VRIJE UNIVERSITEIT BRUSSEL

FACULTEIT PSYCHOLOGIE EN EDUCATIEWETENSCHAPPEN

BRAIN, BODY & COGNITION

Semisupervised category learning

Katleen Vandist

Proefschrift aangeboden tot het verkrijgen van de graad Doctor in de psychologische wetenschappen aan de Vrije Universiteit Brussel en van de graad Doctor in de Psychologie aan de KU Leuven.

2019

Promotoren:
Prof. dr. Tim Vantilborgh
Prof. dr. Eva Van den Bussche
Prof. dr. Gert Storms

KU LEUVEN

FACULTEIT PSYCHOLOGIE EN PEDAGOGISCHE WETENSCHAPPEN

BRAIN & COGNITION

Katleen Vandist. **Semisupervised category learning.**

Dissertation submitted in order to obtain the degree of Doctor in Psychological Sciences and Doctor in Psychology, 2019.
Supervisors: Tim Vantilborgh, Eva Van den Bussche and Gert Storms

---

In the human category learning literature, category learning is typically investigated in a supervised or an unsupervised way. Supervised category learning involves that participants receive feedback after each encounter of a category member, whereas unsupervised learning implies that no information about the category label is ever provided. However, both forms of category learning seem ecological implausible. In the current dissertation, we argue that in real life information about the category is provided occasionally, implying that humans learn in a *semisupervised* way. This semisupervised category learning is explored in several ways.

In chapter II, a semisupervised classification learning paradigm was tested using the information-integration category structure. Participants learned in a semisupervised way and their performance was compared to supervised and unsupervised learners. Almost all semisupervised learners obtained high accuracy performance, comparable to the supervised learners. Our results showed that a sufficient amount of feedback is essential for successful learning. Remarkably, early in the learning process trials that were not followed by feedback did not have an impact.

Chapter III focused on the effect of semisupervised learning late in the learning process, when automaticity develops. All participants were

first trained supervisedly and only expert learners were allowed to proceed to the actual experiment. The results showed that at the end of the experiment the semisupervised learners categorized stimuli significantly faster than the supervised learners, even when the total number of trials in the supervised and semisupervised conditions was identical. Thus, late in learning the no-feedback trials did have an impact: they accelerated the development of automaticity.

In Chapter IV semisupervised learning was investigated in the A not A category structure, early in the learning process. Again, semisupervised learning was successful (resulting in high accuracy rates), when at least 25% of the trials was followed by feedback. The no-feedback trials did not have an impact.

Chapter V provides an overview of the obtained empirical results. Integrating the results of this dissertation, three striking conclusions could be made. First, semisupervised learning is successful. Second, the onset of the no-feedback trials in the learning process seems crucial: when implemented early in learning, these trials have no impact; when implemented late in learning, the no-feedback trials accelerate the learning process. Third, in every study individual differences in category learning were observed. These findings are situated in the broader literature and limitations of the studies are discussed. Finally, future research ideas are suggested.

Katleen Vandist. **Gedeeltelijk gesuperviseerd leren van categorieën.**

Proefschrift aangeboden tot het verkrijgen van de graad van Doctor in de Psychologische Wetenschappen en Doctor in de Psychologie, 2019.
Promotoren: Tim Vantilborgh, Eva Van den Bussche and Gert Storms

---

In de wetenschappelijke literatuur over het leren van categorieën bij mensen valt het op dat het leren van categorieën ofwel gesuperviseerd ofwel ongesuperviseerd onderzocht wordt. Gesuperviseerd leren van categorieën houdt in dat de proefpersoon feedback krijgt na iedere ervaring met een element van een categorie terwijl ongesuperviseerd leren juist inhoudt dat er nooit enige informatie over het categorielabel wordt gegeven. Beide vormen van categorie-leren lijken onwaarschijnlijk in het dagelijks leven. In deze thesis argumenteren we dat in het dagelijks leven informatie over categorieën af en toe wordt gegeven, wat inhoudt dat mensen op een gedeeltelijk gesuperviseerde manier leren. Dit gedeeltelijk gesuperviseerd leren werd onderzocht op verschillende manieren.

In hoofdstuk 2 werd een gedeeltelijk gesuperviseerd leerparadigma ontwikkelt en getoetst in de informatie-integratie categoriestructuur. Proefpersonen leerden onder gedeeltelijke supervisie en hun prestaties werden vergeleken met die van proefpersonen in een gesuperviseerde en een ongesuperviseerde conditie. Bijna alle proefpersonen in de gedeeltelijk gesuperviseerde conditie behaalden een hoge accuraatheid, vergelijkbaar met de gesuperviseerde conditie. Onze resultaten toonden ook aan dat een voldoende hoeveelheid feedback essentieel is om succesvol te leren. Opmerkelijk was wel dat vroeg in het leerproces de

trials die niet gevolgd werden door feedback geen impact hadden op het leren.

Hoofdstuk 3 focuste op het effect van gedeeltelijk gesuperviseerd leren laat in het leerproces, op het moment dat automatisme ontwikkelt. Alle proefpersonen werden eerst gesuperviseerd getraind en enkel zij die een expertniveau bereikten werden toegelaten tot het eigenlijke experiment, waar ze het leren ofwel gedeeltelijk gesuperviseerd of volledig gesuperviseerd verderzetten. De resultaten toonden aan dat op het einde van het experiment de gedeeltelijk gesuperviseerde proefpersonen stimuli significant sneller categoriseerden dan de gesuperviseerde proefpersonen, zelfs indien de totale hoeveelheid trials in beide condities gelijk was. Bijgevolg hadden de trials die niet gevolgd werden door feedback laat in het leerproces dus wel een impact: ze versnelden de ontwikkeling van automatisme.

In hoofdstuk 4 werd gedeeltelijk gesuperviseerd leren onderzocht in de ‘A niet A’ categoriestructuur en dit vroeg in het leerproces. Opnieuw was gedeeltelijk gesuperviseerd leren succesvol (aangetoond door een hoge accuraatheid) op voorwaarde dat minstens 25% van de trials gevolgd werd door feedback. De trials die niet werden gevolgd door feedback hadden geen impact.

Hoofdstuk 5 geeft een overzicht van de bekomen empirische resultaten. Wanneer de resultaten van deze thesis geïntegreerd worden kunnen er drie opmerkelijke besluiten getrokken worden. Ten eerste is gedeeltelijk gesuperviseerd leren succesvol. Ten tweede blijkt dat het moment van aanbieden van trials die niet gevolgd worden door feedback in het leerproces cruciaal is voor hun impact op het leren. Wanneer deze

vroeg in het leerproces aangeboden worden, hebben deze trials geen impact; wanneer ze laat in het leerproces geïmplementeerd worden versnellen ze het leerproces. Ten derde werden in iedere studie individuele verschillen in het leren van categorieën vastgesteld. Deze bevindingen worden gesitueerd in de wetenschappelijke literatuur en de tekortkomingen van de studies worden besproken. Tot slot worden richtlijnen voor toekomstig onderzoek voorgesteld.

Acknowledgments

Finally. I made it.

In the several years that I worked on this PhD-project, it was doubtful whether it would ever be accomplished. But yet, here it is, you are holding the book of my PhD-project into your hands. This work had become a personal conquest and I am grateful for the people who believed in me. Their positive vibes motivated me to continue.

A huge thanks to Eva Van den Bussche for many, many reasons. To name a few: her constant positive encouragements, for providing the necessary funding from time to time, for giving me the opportunity to improve my work by funding an extra experiment,…

The second major thanks is to Gert Storms. I was always impressed by his academic knowledge and his positive and realistic words helped me to regain my academic self-belief.

Thanks to the authors who cited my first publication. The observation that they noticed and appreciated my work motivated me. Thanks to them my confidence grew that my work and ideas were interesting. In particular I want to thank to Ashby for the enthusiastic feedback that I received after submitting a paper and to Maddox to let me participate to the International meeting of the Psychonomic Society in 2009.

Last but not least, I want to thank Bjorn, my partner, for allowing me to take time to find out whether the failure to finish earlier was due to the limits of my academic capacities or the false start of this PhD-project.

This work is dedicated to the many people who inspired me. During the years, I have been most fortunate to have been offered the empathy, sympathy, support, intelligence and perseverance of so many people. I am more grateful than I can say to:

Franky, Anneleen, Johan and Kris from Ghent University
Elke, Charlotte, Bart, Inge and Eva from Vrije Universiteit Brussel and KULeuven
Fem, Marleen, Griet and so many other colleagues for rebuilding my professional self-confidence
The nice moments with several students with (dis)abilities and their perseverance to conquer the unfairness.
My many (old) friends like Sarah, Steven, Jeroen, Marlies, Elise, Liesje, Ilse,…
Zita, Robbe and ☺
My family

I hope that some day this work can serve as an inspiration as well.

Contents

1. Introduction **1**

The importance of categorization 2
About a category 3
Human category learning 5
Category learning paradigms 6
Structure of this dissertation 19

2. Semisupervised category learning: the impact of feedback in learning the information-integration task **29**

Introduction 31
Experiment 1 36
Experiment 2 53
General Discussion 62
Appendix 68

3. Semisupervised category learning facilitates the development of automaticity **75**

Introduction 77
Experiment 1 88
Experiment 2 108
General Discussion 125
Appendix 132

4. Semisupervised category learning: the impact of feedback investigated in the prototype distortion category structure **143**

Introduction 145
Experiment 1 154
Experiment 2 165
Experiment 3 170
Experiment 4 174
General Discussion 178

5. General Discussion **189**
Goal of this dissertation 190
Overview of the empirical results 190
Link to the broader literature 196
Link to semisupervised category learning models 206
Limitations of the studies and future directions 209
Implications of the obtained results for daily life 215

Chapter 1

Introduction

The importance of categorization

Is this plant edible or poisonous? Was that sound made by a predator or by the wind? All animals assign objects and events in the environment to separate classes or categories. This ability is evolutionary important: it allows us to respond adequately and fast in case of danger. Any species lacking this ability would quickly become extinct.

Although there are countless objects and events in the world, the ways that people use to categorize them are restricted. To be of use, categorization has to be highly constrained because the number of ways of partitioning unclassified objects grows exponentially with each additional object. For example, one can partition 3 objects in 5 ways, 4 objects in 18 ways, 5 objects in 56 ways and 10 objects in more than 100.000 ways (e.g., Heit & Bott, 2000; Medin & Ross, 1997). Considering the number of objects in the world, it is clear that our categories constitute only the tiniest subset of all possible partitionings. Moreover, categorization has the amazing advantage that not all features of objects or events have to be processed completely to activate the stored-up knowledge and to react adequately. To see an animal with sharp teeth, for instance, is enough to know it can bite and to keep distance. Categorization is quick and very useful.

Also for us, humans, categorization plays an important role in our daily lives. We will not cross the road if we hear the siren of an ambulance. When we interact with others we ask ourselves if we can trust the other person or not. When we read a handwritten letter, even in a handwriting we've never seen before, we are permanently recognizing and forming letters to meaningful words. *How* we categorize is one of

the most fundamental problems of cognitive science. In fact, categorization is the basis for the construction of our knowledge of the world. Hence, the scientific literature on categorization is enormous and has a long history. The study of categorization has long been at the core of the founding disciplines of the cognitive sciences, like anthropology, computer science, linguistics, neuroscience, philosophy and psychology (Ashby & Maddox, 2005; Markman & Ross, 2003). All these different angles led to a myriad of definitions, theories, models and methods about categorization. It is beyond the scope of this dissertation to discuss them all. The current dissertation will focus on human category learning only.

About a category

According to the classical view, categories can be defined by singly necessary and jointly sufficient features (see Smith & Medin, 1981). An example is the category of triangles. The jointly sufficient features are that triangles are made of three lines and three angles. Remark that this description corresponds to the definition of triangles. Additionally, the classical view postulates that categories are clear-cut: an object belongs to the category (and hence has all the jointly sufficient features) or not. While definitions may work for some categories, most categories do not comply with this criterion. Consider for example the category of dogs. What definition will you use? That dogs have four legs and a tail? But many animals have these features, and still we are able to distinguish a cat from a dog. Moreover, if the dog is unlucky to lose a paw and the definition is not valid anymore, we will still recognize a three-legged dog as a dog. Maybe you could use the DNA-profile as a definition. It will

solve paw-losing situations, but in daily life we do not read DNA. Yet we can classify dogs from non-dogs.

In reply to these issues, the family resemblance theory postulates that natural categories are characterized by members that tend to have many features in common but not all (Rosch & Mervis, 1975). Hence, there is no set of defining properties that any and all examples have. Instead, different individual members in a category are thought to have some features in common with other members of the same category and some features in common with members of contrasting categories. For example, most birds have feathers, fly and lay eggs. However, there are birds that lack some of these features (e.g., an ostrich cannot fly) and some non-birds contain some of these features (e.g., a female crocodile lays eggs and a bat can fly). The shared features of family resemblance category members make (probabilistic) induction about the other features of a category member possible. For instance, observing that an animal has feathers, you might think ‘it is a bird’. Immediately other features of birds pop into your mind like ‘can fly’, ‘lives in trees’ and ‘sings’. Based on our knowledge of the category, we infer other features that are typically correlated with the known features. These inter-feature correlations do not need to be perfect. In principle, a category has positive utility as long as some of its features can be predicted with greater-than-chance reliability. In summary, this view implies a kind of vagueness in categories.

Currently, both the classical view and the family resemblance theory are not free of criticisms (see Harnard, 2005), but yet there exists no all-embracing alternative.

Human category learning

It is obvious that a newborn baby is not able to divide objects into meaningful entities as a grownup does. Most categories are taught during our childhood. We learn, for instance, that mammals are animals that breastfeed their offspring and that birds have feathers. Grownups pick up an apple and tell a toddler it is an apple. They inform children that lettuce is edible and that grass is not. All these categories are learned by an external 'teacher'. If a child picks up an orange and says 'apple', then the child will be corrected. In this way, learning which items belong to which category happens completely supervised. In supervised learning, a learner makes a classification and immediately feedback, in the form of the correct answer, is provided, always and after every single classification. The majority of the literature focuses on supervised category learning.

Nevertheless, in daily life everybody uses unsupervised categories that are not trained by others. There are situations in which we 'feel' that a group of stimuli belongs to the same category. Such an intuition is most obvious in perceptual grouping, where sometimes we have an immediate impression that stimuli can be grouped into clusters (e.g., see Compton & Logan, 1993). Other unsupervised categories are, for example, our idea of the perfect husband/wife or filing of paperwork. These categories develop through personal experiences and observations. There is no external teacher to judge every single categorization as wrong or right. These categories are acquired unsupervised. Little attention has been devoted to unsupervised category learning in the literature (Gureckis & Love, 2003; Love, 2002; Pothos & Chater, 2005) and only a few studies compare supervised to unsupervised learning directly (Ashby, Queller

& Berretty, 1999; Colreavy & Lewandowsky, 2008; Love, 2002; Pothos, Edwards & Perlman, 2011).

