent Exponential Decay Curves. *The Journal of Chemical Physics*, *31*(4), 978–986. doi:10.1063/1.1730560

Garner, W. R. (2014). *The Processing of Information and Structure*. Psychology Press.

Gati, I., & Tversky, A. (1984). Weighting common and distinctive features in perceptual and conceptual judgments. *Cognitive Psychology*, *16*(3), 341–370. doi:10.1016/0010-0285(84)90013-6

Gazda, G. M., & Mobley, J. A. (1981). INDSCAL multidimensional scaling. *Journal of Group Psychotherapy, Psychodrama & Sociometry*, *34*, 54–73.

Geraci, L., & McCabe, D. P. (2006). Examining the basis for illusory recollection: The role of remember/know instructions. *Psychonomic Bulletin & Review*, *13*(3), 466–473. doi:10.3758/BF03193871

Goldstone, R. L. (1994). The role of similarity in categorization: providing a groundwork. *Cognition*, *52*(2), 125–157. doi:10.1016/0010-0277(94)90065-5

Goldstone, R. L. (1995). Effects of Categorization on Color Perception. *Psychological Science*, *6*(5), 298–304.

Gruppuso, V., Lindsay, D. S., & Kelley, C. M. (1997). The process-dissociation procedure and similarity: Defining and estimating recollection and familiarity in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(2), 259–278. doi:10.1037/0278-7393.23.2.259

Hartigan, J. A., & Hartigan, P. M. (1985). The Dip Test of Unimodality. *The Annals of Statistics*, *13*(1), 70–84. doi:10.1214/aos/1176346577

Hunter, R. S. (1987). *The Measurement of Appearance*. John Wiley & Sons.

Indow, T. (1988). Multidimensional studies of Munsell color solid. *Psychological Review*, *95*(4), 456–470. doi:10.1037/0033-295X.95.4.456

Ishihara, S. (1981). *Ishihara's Tests for Colour-Blindness, 24 Plates Edition*. Tokyo: Kanehara.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, *30*(5), 513–541. doi:10.1016/0749-596X(91)90025-F

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666. doi:10.1016/j.patrec.2009.09.011

Jain, A. K., & Dubes, R. C. (1988). *Algorithms for Clustering Data*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc.

Kaufman, L., & Rousseeuw, P. J. (1990). Partitioning Around Medoids (Program PAM). In *Finding Groups in Data* (pp. 68–125). John Wiley & Sons, Inc. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/9780470316801.ch2/summary

Kuehni, R. G. (2001). Color space and its divisions. *Color Research & Application*, *26*(3), 209–222. doi:10.1002/col.1018

Lawler, G. F., & Limic, V. (2010). *Random Walk: A Modern Introduction*. Cambridge; New York: Cambridge University Press.

Lee, M. D. (2001). Determining the Dimensionality of Multidimensional Scaling Representations for Cognitive Modeling. *Journal of Mathematical Psychology*, *45*(1), 149–166. doi:10.1006/jmps.1999.1300

Li, B., Chang, E., & Wu, Y. (2003). Discovery of a perceptual distance function for measuring image similarity. *Multimedia Systems*, *8*(6), 512–522. doi:10.1007/s00530-002-0069-9

Lindsey, D. T., & Brown, A. M. (2009). World Color Survey color naming reveals universal motifs and their within-language diversity. *Proceedings of the National Academy of Sciences*, *106*(47), 19785–19790. doi:10.1073/pnas.0910981106

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, *95*(4), 492–527. doi:10.1037/0033-295X.95.4.492

Macmillan, N. A., & Creelman, C. D. (2004). *Detection Theory: A User's Guide*. Psychology Press.

Maechler, M., original), P. R. (Fortran, original), A. S. (S, original), M. H. (S, maintenance(1999-2000)), K. H. (port to R., Studer, M., & Roudier, P. (2014). cluster: Cluster Analysis Extended Rousseeuw et al (Version 1.15.3). Retrieved from http://cran.r-project.org/web/packages/cluster/index.html

Mahy, M., Van Eycken, L., & Oosterlinck, A. (1994). Evaluation of Uniform Color Spaces Developed after the Adoption of CIELAB and CIELUV. *Color Research & Application*, *19*(2), 105–121. doi:10.1111/j.1520-6378.1994.tb00070.x

Mandler, J. M. (2004). *The Foundations of Mind: Origins of Conceptual Thought: Origins of Conceptual Thought*. Oxford University Press, USA.

