INTERACTION BETWEEN SUPERVISED AND UNSUPERVISED INFORMATION IN HUMAN CATEGORY LEARNING

Sára Jellinek

Central European University
Department of Cognitive Science

In partial fulfilment of the requirements for the degree of Doctor of Philosophy in Cognitive Science

Supervisor: József Fiser

Secondary supervisor: Máté Lengyel

Budapest, March 2020

Declaration of Authorship

I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or which have been accepted for the award of any other degree or diploma at Central European University or any other educational institution, except where due acknowledgement is made in the form of bibliographical reference.

_____

Sára Jellinek

Abstract

In my thesis, I investigate human semi-supervised learning (SSL), the ability to learn simultaneously from labelled and unlabelled data with high efficiency. I argue that the majority of earlier results of human SSL in the literature are straightforward if one assumes that learners adopt a generative learning method when acquiring new categories. Since generative learning means creating a representation of the entire outside world, under natural conditions, this amounts to building rich models based on all the incoming information, which is, by definition has to be a mixture of labelled and unlabelled data. Therefore, I propose that research on SSL should not focus on the typically investigated question of *whether* humans can integrate supervised and unsupervised information, since they evidently can. Rather, the focus of investigation on SSL should be on the question of *how* this integration occurs, namely the behavioral and neural specifics of the interaction between learning from labelled and non-labelled data during the course of knowledge acquisition.

Following my proposal, I present three studies relevant to the field of human category learning. First, I present evidence for automatic generative learning in humans. I show that irrespective of the task at hand and the relevance of the incoming information to solving that task, humans automatically build a generative internal model of the data, even when a much simpler discriminative model would suffice for completing the task perfectly.

In the second study, I investigate whether performing first unsupervised then supervised learning or vice versa during a categorization task makes a difference in learning. The majority of SSL categorization studies in the field investigated the influence of unsupervised information on a category representation that had been built earlier based on supervised information.

However, all these studies used different stimuli, different procedures, and different set of participants so their conclusions are hard to incorporate into one comprehensive framework. There exist no earlier study in the literature addressing the question whether there is any qualitative difference between the emerging representations of SSL depending on the order of training. Characterizing this aspect of the integration process is the first step towards a comprehensive understanding of human SSL.

Finally, in the third study, I investigate the neural correlates of the process by which internal representation of novel categories in the cortex emerge. I analyze the changes of commonly studied neural correlates (P300 ERP, α ERD and θ ERS) of categorization throughout the process of category acquisition. Previous studies addressing these neural correlates lack two important characteristics of human categorization. First, they used highly familiar categories, thus no information was gained about the dynamics of particular neural responses as categories emerged. Second, the commonly used stimuli of these studies were few and highly discrete. This hinders understanding how the structure of the stimulus modulates neural responses. Eliminating these shortcomings by using unfamiliar categories and continuously changing feature sets, I show that commonly studied neural correlates have the potential of reflecting the ongoing emergence of the internal representation, and in addition, they are modulated by the difficulty of the task, and the strength of category membership.

Acknowledgements

I would like to mention some people to whom I am very grateful for their help and support over the past few years.

First, I would like to thank my supervisor, József Fiser for his help, advice and supervision during my studies.

I am grateful to all former and current members of our lab. The friendly, supportive environment they have created, made these few years a period in my life I will be happy to look back on. I owe special thanks to Oana Stanciu, Ádám Koblinger, József Arató, Márton Nagy, Gábor Lengyel, Benő Márkus and Tünde Szabó for all their help and advice on various topics, inspiring discussions and fun chats during coffee breaks.

Outside of our lab first, I would like to thank Barbara Pomiechowska, Eugenio Parise, Gergely Csibra and Ágnes Volein for their guidance and help with my EEG study.

I am thankful to the members of the Department of Cognitive Science, especially Réka Finta for all her hard work and always cheerful, encouraging attitude.

Finally, I would like to thank my family, especially Dóra Jellinek, András Pogány, Edit Fehérvári, Erika Höfler, Ferenc Molnár and most importantly, my husband, Attila Molnár for their constant encouragement and support on so many personal and scientific levels.

Contents

List of Figures

Chapter 1

Introduction

We continuously interpret and characterize incoming sensory information and this process is augmented by abstract latent mental constructs represented in our brain and variably referred to as schemes, scripts, categories, concepts or objects. These conceptual constructs evolve in time due to experience in the current sensory and social contexts and they perpetually bias our perception (Gauthier, James, Curby and M.J.Tarr, 2003; Goldstone, 1994; Goldstone, Lippaa and Shiffrin, 2001; Op de Beeck, Wagemans and Vogels, 2003), information processing, and ultimately our interpretation and interaction with our environment, in other words, our cognition (Canini, Shashkov and Griffiths, 2010; Canini and Griffiths, 2011; Heller, Sanborn and Chater, 2009; Taylor, Devereux, Acres, Randall and Tyler, 2012; Harnad, 2005). Categorization, the action of separating and grouping information based on some similarity measurement and relevance is a central constituent of this process. As a consequence, understanding the mechanisms underlying categorization and the acquisition of categories is crucial for understanding human cognition.

Object categorization has been argued to occur at multiple levels. Bornstein (1984) defined

four types of categorization. The simplest form is *identity categorization* that allows one to recognize the same object presented multiple times as such across a series of many other objects. The second is recognition *equivalence categorization* that enables one to identify the same object even when it is presented with multiple variations in appearance (rotation, modality or dimensionality). The third is *perceptual equivalence categorization*, possibly the most commonly investigated form of categorization in cognitive sciences, also, the main focus of the present thesis, which describes grouping of objects that are physically distinct, but share qualitatively similar attributes. Finally, *conceptual categorization* requires additional knowledge about the objects beyond their perceptual appearance, for example their function or the role they play in different events, to be grouped together.

There are many aspects of perceptual categorization and category learning processes that influence the efficiency of learning and the resulting representation, an internal model of components and structure of our environment. Such aspects can be sorted into two groups based on whether they define the quantity and nature of potentially incoming information (external), or if they are attributes of the emerging representation on the receiving end (internal). Examples of the former group are the number of (relevant) stimulus dimensions and their relation to one another (single or multiple dimensions (Ashby and Maddox, 2011; Ashby and Valentin, 2017), integral or separable stimulus dimensions (Shepard, 1991; Nosofsky and Palmeri, 1996; Maddox and Dodd, 2003)), the presence and type or absence of feedback during acquisition (supervised or unsupervised learning), or the distribution of stimuli in relation to the task (whether the stimulus distribution is suggestive of the category boundary (Ell and Ashby, 2006)). Aspects concerning internal processes of the learner are for instance the learning strategy and the structure of the resulting representation (generative or discriminative learning (Hsu and Grif-

fiths, 2010)) or the conscious accessibility and content of the acquired knowledge (implicit or explicit learning) (Sun, Zhang, Slusarz and Mathews, 2007; Ziori and Dienes, 2012).

To acquire a reliable and exhaustive model of human category acquisition, experimental studies have systematically addressed the role and influence of the above mentioned (mostly external) factors and their effect on one another. These studies utilized various perceptual modalities (visual, auditory (Goudbeek, Smits, Cutler and Swingley, 2005; Liu, Montes-Lourido, Wang and Sadagopan, 2019), haptic (Gaißert, Waterkamp, Fleming and Bülthoff, 2012; Schwarzer, Küfer and Wilkening, 1999), olfactory (Locatelli, Fernandez and Smith, 2016)), a wide range of stimulus complexity (from one dimensional (Hsu and Griffiths, 2010), through 2-3D simple objects (Markant and Gureckis, 2014) to natural scenes (Li, VanRullen, Koch and Perona, 2002)) and a great variety of features of supervision over the learning process (delay (Maddox, Ashby and Bohil, 2003; Stephens and Kalish, 2018), valence (Ashby and O'Brien, 2007), presence or absence of feedback or the use of labels (Ashby, Maddox and Bohil, 2002)).

One particular aspect of categorization, the presence or absence of feedback during category acquisition allows to address two crucial research questions. By adding feedback to the learning process, the capacity and dynamics of human category learning can be investigated. Forming categories in the absence of feedback allows us to see how the unbiased human brain structures incoming information naturally. Although both of these learning types are extremely important, neither of them is realistic in isolation in the light of how natural category formation occurs most of the time in real life. For example, at early age, when children learn about their environment, most of the observation about surrounding objects left unlabelled, that is the learning is unsupervised. However, on some occasions, children do receive labels or corrective feedback when they encounter novel objects in their environment. This begs the question: How

much generalizable are the results of current studies obtained by strictly supervised or strictly unsupervised experimental setups to processes during natural category acquisition?

Intuitively, the learning framework that would best match natural human category acquisition by integrating supervised and unsupervised learning is the method that is called in the machine learning literature semi-supervised learning (SSL). Despite of a wide consensus in the field that SSL is the most ecologically relevant form of human category learning, there are surprisingly few empirical studies addressing any aspects of SSL. In addition, as reviewed below, the usefulness of the existing results is questionable. In the rest of the chapter, I will discuss the most important findings both in supervised and unsupervised learning, and critically review the literature on human semi-supervised learning. I will continue by discussing the most relevant findings on the debate of discriminative *vs.* generative learning in humans. I will argue in favor of generative models as the proper conceptualization of human category learning, and suggest that findings in the literature on SSL are explainable in a completely straightforward manner if one supposes that humans build generative models of their environment.

1.1 Supervised, unsupervised and semi-supervised learning

1.1.1 Supervised learning

Supervised learning in the context of category learning refers to the condition when the learner receives information about the group membership with each incoming piece of information or each stimulus. Most of the experiments conducted in human categorization research use this condition (Gureckis and Love, 2003a). Supervised learning methods have been proposed having virtually no limits with respect to what task they can teach to humans (Pothos, Edwards and

Perlman, 2011). It has also been argued by Goudbeek et al. (2005) that, in the visual domain, humans are unable to learn category structures requiring the integration of multiple stimulus dimensions (i.e. non-trivial categorization rules) without supervision. Supervised experimental procedures differ widely from one another with respect to the valence, type and timing of supervision.

The most typical type of supervision is corrective feedback following and evaluating the category decision. In some studies, only error signals (negative feedback) are provided, while in others setups correct decisions also get acknowledged (positive feedback). Many studies reported that negative feedback facilitated learning more in experimental setups, where participants had to learn a simple, explicit categorization rule that was easy to verbalize (rule-based (RB) learning (Ashby and Maddox, 2011)). However, Ashby and O'Brien (2007) found that the combination of both positive and negative feedback outperformed the conditions where only either positive or negative feedback was used in more complex tasks that required implicit integration of two stimulus dimensions simultaneously at a pre-decisional stage. For example, such advantage was found in tasks where calculation of a weighted linear combination of the two relevant dimensions was necessary (Ashby and Gott, 1988). This method is called information-integration (II) category learning (Ashby and Maddox, 2011)). This result was later challenged by findings of Freedberg, Glass, Filoteo, Hazeltine and Maddox (2017) arguing that only negative feedback was necessary for successful learning. Integrating these previous results, a viable conclusion is that negative feedback is a *necessary*, but in some cases *not sufficient* for achieving the best performance when teaching categories in a supervised manner.

Apart from feedback, providing labels of categories with each stimulus are another form of supervision. Labels commonly either precede the stimulus or presented simultaneously with it.

Learning environments where the learner receives feedback after category decisions is called feedback learning, while procedures involving the presentation of labels are examples of observational learning. As Ashby, Maddox and Bohil (2002) showed, RB and II learning benefit more from different types of supervision. While RB is usually easy enough to be learned with virtually no supervision, the type of feedback does not modulate learning performance significantly. Type II learning, however, benefits most from feedback learning. An even more interesting difference between these types of supervision is that observational learning is supposed to encourage generative learning (see Section 1.2), by allowing the learner to pay more attention to the distribution of the stimuli, while feedback learning restricts the attention of the learner more to the discrimination rule between the categories, as a result it promotes discriminative learning (Levering and Kurtz, 2015).

A major drawback of purely supervised learning is that it appears to be preventing efficient generalization of the acquired category knowledge to novel stimuli (Jones, Love and Maddox, 2005; Patterson and Kurtz, 2018). Ideally, learners should be able to extend (generalize) the knowledge to new potential members of a category they are familiar with. It would be extremely inefficient and costly if we stood puzzled in front of an object with four legs, a seat and a back, if it had an unusual size, color or material than the chairs we got used to in the past. In addition, supervised learning is not very feasible with its assumption that one needs the help in the form of feedback from a knowledgeable other to clarify the hunch that the perceived object is a chair. Unsupervised learning, however, seem to be able to give rise to a more useful category representation, at least considering generalization of category information.

1.1.2 Unsupervised learning

Unsupervised learning has been explored considerably more exhaustively in machine learning (Hinton and Sejnowski, 1999). With respect to humans, their ability for unsupervised category learning seems to be very limited. Research on unsupervised learning asks an essentially different question than research on supervised learning. While supervised learning can explore the nature and limits of human learning capacity by pushing data and task complexity to extremes, or by manipulating different features of the learning environment, unsupervised learning is rather concerned with how categories and structured representations are naturally formed by humans, what the principles driving such information processing are (Pothos and Chater, 2002).

Typical unsupervised classification experiments require participants to group incoming stimuli without providing any constraints or guidance as to how this classification should happen. The number and content of the emerging groups created by participants reflect the natural, automatic processes that guide information processing in humans (Barlow, 1989; Austerweil and Griffiths, 2009). To further refine our knowledge on these processes, researchers can introduce a few constraints (such as predefined number of possible groups) or manipulate features of the incoming data or the learning environment. Such manipulations might affect the mode of stimulus presentation (simultaneous or sequential) or the number and type of stimulus dimensions (single or multi-dimensional stimuli or integral or separable feature dimensions). For instance, simultaneously presented data lead to similar classification solutions irrespective of the scatter and spatial location of the stimuli, while the order of sequentially presented stimuli has a strong effect on the classification strategy (Zeithamova and Maddox, 2009). Also, the relation of stimulus dimensions influence the learner's propensity to take into account more than one stimulus

dimension when they create categories. Integral dimensions are suggested to be processed holistically and are difficult to attend selectively to each dimension, while separable-dimension stimuli are processed analytically and are easy to attend selectively to these dimensions (Maddox and Dodd, 2003). If stimulus dimensions are highly separable, people tend to exhibit a so-called unidimensional bias where they rely on only one feature dimension to create a classification rule (Ashby, Alfonso-Reese, Turken and Waldron, 1998; Ashby, Queller and Berretty, 1999; Ashby and Maddox, 2011; Maddox, Ashby and Pickering, 2004; Vandist, De Schryver and Rosseel, 2009). With integral dimension stimuli learners are more inclined to rely on a combination of two dimensions when creating categories (Handel and Imai, 1972). This effect is supported by studies form the supervised learning literature (Nosofsky and Palmeri, 1996).

Under unconstrained unsupervised learning, seemingly, there can be no wrong solutions to a categorization problem, since forming categories strongly depends on the internal interpretation of the input. Nevertheless, it is an understandable expectation that the emerging internal representation of the stimuli should somehow match or reflect the distribution of external data. Such a matching is often used as the requirement for optimality of unsupervised categorization, especially in contexts, where the learner is expected to utilize the obtained knowledge in future interactions with her environment (Love, 2002; Love, 2003; Gureckis and Love, 2003a; Gureckis and Love, 2003b). In this sense, even without having access to the internal encoding of the input, there are, indeed, better and worse solutions to the classification problem.

The general agreement is that, in an unsupervised categorization context, observers' performance is optimal only if the categorization task is easy, so that the category clusters are strongly suggestive about where the boundary is. If the task is difficult, for instance it requires integration of two separable feature dimensions, the unidimensional bias will prevail, and the

learners will select one dimension and categorize consistently along that dimension (Ashby, Queller and Berretty, 1999). However, it is beyond doubts that focusing on the structure of the input exclusively by neglecting the effects of the internal representation even in unsupervised settings is incorrect, as there are many implicitly applied constraints and priors originating from the internal representation that will influence learning (Goldstone, Lippaa and Shiffrin, 2001; Gregory, 1997; Mitchell, Ropar, Ackroyd and Rajendran, 2005).

1.1.3 Semi-supervised learning

Semi-supervised learning (SSL) is the form of category learning that combines both supervised and unsupervised information throughout the learning process. Although, in the machine learning literature, there is a vast amount of research addressing the computational aspects of SSL (Zhu, 2005), only a handful of studies tried to explore SSL in humans. The very first human study, which simply aimed at establishing the phenomenon in humans was published only in 2007 (Zhu, Rogers, Qian and Kalish, 2007). Unfortunately, the currently existing literature on SSL is far from being systematic and contains often contradictory results. In the following section, I review and critically evaluate this literature.

The effect of unsupervised trials on representations obtained by previous supervised learning

The earliest studies of SSL focused on the role and impact of unsupervised information in category learning (Zhu, Rogers, Qian and Kalish, 2007; Vandist, De Schryver and Rosseel, 2009; Kalish, Rogers, Lang and Zhu, 2011; Lake and McClelland, 2011; McDonnell, Jew and Gureckis, 2012; Kalish, Zhu and Rogers, 2015). Typical experiments presented learners with a few uniformly distributed supervised (feedback) stimuli from two categories defined along one rel-

evant feature dimension. (See Fig. 1.1 for examples) Learners are assumed to infer the category boundary at the geometrical mid point between the presented supervised stimuli. This supervised learning phase is followed by an unsupervised one, where the learners receive a large number of additional stimuli in an unsupervised manner. Crucially, the distribution of unsupervised stimuli suggest a category boundary that is shifted with respect to the one implied in the supervised training phase. A commonly sought hallmark of SSL in these studies is the change in the representation of the categories, that is signalled by the shifted category boundary as a result of additional unsupervised information (Zhu et al., 2007; Kalish et al., 2011; McDonnell, Jew and Gureckis, 2012; Gibson et al., 2015; Kalish, Zhu and Rogers, 2015). A major flaw in the earliest studies (Zhu et al., 2007; Lake and McClelland, 2011; Kalish et al., 2011) as pointed out by McDonnell, Jew and Gureckis (2012) is that, as a result of the one dimensional categorization and the indefinite range of stimulus values, it is unclear whether learners are updating each category representation separately or they respond to a possible global shift in the stimulus space. Somewhat stronger evidence for SSL was provided by McDonnell, Jew and Gureckis (2012) and Kalish, Zhu and Rogers (2015). McDonnell, Jew and Gureckis (2012) taught participants two categories defined in a two-dimensional (line length and orientation) stimulus space in a supervised training setup. The distribution of the supervised stimuli was equally compatible with two distinct, orthogonal category boundaries along one or the other stimulus dimension. It was the distribution of additional unsupervised stimuli that decided which category boundary is the correct one that best matched the data. Kalish, Zhu and Rogers (2015) designed a study where stimuli that were established to belong to one category in the supervised training phase as a result of unsupervised information switched categories and were categorized reliably as belonging to the other category after the unsupervised phase of the study. Such a change in

categorization behavior signals a much more radical update in the internal representation of the learners than a mere shift in the stimulus space. However, these results are almost trivial if one accepts that human learners try to incorporate all available incoming information during learning irrespective of their relevance to the task or the manner of presentation. Such a learning behavior is called generative learning (briefly mentioned by Gibson, Rogers, Kalish and Zhu (2015)), and it is discussed later in this chapter. In Chapter 2, I will argue that humans, indeed, adopt automatically a generative learning strategy when receiving any new information.

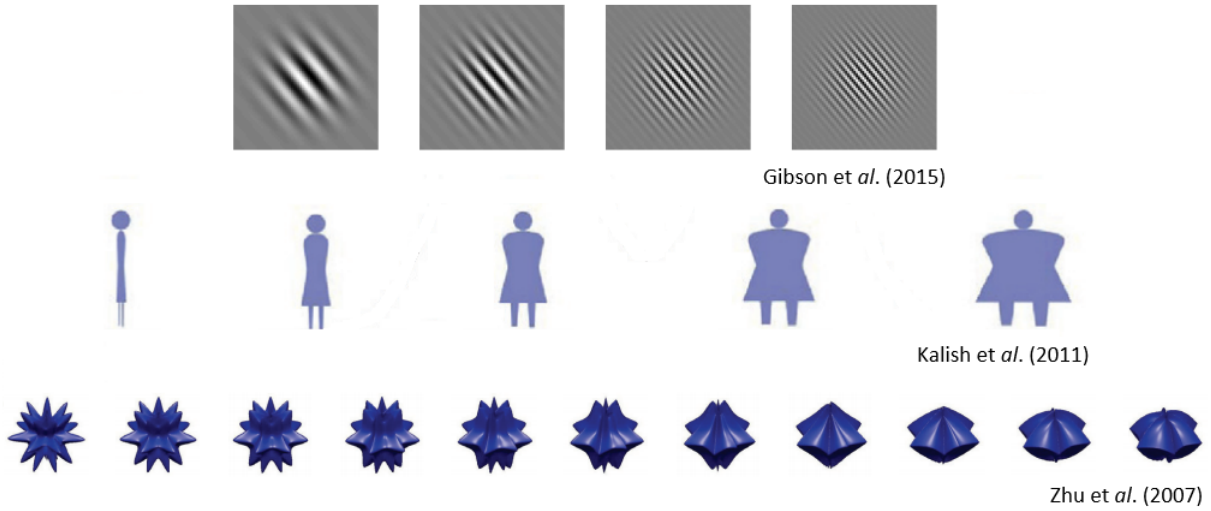


Figure 1.1: Examples of stimuli used in early SSL experiments.

The effect of supervised trials on representations obtained by previous unsupervised learning

While early studies focused on the effect of unsupervised information, Vong, Navarro and Perfors (2015) addressed the usefulness of supervised trials. Not surprisingly, they found that supervision is the most beneficial when unsupervised information is ambiguous and does not lead to an obvious solution to the categorization problem. As opposed to previous studies, they presented all the stimuli simultaneously (not sequentially) to the learners, which makes it difficult to interpret their results in the framework of previous studies, where sequential effects also

biased the emerging internal representation (Jones, Love and Maddox, 2006). Apart from the distribution of the unsupervised stimuli, the timing of supervised information might also have a strong impact on learning. Infants, for instance, were only able to learn two categories of novel creatures when supervised information preceded unsupervised ones, or if all of the stimuli were labeled (LaTourrette and Waxman, 2018).

Gibson et al. (2015) were the first to address the question whether the learners' behavior observed in previous SSL studies was, indeed, explainable only by assuming the integration of supervised and unsupervised information (i.e. SSL), or participants used much less sophisticated heuristics that could result in the same outcome. Fitting different models that assumed SSL or a range of other, much simpler heuristics to human data, they found that, indeed, humans utilized both supervised and unsupervised information when learning about categories in a SSL setup.

The superiority of SSL

After confirming humans' capacity to perform SSL, more recent studies started to focus on different aspects of the learning situation rather than on the basic question of whether humans integrate both supervised and unsupervised information. A hidden premise of these studies is that SSL is superior to both supervised and unsupervised learning. According to this reasoning, if SSL is the most natural form of human category learning, it is expected that the brain is best accommodated to handle information received in a semi-supervised fashion. Consequently, SSL should outperform (i.e. allow for a more accurate internal representation of the incoming information) both strictly unsupervised or supervised learning.

Based on pure reasoning about the unidimensional bias that exist in humans, it is feasible

that the supervised trials could lead to a superiority of SSL over purely unsupervised learning if the category to be acquired is not in line with the learner's internal unidimensional bias. In this case, supervised trials can guide the learner to identify the correct category boundary, while unsupervised trials can allow them to build a more refined representation of the stimulus distribution. However, superiority of SSL in the other condition, is less trivial: Why would unsupervised learning in SSL help to better performance over a purely supervised scenario? As mentioned in Section 1.1.1, a potential benefit of SSL over supervised learning might be a more efficient generalization of category knowledge, as supervised learning often impairs generalization. Indeed, Patterson and Kurtz (2018) conducted a study with relational categories and found that, in the SSL condition with unsupervised stimuli that were highly similar to supervised ones, categorization performance for novel stimuli from the learned categories was the best in SSL. In this sense, SSL outperformed supervised learning.

However, Vandist, De Schryver and Rosseel (2009) failed to show the superiority of SSL over supervised learning in their study. They taught participants categories defined on two dimensions (orientation and frequency of Gabor patches) that required integration of information across dimensions, and varied the number of the added supervised trials (0, 25, 50 or 100%). The results showed no difference in the 25% from the 0% (unsupervised learning) condition, nor between the 50% and the 100% (supervised learning) conditions. Corresponding learning curves and overall accuracy were statistically the same. However, 25% supervision was too little to achieve good performance on such a difficult task, while on the other hand, 50% was enough for participants to learn the categories by the end of the experiment, as it is confirmed from Exp.2. of the study. As a result, the unsupervised trials were no different from redundant filler trials. Moreover, in contrast to previous studies, in this study, supervised and unsupervised

trials were sampled from the same distribution of stimuli implying the same category boundary, so it is hard to quantify the effect of SSL.

In a different study, Vandist, Storms and Bussche (2019) found evidence of facilitated automaticity, defined by the duration of RT in trials, in the SSL condition over the supervised learning condition. They used the same stimuli as in their previous study (Vandist, De Schryver and Rosseel, 2009), and reported more substantial shortening of RTs in the semi-supervised condition compared to the supervised one.

Even though, these studies represent the first steps along an important but so far neglected line of research, they have a couple of shortcomings. First, these studies are very much scattered in terms of the phenomena they address and they do not propose a solid theoretical basis for the mechanisms of SSL. Without a strong common framework, these results have little explanatory power and represent just interesting findings rather than strong building blocks of an emerging model of human SSL. Second, as mentioned above, the two main conclusions of earlier studies are relatively self-evident for two reasons. First, the fact that SSL is the most natural form of learning strongly suggests that humans should be capable of integrating supervised and unsupervised information while learning about categories in their environment. Second, the finding that this integration of supervised and unsupervised information relies on the learner's capacity to learn about the distribution of the stimuli in an unsupervised manner is also expected. An efficient learner should try to retrieve as much information of the incoming stimulation as possible so that the knowledge gained shall be rich, flexible and applicable in other situations as well.

However, there is one important consequence of the finding of the existing SSL studies. If humans, indeed, try to learn about as many aspects of the incoming information as they can (including physical and statistical features as well) to have a rich and flexible knowledge, this requires retrieving and storing statistical and distributional information not only about the strictest discriminating features for the categories but also about auxiliary charateristics of the input. Moreover, any update of the category representation and the implied boundary between categories should also adjusts the relations of these auxiliary charateristics accordingly. Such a way of learning is reminiscent to generative learning as opposed to discriminative learning, and this provides a clearly testable hypothesis: do humans learn in a generative of discriminative manner? In order to investigate this question, first in the following section, I review these two learning methods and their relevance to human category learning.

1.2 Generative and Discriminative learning

Similarly to SSL, the characteristics and distinction between generative and discriminative learning has been widely studied in the field of machine learning (ML) (e.g Ng and Jordan (2001)), while it received much less attention in human research (Hsu and Griffiths, 2010).

There are fundamental differences between the representations the two methods allow for. Generally speaking, a discriminative learner will only focus on aspects of the input that are crucially important and relevant for solving the task at hand. A generative learner on the other hand will try to learn and represent as much of the incoming information as the system possibly can given the list of all potential tasks in the future. The two approaches can be demonstrated via a simple example. Let's say our task is to learn to learn to sort a set of images of dogs and birds into two groups (categorize them) according to species. The discriminative approach will

focus merely on the differences between the two animal groups (e.g. number of legs (2/4), coat (feather/hair), presence of beak (yes/no), position of eyes (side/front)). The generative learner will learn all available information on birds and dogs (size, color, habitat, frequency of different feature values, etc.). By the end of the learning process both learners will be able to solve the categorization task properly. Of course, asymptotically, the discriminative approach will need less training data and will be more precise in categorizing well defined classes.