Category learning paradigms

Given that we can learn in a supervised or in an unsupervised way, category learning paradigms can also be divided into supervised or unsupervised. These paradigms are used to investigate category learning in the laboratory. In supervised category learning, by far the most popular paradigm is the supervised classification learning paradigm. In this paradigm, the learner is given a stimulus, classifies it and he/she receives corrective feedback. (For the basic experiment, see Shepard, Hovland, & Jenkins, 1961). Learning in this paradigm is intentional and there are no (apparent) limits to the complexity of the categories that can be learned (e.g., Maddox, Filoteo, Hejl, & Ing, 2004; McKinley & Nosofsky, 1995). In the supervised observational learning paradigm on the other hand, the learner receives the category label of each stimulus followed by the stimulus itself. The learner just has to confirm the category label, by pressing the appropriate button (for the basic experiment, see Ashby, Maddox, & Bohil, 2002).

In unsupervised category learning, there are two intentional paradigms. In the unsupervised classification learning paradigm, the learner sees a stimulus, classifies it and the trial is over. No feedback, nor category labels are ever provided. (For the basic experiment see Ashby, Queller, & Berretty, 1999 (a).) More popular is the free sorting task. In the basic free sorting task, the learner observes all the stimuli simultaneously. The learner must order the stimuli in a way that seems

natural to her. (For the basic experiment see Imai & Garner, 1965.) The major conclusion from this paradigm is that participants tend to categorize based on one stimulus dimension only (e.g., Ashby et al., 1999 (a); Colreavy et al., 2008; Ahn & Medin, 1992). In these two intentional learning tasks, participants are aware that they are solving a learning task. Contrarily, in incidental paradigms, learners do not know that they participate in a learning task and their attention is directed toward another task. This can be a pleasantness rating of the items, judging the relative positions of parts of the stimulus or rating the similarity between parts of two stimuli. Immediately after this unrelated task, the learner is told that all the presented stimuli belonged to one category. Then new stimuli are shown and the learner has to score the degree of membership to the category. (For the basic experiment, see Love, 2002, 2003.)

Models of category learning

In the past 50 years, many models of categorization have been presented in the literature (Ashby & Maddox, 2005). The goal of categorization models is to explain how a novel item will be classified, how the novel item is stored in our brain or how a participant learns a new category. Based on a model, researchers can develop new hypotheses and run new experiments. The insights of these experiments are used to improve the model. Since this dissertation is focused on category learning, the focus is on models that make a prediction about the learning process.

Most models use the concept of similarity to predict category membership of a new item. Similarity is typically computed as a function of distance in a putative psychological space. A key characteristic is that

the process of category learning may transform the original psychological space, leading to more attention to relevant stimulus dimensions and less to irrelevant dimensions, to support the process of category learning.

Models of categorization can be divided into single category-learning systems and multiple category-learning systems. The early theories virtually all assumed that humans have a single category-learning system. This implies that one system is used to learn all types of categories (see Ashby & Maddox, 1998 for a review; Ashby & Maddox, 2005). Later theories postulated that categorizers have multiple systems at their disposal (e.g., Ashby, Alfonso-Reese, Turken & Waldron, 1998; Ashby, Ennis & Spiering, 2007). Besides this division, categorization researchers have mostly pursued the development of either supervised or unsupervised models of categorization.

Supervised single category-learning models

As noted above, several models assume that there is one underlying category learning system. Perhaps the oldest supervised single category-learning model is the prototype model. The prototype model assumes that category learning is equivalent to learning the category prototype. This prototype is an abstract summary of the category and reflects the central tendency of all the exemplars ever experienced. When a new stimulus is encountered, this stimulus is compared to the prototypes of all (relevant) categories the person possesses. The stimulus will be assigned to the category with the most similar prototype (Homa, Sterling & Trepel, 1981; Posner & Keele, 1968, 1970; Reed, 1972; Rosch, 1973, 1975; Smith & Minda, 1998). On the other hand, the exemplar model

assumes that a category is represented by all the exemplars of the category the participant classified in the past. When an unfamiliar stimulus is encountered, its similarity is computed to the memory representation of every previously seen exemplar from each potentially relevant category. The stimulus is then assigned to the category for which the sum of these similarities is the largest (Brooks, 1978; Estes, 1986, 1994; Hintzman, 1986; Medin & Schaffer, 1978; Nosofsky, 1984, 1986; Storms, 2004). Most empirical studies show that the exemplar model is more successful in explaining human category learning behavior than the prototype model (see Nosofsky, 1992 for an overview). Nevertheless, despite its success, the exemplar model has been criticized by many authors (e.g., Maddox, 1999; Minda & Smith, 2001; Smith & Minda, 2001). One issue is the ecological validity of a full exemplar model: it is difficult to imagine that all individual exemplars ever encountered are stored in memory. This would imply that for a well-known category like cars, a person uses every single previously seen car to calculate the similarity to the category 'cars' and other relevant (contrast) categories before deciding that an unfamiliar stimulus is a car. This is also not in line with the fact that, for well-known categories, our category judgments become faster (e.g., Ashby et al. 2007).

Several mixture models have been proposed (e.g., Rosseel, 2002; Vanpaemel & Storms, 2008), assuming that the prototype and exemplar models are two extremes on a continuum and suggesting clusters of category members, merged in sub-prototypes. A third supervised category learning model, the decision bound model, assumes that subjects partition the stimulus space into response regions. In this view, an object corresponds to a point in a multidimensional space. When presented with an unfamiliar stimulus, the subject determines where the

according point lays in the space, based on similarity. The region in which the point belongs, elicits the associated response. The partition between regions associated with competing responses is called the decision bound. It is a boundary line that optimally distinguishes between two different categories. Hence, category learning is the process of learning the decision bound or learning the regions associated with each response (Ashby & Gott, 1988; Ashby & Lee, 1991; Ashby & Maddox, 1993; Ashby & Townsend, 1986; Maddox & Ashby, 1993).

Unsupervised single category-learning models

In unsupervised category-learning models the principle of similarity plays an important role. For example, the simplicity model of Pothos and Chater (2002), inspired by Rosch (1975), stipulates that the similarities of all pairs of items that belong to the same category should be greater than the similarities of all pairs of items that belong to different categories. The model predicts that the larger the within category similarities of all pairs, the more intuitive the resulting classifications will be. Intuitive classifications, defined as such, are supposed to be learnt more easily.

The first model that provides a single computational framework for both supervised and unsupervised category learning is the Supervised and Unsupervised Stratified Adaptive Incremental Network (SUSTAIN; Love, Medin, & Gureckis, 2004; see also Gureckis & Love, 2003). In SUSTAIN there are two separate, but interlinked, components, one responsible for supervised learning and one responsible for unsupervised learning. Both components function based on the principle of similarity to categorize: in unsupervised categorization, categories emerge for

groups of items that are similar to each other. In supervised categorization, the learning mechanism is based on the exemplar model (e.g., Nosofsky, 1988). The components can interact with each other so that, for example, the learning of a category is influenced by prior perceptions of intuitiveness of the category. In 2007 an update of the model was proposed, tying model operations to brain regions (Love & Gureckis, 2007).

Multiple category-learning systems

In the second half of the 1990's, research in cognitive neuroscience suggested that humans possess multiple category learning systems. Depending on the underlying category structure, participants seem to use other tactics and other factors seem to influence the learning process. Although more research is needed to identify the key qualitative properties and the neutral networks mediating each system, an increasing number of researchers is convinced by the existence of multiple categorization systems (e.g., Ashby, Alfonso-Reese et al., 1998; Brooks, 1978; Erickson & Kruschke, 1998). Based on the current state of behavioral, neuroimaging and neuropsychological patient groups studies, Ashby and O'Brien (2005) distinguished four different category structures that cannot be accounted for by a single categorization strategy. These are the rule-based, information-integration, prototype and unstructured category structures. This distinction is also suggested by Poldrack and Foerde (2008).

In *rule-based category structures*, the category can be learned via some logical reasoning process. The rule that maximizes accuracy (this is the optimal strategy) is easy to describe verbally (Ashby & Maddox,

1998). In the simplest case, only one stimulus dimension is relevant. For example, all the small green objects are 'bushes' and all the big green objects are 'trees'. Conversely, in *information-integration category structures*, the optimal strategy is difficult or even impossible to describe verbally. Accuracy is maximized only if information from two or more stimulus dimensions is integrated at some predecisional stage (Ashby et al., 1988). Perceptual integration could take many forms – from computing a weighted linear combination of the dimensional values to treating the stimulus as a Gestalt. A real-world example of information-integration category structure is judging an X-ray on the presence of tumors. It requires years of training and asked to describe their categorization strategies, experts cannot fully explain exactly how they proceed.

In category structures based on *prototype distortion*, every category has a category prototype (Posner et al., 1968, 1970). The other exemplars of the category are created by randomly distorting the coordinates of the prototype, resulting in stimuli highly similar to the prototype. Mostly this structure is used in an A not A task. In such a task, there is a single central category A and participants are presented with 'A' stimuli (which coordinates are almost identical) or 'not A' stimuli (which coordinates are just random patterns with no coherent structure). The participants have to respond 'yes' when the presented stimulus is a member of category A and 'no' when they think it is not. Repeated presentations of the highly similar stimuli enhance category learning. An example from real-life is a child and a parent walking in the street and the parent tells the child 'look, a car' whenever they encounter a four wheeled object. Cars differ in color, shape etc. but after a while, the child will be able to distinguish a car from non-cars.

Finally, in *unstructured structures*, the category exemplars are defined arbitrarily rather than based on similarity (as in prototype distortion), on the basis of some abstract logical relationship (as in rule-based) or mathematical relationship (as in information-integration). The category of emergency numbers (e.g., 112 in Europe and 911 in the United States) is an example of an unstructured category structure.

The most widely studied structure is the rule-based structure, followed (in order) by the information-integration, the prototype distortion and the unstructured category structures. It is important to note that the name of the category structure does not guarantee that people will solve the task accordingly. It is certainly possible that when faced with an information-integration category structure, a person uses an easy to verbalize rule. In that case the categorization will be inadequate.

To our knowledge, there exists no ubiquitous model explaining categorization in these four different category structures. There is one model, though, that combines two types of category learning structures. The Competition between Verbal and Implicit Systems (COVIS; Ashby, Alfonso-Reese et al., 1998; Ashby & Valentin, 2005; Ashby & Waldron, 1999) model, is a supervised learning model. It postulates two operating systems which compete during category learning: the verbal system, crucial for learning rule-based category structures, and the implicit system, important to learn other structures. Each system operates independently and is differentially influenced by feedback and task conditions. The verbal system operates by testing and selecting explicit rules about category membership. A rule is maintained as long as positive feedback is given. If a decision turns out to be incorrect, then a new candidate rule is tested. The verbal system is considered to be

supported by the prefrontal cortex, the anterior cingulate and the head of the caudate nucleus. By contrast, in the implicit system, also referred to as the procedural-learning system, information from multiple dimensions is integrated predecisionally. Hence, the acquired knowledge is difficult to verbalize. Learning in this system occurs in a slow incremental fashion and is supposed to be highly dependent on reliable and immediate feedback. The implicit system involves the inferotemporal cortex, the tail of the caudate nucleus and the nigrostratial dopamine pathway (Maddox & Ashby, 2004). At the start of a categorization task, the verbal system dominates and this dominance lingers until the implicit system is able to provide more accurate category responses. In general, COVIS has received extensive support from neuroscience (Filoteo, Lauritzen & Maddox, 2010; Nomura, et al, 2007) and experimental studies (e.g., Ashby & Valentin, 2005; Maddox & Ashby, 2004).

Inspired by the COVIS model, another multiple category-learning system was proposed. The Subcortical Pathways Enable Expertise Development model (SPEED; Ashby et al., 2007) extends the implicit system of COVIS to expert or automatic behaviors. The SPEED-model only applies to information-integration category learning. The model assumes two neural pathways mediating response selection, both starting in the sensory association cortex and ending in the premotor area. Both pathways are active during the entire learning process. Early in learning the subcortical pathway dominates. This pathway projects to the premotor area via the striatum, globus pallidus and thalamus. It is a long and slow pathway and feedback (i.e., the release or not of dopamine) is crucial. During the learning process the second pathway, the cortical-cortical pathway, becomes increasingly important. The cortical-cortical pathway projects directly from the sensory association areas to the

premotor cortex. It is fast and independent of feedback (i.e., dopamine). It is supposed to be the dominant pathway for experts in the task. In summary, SPEED assumes that the development of categorization automaticity is a gradual process via which control is passed from the subcortical pathway to the cortical-cortical pathway.

Although the idea of multiple category learning systems gains more and more popularity, it is important to note that different behaviors, such as rule- and similarity-based processing, can be equally successfully simulated in a single neural network (Verguts & Fias, 2009).

Towards semisupervised learning

Taking into account the many paradigms and models on category learning described above, it is striking that human category learning is always split up into supervised and unsupervised category learning, because neither of these types of learning are likely in daily life. In supervised paradigms, the true category label is *always* provided. Translated to daily life, this implies that for every single item we observe, there is somebody or something present informing us what it is. So whenever a toddler encounters a banana (e.g., in the fruit basket, in his/her play kitchen or on television), somebody tells her 'look, a banana'. This is unlikely. Furthermore, supervised learning implicitly assumes that attentional focus and the subsequent feedback are directed towards the same object. This is unlikely as well. For example, leafing through a book together, it can happen that the toddler thinks the parent is talking about the funny monkey instead of the banana.

The other extreme, unsupervised learning, is implausible too. It implies that the true category label is never provided. So during our entire life, nobody would ever tell us what the name of an object is or what objects belong together.

Consequently, we should consider that the natural way of learning falls in between supervised and unsupervised learning. On some occasions, we are informed about the category label. For example, when a parent reads from a book, the yellow things eaten by the monkeys are called bananas. Watching Sesame Street, Bert informs the child that the yellow thing in Ernie's ear is a banana. Or when the parent tells a child it's 'time to eat a banana' and the child observes a yellow arch shaped thing on her plate. However, on other occasions, the parent just asks the child to eat. Or when a toddler leafs through a book on its own, she will see monkeys eating yellow things. No further explanation is provided. In summary, in real-life human category learning seems to happen almost always in a semisupervised way: some stimuli receive feedback (i.e., feedback trials) and other not (i.e., no-feedback trials). This formed the inspiration for the current dissertation.

The main goal of this dissertation is to study how semisupervised category learning operates in humans. The main question is whether people are able to learn in a semisupervised way. Next, we examined what happens to the learning process when people encounter objects with no further explanation (i.e., no-feedback trials) in a stream of similar objects for which information is provided (i.e., feedback trials). There are three possible outcomes: these no-feedback encounters have no impact on the learning process at all, the encounters do have an impact and favor the learning process or the encounters have an impact but delay

the learning process. Furthermore, since several categorization models presume that multiple category learning systems exist where, depending on the underlying category structure, other factors influence the learning process, it is interesting to investigate semisupervised learning in several learning systems.