Mandler, J. M. (2008). On the Birth and Growth of Concepts. *Philosophical Psychology*, *21*(2), 207–230. doi:10.1080/09515080801980179

Maulik, U., & Bandyopadhyay, S. (2002). Performance evaluation of some clustering algorithms and validity indices. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(12), 1650–1654. doi:10.1109/TPAMI.2002.1114856

McDermott, K. C., & Webster, M. A. (2012). Uniform color spaces and natural image statistics. *Journal of the Optical Society of America A*, *29*(2), A182–A187. doi:10.1364/JOSAA.29.00A182

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207–238. doi:10.1037/0033-295X.85.3.207

Nelson, R. (1995). *Probability, Stochastic Processes, and Queueing Theory: The Mathematics of Computer Performance Modeling*. Springer Science & Business Media.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(1), 104–114. doi:10.1037/0278-7393.10.1.104

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57. doi:10.1037/0096-3445.115.1.39

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(1), 87–108. doi:10.1037/0278-7393.13.1.87

Nosofsky, R. M. (1988). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 54–65. doi:10.1037/0278-7393.14.1.54

Nosofsky, R. M. (1989). Further tests of an exemplar-similarity approach to relating identification and categorization. *Perception & Psychophysics*, *45*(4), 279–290. doi:10.3758/BF03204942

Nosofsky, R. M. (1990). Relations between exemplar-similarity and likelihood models of classification. *Journal of Mathematical Psychology*, *34*(4), 393–418. doi:10.1016/0022-2496(90)90020-A

Nosofsky, R. M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception and Performance*, *17*(1), 3–27. doi:10.1037/0096-1523.17.1.3

Nosofsky, R. M. (1992). Similarity Scaling and Cognitive Process Models. *Annual Review of Psychology*, *43*(1), 25–53. doi:10.1146/annurev.ps.43.020192.000325

Nosofsky, R. M. (2011). The generalized context model: An exemplar model of classification. In *Formal approaches in categorization* (pp. 18–39). Cambridge University Press.

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, *104*(2), 266–300. doi:10.1037/0033-295X.104.2.266

Oberauer, K. (2001). Removing irrelevant information from working memory: A cognitive aging study with the modified Sternberg task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(4), 948–957. doi:10.1037/0278-7393.27.4.948

Oberauer, K. (2008). How to say no: Single- and dual-process theories of short-term recognition tested on negative probes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(3), 439–459. doi:10.1037/0278-7393.34.3.439

Pakhira, M. K., Bandyopadhyay, S., & Maulik, U. (2004). Validity index for crisp and fuzzy clusters. *Pattern Recognition*, *37*(3), 487–501. doi:10.1016/j.patcog.2003.06.005

Palmeri, T. J., & Gauthier, I. (2004). Visual object understanding. *Nature Reviews Neuroscience*, *5*(4), 291–303. doi:10.1038/nrn1364

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

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

Richler, J. J., & Palmeri, T. J. (2014). Visual category learning. *Wiley Interdisciplinary Reviews: Cognitive Science*, *5*(1), 75–94. doi:10.1002/wcs.1268

Roberson, D., Davidoff, J., Davies, I. R. L., & Shapiro, L. R. (2005). Color categories: Evidence for the cultural relativity hypothesis. *Cognitive Psychology*, *50*(4), 378–411. doi:10.1016/j.cogpsych.2004.10.001

Roberson, D., Davies, I., & Davidoff, J. (2000). Color categories are not universal: Replications and new evidence from a stone-age culture. *Journal of Experimental Psychology: General*, *129*(3), 369–398. doi:10.1037/0096-3445.129.3.369

Roberson, D., & Hanley, J. R. (2007). Color Vision: Color Categories Vary with Language after All. *Current Biology*, *17*(15), R605–R607. doi:10.1016/j.cub.2007.05.057

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(4), 803–814. doi:10.1037/0278-7393.21.4.803

Roediger, H. L., Watson, J. M., McDermott, K. B., & Gallo, D. A. (2001). Factors that determine false recall: A multiple regression analysis. *Psychonomic Bulletin & Review*, *8*(3), 385–407. doi:10.3758/BF03196177

Santini, S., & Jain, R. (1999). Similarity measures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *21*(9), 871–883. doi:10.1109/34.790428

Schinazi, R. B. (1999). *Classical and Spatial Stochastic Processes* (1999 edition.). Boston: Birkhäuser.