However, an important additional consequence becomes evident if we take our example a step further. Suppose, we get a second task, where we need to select typical examples of the two species. Our generative model will have no problem, since it has learned everything about the two species, including the distribution of features and feature values, so it will be able to tell typical exemplars from rare or extreme ones. The discriminative model however will fail to do anything with our second task as so far it only focused on whether a feature value (e.g. feather as coat) is present or not on an image to solve the first task.

In more formal terms, our first, categorization task in the above example is to estimate to probability of a given animal (a) belonging to a species (s) by trying to learn the distribution $p(s|a)$. A generative model estimates the distribution of the animals in each species, so by the end of the learning process it will be able to approximate the probability of an animal given each of the species, $p(a|s)$. It calculates which species is more probable given an example animal ($p(s|a)$) using the Bayes' rule: $\frac{p(a|s)p(s)}{p(a)}$ (Hsu and Griffiths, 2009; Hsu and Griffiths, 2010). In contrast, the discriminative model does not learn anything about the distribution of the species. It will try to estimate the probability of a species given an exemplar of the species at hand directly. Such an estimation can be done effectively by identifying a few discriminating features that can serve as an abstract boundary between the species (Hsu and Griffiths, 2009;

Hsu and Griffiths, 2010) .

Considering the second task, since the generative model knows about the probability of certain features in each species, making decisions about typicality (high probability) will mean no problem, since in terms of probabilities, non-typical members of the species will have an extremely low probability in any of the species. Lacking knowledge of the feature distributions, the discirminative learner will need to start the learning process anew focusing now on typicality trying to define a boundary between typical and non-typical members in both species.

This explanation can be easily transformed into a description of any sort of categorization task if we change every occurrence of *animal* to *object* (x) and of *species* to *category* (C_i). In these terms, a generative model learns the feature distribution of the stimuli, whereas the discriminative learner will simply map each stimulus to a category without knowing anything about the within-category structure of the stimuli. One particular problem widely explored in perceptual experiments that favours the use of generative models is testing the influence of category variability on categorization behavior (Hsu and Griffiths, 2010; Behbahani and Faisal, 2012). Since a generative model focuses on the distribution of the categories, it will have access to statistical properties describing the categories for instance the mean and the variance of the features. In contrast, this higher-order information is lost or at least it is strongly distorted in a discriminative representation (Figure 1.2).

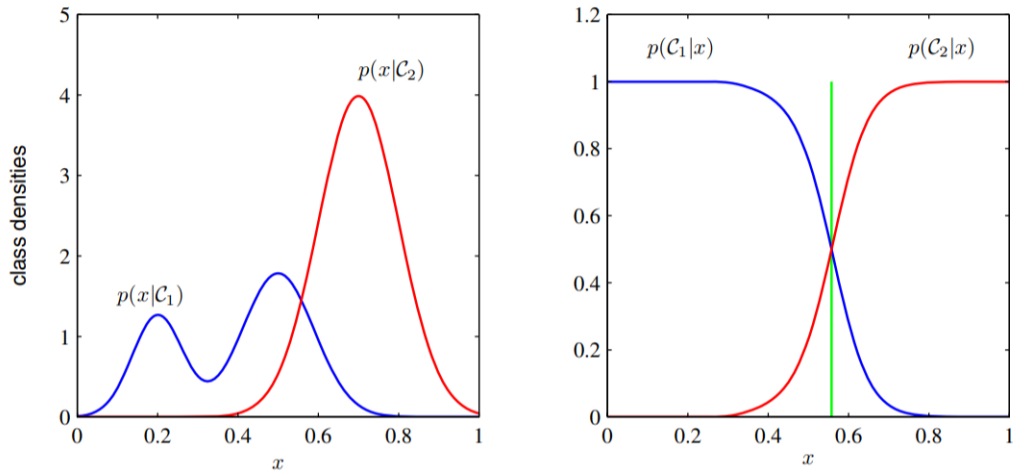


Figure 1.2: Representation of the same stimulus distributions by a generative (left) and a discriminative (right) learner (Bishop, 2006).

From a ML perspective, discriminative models are usually favored since their asymptotic error is lower than that of a generative model, and thus on the long run, a discriminative learner would outperform a generative one (Ng and Jordan, 2001). This is rational if the goal is to design a machine algorithms that is reliable and makes as few mistakes as possible at the limit. However, an important benefit of the generative learner is that initially it outperforms a discriminative learner, since at the beginning of the learning process, a generative learner utilizes initial sparse information better than a discriminative learner. Also, as mentioned before, generative methods end up with a much more flexible knowledge, and a the generative model will perform much better on novel tasks. Considering the typical domains of categorization in machine learning (e.g. images, texts, mail) and the usual conditions of learning (batch processing, data availability), it is not surprising that discriminative learners are more widespread since in these problems, learners typically face a single task, there is a large amount of data available for training, and the long-term performance of the algorithm is crucial (Lasserre, Bishop and Minka, 2006).

However, the question naturally rises whether discriminative models would be the most

appropriate to capture human behavior as well. A discriminative learner is precise and economic in terms of required memory capacity – since it only aims at representing the category boundary – while a generative model allows for a much flexible knowledge and is able to utilize well only a few initial samples. Considering the paramount importance of not requiring hundreds of images of tables and chairs before reliable discrimination between them, a generative model of human learning seems much more plausible. It is equally important that we do not start a learning process from the beginning once we face a new task requiring us to handle the already acquired categories slightly differently. Taking into account these advantages of a generative learner, it would seem logical that humans are adapted to build generative models of incoming information that will provide representation flexibly and efficient adaptation to different tasks and circumstances.

Unfortunately, there exist only a few studies addressing this question of discriminative *vs.* generative models being adequate for humans and even these studies are contradicting. In Chapter 2, I review these studies in detail and present a new study that strongly supports the idea that humans use generative learning.

1.3 The goals of the thesis

In this Introduction, I have briefly reviewed the literature of human learning along two important axes of characterization, labelled as the *un/supervised* and the *generative-discriminative* distinction. In the next three chapters, I will present three studies that furthers our understanding of human learning along these axes.

In Chapter 2, I address the question whether learners automatically adopt a generative approach to process incoming information. Are implicit categories defined by the distribution of

the data represented internally even if the task does not require one to build such categories? Do these categories also emerge if the learner is prompted to build a much simpler, discriminative representation of the data?

Since my findings in Chapter 2 firmly support the view that learners indeed, automatically build generative internal representations of incoming information, in Chapter 3, I will focus on the question *how* human SSL research integrates supervised and unsupervised data while generative learning of categories occurs. Is there a difference whether the same supervised information arrives before or after the unsupervised information? How does supervised data influence the representation built by receiving only unsupervised data, and the other way around? Will supervised information also be integrated into the final representation, or will it overwrite the internal representation built based on unsupervised information? Do learners truly update their representation of learned categories as a result of additional supervised or unsupervised information, even when this update requires them to assign samples to the opposite category at the end of the learning process?

Finally, in Chapter 4, I will investigate the cortical neural correlates of emerging categories during generative category learning. I measure the nature of the emerging neural responses typically associated to the process of categorization as learning of categories progresses. I ask whether and how these neural responses are modulated by the strength of category-membership of the stimulus or the difficulty of the task.

Chapter 2

Evidence for automatic generative learning in humans

2.1 Introduction

In section 1.2, I reviewed the topic of generative and discriminative learning in machines and humans, and argued that generative learning would be more favorable for humans because of their efficiency in early stages of learning and the plasticity of the resulting representation later under new circumstances. Unfortunately, there are only a few empirical studies directly addressing the issue of generative vs. discriminative learning in humans. This is surprising given that the existing studies represent two lines of thoughts that argue for opposing views on how humans learn. Below, I will present and critically evaluate these studies.

Next, I will present my study with a design that specifically addressed the problems I found in the previous studies, and with the results of my study, I provide further evidence in favor of generative learning in humans. Using my paradigm, I show that humans implicitly and

automatically adopt generative learning methods, even if the task is easy enough to prompt a simpler, discriminative approach.

2.1.1 Evidence for a selective use of both models

In two papers, Hsu and Griffith argued for a selective use of generative and discriminative models by humans under different conditions (Hsu and Griffiths, 2009; Hsu and Griffiths, 2010). According to their claim, it is highly context-dependent whether humans adopt one or the other learning strategy. In their 2010 paper, they found that even when the stimuli were the same in two between-subjects conditions, learners adopted either generative or discriminative strategies for learning the categories, and their choice strongly depended on how the learning scenario was framed by the cover story they received before the experiment. In one of the experimental conditions, participants were told that they needed to learn alien signs from two tribes by observing the signs shown by a representative from each alien tribe. This setup was claimed to prompt generative learning, since the task of learning the signs of two distinct tribes from intentionally generated (and possibly highly representative) samples of the categories resonates strongly with the generative concept of learning stimulus distributions separately for each category. Meanwhile, in the discriminative condition, participants were told that the stimuli would not be generated by tribe representatives, but there would be an interpreter alien telling where each sign belongs to.

Crucially, one of the non-overlapping categories had higher variability than the other. (Fig. 2.1) As explained above, such a difference in withing-category distributions will only be accounted for and represented in a generative model that has access to higher-order statistical information about the distribution of the category features. If a learner adopts a generative approach

learning about the categories, such a difference between feature variability will affect how the category boundary is defined by the learner: it will be shifted towards the category with less stimulus variability, allowing for a wider range of stimulus features to be represented in the more variable category. However, such a difference will not affect learners' representations in the discriminative condition, the boundary will be placed equal distance from both categories. (Fig. 2.1)

The results of the study confirmed the authors' prediction: in the condition, where the cover story prompted discriminative learning, participants categorized previously unseen stimuli according to a rule that used a category boundary exactly midway between the two categories. In the condition suggesting generative learning, the boundary was markedly shifted by participants towards the less variable category compared to the boundary defined in the discriminative condition.

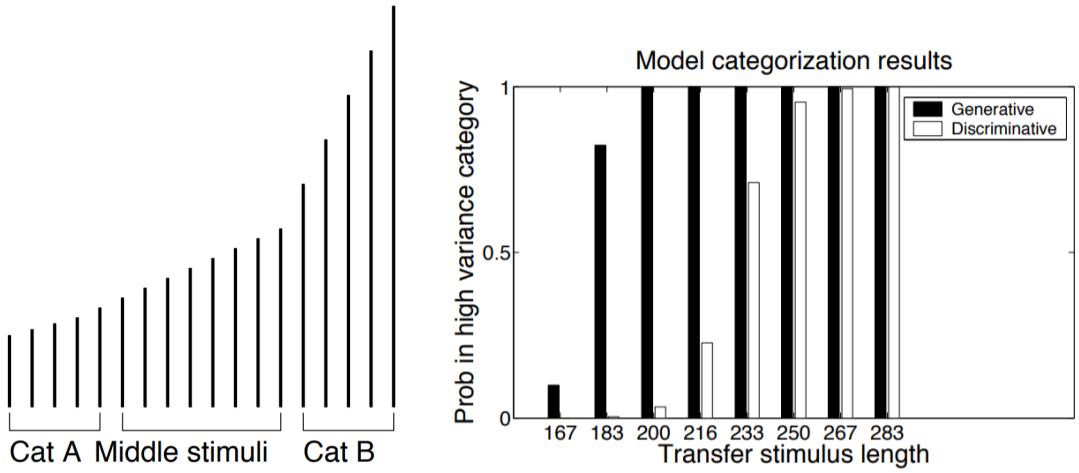


Figure 2.1: Left: Training and transfer stimuli from the experiment of Hsu & Griffiths (2010) Right: Modeled categorization pattern of the transfer stimuli in the two conditions of the same experiment.

Shortcomings of the experimental design and reasoning of the study

Design flaws

In spite of the clear setup and the seemingly straightforward interpretation of the results, the

conclusion of Hsu and Griffiths (2010) is questionable. First, it was not only the cover story that differed in the two experimental conditions. In addition, both the order of stimulus presentation and the adherent labels were slightly modified across the two setups creating a more observation learning-like scenario in the generative condition, and a more feedback-learning type scenario in the discriminative setup. These two kinds of training are known to prompt generative and discriminative learning, respectively (Levering and Kurtz, 2015).

Second, it is not entirely clear what these prompts correspond to in a natural setup or why would it be beneficial for the learner to lose information about the distribution of the data due to such a minimal change between the learning scenarios. It is also unclear whether this contextual alteration is the only type of information that makes humans switching between generative and discriminative learning, or there is a deeper and broader driving force, of which this manipulation is only one indicator.

Finally, even after following the design of the experiment very carefully, I failed to replicate the results in the original, counterbalanced or slightly modified setups. As a minimum, this indicates the very brittle nature of the reported observation.

No distinction between different representations vs. different use of the same representation

More importantly, and apart from the flaws in the design, there is a purely theoretical problem that weakens Hsu and Griffiths' argument for a selective adoption of the learning strategies. As discussed in Section 1.2, a generative model is capable of retrieving all the information contained in a discriminative model. Therefore, a generative representation, if required, can be used for solving a task while showing the hallmarks of either generative or discriminative learning. This of course does not hold the other way around: in general, a discriminative representation

cannot generate features of a generative learning. Unfortunately, in Hsu and Griffiths' experiment, there is no clear distinction between two alternative explanations of the shifting boundary result: an observer cannot be sure whether the results show the typical use of the two different representations or the representation remains the same generative one (containing information about withing-category distribution) across the two cases, but it was used differently in the two tasks.

Considering this issue, it is worth remembering that even if a discriminative model requires much less memory space for storing the acquired knowledge, it seems to be a modest gain for a disproportionately huge price of being forced to learn about the same concepts and same distributions again and again whenever we face a new task. Given this fact, and without additional strong argument about why and how discriminative learning would be more beneficial, it seems more parsimonious to suppose that humans always build generative models of the environment, and use this rich generative representation differently depending on the task at hand.

2.1.2 Evidence for generative representations under all circumstances

Similar considerations led Behbahani and Faisal (2012) to test what humans' natural approach to learning is when they are not prompted to adopt either of the learning strategies. In their experiment, they made participants learn to categorize stimuli from two categories that were defined by two non-overlapping unimodal Gaussian distributions with different means, but same variance in the feature space. The category boundary lied between the Gaussians, at an equal distance from the two category means. After participants reliably learned the categories, there was a second test round, in which participants received an enlarged set of stimuli consisting of the originals, and an additional group of outlier stimuli added to one of the two categor-

ies. This addition increased the variance of the category that included the outliers. Behbahani and Faisal (2012) found that participants' subjective category boundary shifted towards the category with the smaller variance. This happened despite the fact that the categorization task at the second test round could have been perfectly solved by the originally learned boundary, which still allowed for correctly deciding whether a category element fell to the left or the right of the category boundary, irrespective of the additional information about the summary statistics (like variance) of the data. Such a re-coding of internal representations in the light of new information is a hallmark of generative learning.

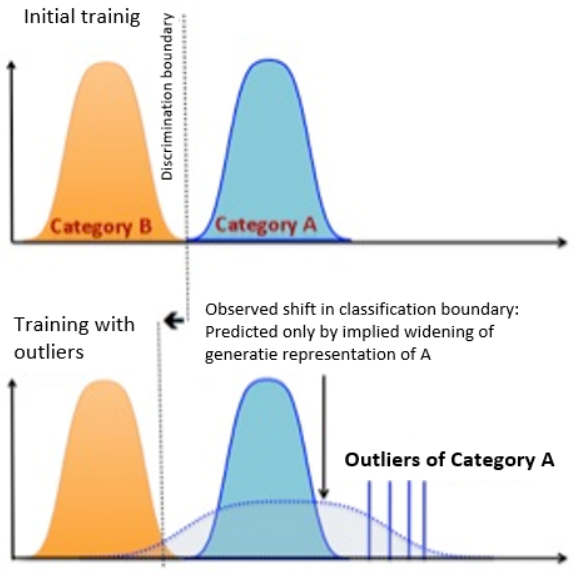


Figure 2.2: Stimulus distribution of experiments by Behbahani and Faisal (2012). The upper figure shows the training data, the lower the second round of training with outliers as well as the shifted category boundary as a result of generative learning.

Shortcomings of the experimental design of the study

Inseparability of data and task

Behbahani and Faisal (2012) replicated their results in many different setups with various stimulus sets. However, they always used stimuli with a single relevant feature dimension and

virtually no other information in the test stimuli. This raises the question whether the generative characteristic of the learning process is, indeed, because humans always learn generatively or because in this task and stimulus design, it is inevitable that the observer will learn "everything" about the stimulus set even though it tries to learn just a restricted task, the boundary between the categories. In other words, due to the extreme simplicity of both the task and the stimuli, it is impossible to disentangle learning about the task and learning about the stimuli in this study. The evidence for *automatic* generative learning would have been far more convincing if participants had learned the distribution of the stimuli even when both the feature and its distribution were perfectly irrelevant for solving the task.

In summary, based on the literature, it seems that although humans do show the hallmarks of both generative and discriminative learning in various conditions, there are no studies convincingly clarifying two fundamental issues of human learning:

- whether humans, indeed, build substantially different representations in different learning situations and not just simply use the same representation differently depending on the context
- whether humans build a (n approximately) full generative model of the environment irrespective of the number and nature of task-relevant feature dimensions.

2.2 Methods and logic of the design

In my study, I searched for evidence supporting automatic context-, task- and stimulus-independent generative learning in humans. To overcome the previously discussed design flaws, I designed

two experiments and a baseline experiment along three goals (outlined below) and tailored my stimuli, task and procedure to achieve those goals. In the following sections, I introduce the three goals and, in parallel, I discuss which aspect of the studies were designed to fulfill the given goals.

2.2.1 Automaticity

As the first goal, the study should demonstrate that the learning strategy humans adopt is automatic, so it is not prompted by the nature of the task and/or the distribution of the data. For this, the task should not rely strongly on the distribution of the tested stimulus feature, rather it should be concerned with the information readily available from another feature of the presented stimulus. This provides a method to avoid the problem in the study by Behbahani and Faisal (2012). Estimating the distribution of a particular feature (e.g. color, size, angle) and its summary statistics (e.g. mean or SD) are tasks that do not require any information from the other features or statistics, and they are readily computable at the current trial, thus there is no need for attention – not to mention representation – of the distributions of other features. In addition, obtaining a representation of summary statistics of e.g. circle ensembles are proven to be a fast, automatic and precise process, and humans are generally good at them without external information beyond what is presently available in the trial (Chong and Treisman, 2003; Chong and Treisman, 2005; Alvarez and Oliva, 2009; Alvarez, 2011). As a result, any knowledge about the distribution of the irrelevant aspect of the data is a result of an implicit, automatic learning process irrespective of the task or the task's relationship to the stimulus distribution.

Stimuli

To fulfill this requirement, I designed my stimuli and task so that automatic emergence of a generative model could be demonstrated.

In all of the experiments in this study, stimuli were ensembles of 2, 4, 5, 6, 8 or 10 circles of varying sizes (radii) that were sampled from a unimodal truncated Gaussian with mean and SD distributed uniformly in a given range (Fig. 2.3).

All experiments had two conditions depending on the task. In one condition, participants had to estimate the mean of the circle sizes in the ensemble (Mean estimation condition), in the other condition, the task was to estimate the standard deviation of the circle sizes (SD estimation condition). A previous pilot indicated that both tasks were easy to perform, and mean estimation was extremely easy for participants. Therefore, mean circle sizes were varying between 16-78 pixels for the SD estimation condition and between 32-62 pixels for the Mean-estimation conditions to make the tasks more balanced and a bit more challenging to the participants. The possible SDs of the ensembles in both conditions ranged from 0 to 13 pixels.

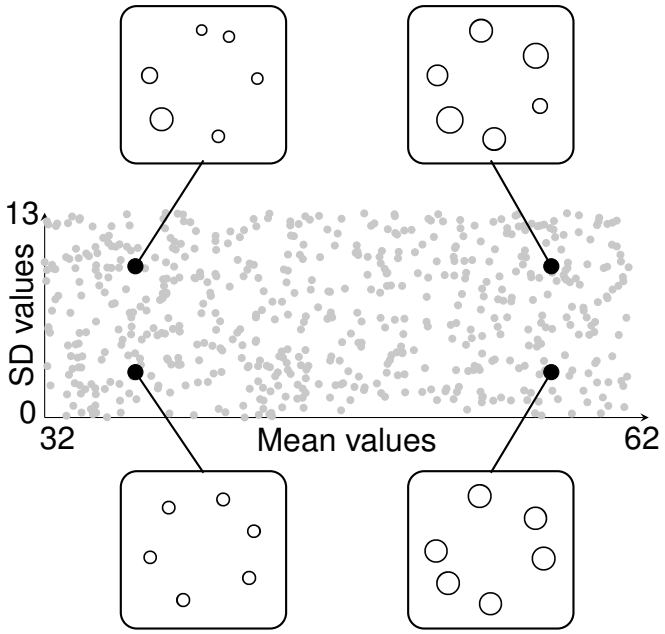


Figure 2.3: Example of the distribution of the joint summary statistics space of mean and SD values used at the Baseline experiment, and four examples of actual stimulus display sampled from different parts of the distribution.

2.2.2 Distinguishability

As the second goal, distribution of the stimuli should allow a clear discrimination between generative and discriminative representations. Changing the variance of one category in a categorization task can be a good candidate, but as argued before, when the stimuli vary along one feature dimension, it is difficult (if not impossible) to detach the representations of stimuli and task, and as a result of this, the chance to prove automaticity gets lost. The solution is to introduce a second, task-irrelevant stimulus dimension, as it only gets accounted for in a generative representation. If the task requires one to attend to only one feature dimension, a discriminative learner will ignore all other feature dimensions if they are irrelevant for solving the task. On the other hand, a generative representation contains information about more than just the most important dimension.

Data distribution

In the Baseline experiment, the distributions were the same, uniformly covering the whole parameter space (Fig. 2.4A).

To test for automaticity, first, in Experiment 1 and 2, two implicit categories were formed in the data in two different ways: either small mean value/small SD and large mean value/large SD defined the two classes, or vice versa, the classes were small mean value/large SD vs. large mean value/small SD (Fig. 2.4D,E). Note that for any one of the parameters (mean or SD), these distributions were uniform, only their joint distribution showed the category structure. Second, for measuring internal representations, I relied on a well-documented phenomenon, the automatically emerging internal biases that naturally accompany category representations. Specifically, I used the bias called *regression towards the mean*, the phenomenon when perceived features of stimuli near the category boundary are distorted in a way to be more similar to the mean of the category (de Haan and Nelson, 1998). Finally, I defined the task of the observer based only on one of the two available features (mean or SD), and I looked for behavioral biases emerging as a result of learning depending on the other feature. An emergence of such a "regression towards the mean" bias would be a hallmark of generative learning, as it would imply that learners' incorporated information about the task-irrelevant stimulus dimension, and the joint distribution of both feature dimensions allowed for the formation of categories in the internal representation of learners.

In the Mean estimation condition, categories in the data were formed in a way that ensembles with small mean also had small SD, while ensembles with large mean had large SD. It was the other way around in the case of the SD estimation condition: small mean ensembles had high SD values and large mean ensembles had small SD values. This difference in the

setup is due to a compensation of a general estimation bias being detected during the pilot studies. Because of slightly different perceptual interaction between mean and variance in the four quadrants of the stimulus space, there exist a relative bias in estimation of mean and also SD at the intermediate values of the two features. These emerging estimation biases are plotted on Fig. 2.5 **(A)** and **(B)**. In both cases, it is clear that one diagonal pairing of two quadrants for defining the two categories leads to more (steep) or less (flat) difference between the estimated values at the two sides of the boundary separating the categories (c.f. yellow and red segments). To make sure that the observers perceptually distinguish stimuli near the implicit category boundary so that different categorical biases could affect them, I chose the implicit categories by using the region-pairs that allow the most distinguishability between implicit categories (red for Means and yellow for SDs). This is why the implicit structures of stimuli in the Mean and SD conditions in this study were different. Notice that this difference does not interfere with the general logic of the tests, since in both conditions, performance is compared to the baseline performance with uniform distribution of data points to show *some* difference, and not to each other to show *relative* differences.

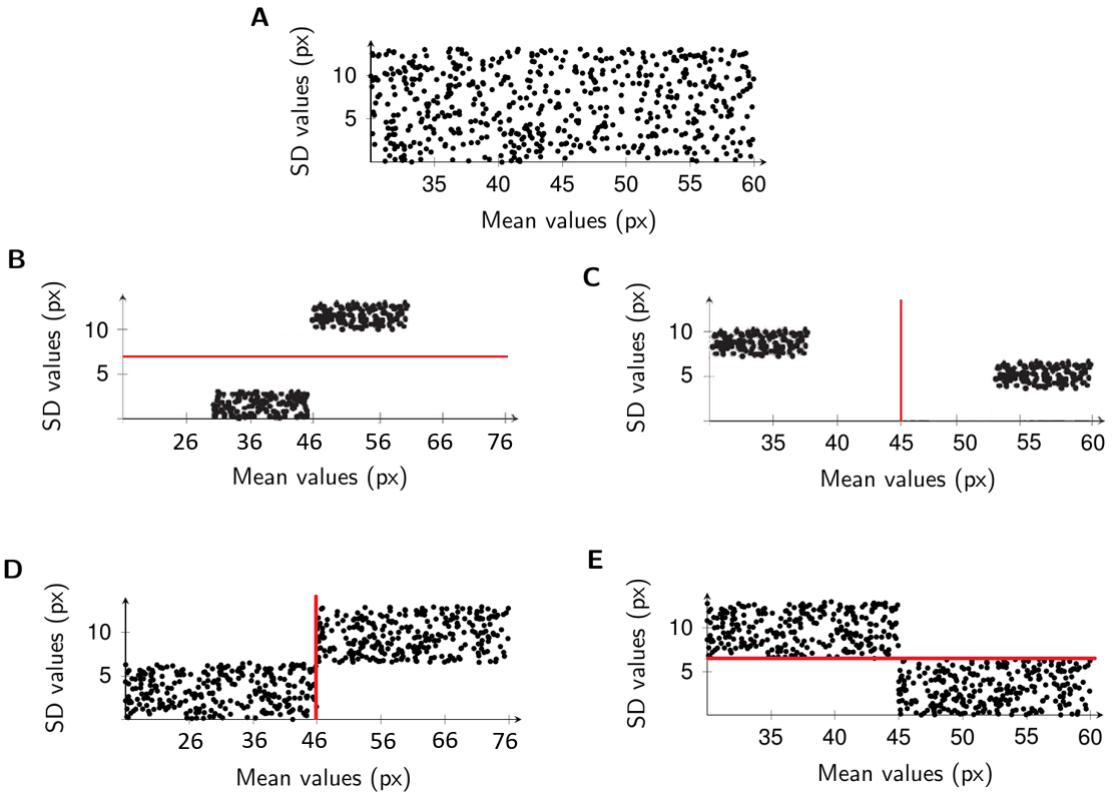


Figure 2.4: Stimulus distributions in the experiments. **A)** Baseline, **B)** Categorization based on SD before Mean estimation, **C)** Categorization based on mean before SD estimation, **D)** Mean estimation, **E)** SD estimation

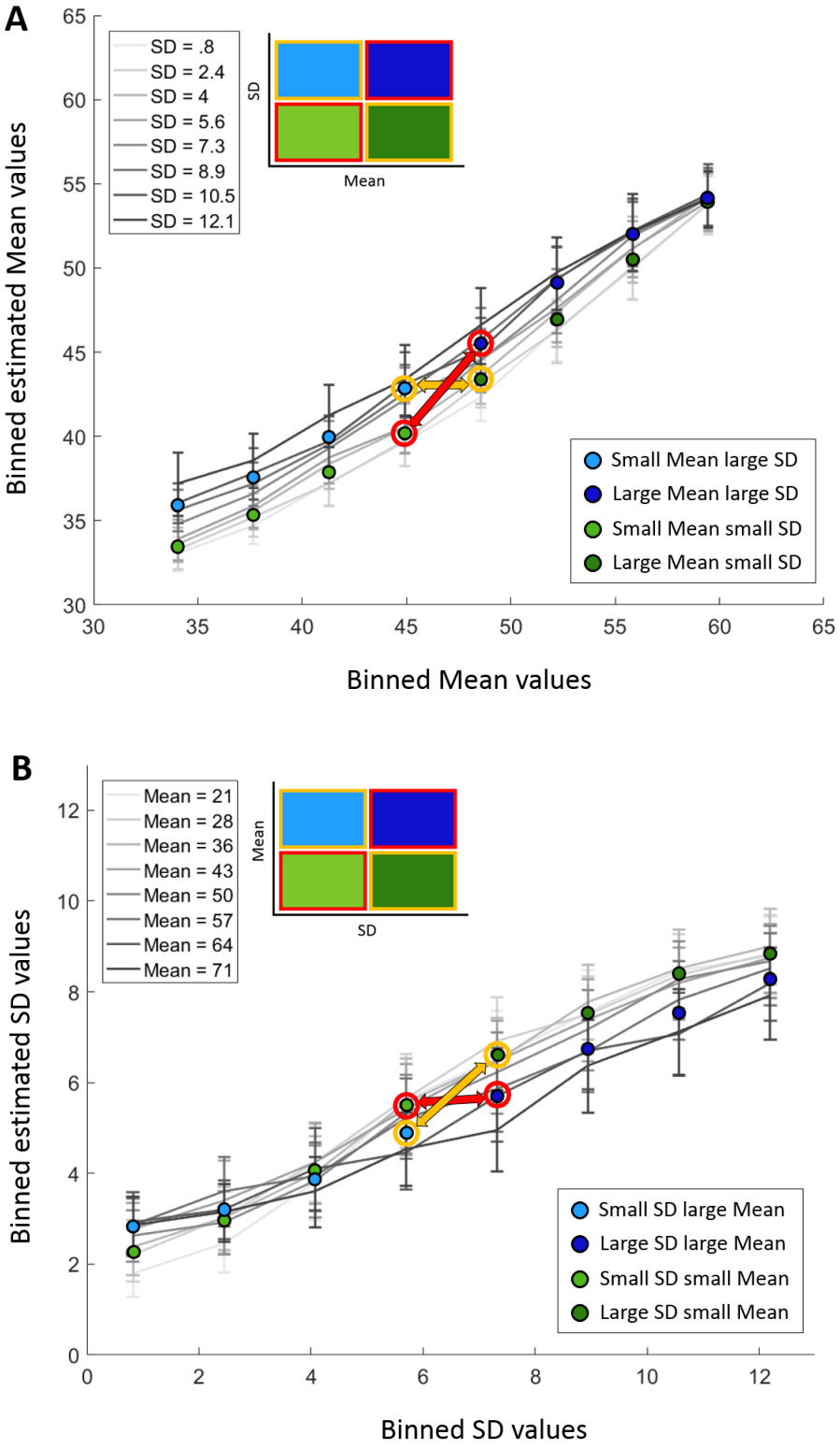


Figure 2.5: Task-relevant parameter estimation biases as a function of task-irrelevant parameter values. Plots show the data gathered in the Baseline experiment with all possible parameter combinations. Line plots follow mean parameter estimates of **A)** the Mean of circle sizes as a function of the task-irrelevant (SD) parameter values or **B)** the SD of circle sizes as a function of the task-irrelevant (Mean) parameter values, with SEM as error bars. Colored circles correspond to the binned mean values of task-relevant parameter estimates grouped by the higher- or lower-than-mean parameter values of the task-irrelevant parameter. Category boundaries suggested by the data are at Mean = 47, and SD = 6.5, for Mean and SD estimation conditions, respectively.