In human category learning, we were among the first to introduce the notion of semisupervised category learning (see Vandist, De Schryver & Rosseel, 2009). Since then, the semisupervised learning idea was picked-up by other researchers as well (Kalish, Zhu & Rogers, 2015; Lake & McClelland, 2011; McDonnell, Jew & Gureckis, 2012; Rogers, Kalish, Gibson, Harrison & Zhu, 2010; Zhu, Gibson, Jun, Rogers, Harrison & Kalish, 2010; Zhu, Rogers, Qian & Kalish, 2007). Meanwhile, in the field of machine learning research in semisupervised learning is booming. Semisupervised learning is even preferred over supervised learning: not only because semisupervised learning is cheaper – in supervised learning every single feedback item must be annotated by a person – also because machines perform better when they are trained in a semisupervised way (Chapelle, Scholkopf & Zien, 2006; Zhu & Goldberg, 2009). In the field of cognitive science, the few findings with regards to semisupervised category learning are mixed.

Four studies showed semisupervised learning effects. All these studies used unidimensional stimuli and a simple (rule-based) category structure. There were two categories. Feedback was provided after a specific subset of stimuli, implying that not all stimuli could potentially be followed by feedback. The semisupervised learning effects were studied on the preferred decision bound of the participant, defined on the unidimensional stimulus representation. The decision bound separates

the two categories. When the participant only uses the feedback trials, a certain decision bound can be expected. When the participant takes all stimuli into account and hence learns semisupervisedly, the decision bound will be shifted. This is due because the categories based on all trials had a different mean and distribution than the categories based on the supervised trials only. Indeed, this seemed to be the case: participants used both the feedback trials and the no-feedback trials (Gibson, Rogers, Kalish, & Zhu, 2015; Kalish, Rogers, Lang & Zhu, 2011; Lake & McClelland, 2011; Zhu et al., 2010). One study did not find an impact of the no-feedback trials (Mcdonnell, Jew, & Gureckis, 2012). Here again, the no-feedback trials covered the whole category and the feedback trials only a subset. Finally, three studies reported that semisupervised learning effects depend on the circumstances. When participants are forced to answer as quickly as possible, semisupervised learning was effective. However, when the response is self-paced, no effects were found (Rogers et al., 2010). Furthermore, when the underlying categories are distinct and the gap between the categories is large, semisupervised effects were found. This was not the case when the underlying categories were ambiguous or the space between the categories was small (Vong, Perfors, & Navarro, 2014). Finally, semisupervised learning seems to depend on the age of the participants: young children (4 to 6 years old) do use no-feedback trials whereas older children (7 to 8 years old) seem to ignore these trials (Kalish, et al., 2015).

In summary, although a few studies on semisupervised category learning exist, it remains unclear under which circumstances semisupervised learning occurs. The goal of this dissertation is to investigate the possibilities and limitations of human semisupervised learning and to gain a better understanding of the contribution of the no-

feedback trials on the learning process. Contrary to previous semisupervised studies, throughout this dissertation the deliberate choice was made to allow every stimulus of the category to be potentially followed by feedback, mimicking real-life category learning as closely as possible.

Structure of the dissertation

In Chapter 2 we developed and tested a semisupervised paradigm in Experiment 1. The semisupervised learning process was compared to supervised and unsupervised learning, using an information-integration category structure. In Experiment 2 the impact of the no-feedback trials on the learning process was investigated.

Whereas Chapter 2 was focused on semisupervised learning early in the learning process, Chapter 3 studied how semisupervised learning functions late in the learning process, when expertise develops. In two experiments, hypotheses based on the SPEED-model of Ashby et al. (2007) were tested and compared to supervised learning. The impact of the no-feedback trials on the automatization process was again studied.

In Chapter 4, the semisupervised paradigm was applied to a different category structure, namely the prototype distortion category structure. This study focused on the early learning process and consists of 4 experiments. The obtained results of semisupervised learners were compared to supervised and unsupervised learners. Consistently, the impact of the no-feedback trials on the learning process was investigated.

Finally, in the last chapter, the most striking findings are recapitulated and related to recent findings in the literature and future directions of research are proposed.

References

Ahn, W.K., & Medin, D.L. (1992). A two-stage model of category construction. *Cognitive Science, 16* (1), 81-121.

Ashby, F.G., Alfonso-Reese L.A., Turken A.U., Waldron E.M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review, 105*, 442–481.

Ashby, F.G., Ennis, J.M. & Spiering, B.J. (2007). A neurobiological theory of automaticity in perceptual categorization. *Psychological Review, 114* (3), 632-56.

Ashby, F.G. & Gott, R.E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. Journal of Experimental Psychology. *Learning, Memory, and Cognition, 14* (1), 33–53.

Ashby, F. G., & Lee, W. W. (1991). Predicting similarity and categorization from identification. *Journal of Experimental Psychology: General, 120*, 150-172.

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology, 37*, 372-400.

Ashby, F.G., & Maddox, W.T. (1998). Stimulus categorization. In Handbook of Perception and Cognition: Measurement, Judgdment and Decision Making, ed. MH Birnbaum, pp. 251-301. New York: Academic.

Ashby, F.G., & Maddox, W.T. (2005). Human category learning. *Annual Review of Psychology, 56*, 149-178.

Ashby F.G., Maddox W.T., & Bohil C.J. (2002). Observational versus feedback training in rule-based and information-integration category learning. *Memory & Cognition, 30*, 666–667.

Ashby, F.G., & O'Brien, J.B. (2005). Category learning and multiple memory systems. *Trends in Cognitive Science, 9* (2), 83-89.

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

Ashby, F.G. & Townsend, J.T. (1986). Varieties of perceptual independence. *Psychological Review, 93* (2), 154-179.

Ashby, F.G., & Valentin, V.V. (2005). Multiple systems of perceptual category learning: Theory and cognitive tests. In Categorization in Cognitive Science. H. Cohen & C. Lefebvre, Eds.: 547–572. Elsevier. New York.

Ashby, F.G., & Waldron, E.M. (1999) (b). On the nature of implicit categorization. *Psychonomic Bulletin & Review, 6*, 363-378.

Brooks, L. (1978). Nonanalytic concept formation and memory for instances. In Cognition and Categorization, ed. E Rosch, BB Lloyd, pp. 169-211. Hillsdale, NJ: Erlbaum.

Chapelle, O., Scholkopf, B. & Zien, A. (Eds.) (2006). Semi-Supervised Learning. MIT Press, Cambridge.

Colreavy, E., & Lewandowsky, S. (2008). Strategy development and learning differences in supervised and unsupervised categorization. *Memory & Cognition, 36*, 762-775.

Compton, B.J., & Logan, G.D. (1993). Evaluating a computational model of perceptual grouping by proximity. *Perception & Psychophysics, 53* (4), 403-421.

Erickson, M.A., & Kruschke, J.K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General, 127* (2), 107-140.

Estes, W.K. (1986). Array models for category learning. *Cognitive Psychology, 18*, 500-549.

Estes, W.K. (1994). Classification and Cognition. New York: Oxford University Press.

Filoteo, J.V., Lauritzen, S., & Maddox, W.T. (2010). Removing the Frontal Lobes: The Effects of Engaging Executive Functions on Perceptual Category Learning. *Psychological Science, 21* (3), 415-423.

Gibson, B.R., Rogers, T.T., Kalish, C.W. & Zhu, X (2015). What causes category-shifting in human semi-supervised learning? Proceedings of the 37th Annual Conference of the Cognitive Science Society (CogSci).

Gureckis, T.M., & Love, B.C. (2003). Towards a unified account of supervised and unsupervised category learning. *Journal of Experimental & Theoretical Artificial Intelligence , 15* (1), 1-24.

Harnard, S. (2005). To Cognize is to Categorize: Cognition is Categorization. In, Lefebvre, Claire and Cohen, Henri (eds.) Handbook of

Categorization. Summer Institute in Cognitive Sciences on Categorisation Elsevier.

Heit, E., & Bott, L. (2000). Knowledge selection in category learning. In D.L. Medin (Ed.), Psychology of learning and motivation (pp. 163-199). San Diego, CA: Academic Press.

Hintzman, D.L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93* (4), 411-428.

Homa, D., Sterling, S., & Trepel, L. (1981). Limitations of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: human learning and memory, 7* (6), 418-439.

Imai, S. & Garner, W.R. (1965). Discriminability and preference for attributes in free and constrained classification. *Journal of Experimental Psychology, 69* (6), 596-608.

Kalish, C. W., Rogers, T. T., Lang, J., & Zhu, X. (2011). Can semisupervised learning explain incorrect beliefs about categories? *Cognition, 120*, 106–118.

Kalish, C.W., Zhu, X., & Rogers, T.T. (2015). Drift in children's categories: when experienced distributions conflict with prior learning. *Developmental Science. 18*(6), 940-956.

Lake, B.M., & McClelland, J.L. (2011). Estimating the strength of unlabeled information during semi-supervised learning. Proceedings of the 33rd Annual Conference of the Cognitive Science Society, 1400–1405.

Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin & Review, 9*, 829-835.

Love, B. C. (2003). The multifaceted nature of unsupervised category learning. *Psychonomic Bulletin & Review, 10*, 190-197.

Love, B., & Gureckis, T. (2007). Models in search of a brain. *Cognitive, affective & behavioral neuroscience, 7*, 90-108.

Love, B.C., Medin, D.L., & Gureckis, T.M. (2004). Sustain: A network model of category learning. *Psychological Review, 111* (2), 309-332.

McDonnell, J.V., Jew, C.A., & Gureckis, T.M. (2012). Sparse category labels obstruct generalization of category membership. *Proceedings of the Annual Meeting of the Cognitive Science Society, 34*, 749-754.

McKinley, S.C., & Nosofsky, R.M. (1995). Investigations of exemplar and decision bound models in large-size, ill-defined category structures. *Journal of Experimental Psychology: Human Perception & Performance, 21*, 128-148.

Maddox, W. T. (1999). On the dangers of averaging across observers when comparing decision bound models and generalized context models of categorization. *Perception & Psychophysics, 61*, 354-374.

Maddox, W.T., & Ashby, F.G (1993). Comparing decision bound and exemplar models of categorization. *Perception & Psychophysics; 53*, 49–70.

Maddox, W.T., & F Ashby, F.G. (2004). Dissociating explicit and procedural-learning based systems of perceptual category learning. *Behavioral Processes, 66* (3), 309-32.

Maddox, W.T., Filoteo, J.V., Hejl, K.D. & Ing, A.D. (2004). Category number impacts rule-based but not information-integration category learning: Further evidence for dissociable category learning systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 227–235.

Markman, A.B. & Ross, B.H. (2003). Category use and category learning. *Psychological Bulletin, 129* (4), 592-613.

Medin, D.L., & Ross, B.H. (1997). Cognitive Psychology (2nd ed). Fort Worth, TX: Harcourt Brace.

Medin, D.L., & Schaffer, M.M. (1978). Context theory of classification learning. *Psychological Review, 85*, 207-238.

Minda, J.P., & Smith, J.D. (2001). Prototypes in category learning: the effects of category size, category structure and stimulus complexity. *Journal of Experimental Psychology: Learning, Memory and Cognition, 27*, 775-799.

Nomura, E.M., Maddox, W.T., Filoteo, J.V., Ing, A.D., Gitelman, D.R., Parrish, T.B., Mesulam, M.M. & Reber, P.J. (2007). Neural correlates of rule-based and information-integration visual category learning. *Cerebral Cortex, 17* (1), 37-43.

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

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

Nosofsky, R.M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory and Cognition, 14* (4), 700-708.

Poldrack, R.A., Foerde, K. (2008). Category learning and the memory systems debate. *Neuroscience and Biobehavioral Reviews, 32* (2), 197-205.

Posner, M.I., & Keele, S.W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77* (3), 353-363.

Posner, M.I., & Keele, S.W. (1970). Retention of abstract ideas. *Journal of Experimental Psychology, 83*, 304-308.

Pothos, E.M., Edwards, D.J. & Perlman, A. (2011). Supervised versus unsupervised categorization: Two sides of the same coin? *Quarterly Journal of Experimental Psychology, 64* (9), 1692–1713.

Pothos, E.M., & Chater, N. (2002). A simplicity principle in unsupervised human categorization. *Cognitive Science, 26* (3) 303 - 343.

Pothos, E.M. and Chater, N. (2005). Unsupervised categorization and category learning. *Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology, 58* (4), 733–752.

Reed, S.K. (1972). Pattern recognition and categorization. *Cognitive Psychology, 3*, 382-407.

Rogers, T.T., Kalish, C., Gibson, B.R., Harrison, J., & Zhu, X. (2010). Semi-supervised learning is observed in a speeded but not an unspeeded 2D categorization task. *In Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, p. 2320–2325.

Rosch, E. (1973). Natural categories. *Cognitive Psychology, 4*, 328-350.

Rosch, E. (1975). Cognitive reference points. *Cognitive Psychology*, *7*, 532-547.

Rosch, E., & Mervis, C.B. (1975). Family resemblances: studies in the internal structure of categories. *Cognitive Psychology, 7* (4), 573-605.

Rosseel, Y. (2002). Mixture models of categorization. *Journal of mathematical psychology, 46* (2), 178-210.

Shepard, R.N., Hovland, C.I., & Jenkins, H.M. (1961). Learning and memorization of classifications. *Psychological Monographs: General and Applied, 75* (13), 1-42.

Smith, E.E., & Medin, D.L. (1981). Categories and concepts. Cambridge, MA: Harvard Univ. Press.

Smith, J.D., & Minda, J.P. (1998). Prototypes in the mist: the early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory and Cognition, 24* (6), 1411-1436.

Smith, J.D., & Minda, J.P. (2001). Journey to the center of the category: the dissociation in amnesia between categorization and recognition. *Journal of Experimental Psychology: Learning, Memory and Cognition, 27*, 984-1002.

Storms, G. (2004). Exemplar models in the study of natural language concepts. *Psychology of Learning and Motivation, 45*, 1-39.

Vandist, K., De Schryver, M., & Rosseel, Y. (2009). Semisupervised category learning: the impact of feedback in learning the information-integration task. *Attention, Perception & Psychophysics, 71* (2), 328–341.

Vanpaemel, W. & Storms, G. (2008). In search of abstraction: the varying abstraction model of categorization. *Psychonomic Bulletin & Review, 15*, 732-749.

Verguts, T., & Fias, W. (2009). Similarity and rules united: similarity- and rule-based processing in a single neural network. *Cognitive Science, 33*, 243-259.

Vong, W. K., Perfors, A., & Navarro, D. J. (2014). The Relevance of Labels in Semi-Supervised Learning Depends on Category Structure. *In Proceedings of the 36th Annual Conference of the Cognitive Science Society*, 1718–1723.