Seaborn, M., Hepplewhite, L., & Stonham, J. (2005). Fuzzy colour category map for the measurement of colour similarity and dissimilarity. *Pattern Recognition*, *38*(2), 165–177. doi:10.1016/j.patcog.2004.05.001

Shepard, R. N. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology*, *1*(1), 54–87. doi:10.1016/0022-2496(64)90017-3

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323. doi:10.1126/science.3629243

Shepard, R. N. (2001). Perceptual-cognitive universals as reflections of the world. *Behavioral and Brain Sciences*, *24*(04), 581–601. doi:10.1017/S0140525X01000012

Simons, D. J., & Rensink, R. A. (2005). Change blindness: past, present, and future. *Trends in Cognitive Sciences*, *9*(1), 16–20. doi:10.1016/j.tics.2004.11.006

Smith, E. R., & DeCoster, J. (2000). Dual-Process Models in Social and Cognitive Psychology: Conceptual Integration and Links to... *Personality & Social Psychology Review (Lawrence Erlbaum Associates)*, *4*(2), 108–131. doi:10.1207/S15327957PSPR0402_01

Stadler, M. A., Roediger, H. L., & McDermott, K. B. (1999). Norms for word lists that create false memories. *Memory & Cognition*, *27*(3), 494–500. doi:10.3758/BF03211543

Sternberg, S. (1966). High-Speed Scanning in Human Memory. *Science*, *153*(3736), 652–654. doi:10.1126/science.153.3736.652

Sturges, J., & Whitfield, T. W. A. (1995). Locating basic colours in the munsell space. *Color Research & Application*, *20*(6), 364–376. doi:10.1002/col.5080200605

Tokunaga, R., & Logvinenko, A. D. (2010). Hue manifold. *Journal of the Optical Society of America A*, *27*(12), 2551–2557. doi:10.1364/JOSAA.27.002551

Tversky, A., & Gati, I. (1978). Studies of Similarity. In E. Rosch & B. Lloyd (Eds.), *Cognition and Categorization* (pp. 1–1978). Lawrence Elbaum Associates.

Tversky, A., & Gati, I. (1982). Similarity, separability, and the triangle inequality. *Psychological Review*, *89*(2), 123–154. doi:10.1037/0033-295X.89.2.123

Tversky, A., & Krantz, D. H. (1970). The dimensional representation and the metric structure of similarity data. *Journal of Mathematical Psychology*, *7*(3), 572–596. doi:10.1016/0022-2496(70)90041-6

Warrington, E. K., & Weiskrantz, L. (1968). A study of learning and retention in amnesic patients. *Neuropsychologia*, *6*(3), 283–291. doi:10.1016/0028-3932(68)90026-2

Winocur, G., & Weiskrantz, L. (1976). An investigation of paired-associate learning in amnesic patients. *Neuropsychologia*, *14*(1), 97–110. doi:10.1016/0028-3932(76)90011-7

Wuerger, S. M., Maloney, L. T., & Krauskopf, J. (1995). Proximity judgments in color space: Tests of a Euclidean color geometry. *Vision Research*, *35*(6), 827–835. doi:10.1016/0042-6989(94)00170-Q

Wyszecki, G., & Stiles, W. S. (2000). *Color Science: Concepts and Methods, Quantitative Data and Formulae* (2 edition.). New York: Wiley-Interscience.

Xiao, Y., Kavanau, C., Bertin, L., & Kaplan, E. (2011). The Biological Basis of a Universal Constraint on Color Naming: Cone Contrasts and the Two-Way Categorization of Colors. *PLoS ONE*, *6*(9). doi:10.1371/journal.pone.0024994

Yoshioka, T., Dow, B. M., & Vautin, R. G. (1996). Neuronal mechanisms of color categorization in areas V1, V2 and V4 of macaque monkey visual cortex. *Behavioural Brain Research*, *76*(1–2), 51–70. doi:10.1016/0166-4328(95)00183-2

Universität Zürich[UZH]

Philosophische Fakultät
Studiendekanat

Universität Zürich
Studiendekanat
Bereich Abschluss
Rämistr. 69
CH-8001 Zürich
Telefon +41 44 634 54 10
www.phil.uzh.ch

Selbstständigkeitserklärung

Hiermit erkläre ich, dass
die Masterarbeit von mir selbst und ohne unerlaubte Beihilfe verfasst worden ist und ich die Grundsätze wissenschaftlicher Redlichkeit einhalte (vgl. dazu: http://www.lehre.uzh.ch/index/LK-Plagiate-Merkblatt.pdf).

Zürich 5.1.2015 .....

Ort und Datum Unterschrift

PERSONAL DATA

Name Alexei Fischer

Born 22.02.1982 in Curitiba, PR, Brazil

Address Langfurren 5
8057 Zürich
Switzerland

Email alexei.fischer@gmail.com