2.2.3 Context independence

The third goal of the project was to show that learners build a generative representation not only during an involved estimation task, but even if the task/context is easy enough for them to completely ignore the task-irrelevant aspects of the distribution. To test this, not only two stimulus dimensions but also two separate tasks were needed. In Task 1, one of the stimulus dimensions would be task-relevant, the other task-irrelevant. Task 1 also had to be extremely easy to prompt the participants for a discriminative approach so that during the discriminative learning, the observer would surely ignore the task-irrelevant dimension. In Task 2 then, the same learner is asked to operate on the until-then task-irrelevant dimension in the same estimation task that had showed the generative bias effect after substantial practice. If the learner built a discriminative representation for Task 1, the corresponding representation should in no way bias their performance in Task 2, and the generative bias should emerge in the same extended period of time as in the condition with no preceding categorization task. However, if the observer performs generative learning even during the easy categorization in Task 1, the learner's acquired knowledge about the joint distribution of the two stimulus/feature dimensions in Task 1 should bias their behavior from the very first trials of Task 2 regardless of the fact that the relevant dimensions in the two tasks are different.

Additional task

In Experiment 2, prior to the parameter estimation task, I inserted an extremely easy categorization task involving the stimulus parameter that later became irrelevant in the estimation task. For example, participants first had to categorize stimuli based on their SD, and then in the second part of the experiment, they had to estimate the means of circle ensembles. Were the

participants showing the same behavioral biases right at the beginning of the Task 2 estimation task (discussed in Section 2.2.2), it would be a sign of generative learning even in the context of a task that would require a much simpler, discriminative representation.

For Task 1, the easy categorization task in Exp. 2, stimuli were generated by sampling the task-relevant (later, at the estimation task, task-irrelevant) parameter values from the extremes of the range, while the task-irrelevant (later relevant) parameter values were sampled near the implicit category boundary (Fig. 2.4B,C). This ensured that Task 1 was easy enough so that participants felt no need to deliberately pay attention to task-irrelevant parameters in search for extra cues that could make the task easier. Meanwhile, sampling the task-irrelevant parameter near the implicit category boundary prevented participants from calling attention to the distribution of the initially task-irrelevant dimension.

2.2.4 Participants, Stimuli, Design and Procedure

Participants

207 participants (mean age 22.3, 127 women) gave informed consent and participated in the experiment. The experimental protocols were approved by the Ethics Committee for Hungarian Psychological Research.

Design

The experiment was created and presented with MATLAB 2014a on an iMac 27" (2560*1440) using Psychophysics Matlab toolbox, conducted in a dimly lit room.

All three experiments had two possible conditions that were defined by the parameter that

the participants had to estimate: *mean* or *SD* of circle sizes. The procedures were the same for both conditions with a slight difference in the structure of the practice trials described below.

Practice

In all experiments, a practice phase preceded the test where participants got familiarized with the response method. Possible mean and SD values were mapped onto a semicircle (Fig. 2.6), and participants gave their estimates by drawing a line on a touchpad with a pen. The angle of their line indicated the estimated value and the length of the line corresponded to their subjective uncertainty about their estimate. This instantaneous and joint response about the value and the uncertainty of the observer is a preferable method over a successive and highly cognitive estimate of inner uncertainty.

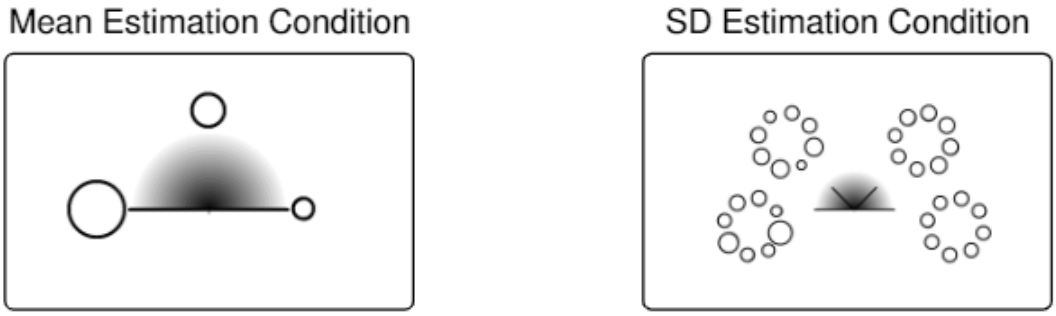


Figure 2.6: Mapping of circle sizes and SD values onto the response semicircle. These images demonstrating the mapping of stimulus parameter values onto angles of the semicircle were presented to participants while they were explained the task. The circles are only for demonstration purposes, they were not presented in the display.

In each of the 100 practice trials, participants saw a fixation cross for 500ms, then a circle ensemble appeared and remained on screen until the end of the trial. 500 ms after the appearance of the ensemble, a response semicircle was added in the middle of the ensemble where participants could provide their estimates. During practice of the mapping from their perception onto the response method of the experiment, corrective feedback was provided after every

trial. For both mean and SD estimate, the participant's response remained visibly superimposed on the repsonse semicircle. In addition, in the mean estimation condition, the feedback was a red circle appearing below the semicircle for 1500ms indicating the size of the circle the subject's response line's angle corresponded to. Based on extensive literature, the true mean size value was not provided as observers are known to have a very precise notion about the correct mean size of an ensemble of circles. In the SD estimation condition, additional corrective feedback was given in the form of a black line superimposed on the semicircle showing the correct response next to the observers answer (see Fig. 2.7). To enhance the precision of subjective uncertainty estimation, I implemented a modified version of the scoring function used by Lengyel, Koblinger, Popovic and Fiser (2015). As confirmed by the observers' behavior, the main function of the practice session was to develop a good mapping from the observers' internal estimate onto the experiment's response metric, and not to "learn" the proper interpretation of the mean and SD information.

Test

Baseline and Experiment 1

Test phase trials had the same procedure (Fig. 2.7 Estimation) both in the Baseline and in Experiment 1. Only stimulus distributions differed: it was uniform in Baseline (Fig. 2.4A) and formed implicit categories in Experiment 1 (Fig. 2.4 D,E). Test trials started with the presentation of a fixation cross for 500ms, then the stimulus was displayed with one of 9 possible presentation times (50, 75, 100, 133, 167, 200, 300, 400 or 600ms) followed by a mask for another 500ms. The mask was replaced by the response semicircle until response. Participants did not get corrective feedback, only the line they drew appeared in the semicircle for 600ms.

Scores were only presented after every 10 trials to maintain the participants' attention. There were 576 test trials with short breaks after every 100 trials.

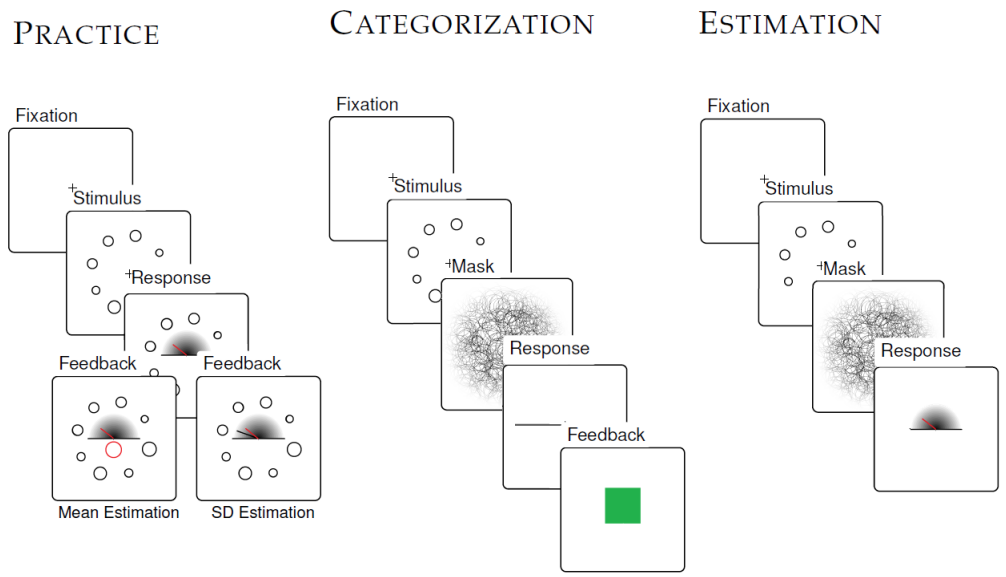


Figure 2.7: Procedures of the experiments. Practice and Estimation is applicable in all experiments, Categorization was only a part of Experiment 2.

Experiment 2

Following the requirement outlined in Section 2.2.3, I introduced a very easy categorization task (Fig. 2.7 Categorization) before the conducting same estimation task as in Experiment 1 (Fig. 2.7 Estimation). To reiterate, the categorization task was performed on the stimulus feature dimension that was irrelevant during the estimation task. For instance, categorization was performed based on the mean sizes of the circle ensembles – irrespective of the SD – , then participants needed to estimate the SDs during the estimation task, for which the mean values of the ensembles were irrelevant. There were three important design characteristics of the categorization task. First, the categorization was performed based on samples that were at the extreme of the distribution of the category feature dimension ensuring an easy categorization task (Fig. 2.4B,C). Second, the samples were near to the orthogonal boundary of the other

feature that was irrelavant for categorization, but became the implicit category boundary for the second estimation task. This ensured that the correlation between different feature dimensions would not be strong during the categorization task. Third, the categorization was performed in the feature dimension that was task-irrelevant in the estimation task. (Fig. 2.4B and D vs C and E) This reduced further the possibility that participants transferred their knowledge about the categories to the feature dimension relevant to the second task.

In the categorization task, trials started with a fixation cross (1000ms), then a circle ensemble appeared for one of the 9 possible presentation times, followed by the same mask as in the estimation task. Participants had to categorize the ensembles based on the simple rule whether the task-relevant parameter of the ensembles was high or low. For instance participants in the mean estimation condition had to categorize ensembles in the first task based on whether the SD of the circles was small or rather it was large. They gave their category decision by drawing a horizontal line pointing with its endpoint either to the left or to the right compared to the starting point. On the first 80 of the 280 categorization trials a green square appeared for 300ms if participant's response was correct and a red square indicated incorrect responses.

2.3 Results

2.3.1 Response method check and exclusion criteria

Given the novelty of the paradigm and the response method, I extensively validated them to make sure that the participants understood the task and learned the response method properly. I ran a baseline experiment with uniform distribution of feature values (Fig. 2.4A), binned

possible parameter values, calculated the mean estimates for the binned datapoints, and plotted the estimates against the true values. Apart from a Weber law-like deviation from the veridical values, participants reliably learned the response method and showed a consistent link between the true values of mean and SD values and their estimates (Fig. 2.8).

The few participants who failed to learn the response method or were unable to consistently solve the task were excluded from the analysis. I calculated Pearson's correlation between estimated and true parameter values throughout the experiment for each participant, and excluded those, who did not reach a correlation of at least $r = .22$. They were all together 14 such participants out of the 207.

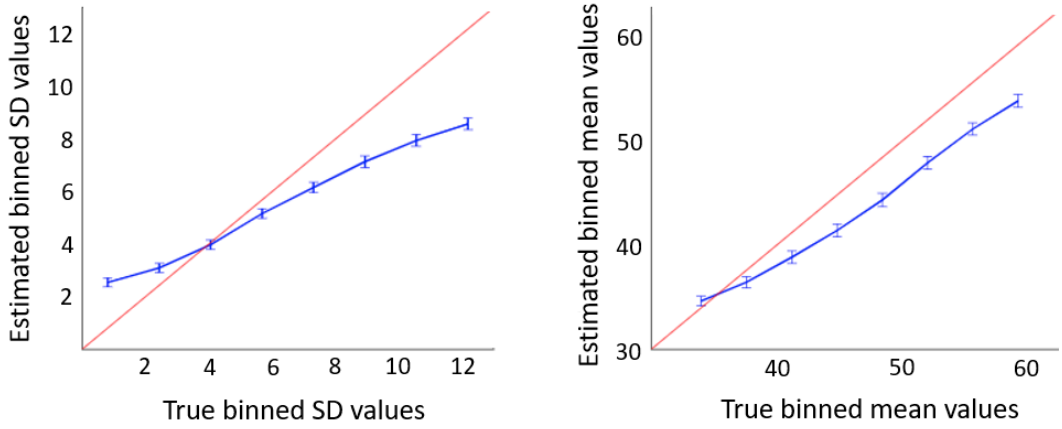


Figure 2.8: Mean estimated parameter values for SD and mean estimates (with SEM as error bar) plotted against true binned parameter values in the Baseline experiment. The red diagonal line indicates veridical correspondence.

2.3.2 Logic of data analysis

My goal in the present study was to demonstrate that humans automatically, implicitly build generative models of their environment even if there is no specific task requiring that or even when solving a task that strongly prompts building a way easier, discriminative model of the data.

If participants indeed build generative representation of the data, in Experiment 1 and 2 they should be represented as two categories, and behavioral biases should emerge that are typical of categorical perception. Such a typical bias is the warping of the similarity space by reducing within-category and increasing between-category perceptual similarity, that specifically affects stimulus regions near the category boundary (Harnad, 2003; Harnad, 2005; Pevtzow and Harnad, 1997). This generates the regression towards the mean effect, when perceived features of stimuli near the category boundary are distorted in a way to be more similar to the mean of the category (de Haan and Nelson, 1998).

Since humans are proved to be extremely quick and precise in estimating summary statistics (Chong and Treisman, 2003; Chong and Treisman, 2005), in the analysis, I focused on trials where the presentation times were short (50, 75, 100 ms). In these cases, sensory information is more noisy, and the observers have to rely more on their internal representation of the stimuli to construct the most likely percept of the stimulus. If participants structured the stimuli into multi-dimensional categories in their internal representation due to generative learning, this representation will serve as a prior when incoming information is noisy or uncertain. Therefore, at short presentation times, perceptual biases such as regression to the mean should emerge.

To enhance the effect of the prior even more, I only analyzed trials where the parameter value in the task-irrelevant implicit dimension during estimation had an extreme value, hence they were clearly diagnostic of the category (given there were indeed implicit category representations), while in the task relevant category the trial was close to the boundary, so that the regression to the mean effect would point to the same direction be. Such data points were compared to the trials with the same parameters sampled from the Baseline experiment. (Fig. 2.9) Since there were two implicit categories in Experiment 1 and 2, one that had smaller than aver-

age feature (mean or SD) values and one that had larger than average feature values, I will refer to these two categories as *Small* and *Large* categories, respectively.

In Small and Large categories, the regression to the mean bias predicts an opposing deviation from the Baseline trials. In the case of the Small category, features should be underestimated, while Large category features should be overestimated. Since trials with exactly the same parameters were compared in the two experiments, any deviation from the Baseline should be caused by a difference in the representation of stimulus distributions.

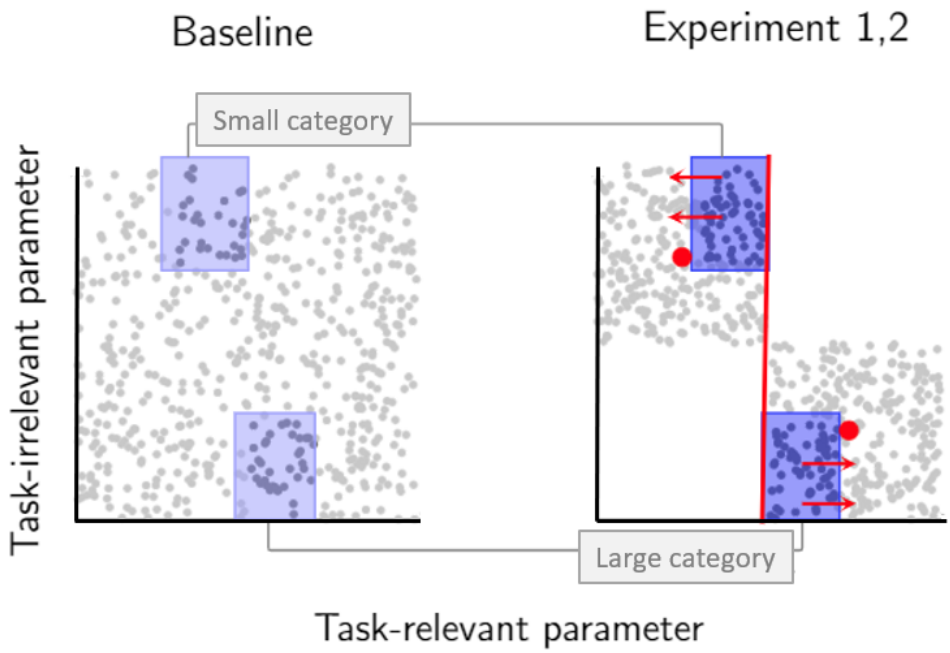


Figure 2.9: Data points sampled for analysis in the Baseline and Experiment 1 and 2. Red circles show the mean of the categories, while red arrows indicate the direction of the expected bias. Pairwise comparisons were conducted between respective groups of data points to reveal deviations in Ex. 1 and 2 from Baseline.

2.3.3 Experiment 1

Experiment I tested whether participants build a generative representation of the environment even if they do not benefit from such a representation while solving the task at hand. To test this hypothesis, I ran Experiment 1, which was virtually the same as the Baseline experiment except

for the distribution of the data in the test phase (Fig. 2.4D,E). Following the logic of analysis described above (Fig. 2.9), I compared data points in Experiment 1 to the corresponding data points in the Baseline experiment. Importantly, I analyzed the data from the first and the second halves of the experiment separately. A significant difference already present at the beginning of the experiment would suggest an artifact in the experimental design rather than a specific effect of a newly formed internal representation. However, if a significant deviation in estimation biases from the baseline emerges only in the second half of the experiment, this would suggest that participants indeed built a generative representation despite its apparent irrelevance in the task.

A two-tailed independent samples t-test applied to the second halves of the two experiments revealed a significant difference both for the Small and Large categories and in the SD estimation [$t_{LargeCategory}(67)=2.55$, $p < .05$, $d =.59$, $t_{SmallCategory}(67)=-3.65$, $p < .001$, $d =-.81$] and Mean estimation conditions alike [$t_{LargeCategory}(55)=2.51$, $p < .05$, $d =.64$, $t_{SmallCategory}(55)=-2.27$, $p < .05$, $d =-.58$]. (Fig. 2.10) This was not true for the comparisons of the first halves of the experiment, where there was no significant difference either in the Mean [$t_{LargeCategory}(55)=-1.82$, $p = .07$, $d =-.47$, $t_{SmallCategory}(55)=-1.13$, $p < .26$, $d =-.3$] or the SD [$t_{LargeCategory}(67)=-1.58$, $p = .11$, $d =-.38$, $t_{SmallCategory}(67)=.09$, $p = .92$, $d =-.37$] conditions. (Fig. 2.10)

In Experiment 1 (as well as in Experiment 2), participants were asked after completing the tasks whether they have noticed any regularities in the data. None of them reported that they would have realized the presence of the two categories, not even after this information was explicitly revealed to them.

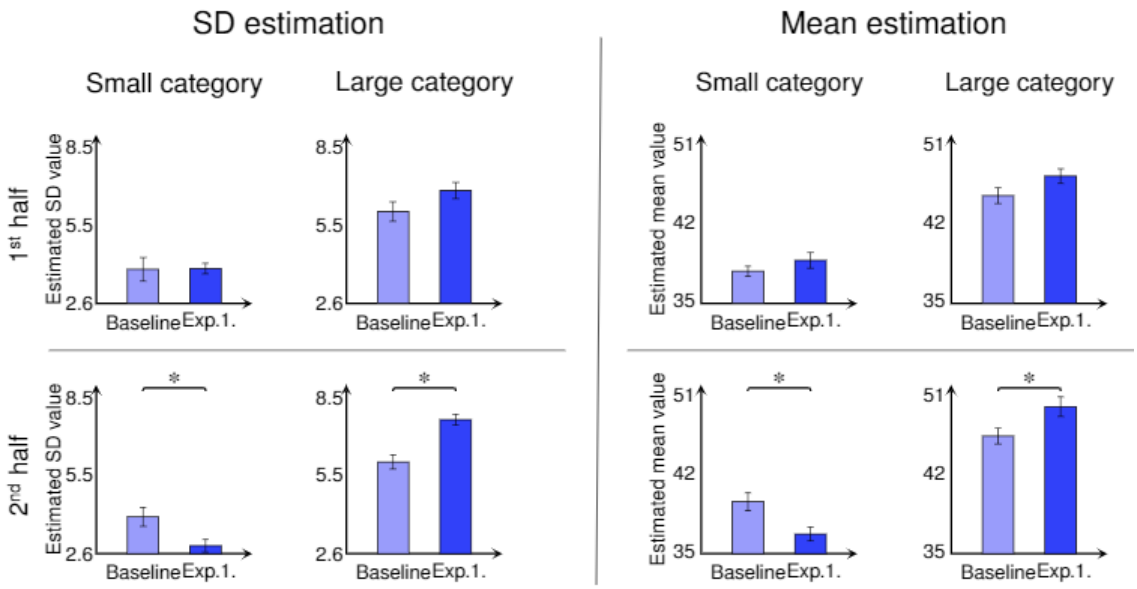


Figure 2.10: Results of Experiment 1. Figures depict the result of an independent samples t-test between respective stimuli (see Fig. 2.9) of the Baseline and Experiment 1 for SD- (left) and Mean estimation conditions (right) as well. Comparative analysis was conducted for the first (upper row) and second halves (bottom row) of the experiment, as well as Small and Large categories, separately. Bar heights correspond to the mean estimated parameter values with SEM as error bars.

2.3.4 Experiment 2

In Experiment 2, participants completed a categorization task before the estimation test and, as expected, they performed very well. In the Mean estimation condition, when the categorization task was based on the SD of sizes of the circle ensembles, average performance was 89.5% (± 10% SD). Performance was slightly better in the other version, when the categorize task was based on the mean circle sizes, yielding an average of 93% (± 13% SD) correct trials. As both versions of the categorization task were very easy, indeed, participants were not forced to search for extra information in the distribution that could help their decision. This supports the assumption that participates likely ignored every other aspect of the stimuli not directly relevant to the task.

The main question of this experiment was whether the regression to the mean bias observed

in the second, but not first half of Experiment 1 would surface already in the first half of Experiment 2 under these conditions. Such an early presence of the expected bias would argue for a generative learning during the categorization task despite of the fact that such a categorization task would require a highly discriminative approach to the problem.

I found a significant deviation from the Baseline already in the first part of Experiment 2 in both the mean [$t_{LargeCategory}(50)=2.05$, $p < .05$, $d = .56$, $t_{SmallCategory}(50)=-2.38$, $p < .005$, $d = -.64$] and the SD estimation [$t_{LargeCategory}(78)=2.67$, $p < .05$, $d = .58$, $t_{SmallCategory}(78)=-2.93$, $p < .005$, $d = -.63$] conditions. These results show that participants indeed formed the expected implicit generative representation of the data during the categorization task.

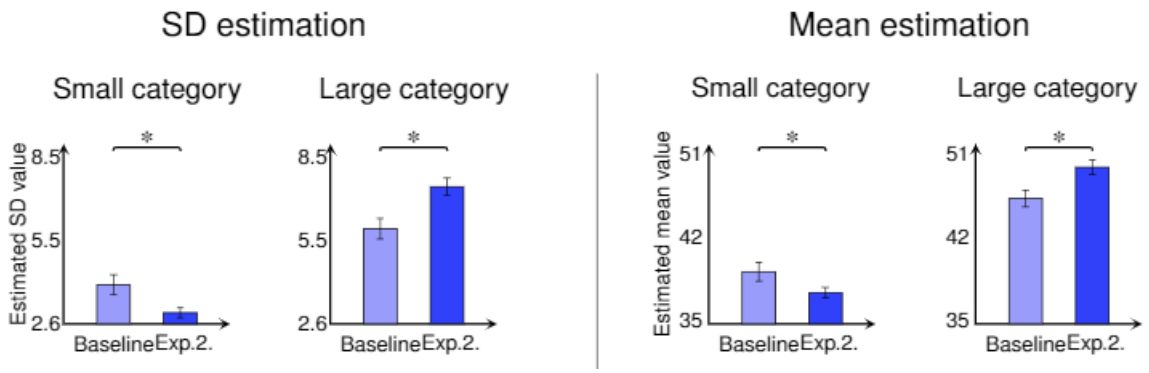


Figure 2.11: Results of Experiment 2. Figures depict the result of an independent samples t-test between respective stimuli (see Fig. 2.9) of the Baseline and Experiment 2. Analysis was conducted for SD- (left) and Mean estimation conditions, as well as Small and Large categories, separately. Bar heights correspond to the mean estimated parameter values with SEM as error bars.