Zhu, X., Gibson, B.R., Jun, K.S., Rogers, T.T., Harrison, J. & Kalish, C. (2010). Cognitive models of test-item effects in human category learning. *In Proceedings of the 27th International Conference on Machine Learning*, 1247–1254.

Zhu, X.; Goldberg, A.B (2009). Introduction to semi-supervised learning. Synth. Lect. *Artificial Intelligent Machine Learning, 3*, 1–130.

Zhu, X., Rogers, T., Qian, R., & Kalish, C. (2007). Humans Perform Semi-supervised Classification Too. *In Proceedings of the 22Nd National Conference on Artificial Intelligence (AAAI'07),* 864–869.

Chapter 2

Semisupervised category learning: the impact of feedback in learning the information-integration task

Vandist, K., Deschryver, M., & Rosseel, Y. (2009). Semisupervised category learning: the impact of feedback in learning the information-integration task. *Attention, Perception & Psychophysics, 71* (2), 328-341.

Abstract

In a standard supervised classification paradigm, stimuli are presented sequentially, participants make a classification, and feedback follows immediately. In this paper, we use a semisupervised classification paradigm in which feedback is given after a pre-specified percentage of trials only. In Experiment 1, feedback was given in 100%, 0%, 25% and 50% of the trials. Previous research reported by Ashby, Queller, and Berretty (1999) indicated that in an information-integration task, perfect accuracy was obtained supervised (100%), but not unsupervised (0%). Our results show that in both the 100% and 50% conditions, participants were able to achieve maximum accuracy. However, in the 0% and the 25% conditions, participants failed to learn. To discover the influence of the no-feedback trials on the learning process, the 50% condition was replicated in Experiment 2, substituting the no-feedback trials by unrelated filler trials. The results indicated that accuracy rates were similar, suggesting no impact of the no-feedback trials on the learning process. Additionally in Experiment 2, the possibility to ever learn in a 25% setting is researched. Using twice as many trials, the results showed that all participants except two succeeded, suggesting that only the total number of feedback-trials is important. The impact of the semisupervised learning results for ALCOVE, COVIS and SPEED-models is discussed.

Introduction

In category learning, there are two supervised learning paradigms: the supervised classification learning paradigm and the supervised observational learning paradigm. Current research is dominated by the supervised *classification* learning paradigm. The typical ingredients of this paradigm are that: 1) participants know in advance the number of contrasting categories; 2) stimuli are presented one at a time; 3) participants classify each stimulus and 4) feedback (showing the correct category label of the stimulus) follows immediately (see Shepard, Hovland, and Jenkins (1961) for a description of the basic experiment). Numerous studies have demonstrated that, using the supervised classification learning paradigm, participants can learn most kinds of complex category tasks (e.g., Ashby & Gott, 1988; Ashby & Maddox, 1990, 1992; Ashby et al., 1999; Medin & Schwanenflugel, 1981). In the supervised *observational* learning paradigm, participants are shown the category label of each stimulus prior to its presentation and they simply confirm the category label with an appropriate button press (see Ashby, Maddox, and Bohil (2002) for a description of the basic experiment). Learning performance is tested in test trials, in which neither feedback nor the category label is given. Results indicate successful learning in simple and in complex category tasks (Cincotta & Seger, 2007). However, using complex category tasks, higher accuracy rates are obtained in supervised feedback learning when training progresses (Ashby et al., 2002).

Few studies have focused on *unsupervised category learning*. In this setting, no feedback (as in the supervised classification learning

paradigm) nor category labels (as in the supervised observational learning paradigm) are given. Unsupervised learning can be intentional or incidental. In incidental learning, the main task does not require to make any category responses. Instead, for example, participants are asked to rate the pleasantness of the stimuli or to judge the relative positions of the stimuli. Yet, these interactions with the stimuli can result in category formation even though participants never received category information (for more details, see Love, 2002, 2003). By contrast, the main task in intentional category learning is the categorization of the stimuli. Several paradigms have been developed to study intentional unsupervised category learning. Perhaps the most used paradigm is the free sorting task in which stimuli are presented simultaneously. The task of the participants is to order the stimuli in a way that seems natural to them (see, e.g. Billman & Davies, 2005; Medin & Wattenmaker, 1987; Spalding & Murphy, 1996; Lassaline & Murphy, 1996). A second paradigm, that we will adopt in our paper, is the *unsupervised classification learning paradigm*. The procedure of this paradigm is similar to the supervised classification learning paradigm: participants are informed of the number of categories, stimuli are shown sequentially and the participant has to classify each stimulus. However, feedback is never given. The main results for both unsupervised learning paradigms are that performance is dominated by the use of unidimensional rules, regardless of the underlying category task (Ashby et al., 1999; Medin & Wattenmaker, 1987; Spalding & Murphy, 1996; Lassaline & Murphy, 1996). These unidimensional rules are easy to verbalize and to apply (e.g., the small stimuli in category A and the long stimuli in category B).

Learning in the supervised and in the unsupervised classification learning paradigm is, among other category tasks, investigated in the

information-integration task (Ashby et al., 1999). This category task is frequently used in research (e.g., Ashby & Ell, 2001; Ashby & Gott, 1988; Ashby & Maddox, 1990, 1992; Ashby & Waldron, 1999; Ashby & Maddox, 2005; Ashby & O'Brien, 2007; Ell & Ashby, 2006). The defining characteristic of the information-integration task is that participants have to combine the perceptual information of both underlying stimulus dimensions simultaneously to obtain good performance. This is difficult to accomplish because the optimal decision bound has no verbal analog (e.g., Maddox, Ashby, Ing, & Pickering, 2004; Ashby et al., 2002; Ashby, Alfonso-Reese, Turken, & Waldron, 1998). Using an information-integration task, Ashby et al. (1999) found that in the supervised condition, participants were able to master the task. In the unsupervised condition, however, learning failed, due to the use of unidimensional rules.

In our view, neither supervised nor intentional unsupervised learning seems ecologically plausible. Translated into daily life, supervised learning implies that our environment acts as an always available, completely reliable, stimulus-by-stimulus teacher. Consequently, for every single object we observe, there is somebody present or there is a nameplate telling you what it is. In addition, this information will always be unambiguously correct. Even in the most scholastic forms of learning, total-feedback (as in supervised feedback learning) or total-information (as in supervised observational learning) is improbable. On the other hand, unsupervised learning implies that we never receive any information about the categories. So during our entire life, nobody would ever tell us what the name of an object is, nor which objects belong together. This is, yet again, very unlikely. In our opinion, the supervised

and the unsupervised classification learning paradigm are two extremes of a continuum and we believe that the daily reality falls somewhere in between. When confronted with (new) objects, sometimes category information is received and sometimes it is not. Therefore, we propose a semisupervised category learning paradigm1. In this paradigm, the pre-specified percentage of trials with feedback can vary, ranging from 0% (resulting in an unsupervised classification learning paradigm) to 100% (resulting in a supervised classification learning paradigm). Category responses following feedback will be referred to as 'feedback-trials', while in the 'no-feedback trial', feedback is not provided. For instance, a 25% semisupervised classification learning paradigm consists of 25% feedback-trials and 75% no-feedback trials.

This investigation fits with the current renewed interest of the impact of feedback in category learning. In the Maddox, Ashby, and Bohil (2003) study, for example, the effect of delayed feedback was investigated. The results showed impaired learning in the information-integration task whereas learning in the simple category task was unaffected. Furthermore, Ashby and O'Brien (2007) examined the efficacy of positive and negative feedback using the information-integration task. Positive feedback signals the current categorization response was correct, whereas negative feedback indicates the opposite. In the full condition, the supervised classification learning paradigm was applied. In the negative feedback condition, feedback followed after 80% of the incorrect responses. In the positive feedback condition and in the partial condition (with positive and negative feedback), the number of feedback-trials was manipulated to be equal to the negative feedback condition; this resulted in approximately 25% feedback-trials out of the

3000 trials. To deal with the higher need of feedback at the start of learning, more feedback-trials were given in the first blocks. The results suggest that both positive and negative feedback is needed for optimal learning in the information-integration task. Interestingly, comparing learning performance after 750 feedback-trials in the supervised and the partial condition reveals that higher accuracy was obtained in the partial condition. Ashby and O'Brien (2007) see two possible explanations for this result: First, in the partial condition, participants benefit from the no-feedback trials and learn from them. Second, more between-session consolidation could occur, due to the distribution of the feedback-trials over five different days.

Although Ashby and O'Brien (2007) show in the partial condition that semisupervised category learning can be successful, some questions remain unanswered. For example, it remains unclear if indeed more feedback-trials are needed in the beginning of learning (as done in the Ashby and O'Brien (2007) study) or not (as investigated in this study). Secondly, we address what contributes to successful semisupervised category learning. Is the percentage of feedback-trials decisive? Or is there a need for a minimum number of feedback-trials? Finally, a third goal deals with the role of the no-feedback trials in the learning process. Are no-feedback trials helpful for learning (as suggested in the Ashby and O'Brien (2007) study)? Do they interfere with the learning process? Or are the no-feedback trials just irrelevant and can they simply be ignored?

Experiment 1

The objective of Experiment 1 is twofold. First, we aimed to replicate the findings of Ashby et al. (1999) in a 100% condition and in a 0% condition. Second, the semisupervised classification learning paradigm was tested in 25% and 50% conditions to find out whether these particular amounts of feedback are sufficient for successful learning.

Method

Participants. In total 34 participants (26 women, average age 22 years) took part in the experiment in return for a small payment.

Design. Experiment 1 consisted of four between-subject conditions (nine participants in the 100% and 25% condition; eight participants in the 0% and 50% conditions). Two conditions, the 100% and the 0%, were replications of Ashby et al. (1999). In the 100% condition, the supervised classification learning paradigm was used; in the 0% condition, the unsupervised classification learning paradigm was used. In the other two conditions, the semisupervised classification learning paradigm was used. Note that the name of the condition indicates the percentage of feedback-trials given. This percentage of feedback-trials was the same in all the blocks. The order of feedback-trials and no-feedback trials was randomized within blocks.

Stimuli and apparatus. In all conditions, gray 300 x 300 square-pixel Gabor patches were shown on a black screen (two examples can be seen in Figure 2.1). In this study, the 'gratings' varied continuously

on two dimensions: the spatial frequency and the spatial orientation. Frequency and orientation are perceptually separable dimensions. The arbitrary stimulus coordinates were converted to physical units using the following transformation: spatial frequency (in cycles/pixel) was converted using $f(y) = 0.01 + (y/1500)$, with y varying between 0 and 100. Spatial orientation was referred to in x degrees, varying between 0 to 100 degrees. These coordinates originated in the information-integration task, as displayed, for example in Figure 2.2. The optimal decision bound, which classified the stimuli perfectly in two categories, was diagonal. Participants viewed the stimuli on a 17" inch LCD monitor with an 800 x 600 resolution, at a distance of approximately one arm's length.

Table 2.1

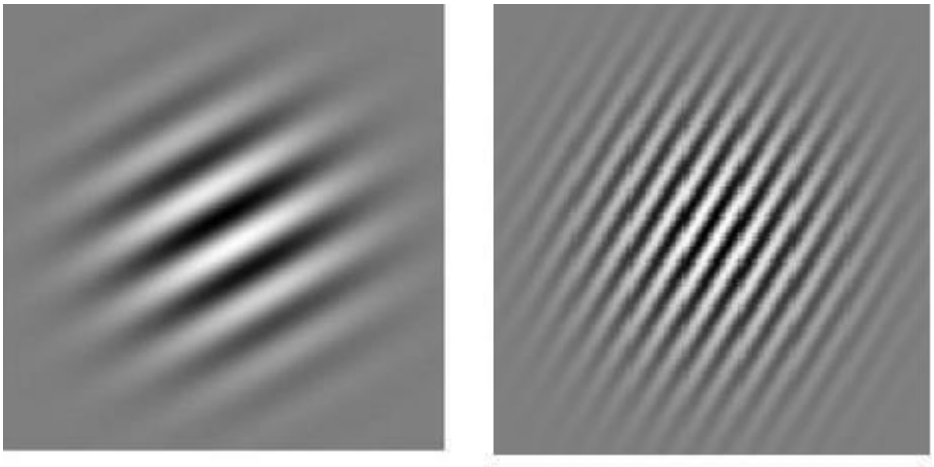
	Frequency		Orientation	
Category	M	SD	M	SD
X*	40	11.88	60	11.88
N*	60	11.88	40	11.88

Parameter values that define the categories.
*Average optimal slope =1; average optimal intercept = 0. *Within-category correlation = .99.*

For every participant 800 stimuli were generated by randomly sampling from the two bivariate normal distributions, leading to 400 'X' stimuli and 400 'N' stimuli. Category X had a different mean than category N did, but the variance and covariance remained the same. Due to the random sampling, the optimal decision bound varied slightly from block to block, although the mean optimal decision bound was $y = x$. The exact parameter values are shown in Table 2.1. As in the Ashby et al.

(1999) study, the mean, the variance and the covariance values were chosen in such a way that a linear decision bound based on one dimension would account for a maximum accuracy of 80%.