In order to ensure the reliability of my interpretation that the found significant deviations can only be attributed to the differences in the acquired internal representations about the stimuli, I calculated absolute changes in parameter estimates between the 1^{st} and 2^{nd} halves of the Baseline and Experiment 1 and 2. Significant changes between the two halves in the Baseline and Experiment 2 would imply the presence of potential biases or artifacts influencing estimation performance I failed to control for. In addition, as previous results would not predict

differences in the magnitude of the found regression to the mean bias between Large and Small categories, I also wished to verify that indeed, the effect is symmetric, there are no consistent further biases towards either direction.

A two-way ANOVA with *experiment* (Baseline, Experiment 1 and 2) and *category* type (Small, Large) as predictors revealed a significant main effect of *experiment* type for both SD [$F(2, 222) = 4.39, p < .05, \eta^2 = .03$] and Mean estimation [$F(2, 151) = 5.58, p < .01, \eta^2 = .06$] conditions. I found, however, no significant interaction between predictors, or significant main effect of *category* type in either conditions [SD: $F(1, 222) = .87, p = .35, \eta^2 = .003$], [Mean: $F(1, 151) = .95, p = .33, \eta^2 = .006$]. Post-hoc paired-samples t-tests on parameter estimates between the first and second halves of each experiment only indicated significant changes in Experiment 1 in both SD [$t(32) = 2.56, p < .05, d = .48$] and Mean estimation [$t(26) = 2.53, p < .05, d = .49$] conditions, but not in the Baseline (Mean: $p = .2$, SD: $p = .86$) or in Experiment 2 (Mean: $p = .78$, SD: $p = 36$).

2.4 Discussion

Whether we are building models of the environment for action, designing experiments or trying to interpret cognitive phenomena, we need to rely on our knowledge of internal representations (Goldstone, Lippaa and Shiffrin, 2001; Gregory, 1997; Mitchell et al., 2005). Therefore, uncovering the nature of representations is of key importance for understanding human cognition.

There are multitude of approaches to investigate representations, but most of them addresses the issue of how much of the externally available information is stored. As it was discussed in section 1.2 generative learning allows for a much richer representation than discriminative learning, which is flexible enough to solve a greater variety of tasks. It would seem obvious to

suppose that humans naturally build adaptable, generative representations and they refine those representations when they face novel tasks.

Yet, only a handful contradictory results exist addressing this issue in human research. Some argue that the nature of learning depends on the learning context which strategy is adopted (Hsu and Griffiths, 2010), while others support the idea of exclusivity of generative learning (Behbahani and Faisal, 2012). Since even the latter type of papers provide questionable evidence as a result of flaws in the experimental designs (see sections 2.1.1 and 2.1.2), I created a new paradigm that aimed to correct these flaws and unequivocally support the theory of context-independent, automatic, implicit generative learning in humans.

In two experiments, I have shown that participants indeed built implicit generative representations of incoming data even if the distribution of the stimuli was completely irrelevant for solving the task, or even if the task strongly prompts a much simpler discriminative representation.

Automatic implicit learning was demonstrated in Experiment 1 by the appearance of certain behavioral biases typical of the presence of categories in the representation as opposed to a Baseline experiment, where there were no such categories formed in the data distribution. Crucially, the measured bias appeared gradually by the end of the experiment, excluding the possibility that the measured bias is inherent to the data and not the learning process or the resulting representation.

Experiment 2 provided further evidence for the fundamental role of generative learning by showing that even in the presence of a task interfering with the overall processing of the incoming input, the emerging representation preserves its generative character. Before assessing the magnitude of the same bias as in Experiment 1, participants were asked to complete an

extremely easy categorization task that typically prompts discriminative learning, while the distribution of the data would still allow for building a similar categorical representation as in Experiment 1. Perceptual biases indicating category formations based on the full structure of the input still emerged. Critically, the behavioral bias that was previously measured in the second half of Experiment 1 was now already present in the first part of the same estimation task.

The lack of significant modulation of estimation biases by time (1^{st} vs 2^{nd} halves of parameter estimation task) or the type of category in the Baseline and Experiment 2 strengthen the current interpretation that attributes the observed regression to the mean bias solely to the structured, generative internal representation of categories defined by the distribution of stimuli in the parameter space.

Thus, we can conclude that even if a task is easy enough so that a simple, discriminative representation suffice for solving it perfectly, humans build a generative internal model of the data that implicitly influence their behavior in any subsequent task.

Chapter 3

Integration of supervised and unsupervised information in SSL

3.1 Introduction

Apart from being a more suitable type of learning than discriminative learning, generative learning is also necessary for SSL. As discussed by Kalish, Zhu and Rogers (2015) and Gibson et al. (2015), processing and integration of supervised and unsupervised information requires a generative approach, i.e. assuming the same underlying generative model that provides samples for both types of learning.

Indeed, available evidence in the SSL literature supports the hypothesis that humans take a generative approach to integrating supervised and unsupervised information when learning about categories. However, a significant amount of the literature fail to provide strong evidence in favor of true SSL, and as a consequence, the generative approach to processing supervised and unsupervised data. As McDonnell, Jew and Gureckis (2012) pointed out (see Section 1.1.3),

a slight shift in the category boundary suggested by supervised vs. unsupervised samples does not require a radical update in the representation, which would be necessary for a strong support in favor of SSL. Lacking such evidence, the generative approach of semi-supervised learning is also unsupported. Kalish, Zhu and Rogers (2015) created a design to provide stronger evidence in favor of SSL by making learners (children) re-categorize certain elements throughout the learning process to the opposite category as a result of the integration of supervised and unsupervised information. Their results imply a developmental shift in children's learning behavior. Older children (around the age of 7-8 years as opposed to approximately 4-5 years) tended to weight supervised information much more, and practically ignore the distribution of unsupervised stimuli. Meanwhile, younger children relied much more on the natural distribution of the stimuli, which resulted in the change of category representation sought by researchers of SSL. This dichotomy raises the question whether their results could be interpreted as a general evidence for the radical update of the internal representation McDonnell, Jew and Gureckis (2012) called for.

A comprehensive assessment of whether SSL operates as generative learning requires the clarification of two questions. First, whether the latter phase of SSL results in a true update of the category representation suggested by initial learning. In other words, one needs to show that, after forming some categories in the initial phase of SSL, the additional new information in the second part of SSL makes the learner re-categorize elements of the acquired categories in a way that it is consistent with the representation gained by combining 'old' and 'new' information. Second, it needs to be shown that supervised information also gets integrated into the representation built by unsupervised information instead of simply overwriting it. There are only three studies in the literature with a design that presents learners with the supervised samples

of SSL after, interleaved, or simultaneously with unsupervised samples, and not strictly before them. Since the stimuli, addressed population and form of presentation in these three studies are all different, it is almost impossible to draw a clear conclusion from them with respect to the questions above.

Based on these studies, the following is known about the impact of supervised information on the representation suggested by unsupervised trials in categorization tasks:

- if the provided unsupervised information lack a clear structure, subsequent supervised information can help the learner orienting their attention to relevant features of the stimuli and creating categories that matches the representation offered by supervised information. However, if the distribution of the unsupervised stimuli is structured enough so that the learner can form clear clusters, subsequent supervised information not contradicting the natural unsupervised clustering of the data will not make much of a difference in the resulting representation in adult learners (Vong, Navarro and Perfors, 2015).

- infants do not benefit from supervised information when learning about categories if the supervised segment follows the segment with unsupervised samples (LaTourrette and Waxman, 2018).

- younger children prefer to create categories along the natural distribution of unsupervised samples, and their representation is not influenced strongly by supervision. In contrast, older children are more inclined to rely completely on the strong, salient supervised information, and this prevents them from incorporating unsupervised data, when supervision is frequently interleaved with unsupervised information (Kalish, Zhu and Rogers, 2015).

In the rest of this chapter, I will present a study that aims at clarifying the above mentioned two missing details in the research of SSL by answering the following three questions:

1. Do learners truly update their representation of learned categories even when this update requires them to categorize particular samples differently at the end of the learning process compared to the beginning?

2. Will supervised information overwrite the initial representation built by unsupervised data or will it be integrated into a joint representation by modifying it only to an extent that will still be compatible with the initial representation?

3. Does the order of the presentation of supervised and unsupervised information matter, or will learners arrive to the same representation irrespective of whether they received supervised or unsupervised samples first?

3.2 Methods

3.2.1 Participants

135 subjects [82 females, mean age = 25 years] gave written informed consent and completed the experiment. 24 subjects were excluded for responding randomly at either of the two test phases, and an additional 25 subjects' data were not analysed as they failed to perform on the supervised training above 65% accuracy by the end of the training phase.

3.2.2 Stimuli

After extensive investigation, I decided to use the stimuli originally created by Op de Beeck, Wagemans and Vogels (2001) for the following four reasons.

First, based on the critique phrased by McDonnell, Jew and Gureckis (2012) of previous studies using primarily one dimensional stimuli, I looked for stimuli varying on multiple feature dimensions. This provides a much richer set of options to select diagnostic features for learning, and thus the category distributions suggested by supervised and unsupervised information could be clearly and obviously distinguishable.

Second, since my goal was to investigate how learners integrate supervised and unsupervised information, especially when the two learning regiments suggest markedly different category structures, I needed a stimulus space where category boundaries defined along different feature axes (or a combination of these) are almost equally easy to learn. Since diagonal boundaries (II learning) are proved to be more difficult to learn with separable dimension stimuli, I chose to work with integral dimension categories where the categorization rule is harder to verbalize, and categorization strategies requiring the integration of different feature dimensions (a.k.a. a diagonal boundary) are easier to adopt (Handel and Imai, 1972; Nosofsky and Palmeri, 1996).

Third, I needed stimuli continuously varying on all of the relevant feature dimensions so that the learner's representation, or category boundary could be reliably assessed and retrieved based on their categorization behavior in a dense stimulus space.

Finally, since I aimed at avoiding possible artifacts resulting from diverse priors the learners might have, I needed a stimulus set of novel shapes that the learners would not be familiar with.

Based on these aims and after an extensive piloting, an appropriate subspace of the Op de

Beeck et al. stimulus space was identified, and 100 novel stimuli (360x250 pixels in size) were generated and ordered in a two dimensional matrix as presented in Fig.3.1.

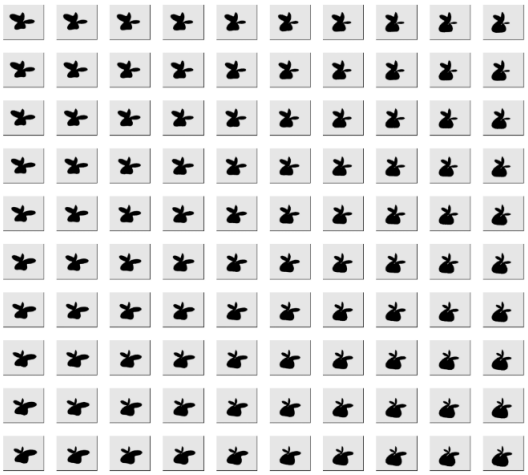


Figure 3.1: Stimuli used in the study ordered in a two dimensional matrix.

3.2.3 Design

There were four conditions of the experiment defined by the order of supervised and unsupervised information and the stimulus dimension along which the initial category boundary was suggested by the data. All four conditions were consisted of two Learning phases followed by Test phases (Fig.3.2).

In *Condition 1*, learners first had to categorize two repetitions of 60 stimuli presented in random order in an unsupervised manner, where the distribution of the data suggested a category boundary along the vertical axis of the stimulus matrix. In the second Learning phase, they received 20 repetitions of 4 stimuli, followed by corrective feedback after each of their category decision. Crucially, the boundary suggested by the stimuli was defined along the horizontal axis of the stimulus matrix, orthogonal to the one suggested by unsupervised trials.

Condition 2 counterbalanced Condition 1 with respect to the suggested boundaries: initial unsupervised trials prompted a boundary along the horizontal axis of the stimulus matrix, while

following supervised trials tried to teach participants a boundary orthogonal to that.

Conditions 3 and 4 counterbalanced the first two conditions with respect to the order of the type of information participants received. Specifically, in Condition 3, they received supervised trials first that suggested a vertical boundary, then they received unsupervised trials with data distribution prompting a horizontal boundary. In Condition 4, the order of vertical, horizontal boundaries was reversed compared to Condition 3.

Participants in Condition 1 and 4 and in Condition 2 and 3 received the same supervised and unsupervised information, respectively, only in different orders. Condition 1 and 4 both included unsupervised information suggesting a boundary along the vertical axis of the stimulus matrix, with supervised information suggesting a boundary orthogonal to that, while in Condition 2 and 3, the pairing between type of learning and orientation was the opposite.

Importantly, the ratio of supervised and unsupervised trials was not as imbalanced as in previous studies, to avoid the situation, in which the second round of Learning phase completely washes out the representation built by the first one as a result of overtraining. Also, as samples did not cover the entire stimulus space, they left a wide range of possible boundaries available for the learners to find, allowing them to build a representation compatible with both supervised and unsupervised information, i.e. a boundary that is diagonally separating the stimulus matrix.

Across all conditions, the Test phases following the Learning phases were identical. During the Test phase, participants were presented with all the stimuli in the stimulus matrix in random order not only the outer 60 shapes, and they had to categorize them without feedback. As the Tests covered the entire stimulus space, it allowed for a precise estimate of the boundary participants used in the stimulus space, while it did not provide any additional information in favor of either of the possible representations suggested by supervised or unsupervised data.

Also, since the two Test phases were identical, it allowed a fair comparison between them, for qualifying the effect of the second Learning phase.

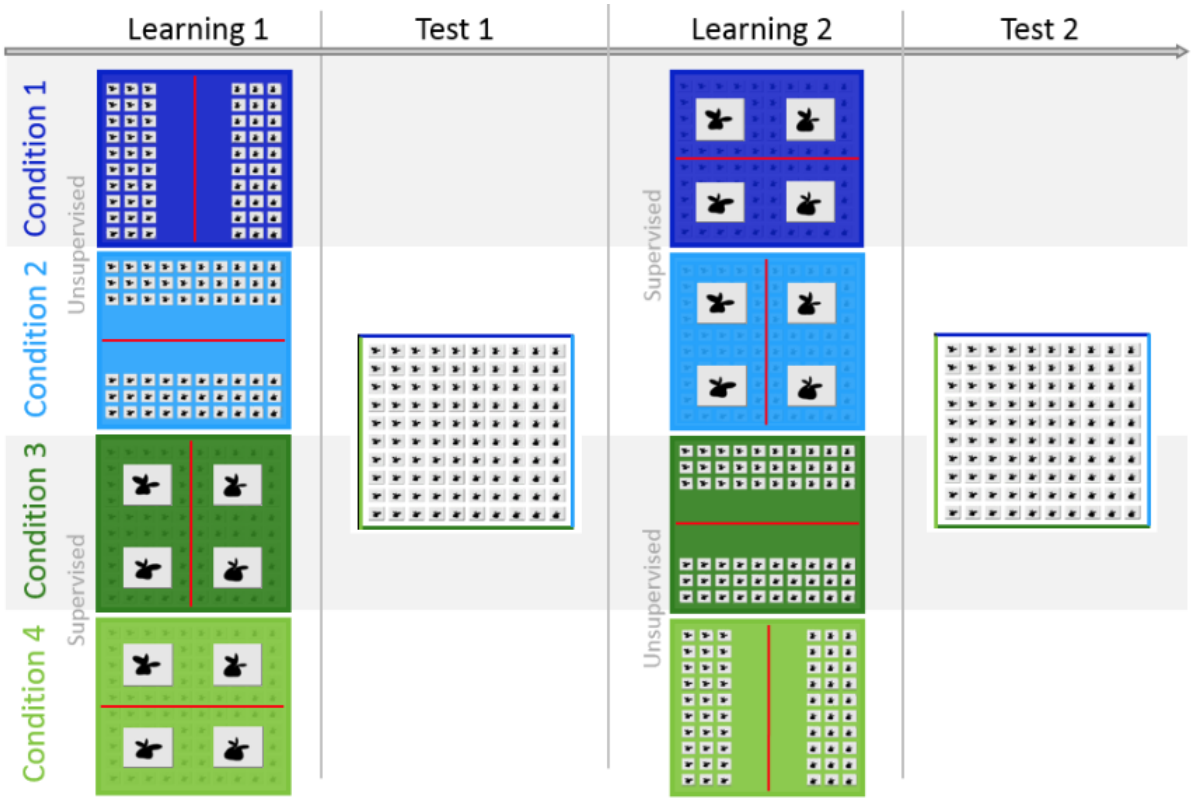


Figure 3.2: Layout of the design of the study.

3.2.4 Procedure

The experiment was created and presented with MATLAB 2014a on an iMac 27" (2560*1440) using Psychophysics Matlab toolbox. It was conducted in a dimly lit and sound attenuated room.

In each condition, a cover story preceded the experimental instructions. Participants were told that biologists discovered a new family of deep sea creatures called by made-up labels of Bitye and Tacok, or Dax and Wug for English speaking participants. These names were also used in earlier studies (Parise, Pomiechowska, Volein, Takács and Csibra, 2018). Participants were also informed that the families consisted of equal number of members. Before the unsuper-

vised learning phases, participants were instructed to categorize members of the two families as they saw fit, while before supervised learning phases, they were told that they would be taught to classify the creatures correctly, and they should try and learn them as fast as possible. There were no additional instructions preceding the Test phases after unsupervised Learning. When the Test phase followed supervised Learning phases, participants were informed that they no longer would receive feedback after their category decisions.

Each trial started by a stimulus appearing in the middle of the screen and it remained there until response. Participants responded by a left or right mouse click. Responses were followed by a blank screen for 500 ms. During the supervised Learning phases, after the blank screen, either a green or red square appeared for another 500 ms indicating whether the category decision of the participant was correct or not. At the end of each trial, learners were asked to indicate their confidence in their decision by adjusting a bar on a scale ranging from 0% to a 100% with moving the mouse on the vertical axis of the screen.

3.3 Results

Participants performed reliably well in all conditions in the first Learning phase with a mean accuracy of 93% in the supervised and 84.5% in the unsupervised training. Their performance was worse (84.5% and 64%) in both types of tasks when those appeared in the second Learning phase (Fig. 3.3). A two-way ANOVA revealed no significant interaction between learning type (supervised/unsupervised) and order (at Learning1 or at Learning2) [$F(1, 74) = .26, p = .6, \eta^2 = .44$]. There was however a significant main effect of learning type ($p < .001$), but not of order ($p = .92$).

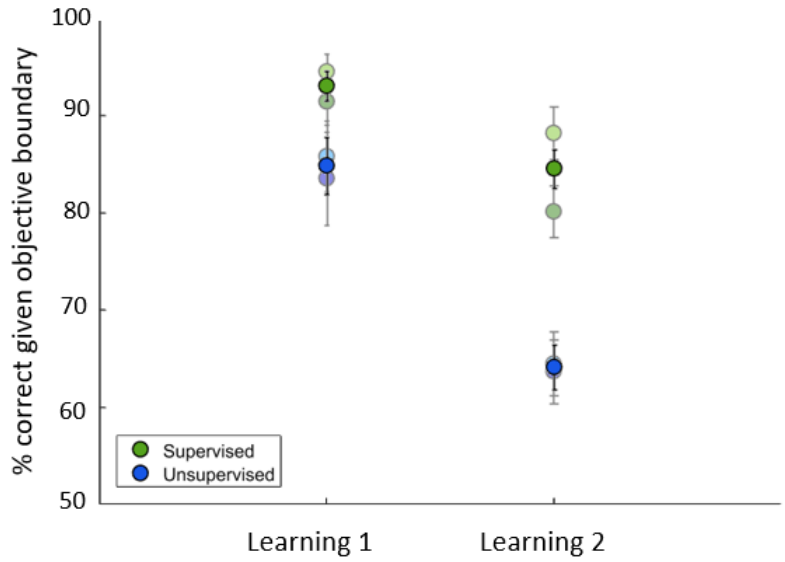


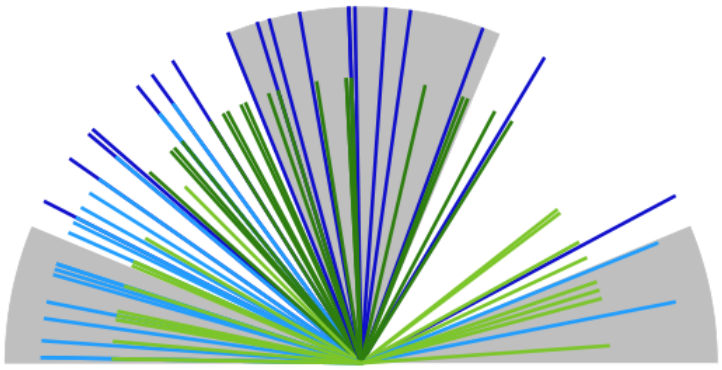
Figure 3.3: Learners' categorization performance relative to the objective category boundary suggested by data distribution of feedback for unsupervised (blue) and supervised (green) tasks, respectively, in both learning phases. Circles represent the mean percent correct performance of each group with SEM as error bars. In lighter colors the same measure is depicted separated by condition.

3.3.1 Defining category boundaries

For all four conditions, data from both Learning and both Test phases were analysed separately. First, the initial 60 and 40 trials of unsupervised and supervised learning phases, respectively, were discarded from the analysis to avoid unnecessary noise due to initial adjustment periods. Response data were transformed into a signed distance value from the inferred boundary and fed to a logistic regression model written in Matlab Stan to infer the angle of the boundary and the slope of the regression function.

While investigating the distribution of the angle of inferred boundaries in the first Test phase, I discovered that not all participants defined a boundary along the cardinal axes of the stimulus matrix (Fig. 3.4). In all conditions, half of the learners categorized incoming stimuli more along a diagonal boundary. For these learners, the distribution of the data in Learning phase 2 suggested a category boundary that was more in line with their close-to-diagonal initial representation

(as inferred from their responses in Test 1), which made it easier for them to interpret and incorporate new incoming data into their already existing representation. Therefore, I decided to handle these participants separately in further analyses from those who inferred a boundary more along the cardinal axes of the stimulus matrix. These two groups will be referred to as the *Diagonal* and *Cardinal* groups, respectively. This left me with sample sizes in each group summarized in Fig. 3.4.



Sample sizes

	Diagonal	Cardinal
Cond.1	10	9
Cond.2	11	10
Cond.3	10	11
Cond.4	11	9

Figure 3.4: Left: Distribution of angles participants defined at Test1. Gray background signals boundary angles that participants defined in the Cardinal group. Right: Sample sizes in each group and condition.

Participants of the Cardinal group in Condition 1 and 2 were all sensitive to the distribution of the data and defined a category boundary along the axis of the stimulus matrix suggested by the distribution of the data. There were only 5 participants in Condition 2, who defined a vertical boundary in spite of the fact that data distribution suggested a horizontal one. These participants were excluded from further analysis, as supervised trials at Learning phase 2 suggested a category boundary that was in line with their initial representation, which defeats the purpose of analyzing the impact of additional supervised trials suggesting a different boundary from the original.

To quantify the accuracy of inferred boundaries, I calculated the percent of trials matching between human data and the inferred boundary for all conditions and all Learning and Test

phases. The inferred boundaries predicted participants' categorization responses reliably well, as on average, the match was over 80% in all conditions. Best matches were achieved in the two Learning conditions, where average model accuracy was above 90%. Since data density around the inferred boundary was higher in the Test phases, noisy human responses might result in lower match between model prediction and human data (Fig. 3.5).

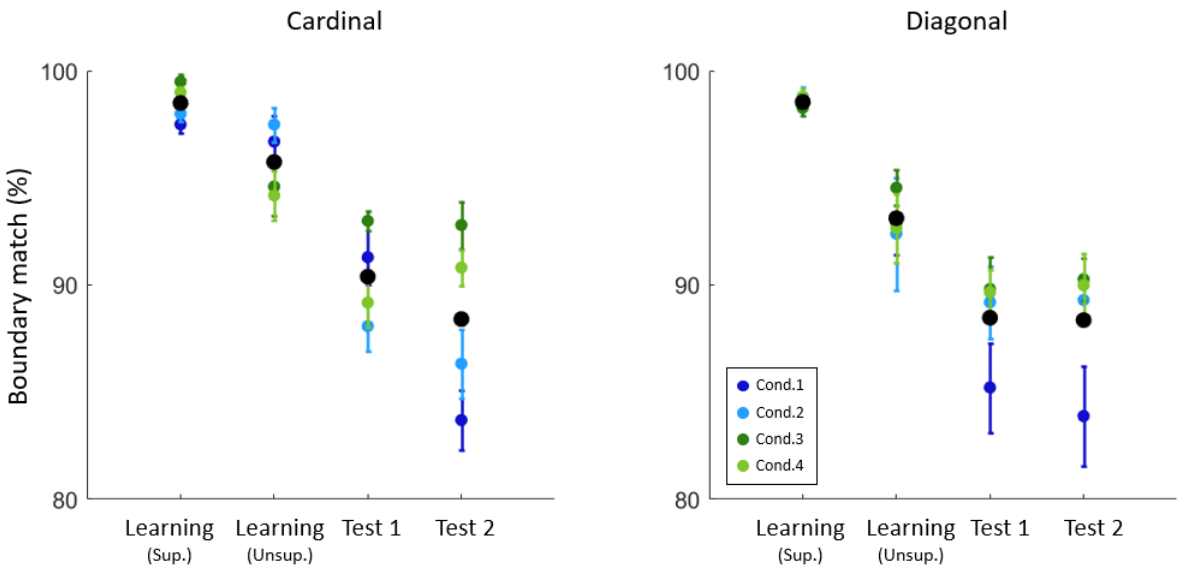


Figure 3.5: Match between inferred boundaries and human data for Cardinal and Diagonal groups. Black dots correspond to mean percent match in each learning and test phases, while colored dots and error bars represent mean and SEM of percent match in each condition for Learning and Test phases.

3.3.2 Changes in the representation caused by additional information

Results of the fitted logistic regression models provided two important measures of participants' representations: the *angle* of the inferred boundary, i.e. how are data clustered in participants' internal representation about them, and the *slope* of the regression function which is an indirect, implicit measure of the learners' confidence towards this representation.

3.3.2.1 Changes in the angle of the boundary between Test phases

To quantify the effect of information gained by learners in the second Learning phase on the initial representation measured at Test 1, I first calculated the absolute angle changes from Test 1 to Test 2 (Fig. 3.6). One sample t-tests indicated significant changes from 0 degrees in all eight groups. Most changes occurred in Conditions 1 (M = 58.83, SD = 31.25) and 2 (M = 56.6, SD = 29.82) with no significant differences between these two conditions in either the Cardinal [$t(19) = .17, p = .87, d = .07$] or Diagonal [$t(17) = -.03, p = .97, d = -.01$] groups.

Interestingly, though angle changes were not significantly different between Conditions 3 and 4 in the Cardinal group [$t(19) = -1.81, p = .09, d = -.79$], a two-sample t test indicated that angle changes were significantly more prominent in Condition 4 (M = 13.11, SD = 11.12) than in Condition 3 (M = 4.35, SD = 3.37) [$t(18) = -2.49, p < .05, d = -1.12$].