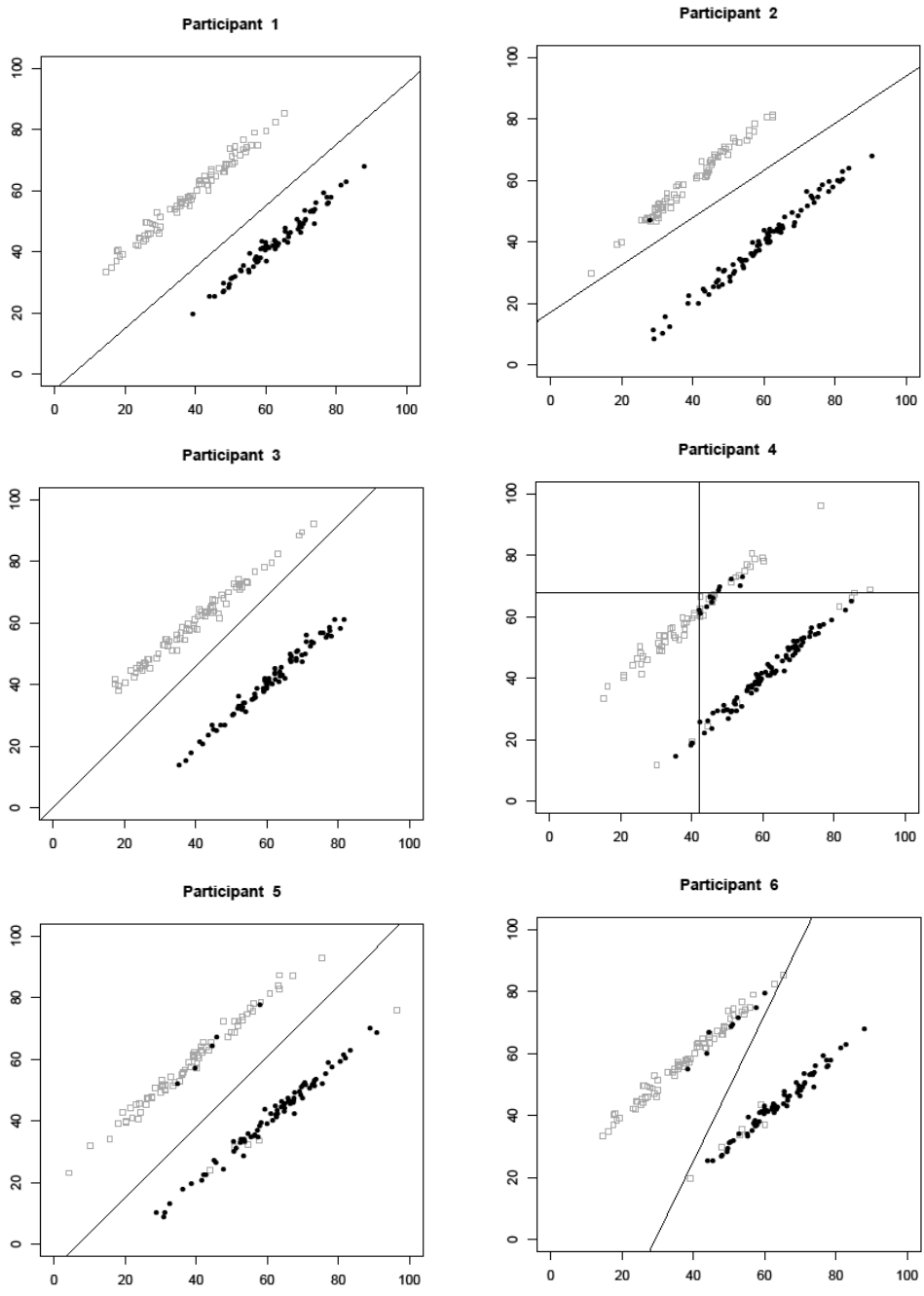
Figure 2.1



Two examples of Gabor patches

Procedure. All participants were tested individually in a dimly lit room. For every participant, 800 trials were divided into 10 blocks, so that an equal number of stimuli of each category was guaranteed in every block. Depending on the condition, the procedure differed. In the 100% condition, participants were told that they would see stimuli one by one, originating from two categories X and N. They were asked to respond by pressing “X” on the keyboard if they believed that the stimulus was an X ,and to press “N” when they believed the stimulus was an N.

Figure 2.2


Participant 1
Participant 2
Participant 3
Participant 4
Participant 5
Participant 6
0
20
40
60
80
100

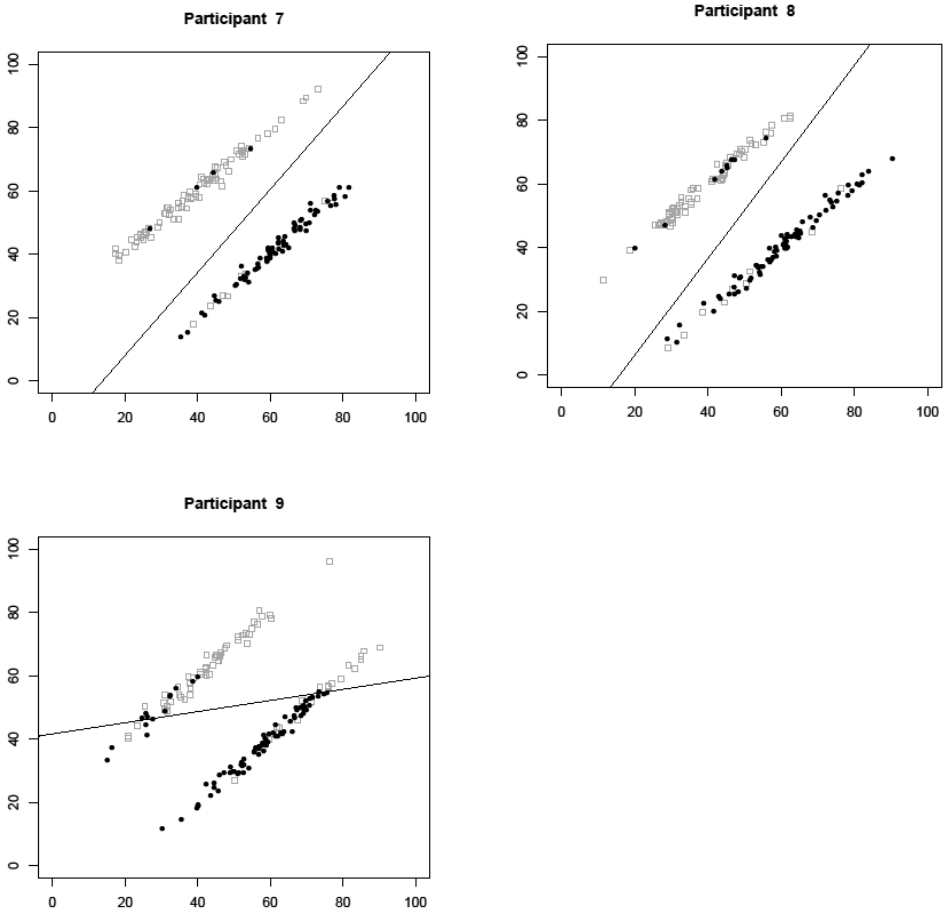


Figure 2.2, Participant 1: in this figure, an example of the underlying category structure is shown. Stimuli of category X are printed in gray squares, stimuli of category N are printed in filled black circles.
Figure 2.2 as a total: Category responses during the last two blocks for each participant in the 100% condition of Experiment 1. Also shown are the decision bounds of the best-fitting models. Y-axis is the Orientation dimension, X-axis the Frequency dimension.

Participants were informed that they would receive feedback (i.e., the true category label) after each category response. At the end of each block, the percentage of the correct responses was shown; consequently,

participants were encouraged to do better in the next block. At the beginning of the experiment, they were informed that it was possible to achieve maximum accuracy. A trial started when the stimulus was projected in the middle of the screen. After the participant responded, the feedback became visible under the stimulus for 1500 milliseconds, followed by an inter-trial interval (ITI) of 1000 milliseconds.

The 0% condition is based on Experiment 1A of Ashby et al. (1999). Two types of blocks were adopted: observation blocks (odd blocks) and response (even) blocks. In the observation block, participants had to watch the stimuli one by one, whereas the procedure of the response blocks was similar to that of the supervised condition, except that feedback was given neither after a category decision nor at the end of the block. The participants were requested to choose the category labels arbitrarily at the start of each response block, but, once a decision was made, participants had to be consistent throughout the entire block.

In the semisupervised conditions with 25% or 50% feedback the participants got the same instructions as in the supervised condition except that feedback might or might not follow the response. At the end of each block, the accuracy of the feedback-trials was displayed in percentages. These feedback-trials were identical to the 100% condition. In the no-feedback trials, the stimulus remained visible after the response for 1500 milliseconds, which was then followed by the ITI.

Results

Two types of analyses were performed. First, the mean accuracy over blocks was calculated for each condition. Second, model-based analyses were run to determine the category decision strategy used by the participant.

Accuracy Analysis. Recall that all participants can achieve perfect accuracy when they use the optimal decision bound – that is, by integrating the information from the two stimulus dimensions at some predecisional stage. On the other hand, using an unidimensional decision rule, the accuracy amounts to a maximum of 80%. This implies that participants with a result of more than 80% probably adopted a (suboptimal) information-integration decision rule. Hence, in this article, category learning is considered as successful when a score higher than 80% is obtained. The accuracy analyses are based solely on the trials with feedback. The procedure for calculating the accuracy in the 0% condition was the same as in Ashby et al. (1999). In this condition, category assignment of the labels 'X' and 'N' was arbitrary; it was assumed that participants assigned the labels to obtain the highest percentage accuracy within each block. Consequently, for each participant, the classification resulting in the highest percentage accuracy was used; it was, therefore, impossible for participants to have an outcome of less than 50%.

Figure 2.3 shows for every condition the average percent correct of each (response) block. In Table 2.2 and 2.3, the accuracy for every participant in the last two blocks is displayed. In the 100% and 50% conditions, learning was successful: the mean accuracy increased from

an average 70% (*SD*=13.86) for the 100% condition and an average 74% (*SD*=11.26) for the 50% condition to almost perfect accuracy in the last block (92% (*SD*=6.64) and 92% (*SD*=8.44), respectively). The individual accuracy rates show that all participants obtained a score higher than 80%, except participant 9 of the 100% condition who failed to reach the learning criterion. Nevertheless, there is no indication that successful learning occurred in the 0% and 25% condition because the average percent correct stays below the 80% accuracy level throughout all blocks. For the 0% condition, the average percent correct for response block 1 is 73% (*SD*=4.71) and in the last response block 72% (*SD*=7.09). All participants score less than 80%. In the 25% condition, the average percent correct for block 1 is 61% (*SD*=16.48) and 74% (*SD*=8.46) for the last block. Participant 1 and 3, however, reached 80% and 88% accuracy respectively in the last two blocks.

Model-based analysis. Model-based analysis are run to find out the category decision strategy used by the participant. The category decision strategy for each participant was calculated on the last two response blocks. Four different decision-bound models were fit to these data (see the Appendix for details). Decision-bound models assume that participants divide the perceptual space into response regions by constructing a decision bound. For each stimulus participants determine the region in the perceptual space and then elicite the appropriate response (see, e.g., Ashby & Gott, 1988; Maddox & Ashby, 1993). Three of the tested models are rule-based, i.e. the horizontal one-dimensional classifier, the vertical one-dimensional classifier and the general conjunctive classifier (see e.g., Maddox & Ashby, 2004). If participants adopt one of these category decision strategies, category learning failed.

The fourth model, the general linear classifier, is information-integration based and can approximate the optimal (diagonal) bound. Only with this category decision strategy, perfect accuracy can be obtained, indicating successful learning.

Table 2.2

Participant	GLC	DIM-O	DIM-F	GCC	% Correct last two blocks
			100%		
1	25.83	166.91	110.63	60.61	99
2	24.13	147.84	155.14	58.97	99
3	15.23	150.51	138.83	43.45	100
4	128.15	197.47	175.16	95.00	88
5	94.57	197.33	170.35	119.04	93
6	90.32	210.50	114.38	97.704	91
7	97.90	190.43	156.52	100.30	93
8	122.91	205.07	166.58	134.20	89
9	132.82	133.88	241.41	137.77	79
			0%		
1	115.19	223.72	115.25	115.11	69
2	71.48	231.53	66.47	71.54	78
3	106.54	234.78	102.88	106.62	72
4	76.41	220.86	82.23	76.11	73
5	37.75	33.21	237.39	38.29	79
6	108.45	226.52	106.54	107.64	71
7	106.13	227.69	102.34	106.15	72
8	146.73	168.30	195.94	144.48	56

Bayesian Information Criterion (BIC) scores for all participants in the 100% and the 0% conditions of Experiment 1

Note – BIC scores corresponding to the best-fitting models are underlined. GLC, general classifier; DIM-O, one-dimensional classifier, orientation; DIM-F, one-dimensional classifier, frequency; GCC, general conjunctive classifier

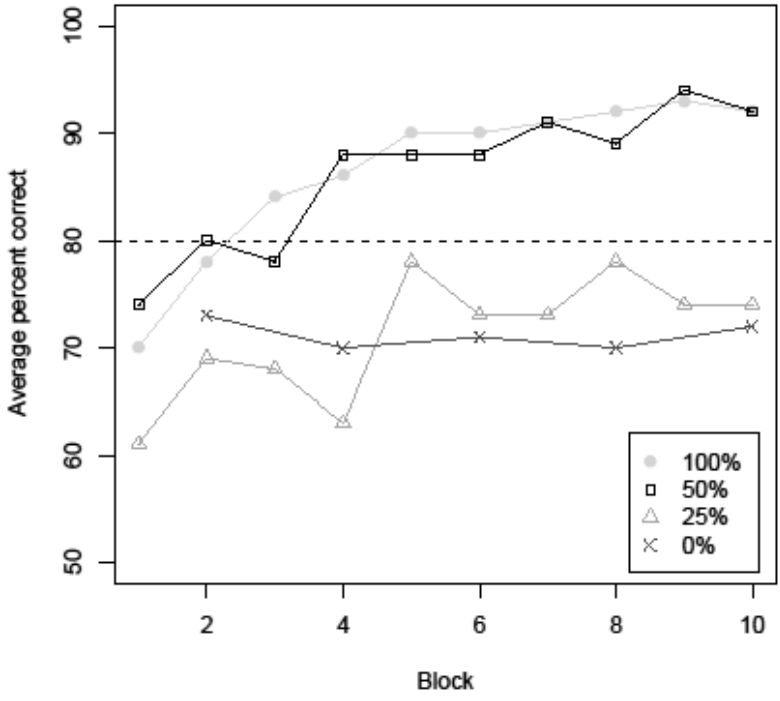
Table 2.3

Participant	GLC	DIM-O	DIM-F	GCC	% Correct last two blocks
			25%		
1	122.34	138.94	226.73	120.79	80
2	56.32	229.80	51.38	56.46	70
3	91.58	212.23	133.84	112.49	88
4	147.40	226.49	145.03	146.96	68
5	146.94	220.93	153.35	145.02	75
6	136.12	131.16	240.03	135.51	75
7	125.15	120.10	239.06	125.01	73
8	175.60	237.27	171.88	175.62	65
9	173.71	240.48	168.71	172.62	75
			50%		
1	105.40	190.13	167.51	102.69	93
2	29.81	144.62	155.96	72.78	99
3	128.79	210.66	163.37	146.11	89
4	50.35	157.71	167.73	85.89	99
5	100.86	186.73	170.69	113.59	93
6	93.03	188.27	160.30	104.52	91
7	86.15	174.72	173.10	112.70	98
8	57.76	168.41	156.77	97.16	96

Bayesian Information Criterion (BIC) scores for all participants in the 25% and the 50% conditions of Experiment 1

Note – BIC scores corresponding to the best-fitting models are underlined. GLC, general classifier; DIM-O, one-dimensional classifier, orientation; DIM-F, one-dimensional classifier, frequency; GCC, general conjunctive classifier