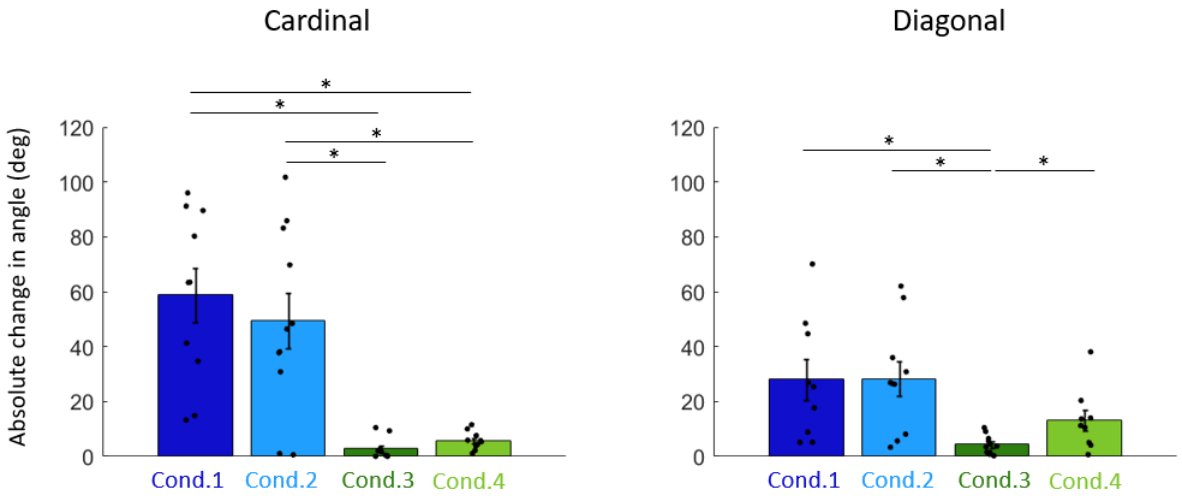


Figure 3.6: Absolute changes in the angle of category boundaries participants defined from Test1 to Test2. Bars depict the mean of changes in each group with SEM as error bars. Black dots represent individual subject's data both in the Cardinal (left) or Diagonal (right) groups.

3.3.2.2 Changes in the slope of the regression function between Test phases

Changes in the angle of the boundary is only one measurement of the internal representation of learners. Knowing how sharp this boundary is, or whether as a result of additional learning,

participants start responding more or less noisy around the boundary at Test 2 is at least as important as knowing where the separating line lies between the categories. As the slope of the fitted regression function is a good approximation of noisiness around the boundary, I calculated changes in the slope at Test 2 *relative* to their initial value measured at Test 1 to avoid misleading magnitudes resulting from individual differences. (Fig. 3.7).

Except for Condition 1 in the Cardinal group, there were no significant changes in the slope of the regression function between the two Test phases. The slope in Condition 1 of the Cardinal group decreased significantly by 43% (SD = 31.44) [$t(9) = 4.37, p < .01, d = 1.38$]. This change was not significantly higher than the relative change of the slope in Condition 2 [$t(19) = 1.02, p = .32, d = .45$], however, it differed from the ones observable in Conditions 3 [$t(18) = 2.59, p < .05, d = 1.16$] and 4 [$t(19) = 3.98, p < .001, d = 1.74$] as well.

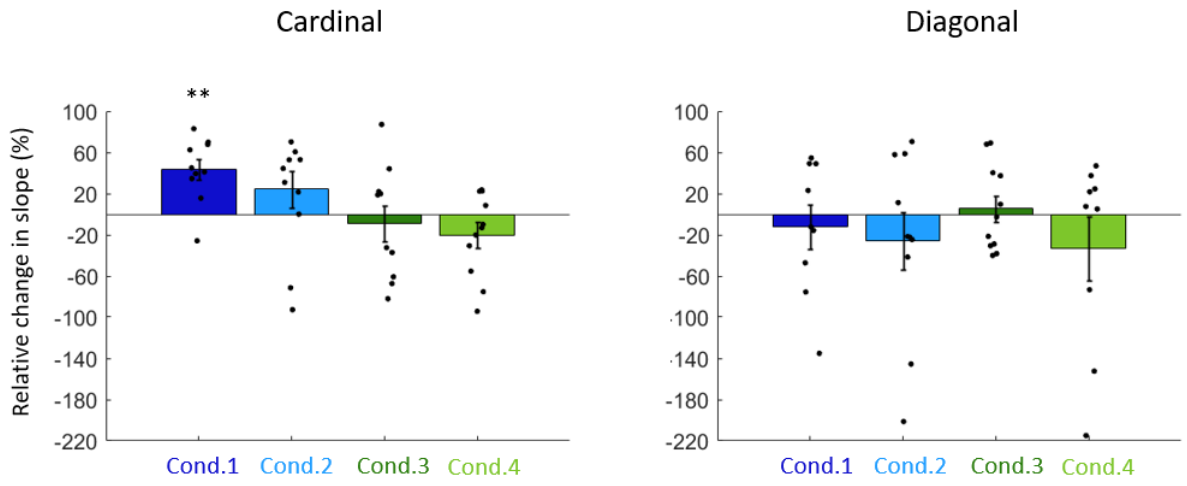


Figure 3.7: Relative changes in the slope of the regression function fitted to learners' category responses. Positive values indicate the increase of noise (i.e. decrease in confidence), while negative values correspond to the sharpening of the boundary by Test 2 relative to Test 1. Bars depict the mean of changes in each group with SEM as error bars. Black dots represent individual subject's data both in the Cardinal (left) or Diagonal (right) groups.

3.3.2.3 Differences in the final representation across groups and conditions

Though boundary and slope changes are appropriate to quantify the effect of the information received at the second Learning phase, it is also important to take a look at the final representations as well. A two-way ANOVA with Conditions and Groups as predictors revealed no significant interaction either for slopes measured at Test 1 [$F(3, 78) = 1.25, p = .29, \eta^2 = .06$] or at Test 2 [$F(3, 78) = 1.77, p = .16, \eta^2 = .04$]. Test results indicated no main effects at Test 1 either for Condition [$F(3, 78) = 1.93, p = .13, \eta^2 = .13$] or Group [$F(1, 78) = 1.64, p = .2, \eta^2 = .008$]. For Test 2, however, it revealed a significant main effect, though with small effect size, for Condition [$F(3, 78) = 3.7, p < .05, \eta^2 = .07$], but not for Group [$F(1, 78) = .65, p = .42, \eta^2 = .02$]. Such a significant main effect can be explained by the already reported significant decrease in the slope summarized at Fig. 3.7.

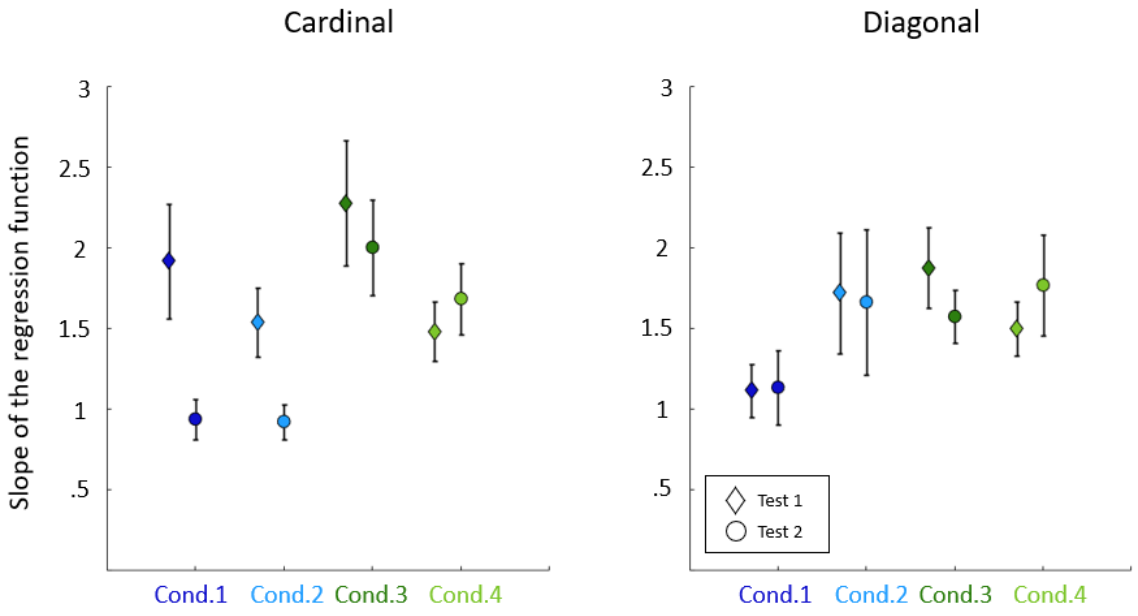


Figure 3.8: Mean slopes of the regression function inferred from the data gathered at Test1 (diamonds) and Test2 (circles) phases, for the Cardinal (left) and Diagonal (right) groups separately, with SEM as error bars.

Considering the angle of the boundaries defined by participants, as the two extremes of a possible representation is either a horizontal or a vertical boundary (a line with an angle of

0/180 or 90 degrees), I transformed all inferred angles at Test 1 and Test 2 into a value on a scale ranging from 0 and 1, corresponding to a horizontal and vertical line, respectively. Such relative angle values allow me to define if a boundary line is closer to a vertical or a horizontal boundary irrespective of whether a vertical line is approximating 90 degrees with an acute or obtuse angle, and similarly, whether a close-to-horizontal boundary is closer to a 0 or a 180 degree angle.

I discovered a markedly different effect of Learning 2 signaled by the changes from initial (boundary angle at Test1) to final (boundary angle at Test2) representations between the Cardinal and Diagonal groups. (Fig. 3.9) A two-way ANOVA indicated a significant interaction between the type of information received at Learning 2 (i.e. supervised in Conditions 1 and 2 or unsupervised in Conditions 3 and 4) [$F(1, 38) = 8.44, p < .01, \eta^2 = .09$]. Simple main effects analysis revealed a significant difference between both factors. Supervised information at Learning 2 had a much larger impact on the boundary than unsupervised information ($p < .001$), and such changes were much larger on average in the Cardinal group than in the Diagonal group ($p < .01$).

Paired samples t-tests for all Conditions in both Groups revealed a significant change of the angle of the boundary in Condition 1 [$t(9) = 5.37, p < .001, d = 2.67$] and Condition 2 [$t(10) = -6.7, p < .001, d = -3.12$] of the Cardinal group, and in Condition 2 [$t(8) = -3.7, p < .01, d = -1.53$], but not in Condition 1 of the Diagonal group.

It is also interesting to examine how the final representations differ from one another across Groups and Conditions. A one-way ANOVA failed to reach significance when comparing the final relative boundaries of Conditions 1 and 4 in the Cardinal and Diagonal groups against each other [$F(3, 35) = 1.98, p = .19, \eta^2 = .14$]. However, such a comparison revealed significant

differences across Conditions 2 and 3 in the Cardinal and Diagonal groups [$F(3, 38) = 4.01, p < .05, \eta^2 = .24$]. Post hoc two-sample t tests were only significant when comparing Condition 3 of the Cardinal group to any of the Condition 2 (Cardinal) ($p < .05$) or Condition 2 ($p < .05$) or Condition3 (Diagonal) ($p < .001$) groups.

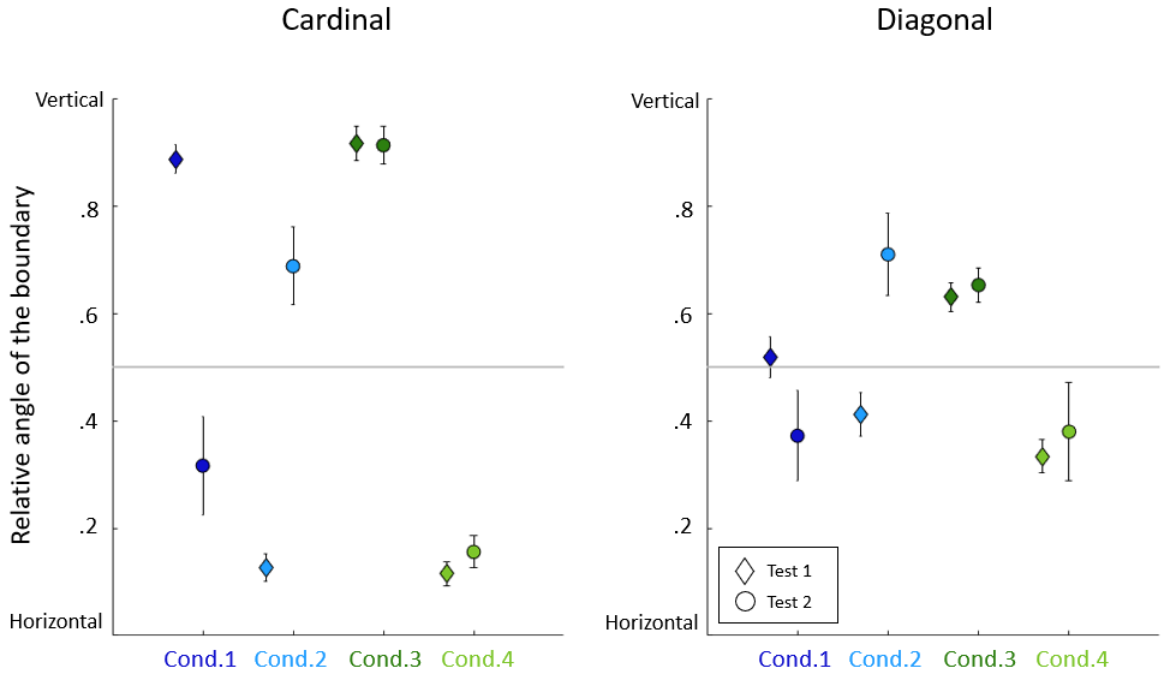


Figure 3.9: Relative angles of the boundaries inferred at Test1 and Test2 for all conditions in both Cardinal (left) and Diagonal (right) groups. The closer a value is to 1, the more vertical is the angle of the inferred boundary, while the closer it is to 0, the closer it is to a horizontal line. Diamonds correspond to the mean of the relative angles at Test1, circles stand for the mean of these at Test2, with SEM as error bars.

Considering the significant changes in the angles of category boundaries in Condition 1 and Condition 2, a strong effect of supervised training in Learning 2 seems obvious. If we hypothesize that as a result of generative learning participants try to incorporate new information into their already existing representation, their learning performance should be worse in Condition 1 relative to Condition 4 or Condition 2 relative to Condition 3. In Conditions 1 and 2 participants already have a representation built by unsupervised information, as a result, they already have an articulated prior on the angle of the category boundary, compared to Conditions 3 and 4

where no task precedes supervised learning, so learners' prior about the boundary should be assumed almost flat. Assuming generative learners, who aim to incorporate new information into their already existing representation, the mismatch between the boundary inferred from unsupervised trials and the one suggested by supervised information should also modulate performance on the supervised training. The larger the mismatch, the worse learners' performance should be. As a result, performance in Condition 1 or 2 of the Diagonal group should be better than performance of Condition 1 or 2 of the Cardinal group.

Analyzing performance at supervised training in Conditions 1 and 4 should be handled separately from the ones in Conditions 2 and 3. In the earlier case participants received unsupervised information suggesting a vertical boundary, while supervised training aimed to teach them a horizontal one, so the type and distribution of information was the same, only in different order. This was the other way around in Conditions 2 and 3.

First, I split trials at supervised Learning into 8 bins of equal size to get a learning curve and see trends in improvement. A two-way ANOVA revealed significant main effects of both bins and Condition ($p > .001$) as predictors for performance at supervised training, both at comparing Conditions 1 and 4 [time bins: $F(7, 208) = 14.22, p > .001, \eta^2 = .32$], [Condition: $F(2, 208) = 41.48, p > .001, \eta^2 = .28$] and Conditions 2 and 3 [time bins: $F(7, 208) = 7.66, p > .001, \eta^2 = .19$], [Condition: $F(2, 208) = 22.91, p > .001, \eta^2 = .1$]. To further analyze learning trends, two one-way ANOVAs were run separately for the early and late stages of the learning process by collapsing the fist two and the last two bins of data in all groups and conditions. Results revealed no significant differences between Conditions at the early stages of learning for the comparison of Conditions 1 and 4 [$F(2, 57) = 2.14, p = .12, \eta^2 = .06$], nor when comparing Conditions 2 and 3 [$F(2, 61) = .67, p = .51, \eta^2 = .01$]. By the end of the training session, performances became

significantly different comparing Conditions 1 and 4 [$F(2, 57) = 20.75, p < .001, \eta^2 = .43$], as well as comparing Conditions 2 and 3 [$F(2, 61) = 9.25, p < .001, \eta^2 = .23$]. Post hoc two-sample t tests revealed significant differences between Condition 1 and 2 (Cardinal groups) ($p < .001$), Condition2 of the Cardinal and Condition1 of the Diagonal ($p < .01$) as well as between Condition 1 of the Cardinal and Condition 1 of the Diagonal group ($p < .01$). Similarly, I found significant differences between Conditions 2 and 3 of the Cardinal group ($p < .001$), and even between Condition2 of the Cardinal group and Condition 2 of the Diagonal group ($p < .05$), but differences did not reach significance between Condition 3 of the Cardinal group and Condition 2 of the Diagonal group ($p = .07$).

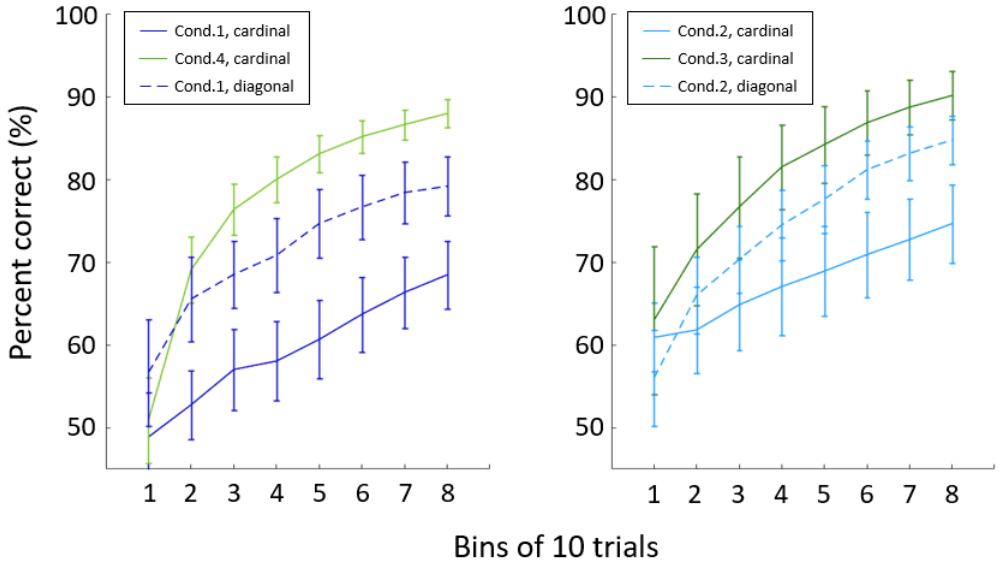


Figure 3.10: Comparison of changes in performance over supervised learning phases of different groups and conditions. Trials were split into 8 bins of equal size, and mean performance plotted with SEM as error bars. Conditions 1 and 4 of the Cardinal group and Condition1 of the Diagonal group were compared (left), and conditions 2 and 3 of the Cardinal group and Condition2 of the Diagonal group were compared (right), separately.

3.4 Discussion

In this study, I investigated two important attributes of SSL: the nature of the update of internal representations as a result of additional information (addressing the problem raised by McDonnell, Jew and Gureckis (2012)), and the effect of the order of presentation of supervised and unsupervised information, with a special emphasis of the integration of supervised information into the internal representation formed by unsupervised trials. To this end, I designed a study, where in four conditions, participants received almost equal amount of supervised and unsupervised trials suggesting category boundaries in the stimulus space that were orthogonal to each other.

Even though learners' good initial performance, estimation of the angle of the boundaries they used to categorize incoming stimuli revealed that half of them defined a category boundary that separated the stimulus space diagonally instead of the cardinal one that would be *a priori* assumed based on the findings of previous studies, where participants inferred the category boundary to be at the mid point between presented (supervised) stimuli (Zhu et al., 2007; Lake and McClelland, 2011; Kalish et al., 2011; Kalish, Zhu and Rogers, 2015; Gibson et al., 2015) (Fig. 3.4). The fact that learners performed well at Learning1 and still, they built a representation with a diagonal boundary ensures that such a boundary is a valid compromise when one intends to integrate information suggesting orthogonal boundaries at two separate Learning phases.

Supervised and unsupervised information in Learning 2 had markedly different effects on the initial representation built at Learning 1. Examining only the changes in the angle of the boundary inferred from data at Test 1, the initial, and Test 2, the final representation, lack of significant changes in Conditions 3 and 4 might imply that unsupervised data had no effect on

the representation built by supervised data at Learning phase 1. (Fig. 3.9) Supervised information, however caused significantly higher changes in the representation built by unsupervised data, at least in the Cardinal group. Considering the absolute changes in the angle of the category boundary (Fig. 3.6), interestingly, I found no significant differences between Conditions 1, 2 and 4 in the Diagonal group. Which implies that unsupervised information in the second Learning phase had a similar effect on the representation built by supervised data than the other way around in Conditions 1 and 2.

An effect of unsupervised information on the forming final representation can also be observed by comparing participants' performances on Learning 1 and Learning 2. Participants' responses at Learning phase 1 suggest that they acquired the categories well relative to the category boundary suggested to them by the distribution of the data at unsupervised Learning phases or corrective feedback at supervised Learning phases. (Fig. 3.3) However, when received the same type (supervised or unsupervised) of data at Learning phase 2, their performance on both types of tasks dropped significantly both for supervised [$t(80) = 3.04, p < 0.001, d = .75$] and unsupervised [$t(80) = 5.55, p < 0.001, d = 1.22$] tasks, which signals an interference with the initial representation.

A closer examination of participants' performance at supervised training at Learning phase 2 revealed that their performance at the supervised training is modulated by the proximity of the boundary inferred at Learning 1 to the one suggested by Learning 2. (Fig. 3.10 dashed blue vs solid blue line for Conditions 1 and 4 (left), and dashed green vs solid green line for Conditions 2 and 3 (right)). Such a behavior can be interpreted as an attempt to integrate new incoming data into their initial representation of the categories (formed by Learning 1). Such an attempt can be considered as a hallmark of generative learning, where both supervised and unsupervised

information are considered to be generated by the same underlying model.

Comparing relative angles of the inferred boundaries at Test 1 and Test 2 is also in favor of the generative hypothesis of learners (Fig. 3.9). Supervised information did not overwrite the initial representation, only modify it drawing it more towards a final representation that is compatible with the information possibly gained from both supervised and unsupervised samples. A similar tendency is observable in the Diagonal group, however, significant changes are only induced in Condition 2 by the supervised Learning phase (Diagonal group of Fig. 3.9). The fact that supervised information did not overwrite completely the initial representation to an extent that participants abandoned the boundary they inferred from unsupervised data can be considered as a hallmark of generative learning, where the learners' goal is to build a model of their learning environment that can fit all incoming information irrespective of the presence or absence of supervision.

However, supervised information seemed to influence the representation built by unsupervised samples more than the other way around, especially when the suggested boundaries markedly differed from one another (Cardinal group of Fig. 3.9). Such a behavior of learners is not surprising given the findings of Zeithamova and Maddox (2009). They found that even without explicit supervision, if the sequence of incoming unsupervised samples make one feature dimension more salient then another in a two dimensional feature space, participants will be more likely to categorize along this salient dimension, even if the distribution of the data clearly favors a feature dimension orthogonal to that. In addition, since supervised samples are considered to be stronger and more salient than unsupervised ones (Kalish, Zhu and Rogers, 2015), it is not surprising that learners would not abandon the boundary suggested by supervised information. This strategy of learners is reflected in the insignificant changes in the angle of the

boundary both in the Cardinal and Diagonal groups. Such a learning behavior is also consistent with the generative hypothesis of learners' strategy of processing incoming information. In the Diagonal group (as explained in Section 3.2.2) the distribution of unsupervised information allows for a possible diagonal boundary as well, so the initial representation can be considered as being in line with the one suggested by unsupervised data. Considering the Cardinal group, though learners might be able to incorporate the gap in the distribution of unsupervised samples into their representation, but the boundary is not necessarily defined along the midpoints of the gaps in the distribution of stimuli in such a two dimensional stimulus space (Zeithamova and Maddox, 2009).

The lack of significant differences in the final representation indicated by the angle of the inferred boundaries between the Cardinal and Diagonal groups also supports the generative hypothesis of learners. Irrespective of the magnitude of the change necessary for finding a representation compatible with supervised and unsupervised data, learners arrive to the same results. The fact that these representations in Conditions 1 and 4 as opposed to Conditions 2 and 3 differ from one another imply that supervised information is indeed stronger, drawing the final representation towards the direction that favors supervised training, i.e. towards the vertical axis in Conditions 1 and 4, while towards the horizontal axis in Conditions 2 and 3.

The fact that I found significant absolute changes in the angle of the category boundary for Condition 4, but not for Condition 3 in the Diagonal group (Fig. 3.6), a significant decrease in the slope of the regression function at Condition 1, but not Condition 2 in the Cardinal group (Fig. 3.7), or the fact that there were 5 participants preferring to define the category boundary on the vertical axis, while they were receiving data suggesting a horizontal boundary at Condition 2 all point to the direction that there is are more factors defining participants' internal representa-

tion than only the type (supervised, unsupervised) and distribution of the data, or the order of the presentation of supervised and unsupervised information. However, a pilot study failed to indicate firm preferences for any of the features of the stimuli that could be considered to be more salient than others, these results suggest, that there might be an implicit preference for defining the boundary along the vertical axis of the stimulus matrix. Such, possibly implicit preferences seem to bias the resulting internal representation of the categories just like parameters of the learning environment that are transparent to an experimenter. To gain a deeper understanding of the processes at play while the acquisition of novel categories, researchers should aim for retrieving as detailed information about learners' internal representation throughout the learning process as possible, so that they could account for hidden influencing factors.

Finally, in line with previous critique discussed at the introduction of this chapter, in this study all significant changes in the angle of the boundary from Test 1 to Test 2 inherently required the re-categorization of certain elements in the stimulus space. Any significant changes can be considered to be a result of true update of the internal representation.

Based on the above findings, we can conclude that learners indeed update their representation as a result of additional incoming information in SSL. Such a change in representation can be elicited both by supervised and unsupervised information. To be able to quantify such radical updates, it is worth using high density multidimensional stimuli.

Supervised information does not overwrite the initial representation built based on unsupervised samples. The nature of the integration of additional information is such that is seeks for a hypothesis in the space of possible boundaries which will be consistent with both supervised and unsupervised information provided.

Taken together the present results with that of previous studies, it seems that the order of presentation of supervised and unsupervised information does not matter, as long as they do not contradict one another to an extent that makes their handling as part of the same generative model impossible.

Implications to previous studies

The findings and conclusions discussed above have important implications to at least two previous studies. In their conclusions, Zhu et al. (2007) failed to explain an apparent flattening of decision curves of each participant after the exposure to unsupervised samples. Such a flattening is consistent with the generative hypothesis discussed above. The flattening of learning curves might be a result of an attempt for not completely abandoning the initial hypothesis about the category boundary suggested by supervised information, but trying to incorporate it into the representation favored by unsupervised samples by increasing the variance of the distribution of categories in the learners' internal representation.

The failure of finding evidence of SSL in McDonnell et al.'s (2012) study (See section 1.1.3) might also be explained, or at least re-framed in light of the evidence presented above. The distribution of stimuli used in their study – similarly to ours – allowed for the inference of multiple category boundaries consistent with the distribution of the data. The two most frequently used category boundaries by participants (i.e. the Bimodal (Cardinal, or rather vertical in our case) and the Two-Dimensional (Diagonal)) are both consistent with the distribution of supervised and unsupervised samples alike. As in their design supervised samples were interleaved with unsupervised ones, and they did not have test phases covering a greater area of the stimulus space that would allow them to retrieve learners' precise internal representation of the category

boundaries, it is difficult to quantify the effect of supervised and unsupervised information on the final representation. Given the generative theory of SSL, they did not fail to find hallmarks of SSL, since learners' strategy of categorizing stimuli suggest category boundaries that are consistent with both supervised and unsupervised information as well.

For a deeper understanding of cognitive processes in work while categorization or acquisition of categories, I should also investigate the neural mechanisms underlying such processes. In the upcoming chapter, I will present a study where I used EEG, a non-invasive, widely applied method for recording neural activity in the brain while teaching participants about novel visual categories.