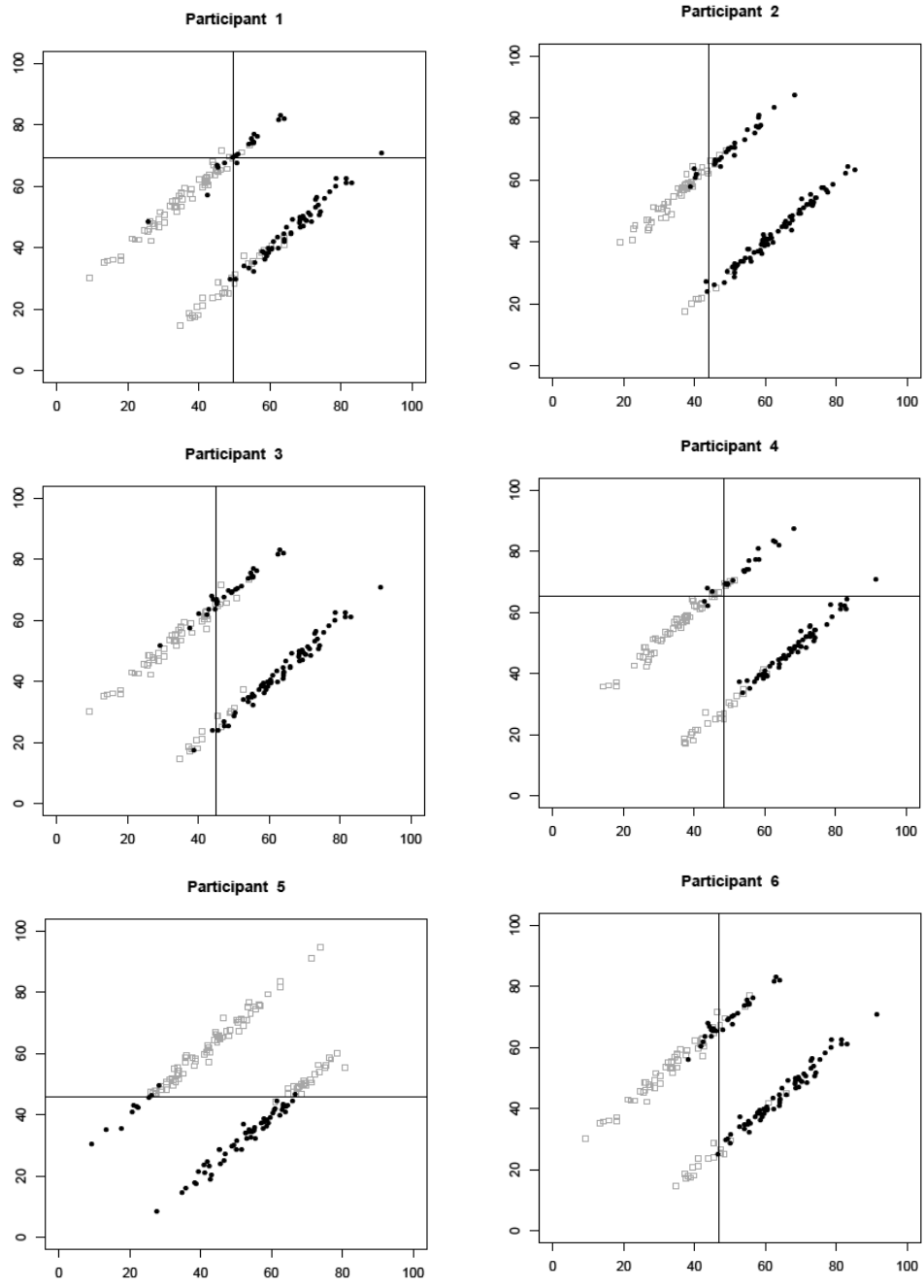
Figure 2.3

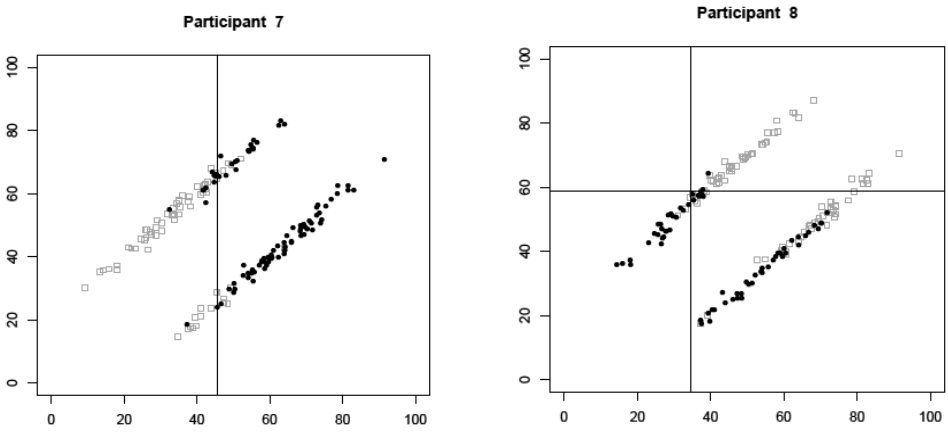


Mean percent correct by block in each (semi)-supervised condition of Experiment 1.

The participants' responses of the last two blocks along with the decision bound of the best fitting model are shown in Figure 2.2, 2.4, 2.5 and 2.6 for the 100%, 0%, 25% and 50% condition respectively. The BIC scores corresponding to the four decision bound models are depicted in Table 2.2 and 2.3.

Figure 2.4

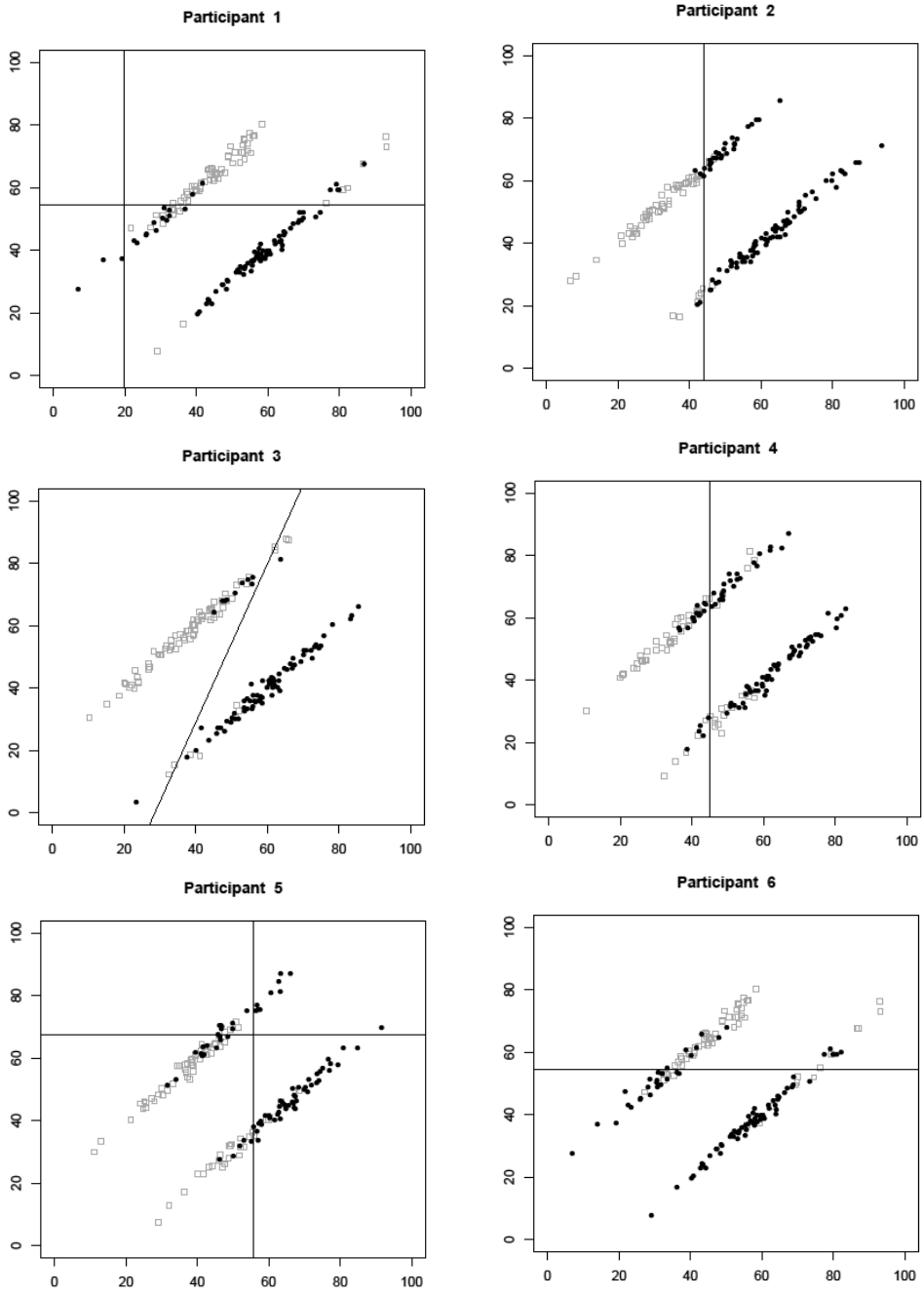

Participant 1
Participant 2
Participant 3
Participant 4
Participant 5
Participant 6
0
20
40
60
80
100

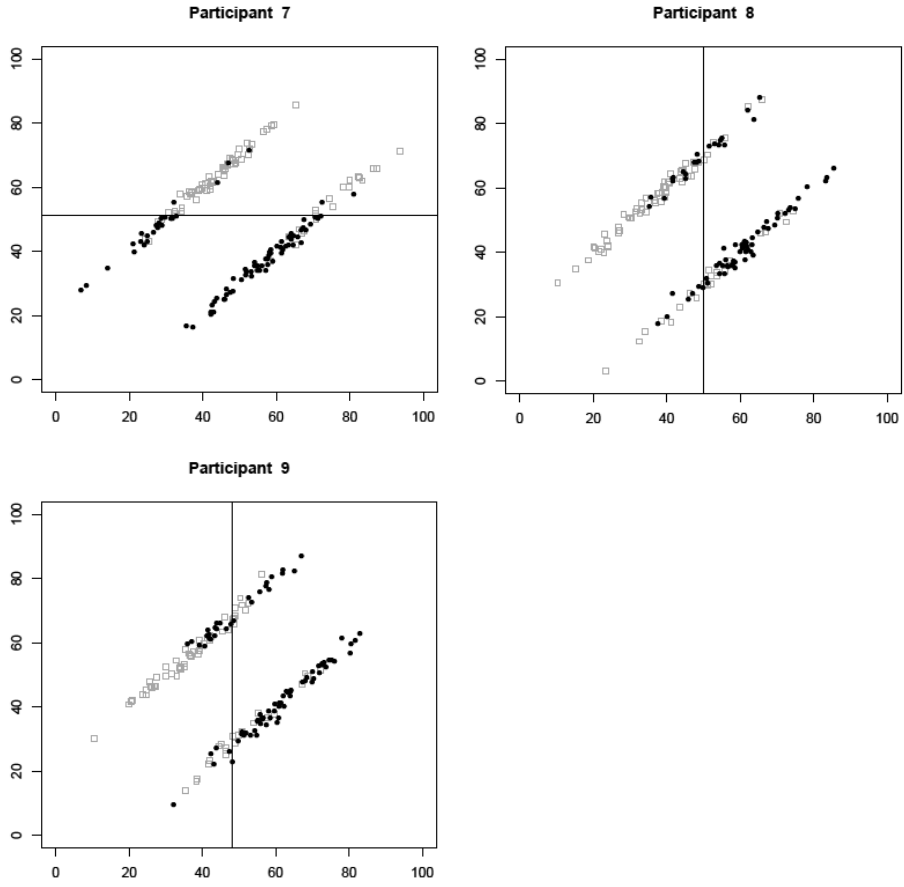


Category responses during the last two blocks for each participant in the 0% condition of Experiment 1. Also shown are the decision bounds of the best-fitting models. Y-axis is the Orientation dimension, X-axis the Frequency dimension.

In the 100% condition, all participants but one used an information-integration based strategy, indicating that they had learned the category task. In the 0% condition all participants failed to learn the category task. Participants 2, 3, 6 and 7 used a one-dimensional decision bound based on the frequency, while participant 5 used a one-dimensional decision bound based on the orientation. The responses of participant 1 and 8 were best fitted by the general conjunctive classifier. In the 25% condition, several category decision strategies were used. Participants 2, 4, 8 and 9 applied a one-dimensional model based on the frequency whereas the data of participant 6 and 7 were best fitted with a one-dimensional model based on the orientation. Participant 1 and 5 used a general conjunctive strategy.These results suggest that they were not able to learn the category task. However, the responses of participant 3 were best fitted by the general linear classifier. It seems that this participant learned the category task.

Figure 2.5

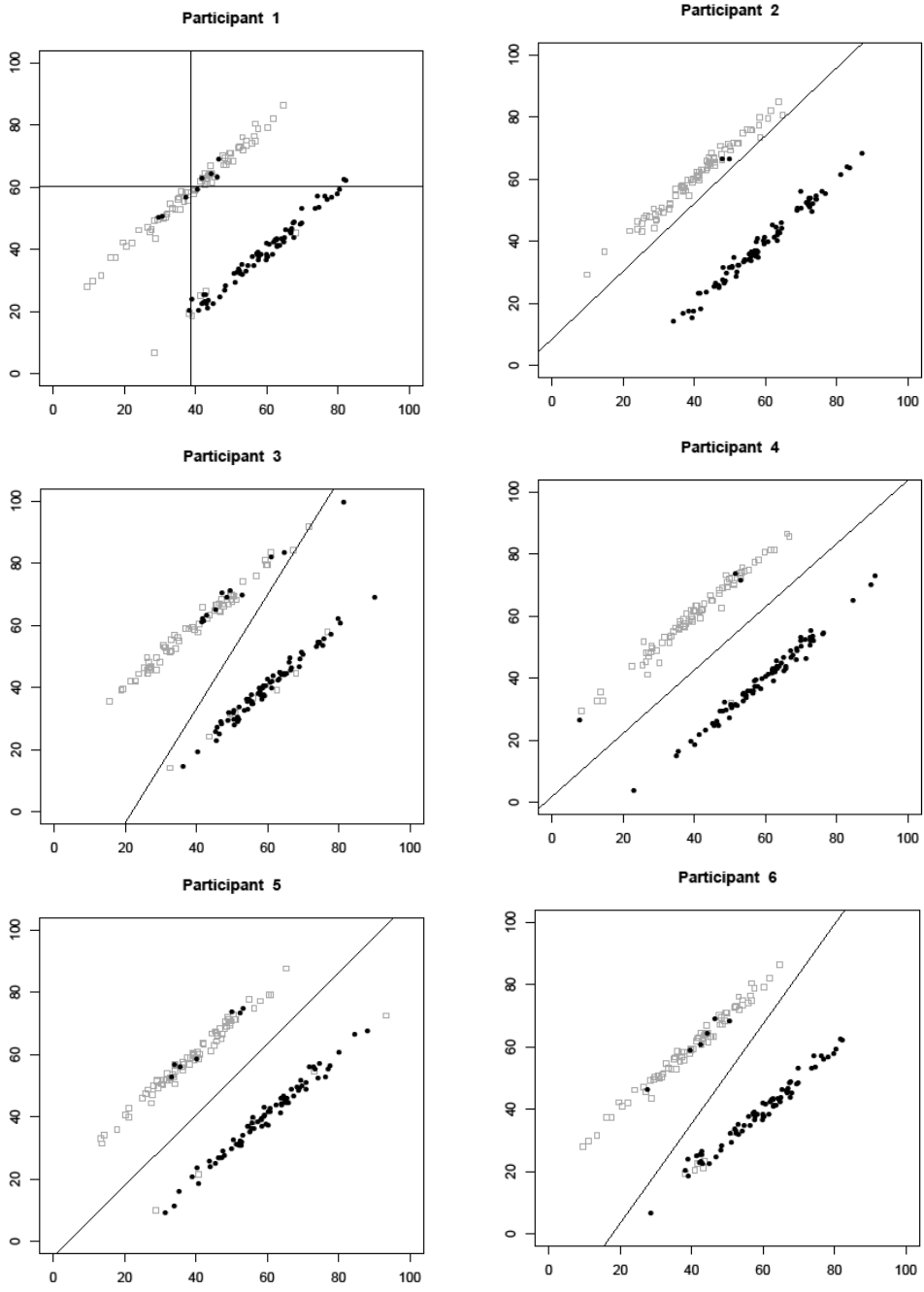

Participant 1
Participant 2
Participant 3
Participant 4
Participant 5
Participant 6
0
20
40
60
80
100