Chapter 4

Neural correlates of emerging representations of novel categories

4.1 Introduction

Two different domains of neural activity most commonly studied with EEG are event-related potentials (ERP) and changes in ongoing neural oscillations. ERPs are neural responses that are directly elicited by some sensory, motor or cognitive event, stimulus (Luck, 2005). Frequency analysis is concerned with the changes in ongoing neural oscillatory activity in response to an event. A common way of frequency analyses is to define and quantify the magnitude of synchronization or desynchronization of oscillatory activities of neural ensembles (Sanei and Chambers, 2007). Neurons' coherent electrical activity in a neural ensemble can increase (they synchronize their firing patterns) or decrease (desynchronize) locally as a response to incoming stimulation. The pattern of rhythmic activity across neurons defines the frequency band, in which neurons oscillate. One of the most commonly studied frequency bands is the alpha band,

where neuron ensembles oscillate at around 8-12Hz. In this chapter, I am presenting a study on neural correlates of the ongoing acquisition of novel categories.

For an investigation of internal representation of categories, I needed a neural signal that differentially responded to separate categories. For the sake of simplicity, I chose to work with only two categories at once. To draw general conclusions about categorization, I looked for a neural signal that was modality- and feature-independent as much as possible. These neural responses would be most likely modulated by the modality- and feature-independent attributes of the investigated categories. Such an attribute of a category could be its occurrence probability. Fortunately, both in the domain of event-related potentials and frequency analysis, there are neural signals that are sensitive to frequency of occurrence. The P300 ERP, alpha band ERD and theta band ERS are all sensitive to frequency differences between alternating stimuli (Sochurková, Brázdil, Jurák and Rektor, 2006; Yordanova and Kolev, 1998a; Yordanova, Kolev and Polich, 2001; Peng, Hi, Zhang and Hu, 2012; Klimesch, 1999). All three neural signatures are widely investigated, they can be elicited by stimuli from many different modalities (Peng, Hi, Zhang and Hu, 2012; Peng, Hu, Mao and Babiloni, 2015). In addition, the same paradigm, called the oddball paradigm is used to establish frequency differences between stimuli or categories in studies involving P3 ERP and power changes in different frequency bands, and their relationship is also investigated (Sochurková et al., 2006; Yordanova and Kolev, 1998a; Harper, Malone and Iacono, 2017; Yordanova, Kolev and Polich, 2001; Peng et al., 2015).

4.1.1 The oddball paradigm

The oddball paradigm is an experimental research design that was first used by Squires, Squires and Hillyard (1975) to investigate ERPs. The most essential feature of the paradigm is the al-

ternating sequence of two stimuli where one of the stimuli has much less appearance probability than the other. The relative frequencies of the rare and frequent elements are typically 0.2 and 0.8, though it is possible to elicit similar effects with less articulated frequency differences (e.g. (Parise et al., 2018)). The paradigm has been adapted to diverse sensory modalities. Auditory stimuli are usually two different tones of different frequencies (e.g. 500 and 1000Hz), visual stimuli might be two different shapes (e.g. a circle and a triangle), somatosensory and noxious stimulation can be electric pulses of different, well distinguishable intensities (Peng et al., 2012).

More importantly for my current goal, frequency differences can also be created "semantically", by grouping different number of stimuli, each with the same appearance frequency together. One group might contain 3-4 times more elements than the other one. Once a semantic bound (category) is formed across the grouped stimuli, even though the appearance probability of each individual element is the same, the relative frequency of the category containing more elements will be higher (3-4 times more in the current example) than the other's.

Using this version of the oddball paradigm, Parise et al. (2018) grouped four objects (fork, spoon, knife, hammer) as stimuli with the same appearance probability across objects, though conceptually the fork, knife and spoon belong to the same category of CUTLERY. Even though none of the four objects were shown more frequently than the others, participants evidently grouped CUTLERY into one category, which was reflected by the neural signals, as both in ERPs and oscillatory power changed their neural signatures typical for the frequent stimuli in the oddball paradigm. In the case of novel stimuli, substantial differences in neural response can only be expected though, if separate representations are formed of the newly acquired categories (Parise et al., 2018).

The oddball effect (significantly different response to rare stimuli than to frequent ones) can be observed in pupil dilation (Kamp and Donchin, 2015; Liao, Yoneya, Kidani, Kashino and Furukawa, 2016), subjective perception of duration (Schindel, Rowlands and Arnold, 2011) as well as in differences of neural responses of the P3 ERP and in changes of power in different frequency bands.

4.1.2 The P300 ERP

The P300 event-related potential is a neural response most commonly investigated in an oddball paradigm and elicited by the stimulus approximately 300ms after its presentation. The P3 is a positive-going amplitude in the EEG signal that has two components differing in latency and that are elicited by different stimulus characteristics. The P3a, or novelty P3 is an earlier component that results in higher amplitude to novel, surprising stimuli over frontal or central sites with a peak latency around 250-280 ms, signaling attention orientation towards task-relevant or novel stimuli (Harper, Malone and Iacono, 2017).

In contrast, the target P3 or P3b component peaks later as a result of stimulation by task-relevant, rare target stimuli, and it reflects categorization and context updating. The latency of P3b varies between 250-500ms, and it is more prominent at parietal areas (Polich, 2007; Harper, Malone and Iacono, 2017). To elicit a P3b, the participant must be engaged in some sort of task that requires them to react to incoming stimuli. For example, an auditory oddball only elicits a P3 response if the stimulus is presented at the attended ear (Picton, 1992). It is not only attention that influences the occurrence of P3. Since it is sensitive to workload (Donchin, 1981), it is possible to habituate P3, and on the long run, to decrease the neural response even to rare stimuli (Kok, 2001). In addition, as the task gets easier – for example, as a result of

learning –, the amplitude of the P3 will get smaller (Picton, 1992). Based on these findings, many interpret the P3 as a signal of the degree of top-down cognitive attentional involvement in a task (Debener, Kranczioch, Herrmann and Engel, 2002).

Once elicited, the P3 amplitude varies with the improbability of the stimulus. It is not only the overall relative frequency of the rare and frequent elements that modulates P3 amplitude, but also within-experiment temporal aspects of stimulus presentations. Increased temporal delay between the appearance of two rare elements results in higher P3 response amplitude. The latency of maximum peak can also vary, and it is modulated by the difficulty of the task. Its peak amplitude is around 300ms after stimulus presentation only if the task requires a simple discrimination decision. The more cognitive effort is required for solving a task, the later the maximum amplitude is expected. Crucially, motor task demands do not affect P3, it depends mainly on perceptual resources (Picton, 1992).

4.1.3 The alpha ERD

Stimulus-evoked and ongoing neural oscillations are both hypothesized to be influencing information processing in the brain (Engel, Fries and Singer, 2001). For example, the phase of ongoing EEG activity can hinder or support the processing of incoming information (Van Rullen, Busch, Drewes and Dubois, 2011), or event-related modulation (increasing or decreasing synchrony of neural firing) of ongoing oscillation can mediate the activation of different areas involved in information processing (Klimesch, Sauseng and Hanslmayr, 2007). In my study, I am focusing only on event-related changes in the alpha power, more specifically on deychronization in the alpha band.

Alpha ERD is a decrease in the amplitude of neural activity in the alpha frequency band

(7.5-12.5 Hz) as a response to incoming stimulation (Klimesch, 1999). There are alpha systems suggested in the brain that control and influence changes in neural responses in the alpha band, and as a result mediate the processing of incoming information (Kolev, Yordanova, Schürmann and Batar, 1999). Similar to P3, alpha ERD is mostly considered to be a reflection of controlled attention allocation and memory updating that can be achieved by selective inhibition of unnecessary areas and timing of ongoing oscillation (Klimesch, Doppelmayr, Russegger, Pachinger and Schwaiger, 1998; Klimesch, 2012; Keller, Payne and Sekuler, 2017). According to the inhibition timing hypothesis of alpha oscillations, strong alpha power reflects the inhibition of areas that are not used in solving the current task or not involved in processing the incoming information. Desynchronization in the alpha band, in turn, aids processes that are needed for the current task (Klimesch, Sauseng and Hanslmayr, 2007). As a result of learning and practice, the system will know what processes to inhibit or aid for a better task performance, and the dynamics of synchronization and desynchronization can change throughout an experimental session (Bays, Visscher, Le Dantec and Seitz, 2015).

Desynchronization in the alpha band is not an all-or-none response. Spatial, temporal, frequency and power differences can be linked to different features of the task and stimuli. Different frequency bands within the alpha range seem to reflect different cognitive processes. Upper alpha ranges (above 10 Hz) are associated with the processing of task specific, sensory semantic information, as well as controlling knowledge access and semantic (long-term memory) update. Meanwhile, desynchronization in the lower alpha band (below 10 Hz) seem to reflect attentional processes (Klimesch et al., 1998; Klimesch, Doppelmayr and Hanslmayr, 2006). Task-specific, higher alpha ERD responses are most prominent in task-specific areas topographically, while lower alpha responses are topographically more widespread and mostly

reflect basic attentional processes (Peng et al., 2015). Similar to the latency of P3 ERP, the magnitude of alpha ERD reflects the amount of mental effort required by the task (Sutoh, Yabe, Sato, Hiruma and Kaneko, 2000). It is also sensitive to the frequency or surprise value of incoming stimuli: it is more articulate after the presentation of rare stimuli than after frequent stimuli. As a result, alpha ERDs can also be investigated using the oddball paradigm just like P3 ERPs (Peng et al., 2015)(Vázquez-Marrufo, Galvao-Carmona, Lugo, Ruíz-Pena, Guerra and Ayuso, 2017).

4.1.4 The theta ERS

Theta (4-7.5 Hz in humans) is the dominant rhythm in the hippocampus of lower mammals, and it is mostly investigated in the hippocampus of humans as well (Klimesch, 1999). As a result, the function and role of human cortical theta responses in cognitive processes is less clear, though its involvement in cortico-hippocampal interaction seems to be supported by many studies (Keller, Payne and Sekuler, 2017; Basar, Basar-Eroglu, Karakas and Schurmann, 2001).

In the cortex, theta ERS is found to be a common response for a variety of tasks involving working memory, especially memory retrieval and cognitive control (Cavanagh and Frank, 2014; Klimesch, Doppelmayr and Hanslmayr, 2006). It is hypothesized to reflect mainly novelty detection, and to be involved in top-down control of memory encoding, especially of episodic memory. Intensity of theta ERS can be modulated by attentional demands, task difficulty, and cognitive load – similar to alpha ERD. However, while alpha desynchronizes, theta synchronizes as a response to incoming stimuli under similar circumstances (Klimesch, 1999). Theta response is also modulated by the novelty or surprise effect of the stimulus. Investigating it with and oddball paradigm (either auditory or visual), theta ERS is significantly more

enhanced for infrequent or unexpected, novel stimuli.

We need to distinguish between early (0-300ms after stimulus onset) and late (between 300-600ms after stimulus onset) theta responses (Yordanova and Kolev, 1998b). Early theta responses are assumed to be modality specific as they are most prominent over the vertex and occipital areas for auditory and visual stimulation, respectively. At frontal and parietal areas, it is the most enhanced for expected, predictable stimuli. As opposed to this, late theta seems to be modality independent, and it is topographically restricted mainly to midfrontal scalp regions.(Harper, Malone and Iacono, 2017) Also, late but not early theta gives more articulate responses for oddball as opposed to predictable or passive stimuli. The above listed attributes are primarily true for late theta responses.

4.1.5 P3 ERP, Alpha ERD and theta ERS

Neural responses in a variety of frequency bands play a role in diverse cognitive processes like attention, learning or memory update involving many sensory and cognitive levels (Basar et al., 2001). By now, some similarities between the processes associated with P3 ERP, upper alpha ERD and late theta ERS might be apparent. They are all modality independent (but only in reaction to target stimuli – alpha ERD showed modality-specific responses for non-target stimuli), they reflect task-related high cognitive activation and attention, they are involved in similar processes, i.e. categorization or memory updating, and they are all sensitive to stimulus frequency reflected by the more articulated responses for rare stimuli (Sochurková et al., 2006; Peng et al., 2012). All these similar characteristics of the above discussed EEG responses justify their involvement in the present study addressing neural correlates of the acquisition of novel categories.

4.1.6 Goal of the present study

As discussed above, creating appearance frequency differences between categories or stimuli in an EEG experiment allows one to elicit differential neural responses for stimuli from one category and the other. In such studies, categories are most commonly defined by one target and one non-target stimulus; one target vs many non-target stimuli (Azizian, Freitas, Watson and Squires, 2006b; Azizian, Freitas, Parvaz and Squires, 2006a); discrete stimuli that belong to categories already familiar to the observer; or novel, discrete stimuli that participants are excessively trained to learn to categorize before the EEG recording (Parise et al., 2018).

There are, however, two important aspects of categorization that cannot be addressed by experimental setups and stimulus sets commonly used in previous studies. First, the neural correlates of the ongoing acquisition of categories are not accounted for, and second, it is still unknown how the within-category structure modulates neural responses. Studies by Azizian et al. (2006b) and Azizian et al. (2006a) might serve as a good starting point to investigate these issues. In their study, the authors used one target stimulus and created the non-targets in a way that they differed from the target in different degrees in similarity. According to their results, P3 ERP was sensitive to perceptual similarity of non-targets to the target stimulus, and more perceptual similarity predicted more articulate P3 responses. Nevertheless, it is still unknown if this result holds for scenarios where the similarity is established between two categories (more than one element composing each category), and for stimulus sets, where differences between targets and non-targets or between frequent and rare categories are less perceptually prominent. In addition to these issues, I also asked if alpha ERD and theta ERS showed similar patterns of modulation to those of P3 ERP in such complex stimulus environments.

In my study, I altered the commonly used oddball paradigm in a way that it allowed me to

investigate the following two major questions:

1. *Is the ongoing acquisition of categories traceable by neural responses recorded with EEG?*

   If P3, alpha ERD or theta ERS responses reliably reflected (for instance by a gradual emergence or more and more articulated differences in response to rare and frequent category elements) the process of category acquisition, they might be a good tool for tracing more complex or less transparent learning scenarios such as implicit learning.

   To be able to track the emergence of internal representations, I need to make participants learn to categorize novel stimuli and start the EEG recording already at the very first trial of this learning.

2. *Can any of these neural responses provide us with more sophisticated information about the developed category representation other than plain category membership?*

   Finding evidence for a modulation of neural responses by different aspects of the representation – like uncertainty or within-category structure – could provide a tool for mapping internal representations of categories, as well as refine the theory on the function of each of the neural signals involved.

   To see how the structure of the category influences neural responses – if at all – I need stimuli varying on continuous feature dimensions, so that I can have response measurements to stimuli that are more or less typical to the category considered.

4.2 Methods

4.2.1 Participants

Thirty-one right handed subjects [18 females, mean age = 24 years] gave written informed consent and completed the experiment. Data from 6 subjects were excluded from analysis due to excessive EEG artifacts.

4.2.2 Stimuli

Following the above outlined goals, I needed stimuli that varied on a continuous dimension to ensure within-category structural differences among stimuli. I also needed the categorization task to be somewhat difficult so that participants could not learn the category boundary immediately after a couple of trials, hence, the gradual emergence of category representations could be traced. For this reason, I needed novel stimuli, for which any priors on categories, and hence, priors on category boundaries would not bias the emerging representation unpredictably. In addition, the stimuli had to vary along more than one clearly identifiable dimension to make the task less trivial.

I used the parametrically tunable stimuli created by Op de Beeck, Wagemans and Vogels (2001) (See 4.1). These shapes varied continuously along integral stimulus dimensions between the extremes of the range. In order to use the oddball paradigm, I needed unequal relative frequencies of the two categories. There are two common ways of establishing frequency differences between two categories. One can either have equal number of elements in both categories and increase the appearance probability of elements in one category, or the numerosity of stimuli in one of the categories can be increased while the appearance probability of each element

could stay identical. In the first case, an enhanced neural responses can be interpreted as a surprise effect due to the fact that stimuli in the rare category are less familiar, less expected. This is a response that does not require one to form an internal representation of the stimuli that would group a subset of them into one category and the rest in another. In the meantime if one establishes unequal appearance probabilities between categories by shifting the category boundary on the stimulus range to include less elements into the category intended to be rare, but the relative frequency of all the stimuli are the same, more articulated neural response to samples from the rare category would unequivocally signal the presence of emerged internal representation of the categories, as frequency differences are only interpretable on the category but not the stimulus level. Following the latter strategy, I sampled 12 stimuli from the generated stimulus space out of which 3 belonged to one category and 9 to another. (Fig. 4.1) Stimuli were presented in 360x250 pixels in size.

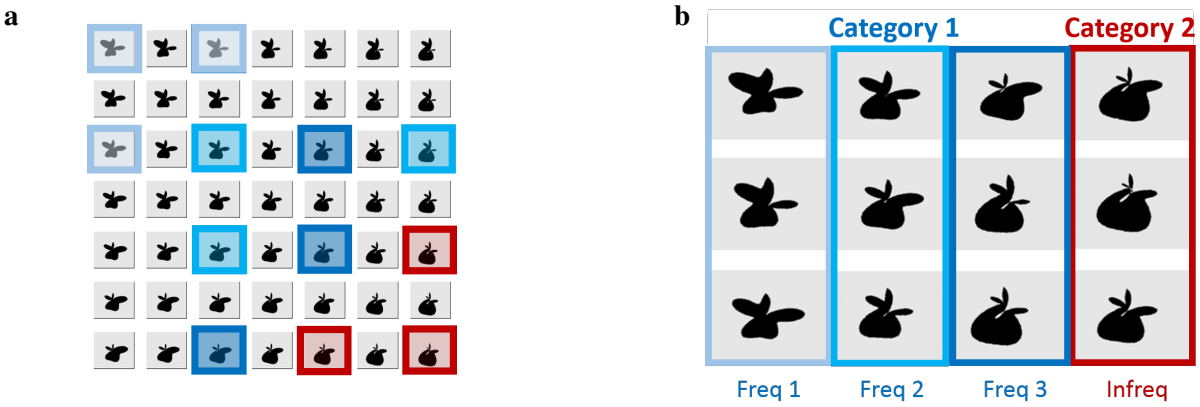


Figure 4.1: a) Stimulus space with sampled stimuli used in the study. **b)** Stimuli comprising the frequent (blue) and the infrequent (red) categories. Stimulus groups in the frequent category are defined by participants categorization behavior. Worse performance implying more similarity to the other, infrequent category.

Differences in neural response reflecting within-category structure is expected to be apparent across three stimulus groups in the frequent category, where the three groups are defined by stimuli that are gradually diverge in similarity from stimuli in the rare category. All three

within-category stimulus groups are consisted of 3 stimuli so that they could be compared to other within-category groups or to the rare category with equal sample size. Stimuli consisting the rare and frequent categories were counterbalanced across the study.

4.2.3 Procedure

The experiment was created and presented with Matlab 2014a using Psychophysics Matlab toolbox. Participants were seated in a dimly lit room in front of a computer screen. According to the experimental instructions, they were about to learn classifying two deep sea animal species called by made-up labels also used in similar studies: Bitye and Tacok, or Dax and Wug for English speaking participants (Parise et al., 2018). After the instructions, participants were provided 2 random examples from both categories with the corresponding labels. Later, they were asked to classify each figure they saw. Stimuli appeared at the center of the screen for 800ms following a fixation cross of random duration between 400-600ms. As the stimuli disappeared, a response screen followed with the labels prompted on the left and right side of the screen. Participants made their category decision with pressing left and right side buttons on a gamepad. Correct classification decisions were followed by a green circle while a red cross signalled incorrect responses for 500ms. In addition to choosing the category, participants were asked to provide a confidence judgments after each of their responses on a scale from 0 to 100%.

The experiment consisted of 3 experimental blocks separated by breaks of a few minutes. In each block the 12 stimuli were presented 20 times in random order.

EEG recording and analysis

High-density, continuous EEG was recorded using Hydrocel Geodesic Sensor Nets (Electrical Geodesics Inc., Eugene, OR, USA) including 128 channels equally distributed on the scalp, referenced to the vertex (Cz). The sampling rate was 500 Hz with a low-pass filter of 200 Hz.

EEG was band-pass filtered between 0.3 and 30 Hz. Continuous data were segmented into 12 groups: 4 stimulus groups (Freq1, Freq2, Freq3 and Infreq) × 3 blocks of the experiment (Block1, Block2, Block3). Segments were defined 600 ms before and 1200 ms after stimulus onset. Epochs were classified as artifacts whenever the average amplitude of a 80 ms sliding window exceeded 55 μV at horizontal EOG channels, 140 μV at vertical EOG channels, and 80 μV at any other channel. Bad channels were automatically interpolated in epochs in which ≤ 10% of the channels contained artifacts; epochs in which > 10% of the channels within a -200 and 800 ms window around stimulus onset contained artifacts were automatically rejected.

Wavelets

Retained segments were imported into Matlab using EEGLAB (v9.0.5.6b) and re-referenced to average reference. After referencing, epochs were convoluted by complex Morlet wavelets within the frequency band of 5-15 Hz with 1 Hz resolution using a custom-made script collection, WTools. Epochs then were baseline corrected to a 200 ms interval immediately preceding stimulus onset. I defined sparate ROIs for expected upper alpha ERD and theta ERS, as based on previous literature they are expected to be most prominent at parietal (Klimesch, Doppelmayr and Hanslmayr, 2006; Peng et al., 2012; Yordanova and Kolev, 1998a; Yordanova, Kolev and Polich, 2001) and midfrontal (Harper, Malone and Iacono, 2017; Yordanova and Kolev, 1998b)

regions, respectively. All epochs were baseline corrected to a 200 ms long interval immediately preceding the onset of the stimulus. Absolute values of complex coefficients were computed at the ROIs within the time window of 300-800 ms(Parise et al., 2018) and 200-600 ms(Harper, Malone and Iacono, 2017) and frequency range of 8-12 Hz and 3-8 Hz for alpha ERD and theta ERS, respectively.

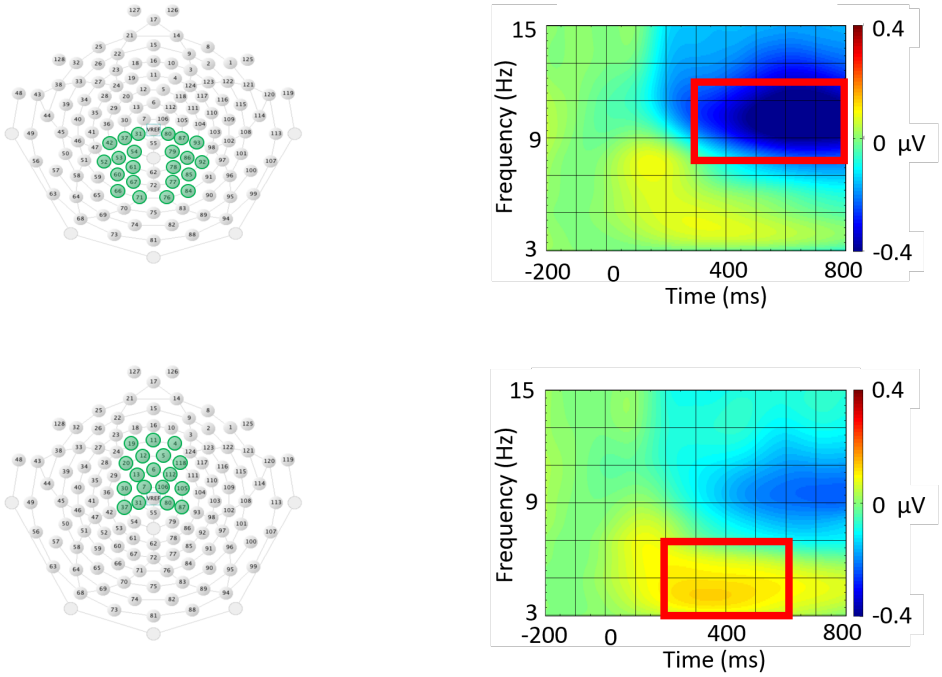


Figure 4.2: ROIs **a)** and **c)** and time window with frequency band **b)** and **d)** defined for α (upper) and θ (lower) response analysis, respectively.

ERPs

Epochs were baseline corrected to the 200 ms interval preceding stimulus onset. Bilateral, symmetric ROIs were defined parietally following Parise et al. (2018). P300 ERDs were quantified as mean amplitude within the time window of 250 and 500 ms after stimulus onset.

4.3 Results

4.3.1 Behavioral results

Participants successfully acquired the categories by the end of the experiment. Average performance across stimulus groups in Part3 was 96.6%. A gradual increase could be observed part-by-part throughout the experiment. Improvement in behavioral performance was closely followed by an increase in subjective confidence across the three experimental blocks. (Fig. 4.3)

However, the within-category performance in the frequent category was not uniform. [$F(2, 48) = 14.34, p < 0.001, \eta^2 = .37$] Categorization performance for stimuli in stimulus group Freq3 was significantly worse than in Freq1 [$t(24) = 4.11, p < 0.001, d = 7.3$] or Freq2 [$t(24) = 3.58, p < 0.01, d = 2.2$]. These significant differences also held between Freq1 [$t(24) = 5.37, p < 0.001, d = 11.7$], Freq2 [$t(24) = 4.74, p < 0.001, d = 3.4$] and the Infreq group. There were no significant differences between stimulus groups Freq1 and Freq2 [$t(24) = 1.95, p = .06, d = .37$] or between Freq3 and Infreq [$t(24) = 1.91, p = .67, d = .47$].

The same pattern of differences could be observed in the case of decision confidence. There were significant differences between stimulus groups Freq1, Freq2 and Freq3; and between stimulus groups Freq1, Freq2 and Infreq. However, there was no significant difference either between Freq1 and Freq2 or between Freq3 and Infreq in terms of subjective decision confidence.

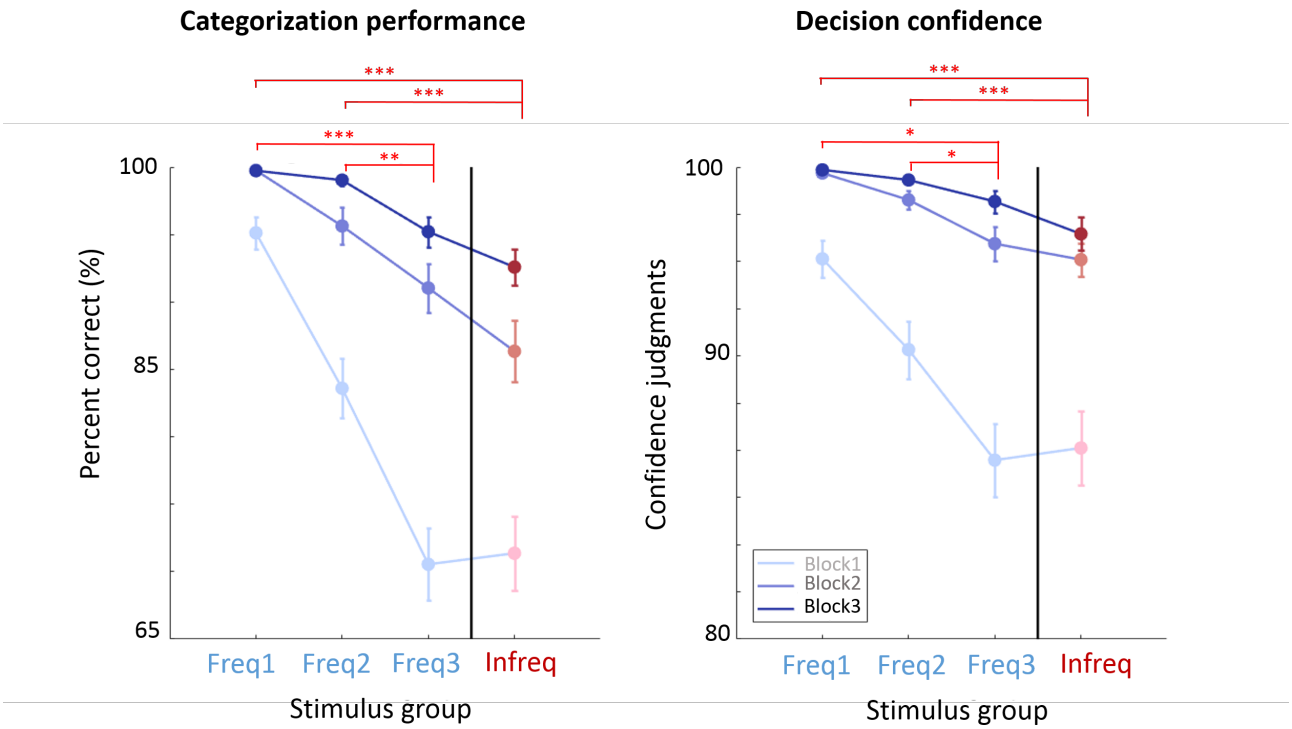


Figure 4.3: Behavioral results of the study: changes in categorization performance (left) and subjective decision confidence (right) as a function of practice – the number of completed blocks – and stimulus groups.

4.3.2 Neural results

4.3.2.1 P300 ERP

A repeated measures two-way ANOVA including the four Stimulus groups and the three Blocks as factors did not reveal significant main effect of Stimulus groups [$F(3, 72) = .82, p = .48, \eta^2 = .03$], nor an interaction between the factors [$F(6, 114) = .35, p = .9, \eta^2 = .01$]. However, I found a significant main effect of Block [$F(2, 48) = 3.88, p < 0.05, \eta^2 = .13$]. Post-hoc t-tests indicated a significant increases in P3 ERD amplitude from Block2 to Block3 [$t(24) = 3.18, p < .01, d = .18$].

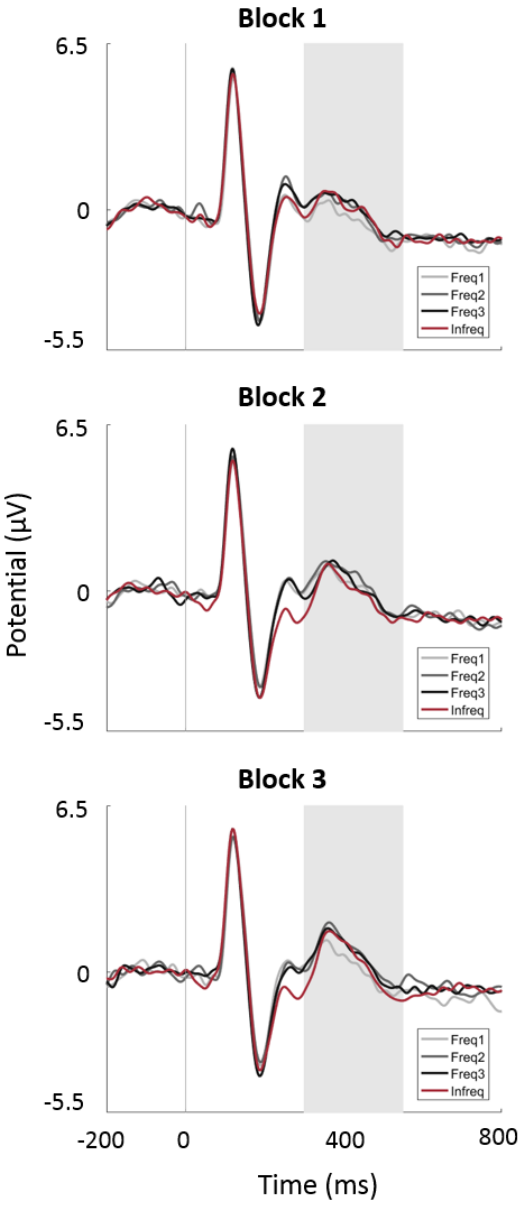


Figure 4.4: Changes in P300 ERP amplitude across the three experimental blocks. Gray area signals the time window of analysis.

4.3.2.2 α ERD

Expected α ERD responses on the left hemisphere developed gradually following the consolidation of categories. In a repeated measures two-way ANOVA, I found a significant interaction between the four stimulus groups and the three blocks in the experiment [$F(6, 144) = 2.96, p < .01, \eta^2 = .12$]. Apart from this interaction, there was also a significant main effect of stimulus

groups [$F(3, 72) = 3.92, p < .05, \eta^2 = .13$]. While in Block1 α power was almost the same in all four stimulus groups, by Block3 significant α ERD differences emerge between stimulus groups on the left hemisphere.

The pattern of these differences closely followed the pattern of behavioral performance of the subjects. Significant differences were found between Freq3 and Freq1 [$t(24) = 2.56, p < .05, d = .39$] as well as Freq2 [$t(24) = 2.7, p < .05, d = .39$] groups, ad similarly between Infreq and Freq1 [$t(24) = 2.66, p < .05, d = .45$] or Freq2 [$t(24) = 2.77, p < .05, d = .38$] stimulus groups. There was no significant difference, however, between α responses for Freq1 and Freq2 [$t(24) = .45, p = .6, d = .08$] or Freq3 and Infreq [$t(24) = .3, p = .7, d = .03$] groups.

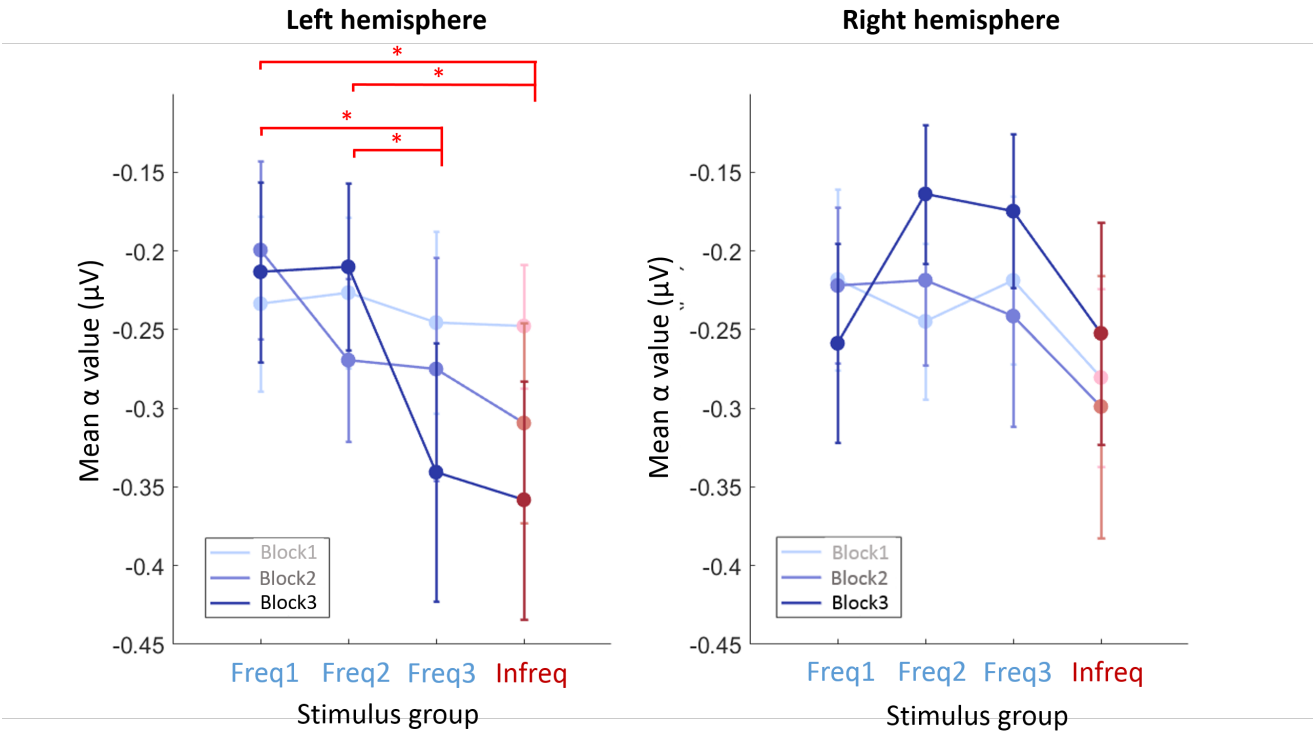


Figure 4.5: Changes in the desynchronization patterns in the α band as a function of practice (number of blocks) and stimulus groups on the left and right hemispheres. Plot depicts mean measured power in the α band with SEM as error bars.

4.3.2.3 θ ERS

The θ responses were markedly different from α ERD responses. Results of a repeated measures two-way ANOVA on the effects of stimulus groups and blocks showed no significant interaction, but indicated a main effect of stimulus groups. [$F(3, 72) = 5.25, p < .01, \eta^2 = .15$]

By Block3, significant θ differences emerged between the Infreq group and stimulus groups Freq1 [$t(24) = 3.57, p < .01, d = .77$], Freq2 [$t(24) = 3.79, p < .01, d = .75$] and Freq3 [$t(24) = 2.39, p < .05, d = .42$]. Unlike in the case of α responses, θ ERS for stimulus groups within the frequent category gradually increased in proportion to the proximity of the stimulus group to the category boundary, but this difference did not reach significance. [$F(2, 48) = 2.028, p = .14, \eta^2 = .08$]

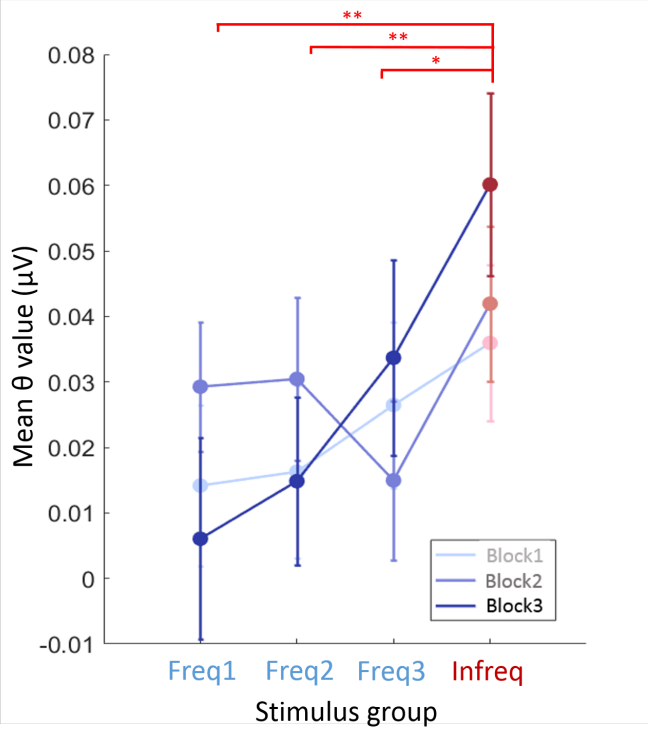


Figure 4.6: Changes in synchronization patterns in the θ band as a function of practice (number of blocks) and stimulus groups. Plot depicts mean measured power in the θ band with SEM as error bars.

4.4 Discussion

In my study, I created a modified version of a widely used paradigm to test if commonly investigated neural correlates of categorization behavior (P300 ERP, α ERD and θ ERS) could be meaningfully interpreted in more complex, more natural scenarios.

Participants gradually built a stable representation about the categories block by block throughout the experiment. Categorization performance was not ceiling for all stimulus groups even in Block3. Stimuli in the Infreq and the Freq3 stimulus groups that was the closest to the category boundary proved to be significantly more difficult for participants to categorize than the rest of the stimuli. This conclusion is supported by the significantly lower confidence for these stimulus groups. Performance and confidence differences between these two groups were not significant, they seemed to be equally difficult.

Acquisition of the categories across blocks was followed by the gradually increasing amplitude of P300 ERPs. Significant P3 ERPs did not emerge between any of the stimulus groups even by the last block. There are two possible explanations for the lack of differences. First, it is possible that more pronounced frequency differences are necessary to elicit P3 differences in such complex tasks. Usual frequency differences in oddball paradigms are 1:9 or 2:8 (Yordanova and Kolev, 1998b; Yordanova, Kolev and Polich, 2001; Yordanova and Kolev, 1998a; Sochurková et al., 2006; Peng et al., 2015; Harper, Malone and Iacono, 2017), etc., and as indicated above, the amplitude of P3 ERP can be modulated by frequency differences between expected and infrequent stimuli. Another possible explanation is that P3 is a neural signal that needs a longer consolidation period of the acquired knowledge. Parise et al. (2018) in their

study during a training phase taught participants two categories consisting of 3-3 novel objects. A test phase followed the training, where participants had to categorize the elements of the recently acquired categories. Crucially, one category contained all 3 elements practiced during the training phase, while in the other only one member of the other category appeared with equal probability to any of the first category's 3 elements. This created a frequency oddball with a 1:3 ratio. The authors found no P3 differences for the newly acquired categories, while in a previous experiment they managed to elicit significant P3 ERD differences presenting familiar objects with the same frequency ratio.

As opposed to P300 ERP, significant α ERDs by Block3 indicated that α ERD was a neural response sensitive even to newly acquired categories. In addition, its gradual emergence makes it a useful tool for tracking the emergence of novel categories. Since the oddball effect can only be interpreted on the category, but not the stimulus level, forming internal representations of the categories is inevitable for eliciting such a difference in neural response. Apart from its gradual emergence, α ERD is not an all-or-none response to different categories. Since the pattern of desynchronization closely followed behavioral categorization performance, we conclude that α ERD is modulated by task difficulty. Crucially, this modulation does not only reflect task difficulty, as in this case, we should have observed the same pattern already in Block1 that we found in Block3. To support this interpretation, I ran an additional analysis by calculating categorization performance and average α power for each individual stimulus in each block. To correct for large individual differences, I normalized both measures for each participant. I found a strong and significant correlation between performance and alpha responses in Block3 [$r(10) = .64, p < .05$].

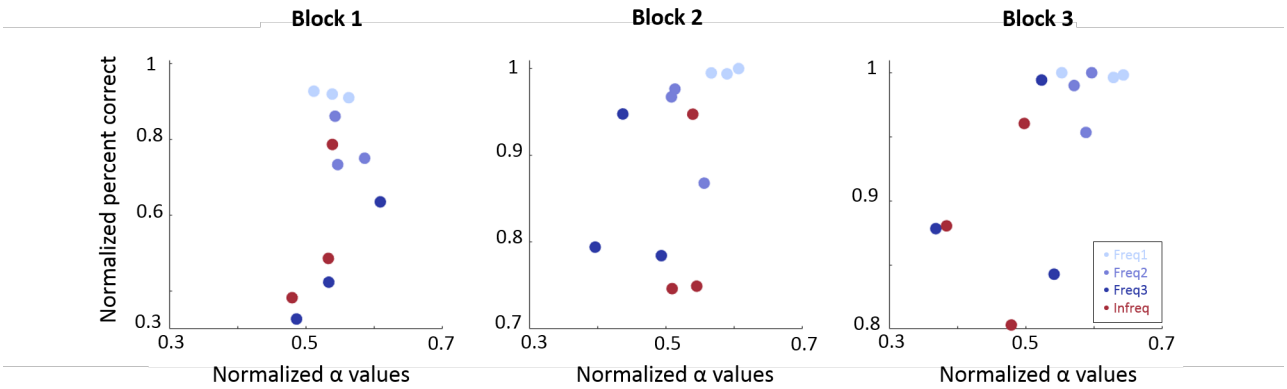


Figure 4.7: Correlation between normalized categorization performance and mean α power for the 12 stimuli used in the experiment. Correlations were calculated for each block separately.

This finding leads to the conclusion that α ERD is a neural signal that is suitable for tracking the ongoing emergence of novel categories, and once the categories are consolidated, it reliably reflects the subjective difficulty of the task. This conclusion is also supported by the previous literature discussed above.

Similar to α ERD, θ ERS responses emerged gradually throughout the experiment, following the course of learning. Significant differences in θ ERS between members of the frequent and infrequent categories clearly signaled the acquisition of categories. In addition, θ ERS seemed to be sensitive to the strength of category membership. Within the frequent category, θ synchronization increased with the proximity of stimulus groups to the category boundary. As stimulus groups showed less similarity to the stimulus group containing the most characteristic stimuli of the frequent category, neural responses tended to be more similar to the ones elicited by the other, infrequent category. Again, the emergence of such a pattern requires the initial consolidation of category representation so that similarity within- and across categories could be interpreted. This interpretation can also be based on previous literature implying that θ ERS is modulated by the novelty or surprise value of a stimulus. The more similar a stimulus is to a

surprising stimulus, the easier it is to confuse one to another, hence, the elicited neural response might also be more similar for the two stimuli

To sum up, just like α ERD, θ ERS emerges gradually following the course of learning, and once the category representations are consolidated, it reflects the strength of within-category membership.

Based on the results of the presented study, we can conclude that P300 ERP is not a useful signal for tracking ongoing emergence of category representations of members with continuously varying feature dimensions. α ERD and θ ERS, however not only follow the process of category acquisition, but they respond to different, but equally important aspects of categorization. α ERD is modulated by subjective task difficulty, and θ ERS reflects the strength of category membership.

In this study, I have demonstrated that commonly investigated neural signals associated with the process of categorization can meaningfully respond to more complex, more natural category structures than the ones previously used in similar studies. These results provide an implicit tool for mapping the emergence, and once consolidated, the structure of category representations in humans. Also, these results support and possibly refine previous hypotheses on the functions of the investigated neural responses.

Chapter 5

General Discussion

In the first part of this thesis, I presented two behavioral studies, each addressing important but surprisingly neglected questions in the line of research on the acquisition of novel categories: whether humans use generative vs. discriminative learning and how they combine unsupervised and supervised information during semi-supervised learning. In the last part of the thesis, I investigated neural signatures of the emergence of novel categories during learning. The results of these studies fit into a coherent view on how information is acquired and represented in the human brain.

5.1 Implications of presented findings

5.1.1 Generative learning

Chapter 2 provided evidence that humans use implicit, automatic generative learning even under circumstances that strongly favor an easier and simpler discriminative learning approach.

The importance of this issue can hardly be overstated. Due to the nature of behavioral experimentation, human behavior is typically investigated scientifically under very constrained and simplified conditions. The result of this practice is a strong overemphasis on humans' ability to solve very specific problems resulting in a distortion of our understanding about what the fundamental problem is that living creatures including humans face on a daily basis. Spectacular and apparently idiosyncratic performances in more complex learning situations have been dismissed as examples that should be considered at a later time when the "basic" learning behavior is clarified based on features of learning identified in simple categorization tasks. As a consequence, discriminative performance has not been interpreted as a special version of a more complex learning apparatus, but as a true building block of human learning. In contrast, the conclusions of Chapter 2 suggest that despite its "higher initial expense", the generative approach is dominant in human learning and solving highly specialized tasks in a discriminative manner builds organically on this basis.

These results also scaffold the necessary predictions and the derived interpretation of the results in the SSL study in Chapter 3. Although, there exist two previous studies suggesting that generative learning is a necessary precursor to SSL (Kalish, Zhu and Rogers, 2015; Gibson et al., 2015), these suggestions remained only hypotheses given the lack of definite evidence supporting these suggestions. Presented results provide evidence that this theoretically conceived prerequisite condition is satisfied, indeed.

Furthermore, the present results of Chapter 2 have two important implications to past and future studies in SSL, respectively. Considering previous studies, if our conclusion, the generative hypothesis saying that humans always learn generatively is correct, it has the potential and obligation to explain so far open questions and correct mistaken conclusions drawn based on

the heavy bias on a simple discriminative interpretation of learning (see Section 3.4). Considering future studies, if experiments are designed so that they provide learners with multitude of incoming information, this would allow future research to proceed to consider more important questions beyond the necessary first step of *whether* learners integrate supervised and unsupervised information, and focus on the true nature of this integration process.

5.1.2 Semi-supervised learning

As a first step on the path opened by results in Chapter 2, Chapter 3 presented the first integrated study on examining the effects of both supervised and unsupervised information on the already existing internal representation of categories. I found that both supervised and unsupervised information get integrated into the final representation. Importantly, although information delivered in a supervised fashion is considered to have a stronger impact on learning than what is learned in an unsupervised manner, supervised information will not overwrite, only modify the information stored in the representation to the extent that is compatible with both old and new information. This is consistent with the generative hypothesis of category learning saying that even supervised learning is essentially unsupervised and thus everything is added up the same way in the resulting representation instead of canceling out.

In addition, this study provides the first definite evidence of a true update of the internal representation as a result of SSL signaled by the re-assignment of certain stimuli into a different category. Although, there exists an earlier attempt for achieving this by Kalish, Zhu and Rogers (2015), the results in their study cannot be generalized to the entire population due to a significant developmental component related to the reported effects. Specifically, the study found that only young children could re-categorize certain stimulus samples, while older participants

failed to update their representations in a way that would signal the incorporation of both supervised and unsupervised information. However, this would suggest that the ability of combining the two types of learning is lost by the time learners arrive to adulthood. In contrast, I suggest that such a combination not only possible in young adults, but it is an essential feature of all human learning.

5.1.3 Neural correlates of category acquisition

Chapter 4 investigated another aspect of complex learning processes, the neural correlates of categorization. My goal with this study was to build on the vast literature of neural correlates (P300 ERP, α ERD and θ ERS), which typically investigated the issue in very simple examples of categorization using a few number of discrete stimuli, and elevate the complexity of the investigated setup to the level of the other two chapters and hence to approximate more natural learning processes. Using multiple samples varying on a continuous feature dimension, separated into two distinct groups allowed for an investigation of not only between category differences, but also for within-category structures (θ ERS), task difficulty (α ERD) and the emergence of the internal representation of the categories being acquired (both).

Based on my results, future studies might use the investigated neural correlates as implicit measures of different descriptive parameters of the internal representation (within-category structure) and the learning environment (task difficulty).

Continuous EEG recording throughout the ongoing acquisition of categories also has important implications to the research of the investigated neural signals. Gradual articulation of α ERD and θ ERS implies that these signals reflect cognitive mechanisms that are involved from the very beginning of the learning process. The lack of P300 ERP modulation both in the

present study and in the one by Parise et al. (2018) for recently acquired, discrete categories suggest that eliciting P300 ERP requires a strong consolidation of categories, hence, it might be involved in cognitive processes related to the processing of incoming information of highly familiar categories.

This lack of significant modulation of P300 ERP might also be relevant to the line of research addressing the causal relationship between – or at least a common modulator of – α ERD and P300 ERP (Peng et al., 2012; Sochurková et al., 2006; Yordanova and Kolev, 1998a; Yordanova, Kolev and Polich, 2001). Using the canonical oddball paradigm with few, highly discrete stimuli, researchers found similar response patterns of the two neural signals. Source and connectivity analysis suggested common neural generators for these neural responses (Peng et al., 2012). The lack of significant P300 ERP modulation, however might put previous results on the suggested connectivity into a new perspective. Presented findings do not refute the existing results on the currently presumed relationship between α ERD and P300 ERP, but they amplify the need for refining the presently accepted theory about what they indicate. In addition, these results confirm and strengthen the proposal that the investigation of these neural correlates require new experimental designs and stimuli, which better capture human category learning under natural circumstances.

5.2 Future directions

5.2.1 Generative learning

The fundamental argument running through this thesis is that learning and using generative models for knowledge acquisition and application is the most suitable method for capturing

human natural cognition. This argument is based on commonly agreed advantages of generative models of the environment, such as the fact that learners would be able to solve multiple tasks using the same, already developed representation about the world/categories. These advantages are supported by multiple lines of research on the transfer of knowledge about categories across different aspects of a learning scenario: the environment (Kole, Healy, Fierman and Bourne Jr., 2010), tasks (Helie and Ashby, 2012), modalities (Wallraven, Bülthoff, Waterkamp, van Dam and Gaißert, 2013) or categories (Qi, Aggarwal, Rui, Tian, Chang and Huang, 2011). However, this argument also comes with a number of further consequences that needs to be integrated in a general framework of human category learning and cognition, in general, to provide a viable model. Specifically, implementing the computational framework of generative models in a biologically feasible manner requires adequate solutions for representing and computing with complex information that encompasses uncertainty (Fiser, Berkes, Orbán and Lengyel, 2010). There are several proposals how this could be done in the brain (Knill and Pouget, 2004; Pouget, Beck, Ma and Latham, 2013; Orbán, Berkes, Fiser and Lengyel, 2016), but very few specific studies that would investigate the consequences and predictions of either of those schemes on human learning either in the context of categorization or decision making.

A related topic is a different type of generalization, automatic re-calibration across various tasks. Categories emerge based on perceptual experience, as mentioned above, are influenced by aspects of the learning scenario. There is also multiple evidence for how categorization (task) affects the perception of stimuli (Goldstone, 1994; Goldstone, Lippaa and Shiffrin, 2001; Op de Beeck, Wagemans and Vogels, 2003; Gauthier et al., 2003). However, little is known about how the initial category representation would change as a result of an additional task. For example, if participants acquired categories with equal sample sizes, would an additional target detection

task on the same stimuli with unequal sample sizes (or increased variance in one of them) of those categories bias the resulting representation about the frequency distribution of previously learned categories? The generative hypothesis would predict so, since an increased exposure to elements from one category should alter the model of the entire learning environment to match expected occurrence probabilities of each category.

The assumption that humans build generative models of their environment, and use these rich, multidimensional representations to solve different tasks emphasizes the importance of research on methods, by which more complex and comprehensive models can be built about the learners' internal representation. As discussed in Section 2.1.1, the same complex, generative internal representation can be used in several different ways that are tailored to the task to be solved at hand. If researchers attempt to retrieve the internal representation only based on how participants solve a single task, results can be misleading. Therefore, a meaningful research program on any learning process can be conducted meaningfully only if the nature of the resulting representation is mapped out to a sufficient degree. Recently, there were several methods developed for retrieving the complex internal model learners' build based on incoming information that can potentially be suitable for this purpose, such as Cognitive Tomography (Houlsby, Huszár, Ghassemi, Orbán, Wolpert and Lengyel, 2013) or d-MCMCP (Hsu, Martin, Sanborn and Griffiths, 2019). However, behavioral experimental designs taking advantage of these methods are still rare, thus, it is imperative that future research put even more emphasis on using and developing such methods.

5.2.2 Semi-supervised learning

Since behavioral studies of SSL represent a relatively recent line of research, there are several relevant aspect of this learning form still virtually unexplored. One of the most important directions would address the developmental aspect of SSL. Arguably, infant and small children know less about the world than adults, but it is unclear whether their different performance in various tasks is simply a result of a difference in knowledge or the method by which stored knowledge is used. To answer this question, a systematic exploration is required to clarify whether unsupervised and supervised elements of SSL are available throughout the developmental process and how those elements interact.

From the moment of birth (and likely even before), humans process incoming information with an attempt to build structured representations about the stream of incoming stimulation. Previous studies established that non-verbal infants are already able to utilize statistical features of the auditory (Saffran, Aslin and Newport, 1996; Saffran, Johnson, Aslin and Newport, 1999) or visual (Fiser and Aslin, 2002) input to extract statistical features that allow them to efficiently build representations about their environment. Such results imply that from an early age, humans are prepared to utilize unsupervised information to gain knowledge about their surroundings. Regarding the supervised component, Csibra and Gergely (2009) argued that infants are already equipped to receive and efficiently process information from a teacher or a supervisor. They are ”sensitive to ostensive signals”, ”develop referential expectations in ostensive contexts” and are ”biased to interpret ostensive-referential communication as conveying information that is kind-relevant and generalizable”. Thus the components of SSL seem to be available from birth.

However, the findings on the interaction between these components are mixed. For example,

LaTourrette and Waxman (2018) reported that infants were able to perform SSL, but only supervised labels could initiate category formation. An important difference potential responsible for this superior effect of supervised samples is that in LaTourrette and Waxman (2018)'s study infants have no access to stimuli from both categories throughout the learning phase, as a result, they did not have access to perceptual and statistical features that would provide them with the opportunity to separate two clusters of stimuli in an unsupervised manner. Instead, infants were trained on multiple samples from one category, and they encountered the sole sample from the other category only during the test.

In contrast, a different study by Kalish, Zhu and Rogers (2015) using a participant pool of older children found the opposite result. Younger children tended to assign less weight to supervised information, and prefer the distribution of the data to determine their final internal representation of the categories. Meanwhile, older children preferred to be lead by supervised information in their categorization behavior. These examples highlight the need for additional systematic research to sort out the apparent controversies between results addressing the developmental aspects of SSL.