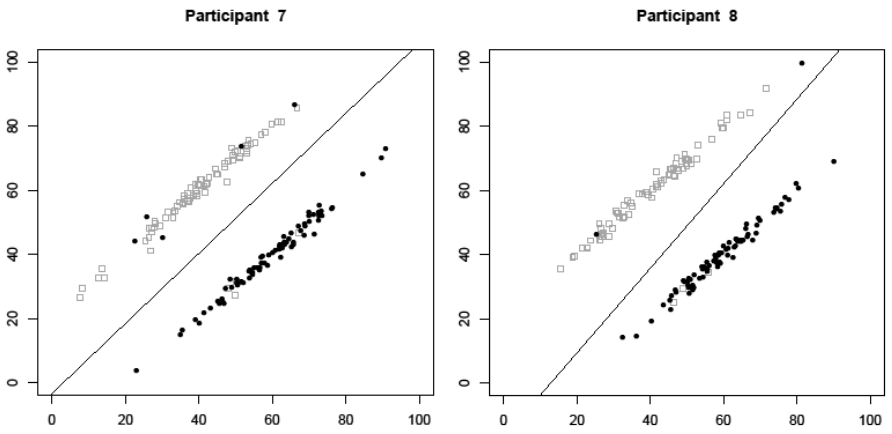


Category responses during the last two blocks for each participant in the 25% condition of Experiment 1. Also shown are the decision bounds of the best-fitting models. The Y-axis is the orientation dimension, the X-axis the frequency dimension.

Finally, in the 50% condition, all participants (except participant 1) succeeded to learn the category task because their responses were best fitted by the general linear classifier. These results suggest that it is possible to learn the information-integration task when feedback is randomly given after 50% of the responses only.

Figure 2.6


Participant 1
Participant 2
Participant 3
Participant 4
Participant 5
Participant 6
0
20
40
60
80
100



Category responses during the last two blocks for each participant in the 50% condition of Experiment 1. Also shown are the decision bounds of the best-fitting models. The Y-axis is the orientation dimension, the X-axis the frequency dimension

.

Discussion

The objective of Experiment 1 was to replicate the findings of Ashby et al. (1999) using a information-integration structure in the 100% (supervised) and the 0% condition (unsupervised). The results show that this replication is successful in both conditions. The mean accuracy of the 100% condition in the last block is close to perfect and, moreover, all participants but one used an information-integration decision bound close to the optimal bound. By contrast, in the 0% condition, the mean accuracy stays far below the 80% and the model-based analyses show the ubiquitous use of rule-based decision strategies. These results indicate that participants succeed in learning the category structure in the 100% condition, whereas in the 0% condition, participants failed to learn the category task.

In the semisupervised condition with 50% feedback, the high accuracy and the information-integration based decision strategies of eight participants indicate successful learning. This result suggests that feedback is not required after each response to learn the information-integration task. Yet, participants apparently fail to learn when the feedback is limited to 25 percent. In this condition, the responses of eight participants were best fitted by a rule-based decision bound.

Experiment 2

In Experiment 2, two follow-up conditions are run. Firstly, we extend the 50% condition of Experiment 1 to study the effect of the no-feedback trials. Secondly, the possibility to ever learn in a 25% semisupervised classification learning paradigm is studied.

The impact of the no-feedback trials. In the 50% condition of Experiment 1, participants could master the category task, although feedback was given only after half of the trials. Even so, the role of the no-feedback trials in the learning process remains unclear. Therefore, the influence of the no-feedback trials, or lack thereof, was studied by comparing the 50% condition of Experiment 1 to a new condition in which the no-feedback trials were replaced by unrelated filler-trials. We envisioned three possible ways the no-feedback trials I which the no-feedback trials might influence the learning process. A first possibility is that the no-feedback trials benefited the learning process, perhaps compensating for the lack of feedback on half of the trials. If this was indeed true, we expect learning to be slower in Experiment 2. A second

possibility is that the no-feedback trials harmed the learning process. In this case, learning would be faster in Experiment 2. A final alternative was that the no-feedback trials did not interfere with the learning process at all. In this case, we would expect similar performance in both conditions.

25% feedback learning possibilities. When the 25% condition of Experiment 1 ended at block 10, it seemed most participants seemed unable to learn the category structure. The reason for this failure could be that to learn this type of category task one needs more than 25% feedback trials This hypothesis will be referred to as the *percentage-feedback-hypothesis*. However, the responses of 1 participant were best fitted by an information-integration decision bound model (i.e., the general linear classifier). Participants might have been able to learn the category task but the experiment might have stopped too soon, leaving open the possibility that, with more blocks in total, the category task could be learned after all. This *absolute number hypothesis* suggested that the absolute number of feedback-trials is indicative for learning, and that learning would occur if participants received more feedback trials in total. The objective of this condition was to find out which was correct: the percentage-feedback-hypothesis or the absolute-number-hypothesis. Therefore, the 25% condition of Experiment 1 was replicated, but this time the participants were subjected to two sessions of 800 trials.

Method

Participants. Seventeen participants (twelve women, average age 21.1 years) participated in the experiment in return for a small payment.

Procedure, stimuli and apparatus. The procedure, the apparatus, the stimuli and the category task were identical to those of the 50% condition as described in Experiment 1. The difference between the 50% condition of Experiment 1 and this condition was that Gabor patches are shown on half of the trials on which, after response, subjects were provided with feedback (i.e. feedback-trials). These trials were randomly alternated with unrelated filler-trials in which the stimuli were the characters 'X' (odd trials) and 'N' (even trials). These fillers had the same size as the Gabor patches but did not disclose anything about the underlying category task or cue the outcome of the following feedback-trial.

For the 25% feedback learning trials, the procedure, the stimuli and the apparatus were exactly the same as those described in the 25% condition of Experiment 1, except that participants received two sessions of 800 trials. The two sessions took place on two consecutive days.

Results

Impact of the no-feedback trials. For this condition, only the feedback-trials were analyzed. The accuracy analysis indicated successful learning: the average percent correct in the last block was 95% (*SD*=4.52) (see Table 2.4 for the individual accuracy rates for the

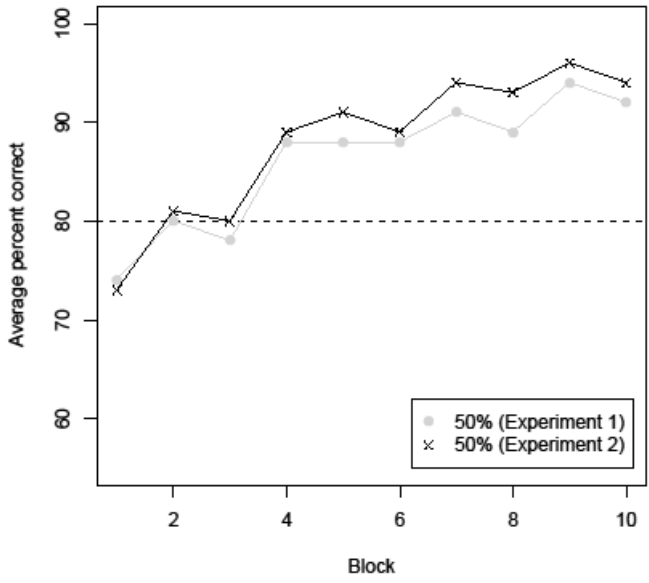
last two blocks). Figure 2.7 shows the average percent correct over the blocks. Note that the learning process observed was similar to that in the 50% condition of Experiment 1. Next, the best-fitting decision bounds models were calculated on all responses of the last two blocks (see Table 2.4 and Figure 2.8). The results show that all participants used an information-integration decision strategy.

Table 2.4

Participant	GLC	DIM-O	DIM-F	GCC	% Correct last two blocks
		50% (filler)			
1	13.15	69.64	79.80	29.36	100
2	23.37	82.58	60.02	34.33	99
3	27.50	67.37	95.18	40.52	98
4	28.49	81.01	73.60	41.92	98
5	52.33	139.38	81.48	79.64	94
6	15.23	141.02	75.68	27.95	100
7	71.42	136.12	79.12	78.63	83
8	69.65	164.88	96.67	84.76	89
		25% (two sessions)			
1	24.76	128.15	173.95	68.71	98
2	122.03	118.55	247.00	122.64	63
3	29.63	180.74	143.99	50.49	95
4	15.23	144.26	140.52	58.08	100
5	50.51	171.07	154.19	91.82	95
6	42.96	167.00	149.23	68.80	95
7	15.23	178.55	123.10	70.69	100
8	136.70	183.97	201.43	146.91	88
9	84.51	232.20	79.99	85.01	63

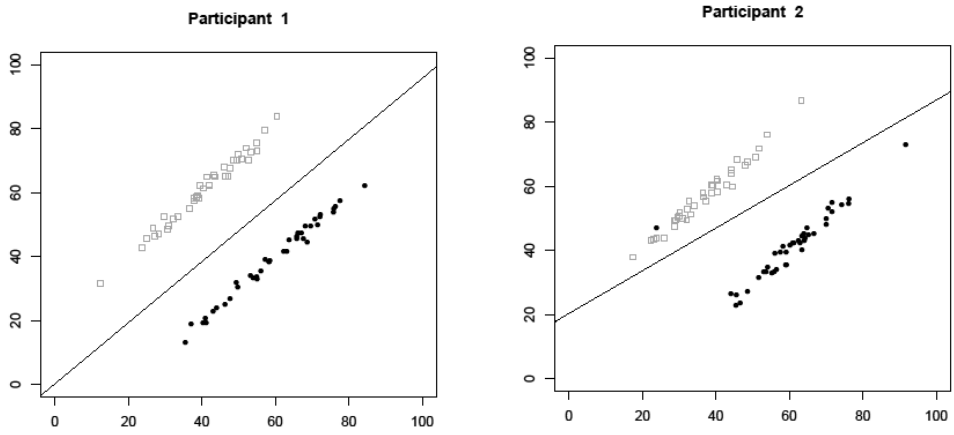
Bayesian Information Criterion (BIC) scores for all participants in the 25% and the 50% conditions of Experiment 2
Note – BIC scores corresponding to the best-fitting models are underlined. GLC, general classifier; DIM-O, one-dimensional classifier, orientation; DIM-F, one-dimensional classifier, frequency; GCC, general conjunctive classifier

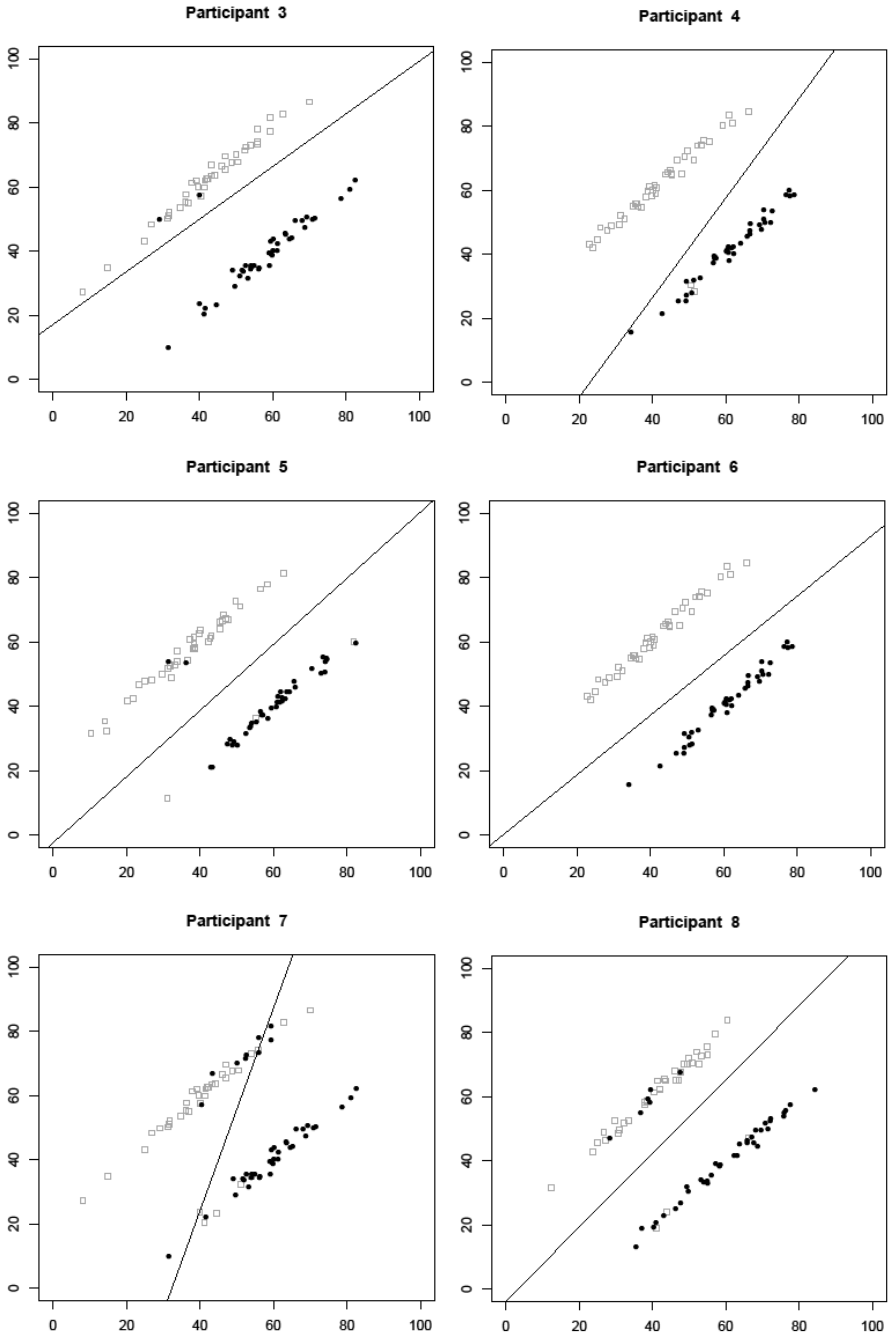
Figure 2.7



Mean percent correct by block in the conditions with 50% feedback

Figure 2.8

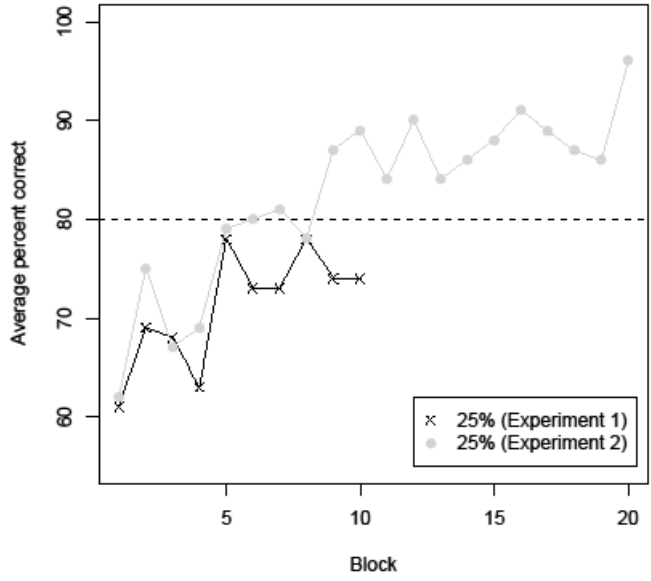




Category responses during the last two blocks for each participant in the impact of the no-feedback trials condition of Experiment 2. Also shown are the decision bounds of the best-fitting models. The Y-axis is the orientation dimension, the X-axis the frequency dimension.