A second, related line of research has to target the effect of pedagogical teaching/learning of supervised trials. It has been argued by Shafto, Goodman and Griffiths (2014) that learners tend to infer significantly different hypotheses about the subject to be learned (rule-based, prototype or causally structured concepts), when they assume the incoming information to be sampled by a knowledgeable teacher (pedagogical learning) as opposed to the alternative scenario of simple random sampling, where no such assumption is made. This theory implies that pedagogically sampled supervised information would have a much stronger effect on the final internal representation of concepts than randomly sampled ones. However, a number of questions need

to be clarified before this implication can be tested. What does pedagogical sampling mean computationally? Is it just the signalling of pedagogical attempt or a different type of information as well? If the latter, do these effects add up or interact? On an objective side of the same question, (given the same objective – accuracy, speed, efficiency of generalization, etc.), which are the most informative samples to select when unsupervised data are also available, and what is theoretically the most beneficial order of presentation of such supervised information (Khan, Mutlu and Zhu, 2011; Bengio, Louradour, Collobert and Weston, 2009)? Only after these questions are answered, one can target the issue of how the level of agreement between supervised and unsupervised information about the structure of concepts modulates the final internal representation.

Finally, an interesting area of research is the interaction between the observers goals and utilisation of the knowledge acquired by SSL (i.e. the adaptivity of generalization). The cornerstone argument of the present thesis is that humans make decisions most of the time based on sparse information of previous encounters with certain categories, and hence generalization is the important aspect of human category learning (Segel and Peterson, 2013; Jones, Love and Maddox, 2005). As discussed in Section 1.1.1, a potential benefit of SSL over supervised learning is that it allows for a greater level of generalization than supervised training and discriminative learning alone does. Patterson and Kurtz (2018) have already supported this claim in their study with relational categories. They found that participants were more inclined to more widely generalize the category knowledge they just acquired to newly observed sample and link the sample to earlier ones, when the test followed a training with SSL rather than purely supervised learning. This result is in agreement with numerous earlier observations that the number of supervised samples negatively correlate with participants' willingness to generalize

category information to new samples. It also suggests that the richer representation presumably developed by the unsupervised component of SSL provides the bridge for making generalization through dimensions that were not directly relevant in building the discriminative features in the supervised component. However, all these observations are tied to the sensory input of the situation using only categorization as the task/goal component of the setup. It is likely that similar to the sensory part, the goal/task part of the setup is also a domain where generalization occurs in the same manner but providing more abstract links for what is "in" and what is "out" in the new categorization task. Nevertheless, this component is presently under-explored in the literature.

5.2.3 Neural correlates of categorization

The potentially different sources of generalization also raise the question of how various neural correlates are linked to these sources. Would the values of these neural signals measured with a novel stimulus and with a stimulus that is similar to the novel one but was already represented as a member of an acquired category be similar? The difficulty of answering this question in an oddball paradigm comes from the fact that the oddity of the new stimulus can come from two sources. Either because the stimulus is novel, hence it is odd, or because it is coming from the rare category. Based on previous findings about how α ERD or θ ERS responds to unexpected stimuli in an oddball paradigm, one might expect the neural responses to infrequent stimuli to be similar to the ones elicited by members of the infrequent category, irrespective of their similarity to either the previously learned frequent or infrequent categories. Alternatively, the generalization of category information could be so strong that the neural responses to the novel stimulus that is nevertheless interpreted as a members of the frequent category will be defined

by the previously established expected frequency. It might also be possible that such novelty effects would even elicit a P300 ERP response that I failed to find during the acquisition of the categories in my study.

An important shortcoming of my study related to the above issue is that it cannot disentangle task difficulty and decision confidence (See similar behavioral patterns at Fig. 4.3 left and right) during the process of category learning. Detecting this distinction is necessary for answering whether sensory or categorical factors determine the change in neural signals. Future studies should aim to address these two factors separately, for instance by imposing noise or reducing presentation times for different samples ranging across the stimulus space.

Considering the potential of such neural signals to map within-category structures, it would also be interesting to investigate, whether and how α ERD and θ ERS would reflect the shape of non-uniform distributions of samples in the internal representation of categories. With such a knowledge, it would be possible to further distinguish between two alternatives of how conceptual categories and neural signals are related. For example, one possibility is that if the two categories are defined by non-overlapping Gaussians, the expected probability of the stimuli would modulate the strength of the investigated neural signals. An alternative prediction is that only simple proximity to the category boundary matters and it is the sole predictor of the magnitude of neural responses.

5.3 Conclusions

In the present thesis, I investigated the nature of human knowledge acquisition through the paradigm of semi-supervised learning. My results support the idea that humans use a generative learning strategy in all situations, that labelled and unlabelled information gets continuously and seamlessly integrated in this framework, and that neural correlates commonly used for the investigation of categorization in humans have the potential to reflect much more sophisticated processes relevant to categorization than generally recognized before. These results set the stage for investigating human learning under more complex and more naturalistic conditions.

Bibliography

Alvarez, G.A. (2011). 'Representing multiple objects as an ensemble enhances visual cognition'. In: *Trends in Cognitive Sciences* 3, pp. 1364–6613.

Alvarez, G.A. and A. Oliva (2009). 'Spatial ensemble statistics are efficient codes that can be represented with reduced attention'. In: *PNAS* 18, 7345–7350.

Ashby, F. G. and R. E. Gott (1988). 'Decision rules in the perception and categorization of multidimensional stimuli'. In: *Journal of Experimental Psychology. Learning, Memory, and Cognition* 14, 33–53.

Ashby, F. G., S. Queller and P. M. Berretty (1999). 'On the dominance of unidimensional rules in unsupervised categorization'. In: *Perception & Psychophysics* 61, pp. 1178–1199.

Ashby, F. G. et al. (1998). 'A neuropsychological theory of multiple systems in category learning'. In: *Psychological Review* 105, pp. 442–481.

Ashby, F.G. and W.T. Maddox (2011). 'Human Category Learning 2.0.' In: *Ann N Y Acad Sci.* 1224, 147–161.

Ashby, F.G., W.T. Maddox and C.J. Bohil (2002). 'Observational versus feedback training in rule-based and information-integration category learning'. In: *Memory and Cognition* 30, pp. 666–677.

Ashby, F.G. and J.R.B. O'Brien (2007). 'The effects of positive versus negative feedback on information-integration category learning'. In: *Perception & Psychophysics* 69.

Ashby, F.G. and V.V. Valentin (2017). 'Multiple Systems of Perceptual Category Learning: Theory and Cognitive Tests'. In: 2nd ed., pp. 157–188.

Austerweil, J.L. and T.L. Griffiths (2009). 'The effect of distributional information on feature learning'. In: *Proceedings of the 31st Annual Conference of the Cognitive Science Society*. Ed. by Yolande Berbers and Willy Zwaenepoel.

Azizian, A. et al. (2006a). 'Beware misleading cues: Perceptual similarity modulates the N2/P3 complex'. In: *Psychophysiology* 43, pp. 253–260.

Azizian, A. et al. (2006b). 'Electrophysiological correlates of categorization: P300 amplitude as index of target similarity'. In: *Biological Psychology* 71, pp. 278–288.

Barlow, H. B. (1989). *Neural Computation*. Cambridge: MIT Press.

Basar, E. et al. (2001). 'Gamma, alpha, delta, and theta oscillations govern cognitive processes'. In: *International Journal of Psychophysiology* 39, pp. 241–248.

Bays, B.C. et al. (2015). 'Alpha-band EEG activity in perceptual learning'. In: *Journal of Vision* 15.10, pp. 1–12.

Behbahani, F.M.P. and A.A. Faisal (2012). 'Human category learning is consistent with Bayesian generative but not discriminative classification strategies'. In: doi: 10.3389/conf.fncom.2012.55.00095. Bernstein Conference 2012.

Bengio, Y. et al. (2009). 'Curriculum learning'. In: *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 41–48.

Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Bornstein, M.H. (1984). 'A descriptive taxonomy of psychological categories used by infants'. In: *Origins of cognitive skills*. Ed. by C. Sophian. Hillsdale, NJ: Erlbaum, 313–338.

Canini, K.R. and T.L. Griffiths (2011). 'A nonparametric Bayesian model of multi-level category learning'. In: *Proceedings of the 25th AAAI Conference on Artificial Intelligence*. AAAI.

Canini, K.R., M. M. Shashkov and T.L. Griffiths (2010). 'Modeling Transfer Learning in Human Categorization with the Hierarchical Dirichlet Process'. In: *Proceedings of the 27th International Conference on Machine Learning*. ICML.

Cavanagh, J.F. and M.J. Frank (2014). 'Frontal theta as a mechanism for cognitive control'. In: *Trends in Cognitive Sciences* 18, pp. 414–421.

Chong, S.C. and A. Treisman (2003). 'Representation of statistical properties'. In: *Vision Research* 43, 393–404.

— (2005). 'Statistical processing: computing the average size in perceptual groups'. In: *Vision Research* 45, 891–900.

Csibra, G. and Gy. Gergely (2009). 'Natural Pedagogy'. In: *Trends in Cognitive Science* 13.4, pp. 148–153.

de Haan, M. and C.A. Nelson (1998). 'Discrimination and categorization of facial expressions of emotion during infancy'. In: *Perceptual Development: Visual, Auditory, and Lan-*

guage Development in Infancy. Ed. by A. Slater. London: University College London Press, pp. 287–309.

Debener, S. et al. (2002). 'Auditory novelty oddball allows reliable distinction of top-down and bottom-up processes of attention'. In: *International Journal of Psychophysiology* 46, pp. 77–84.

Donchin, E. (1981). 'Presidential Address, 1980: Surprise!...Surprise?' In: *Psychophysiology* 18.5, pp. 493–513.

Ell, S.W. and F.G. Ashby (2006). 'The effects of category overlap on information-integration and rule-based category learning'. In: *Perception & Psychophysics* 68, 1013–1026.

Engel, A.K., P. Fries and W. Singer (2001). 'Dynamic predictions: Oscillations and synchrony in top-down processing'. In: *Nature Reviews Neuroscience* 2, pp. 704–716.

Fiser, J. and R.N. Aslin (2002). 'Statistical learning of new visual feature combinations by infants'. In: *PNAS* 99.24, pp. 15822–15826.

Fiser, J. et al. (2010). 'Statistically optimal perception and learning: from behavior to neural representations'. In: *Trends in Cognitive Science* 14.3, pp. 119–130.

Freedberg, M. et al. (2017). 'Comparing the effects of positive and negative feedback in information-integration category learning'. In: *Memory & Cognition* 45.1, 12–25.

Gaißert, N. et al. (2012). 'Haptic Categorical Perception of Shape'. In: *PlosOne* 7, pp. 1–7.

Gauthier, I. et al. (2003). 'The influence of conceptual knowledge on visual discrimination'. In: *Cognitive Neuropsychology* 20, pp. 507–523.

Gibson, B.R. et al. (2015). 'What causes category-shifting in human semi-supervised learning?' In: *Proceedings of the 34th Annual Conference of the Cognitive Science Society*. Ed. by N. Miyake, D. Peebles and R.P. Cooper. AAAI. Austin, TX: Cognitive Science Society.

Goldstone, R. (1994). 'Influences of categorization on perceptual discrimination'. In: *Journal of Experimental Psychology: General* 123, pp. 178–200.

Goldstone, R.L., Y. Lippaa and R.M. Shiffrin (2001). 'Altering object representations through category learning'. In: *Cognition* 78, pp. 27–43.

Goudbeek, M. et al. (2005). 'Acquiring auditory and phonetic categories'. In: *Handbook of categorization in cognitive science*. Ed. by C. Lefebvre and H. Lefebvre. Amsterdam: Elsevier, pp. 497–513.

Gregory, R.L. (1997). 'Visual illusions classified'. In: *Trends in Cognitive Sciences* 1.5, pp. 190–194.

Gureckis, T.M. and B.C. Love (2003a). 'Human supervised and unsupervised learning as a quantitative distinction'. In: *International Journal of Pattern Recognition and Artificial Intelligence* 17.5, pp. 885–901.

— (2003b). 'Towards a unified account of supervised and unsupervised category learning'. In: *Journal of Experimental & Theoretical Artificial Intelligence* 15.1, pp. 1–24.

Handel, S. and S. Imai (1972). 'The free classification of analyzable and unanalyzable stimuli'. In: *Perception & Psychophysics* 12, 108–116.

Harnad, S. (2003). 'Categorical Perception'. In: *Encyclopedia of Cognitive Science*. Ed. by L. Nadel. Nature Publishing Group: Macmillan, pp. 497–513.

— (2005). 'To cognize is to categorize: Cognition is categorization'. In: *Handbook of categorization in cognitive science*. Ed. by H. Cohen and C. Lefebvre. Amsterdam: Elsevier.

Harper, J., S.M. Malone and W.G. Iacono (2017). 'Theta- and delta-band EEG network dynamics during a novelty oddball task'. In: *Psychophysiology* 54, pp. 1590–1605.

Helie, S. and F.G. Ashby (2012). 'Learning and transfer of category knowledge in an indirect categorization task'. In: *Psychological Research* 76, pp. 292–303.

Heller, K.A., A. Sanborn and N. Chater (2009). 'Hierarchical Learning of Dimensional Biases in Human Categorization'. In: *In Neural Information Processing Systems*.

Hinton, G. and T.J. Sejnowski (1999). *Unsupervised Learning*. Cambridge, Massachusetts: MIT Press.

Houlsby, N.M.T. et al. (2013). 'Cognitive Tomography Reveals Complex, Task-Independent Mental Representations'. In: *Current Biology* 23.21, pp. 2169–2175.

Hsu, A.S. and T.L. Griffiths (2009). 'Differential Use of Implicit Negative Evidence in Generative and Discriminative Language Learning'. In: *Advances in Neural Information Processing Systems*. Vol. 22.

— (2010). 'Effects of generative and discriminative learning on use of category variability'. In: *Proceedings of the 32nd annual conference of the Cognitive Science Society*. Ed. by R. Catrambone S. Ohlsson. Vol. 22. Austin, TX: Cognitive Science Society.

Hsu, A.S. et al. (2019). 'Identifying category representations for complex stimuli using discrete Markov chain Monte Carlo with people'. In: *Behav Res.* 51, 1706–1716.

Jones, M., B.C. Love and W.T. Maddox (2005). 'Stimulus Generalization in Category Learning'. In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 27. Cognitive Science Society, pp. 1066–1071.

Jones, M., B.C. Love and W.T. Maddox (2006). 'Recency effects as a window to generalization: Separating decisional and perceptual sequential effects in category learning'. In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 32.2, pp. 316–332.

Kalish, C.W., X. Zhu and T.T. Rogers (2015). 'Drift in children's categories: When experienced distributions conflict with prior learning'. In: *Developmental Science* 18.6, pp. 940–956.

Kalish, C.W. et al. (2011). 'Can semi-supervised learning explain incorrect beliefs about categories?' In: *Cognition* 120.1, 106–118.

Kamp, S-M. and E. Donchin (2015). 'ERP and pupil responses to deviance in an oddball paradigm'. In: *Psychophysiology* 52.4, pp. 460–71.

Keller, A.S., L. Payne and R. Sekuler (2017). 'Characterizing the roles of alpha and theta oscillations in multisensory attention'. In: *Neuropsychologia* 99, pp. 48–63.

Khan, F., B. Mutlu and J Zhu (2011). 'How Do Humans Teach: On Curriculum Learning and Teaching Dimension'. In: *Advances in Neural Information Processing Systems 24*. Ed. by J. Shawe-Taylor et al. Curran Associates, Inc., pp. 1449–1457.

Klimesch, W. (1999). 'EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis'. In: *Brain Research Reviews*, pp. 169–195.

— (2012). 'α-band oscillations, attention, and controlled access to stored information'. In: *Trends in Cognitive Sciences* 16, pp. 606–617.

Klimesch, W., M. Doppelmayr and S. Hanslmayr (2006). 'Upper alpha ERD and absolute power: their meaning for memory performance'. In: *Prog. Brain Res.* 159, pp. 151–165.

Klimesch, W., P. Sauseng and S. Hanslmayr (2007). 'EEG alpha oscillations: The inhibition-timing hypothesis'. In: *Brain Research Reviews* 53, pp. 63–88.

Klimesch, W. et al. (1998). 'Induced alpha band power changes in the human EEG and attention'. In: *Neuroscience Letters* 244, pp. 73–76.

Knill, D.C. and A. Pouget (2004). 'The Bayesian brain: the role of uncertainty in neural coding and computation'. In: *Trends Neurosci.* 27.12, pp. 712–79.

Kok, A. (2001). 'On the utility of P3 amplitude as a measure of processing capacity'. In: *Psychophysiology* 38, pp. 557–577.

Kole, J.A. et al. (2010). 'Contextual memory and skill transfer in category search'. In: *Memory and Cognition* 38.1, pp. 67–82.

Kolev, V. et al. (1999). 'Event-related alpha oscillations in task processing'. In: *Clinical Neurophysiology* 110, pp. 1784–1792.

Lake, B. and J. McClelland (2011). 'Estimating the strength of unlabeled information during semi-supervised learning'. In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Cognitive Science Society, pp. 1400–1405.

Lasserre, J.A., C.M. Bishop and T.P. Minka (2006). 'Principled hybrids of generative and discriminative models'. In: *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Washington, DC, USA: IEEE Computer Society, 87–94.

LaTourrette, A. and S.R. Waxman (2018). 'A little labeling goes a long way: Semi-supervised learning in infancy'. In: *Developmental Science* 22.1, pp. 1–9.

Lengyel, M. et al. (2015). 'On the role of time in perceptual decision making'. In: *ArXiv*.

Levering, K.R. and K.J. Kurtz (2015). 'Observation versus classification in supervised category learning'. In: *Memory & Cognition* 43.2, 266–282.

Li, F.F. et al. (2002). 'Rapid natural scene categorization in the near absence of attention'. In: *PNAS* 99.14, pp. 9596–9601.

Liao, H-I. et al. (2016). 'Human Pupillary Dilation Response to Deviant Auditory Stimuli: Effects of Stimulus Properties and Voluntary Attention'. In: *Front Neurosci* 10, p. 43.

Liu, S.T. et al. (2019). 'Optimal features for auditory categorization'. In: *Nat Commun* 10, p. 1302.

Locatelli, F.F., P.C. Fernandez and B.H. Smith (2016). 'Learning about natural variation of odor mixtures enhances categorization in early olfactory processing'. In: *Journal of Experimental Biology* 219, pp. 2752–2762.

Love, B.C. (2002). 'Comparing supervised and unsupervised category learning'. In: *Psychon Bull Rev* 9.4, pp. 829–835.

— (2003). 'The multifaceted nature of unsupervised category learning'. In: *Psychon Bull Rev* 10, 190–197.

Luck, S. J. (2005). *An Introduction to the Event-Related Potential Technique*. The MIT Press.

Maddox, W.T., F.G. Ashby and C.J. Bohil (2003). 'Delayed feedback effects on rule-based and information-integration category learning'. In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 29.4, pp. 650–662.

Maddox, W.T., F.G. Ashby and A.D. Pickering (2004). 'Disrupting feedback processing interferes with rule-bsed but not information-integration category learning'. In: *Memory & Cognition* 32, pp. 582–591.

Maddox, W.T. and J.L. Dodd (2003). 'Separating perceptual and decisional attention processes in the identification and categorization of integral-dimension stimuli'. In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 29.3, 467–480.

Markant, D.B. and T.M. Gureckis (2014). 'Is it better to select or to receive? Learning via active and passive hypothesis testing'. In: *Journal of Experimental Psychology: General* 143.1, pp. 94–122.

McDonnell, J.V., C.A. Jew and T.M. Gureckis (2012). 'Sparse category labels obstruct generalization of category membership'. In: *Proceedings of the 34th annual conference of the cognitive science society*.

Mitchell, P. et al. (2005). 'How perception impacts on drawings'. In: *J Exp Psychol Hum Percept Perform* 31.5, pp. 996–1003.

Ng, A.Y. and M. Jordan (2001). 'On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes'. In: *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press.

Nosofsky, R.M. and T.J. Palmeri (1996). 'Learning to classify integral-dimension stimuli'. In: *Psychonomic Bulletin & Review* 3.2, 222–226.

Op de Beeck, H., J. Wagemans and R. Vogels (2001). 'Inferotemporal neurons represent low-dimensional configurations of parameterized shapes'. In: *Nature Neuroscience* 4, pp. 1244–1252.

— (2003). 'The Effect of Category Learning on the Representation of Shape: Dimensions Can Be Biased but Not Differentiated'. In: *Journal of Experimental Psychology: General* 132.4, pp. 491–511.

Orbán, G. et al. (2016). 'Neural Variability and Sampling-Based Probabilistic Representations in the Visual Cortex'. In: *Neuron* 92.2, pp. 530–543.

Parise, E. et al. (2018). 'Label-induced categorization of unrelated objects in adults and preverbal infants'. In: *Unpublished manuscript*.

Patterson, J.D. and K.J. Kurtz (2018). 'Semi-supervised learning: A role for similarity in generalization-based learning of relational categories'. In: *Proceedings of the 40th Annual Conference of the Cognitive Science Society*. Ed. by C. Kalish et al. Austin, TX: Cognitive Science Society, pp. 2211–2217.

Peng, W. et al. (2012). 'Causality in the Association between P300 and Alpha Event-Related Desynchronization'. In: *PLoS ONE* 7.4.

Peng, W. et al. (2015). 'Widespread cortical alpha-ERD accompanying visual oddball target stimuli is frequency but non-modality specific'. In: *Behavioural Brain Research* 295, pp. 71–77.

Pevtzow, R. and S. Harnad (1997). 'Warping similarity space in category learning by human subjects: the role of task difficulty'. In: *Proceedings of SimCat 1997: Interdisciplinary Workshop on Similarity and Categorization*. Ed. by M. Ramscar et al., pp. 189 –195.

Picton, T.W. (1992). 'The P300 Wave of the Human Event-Related Potential'. In: *Journal of Clinical Neurophysiology* 9.4, pp. 456–479.

Polich, J. (2007). 'Updating P300: An integrative theory of P3a and P3b'. In: *Clinical Neurophysiology* 118.10, pp. 2128–2148.

Pothos, E.M. and N. Chater (2002). 'A simplicity principle in unsupervised human categorization'. In: *Cognitive Science* 26, pp. 303–343.

Pothos, E.M., D.J. Edwards and A. Perlman (2011). 'Supervised vs. unsupervised categorization: Two sides of the same coin?' In: *Quarterly Journal of Experimental Psychology* 64, pp. 1692–1713.

Pouget, A. et al. (2013). 'Probabilistic brains: knowns and unknowns'. In: *Nature Neuroscience* 16, pp. 1170–1178.

Qi, G.-J. et al. (2011). 'Towards Cross-Category Knowledge Propagation for Learning Visual Concepts'. In: *Proc. IEEE Intl Conf. Computer Vision and Pattern Recognition*, 897–904.

Saffran, J.R., R.N. Aslin and E.L. Newport (1996). 'Statistical Learning by 8-Month-Old Infants'. In: *Science* 274, pp. 1926–1928.

Saffran, J.R. et al. (1999). 'Statistical learning of tone sequences by human infants and adults'. In: *Cognition* 70.1, pp. 27–52.

Sanei, S. and J.A. Chambers (2007). *EEG signal processing*. John Wiley & Sons. Chichester: England.

Schindel, R., J. Rowlands and D.H. Arnold (2011). 'Effects of aging on event?related brain potentials and reaction times in an auditory oddball task'. In: *Journal of Vision* 11.2.

Schwarzer, G., I. Küfer and F. Wilkening (1999). 'Learning categories by touch: On the development of holistic and analytic processing'. In: *Memory & Cognition* 27, pp. 868–877.

Segel, C.A. and E.J. Peterson (2013). 'Categorization = Decision Making + Generalization'. In: *Neurosci Biobehav Rev* 37.7, pp. 1187–2000.

Shafto, P., N.D. Goodman and T.L. Griffiths (2014). 'A rational account of pedagogical reasoning: Teaching by, and learning from, examples'. In: *Cognitive Psychology* 71, 55–89.

Shepard, R.N. (1991). 'Integrality versus separability of stimulus dimensions: From an early convergence of evidence to a proposed theoretical basis'. In: *The perception of structure: Essays in honor of Wendell R. Garner*. Ed. by G.R. Lockhead and J.R. Pomerantz. Washington, DC, US: American Psychological Association, pp. 53–71.

Sochurková, D. et al. (2006). 'P3 and ERD/ERS in a Visual Oddball Paradigm'. In: *Journal of Psychophysiology* 20, pp. 32–39.

Squires, N.K., K.C. Squires and S.A. Hillyard (1975). 'Two varieties of long-latency positive waves evoked by unpredictable auditory stimuli in man'. In: *Electroencephalogr Clin Neurophysiol* 38.4, pp. 387–401.

Stephens, R.G. and M.L. Kalish (2018). 'The effect of feedback delay on perceptual category learning and item memory: Further limits of multiple systems'. In: *Journal of Experimental Psychology: Learning Memory and Cognition* 44, pp. 1397–1413.

Sun, R. et al. (2007). 'The interaction of implicit learning, explicit hypothesis testing learning and implicit-to-explicit knowledge extraction'. In: *Neural Networks* 20, pp. 34–47.

Sutoh, T. et al. (2000). 'Event-related desynchronization during an auditory oddball task'. In: *Clinical Neurophysiology* 111, pp. 858–862.

Taylor, K.I. et al. (2012). 'Contrasting effects of feature-based statistics on the categorisation and basic-level identification of visual objects'. In: *Cognition* 122, pp. 363–374.

Van Rullen, R. et al. (2011). 'Ongoing EEG phase as a trial-by-trial predictor of perceptual and attentional variability'. In: *Front. Psychol.* 2.30.

Vandist, K., M. De Schryver and Y. Rosseel (2009). 'Semisupervised category learning: The impact of feedback in learning the information-integration task'. In: *Attention, Perception, & Psychophysics* 71.2, 328–341.

Vandist, K., G. Storms and E. Van den Bussche (2019). 'Semisupervised category learning facilitates the development of automaticity'. In: *Attention, Perception, & Psychophysics* 81, 137–157.

Vázquez-Marrufo, M. et al. (2017). 'Retest reliability of individual alpha ERD topography assessed by human electroencephalography'. In: *PLOS ONE*, pp. 1–15.

Vong, W.K., D.J. Navarro and A. Perfors (2015). 'The helpfulness of category labels in semi-supervised learning depends on category structure'. In: *Psychonomic Bulletin & Review*, pp. 1–9.

Wallraven, C. et al. (2013). 'The eyes grasp, the hands see: Metric category knowledge transfers between vision and touch'. In: *Psychonomic Bulletin & Review* 21, 976–985.

Yordanova, J. and V. Kolev (1998a). 'Event-related alpha oscillations are functionally associated with P300 during information processing.' In: *NeuroReport* 9, pp. 3159–3164.

— (1998b). 'Single-sweep analysis of the theta frequency band during an auditory oddball task'. In: *Psychophysiology* 35, pp. 116–126.

Yordanova, J., V. Kolev and B. Polich (2001). 'P300 and alpha event-related desynchronization (ERD)'. In: *Psychophysiology* 38, pp. 143–152.

Zeithamova, D. and W.T. Maddox (2009). 'Learning Mode and Exemplar Sequencing in Unsupervised Category Learning'. In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 35.3, pp. 731–741.

Zhu, X. (2005). 'Semi-Supervised Learning Literature Survey'. In: *Proceedings of the 21st conference on artificial intelligence*. Ed. by R.C. Holte and A. Howe. Menlo Park, CA: The AAAI Press, 864–870.

Zhu, X. et al. (2007). 'Humans perform semi-supervised classification too'. In: *AAAI*. The AAAI Press, 864–870.

Ziori, E. and Z. Dienes (2012). 'The time course of implicit and explicit concept learning'. In: *Consciousness and Cognition* 21, pp. 204–216.