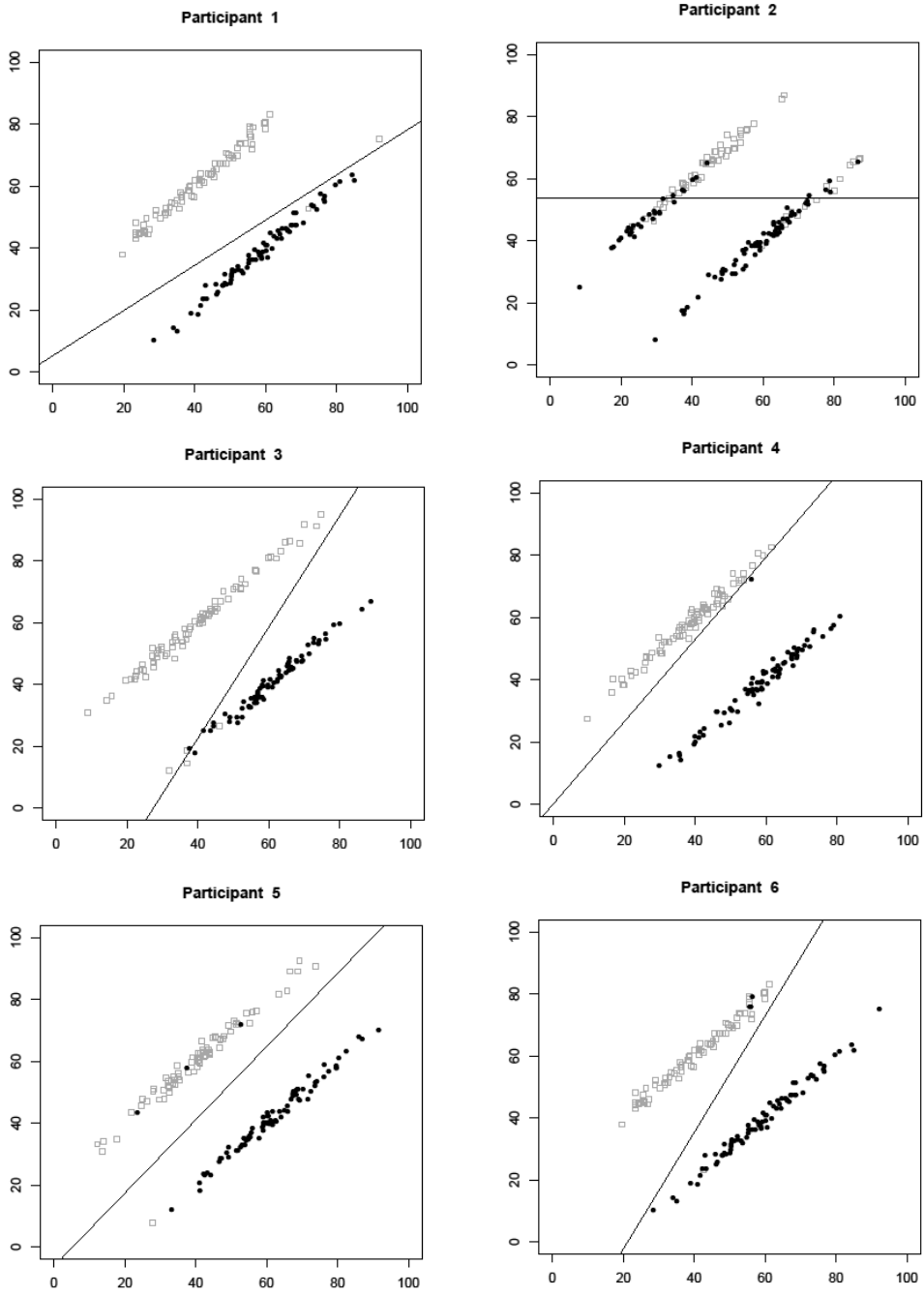
The 25% feedback learning possibilities. The accuracy analysis suggested successful learning: the average percent correct in the last block was 96% (*SD*=6.82) (see Table 2.4 for the individual accuracy rates for the last two blocks, see Figure 2.9 for the average percent correct over the blocks). Similarly, the model-based analysis revealed the use of information-integration decision strategies, except for participant 2 and 9. The results are shown in Figure 2.10 and Table 2.4.

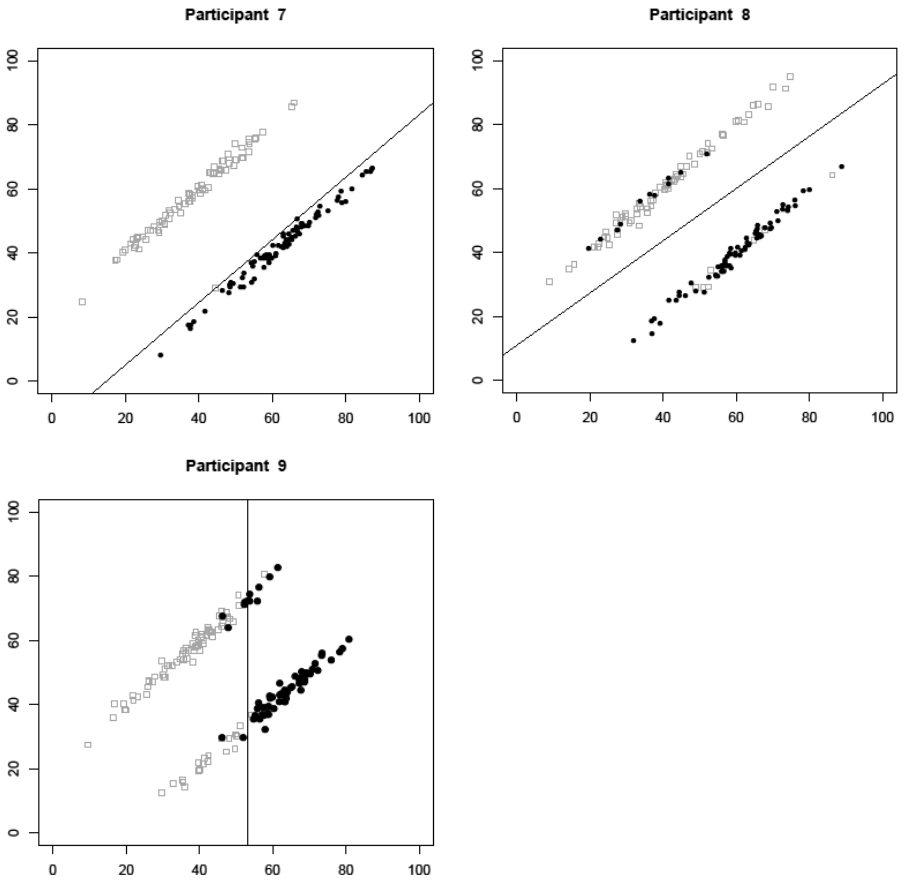
Figure 2.9



Mean percent correct by block in the condition with 25% feedback

Figure 2.10


Participant 1
Participant 2
Participant 3
Participant 4
Participant 5
Participant 6
0
20
40
60
80
100



Category responses during the last two blocks for each participant in the 25% feedback trials learning possibilities condition of Experiment 2. Also shown are the decision bounds of the best-fitting models. The Y-axis is the orientation dimension, the X-axis the frequency dimension.

Discussion

In Experiment 2, two additional conditions were run for a better comprehension of semisupervised feedback learning. First, the effect of the no-feedback trials on the learning process was addressed. The results demonstrated no loss or gain in performance in this unrelated filler-

condition as indicated by a high average percent correct, similar to the 50% condition of Experiment 1. Additionally, eight participants appeared to have used an information-integration based decision strategy. In conclusion, these results suggest that when no feedback is given, observing and responding to relevant stimuli does not influence the learning process in this task. Second, the ambiguous results of the 25% condition of Experiment 1 were clarified. Two feasible hypotheses were tested: the percentage-feedback-hypothesis and the absolute-number-hypothesis. The results reveal that by giving two sessions instead of one, all but 2 participants mastered the category task. This suggests that the total number of feedback-trials is indicative of learning or failure to learn. Interestingly, after one session, the accuracy rates were much higher than in the 25% condition. We believe that this result was due to the greater commitment of the participants in Experiment 2. Perhaps because the participants knew in advance that they would return and perform the same task, they tried harder to learn the categories.

General Discussion

In the category learning literature, the focus was mainly on strictly supervised learning or on strictly unsupervised learning. In the supervised category learning paradigm, feedback followed immediately after each classification response whereas in the unsupervised category learning paradigm, feedback was never provided. In this article we argued that both these paradigms are ecologically implausible. In daily life, we argue sometimes get feedback about the category membership of an

object, and at other times we do not. It is important that category learning researchers should take this into account. Therefore, in this article, a semisupervised category learning paradigm was used in which the percentage of feedback-trials was manipulated. In order to facilitate comparison with earlier work, we used the information-integration task. A first finding is that participants were able to learn the information-integration task when feedback followed after half of the trials (Experiment 1). This finding is in line with the partial condition of Ashby and O'Brien (2007). Nevertheless, this result is surprising too since Maddox et al. (2003) found that supervised category learning in the information-integration task was impaired when feedback was delayed. In the 50% condition, feedback was even omitted after half of the trials and still no interference in the learning process was observed. Secondly, we found that learning even occurred when feedback was provided after only a quarter of the trials, but only if a sufficient number of feedback-trials were presented (Experiment 2). The major conclusion is that participants do not need feedback after every single response to master the information-integration task as long as the absolute number of feedback-trials is sufficiently high enough. Additionally, since the percentage of feedback-trials was equal across blocks, it seems that there is no need for a higher percentage of feedback-trials in the first blocks in order to achieve successful learning as suggested by Ashby and O'Brien (2007). Further research is needed to find out whether the same conclusions hold for other complex category tasks and other (semisupervised) paradigms.

This article is the first in category learning literature which directly investigate the nature of the no-feedback trials during the learning process (Experiment 2). Clearly, adopting filler-trials instead of no-

feedback trials did not influence learning. Although participants had to respond to the stimuli (and presumable this decision involved a lot of brain processing), they seemed to simply ignore these trials when no feedback followed. This finding rebutted the suggestion made by Ashby and O'Brien (2007) that no-feedback trials may benefit the learning process.

To the best of our knowledge, no category learning models make explicit predictions about semisupervised category learning (using continuous stimulus dimensions). In order to relate our results to the broader literature, three (supervised) category learning models are discussed: the ALCOVE model, the COVIS model and the SPEED model. The ALCOVE model was developed by Kruschke (1992). According to the 'Attention Learning Covering Map' model (ALCOVE), learning is error driven. Kruschke explicitly stipulated that 'When there is no error, nothing changes (p. 25)'. Due to the absence of any feedback in the no-feedback trials, it was structurally impossible to obtain an error. Consequently and in line with our results, the ALCOVE model predicts no influence of the no-feedback trials on the learning process.

On the other hand, the COmpetition between Verbal and Implicit Systems model (COVIS) predicts interference of the no-feedback trials. COVIS was developed by Ashby et al. (1998) and postulated two systems which compete during learning. Learning the information-integration task relies on the implicit system (e.g., Ashby et al., 1998; Ashby & Waldron, 1999), and only this system will be explained. Critical to learning are the synapses between the visual cortical cells and the medium spinal cells in the caudate. To strengthen these synapses, and enhance learning, three factors are required: (1) a strong presynaptic

activation, (2) strong postsynaptic activation, and (3) dopamine release. Factor 3, dopamine, serves as a reward-mediated signal. Hence, it will only be released when positive feedback is given. Although there is large empirical support for COVIS (for recent reviews, see Ashby & Maddox, 2005; Ashby & Valentin, 2005; Ell & Ashby, 2006; Maddox & Ashby, 2004; Maddox & Filoteo, 2005), the semisupervised conditions seem incompatible. According to COVIS, the feedback-trials in which a correct classification was made, would strengthen the synapses and boost learning. On the contrary, the absence of factor 3 in the no-feedback trials and the negative feedback-trials would reduce the strength of the active synapses and learning would decrease (Ashby et al., 1998). To our knowledge, there is no specific prediction at what rate active synapses are enhanced (by positive feedback) or diminished (by absence of feedback or negative feedback). But at least a disturbance in learning may be expected in the semisupervised conditions. Especially in the 25% condition where a theoretically maximum 25% of the category responses can lead to strengthening and a minimum 75% of the responses will result into a weakening of the synapses, the interference effects should be detectable. Yet successful learning did occur in the 25% condition of Experiment 2. Apparently, the weakening of the synapses in more than 75% of the trials did not impede learning. Moreover, when we compare the learning processes of the 50% and the 100% conditions of Experiment 1, no difference can be found, despite that in the 50% condition, more than half of the trials should have disturbed learning according to COVIS. Finally, Experiment 2 demonstrated no advantage in the learning process when unrelated (and hence no-disturbing) fillers were used.

Recently, the implicit learning system of COVIS was extended by the Subcortical Pathways Enable Expertise Development model (SPEED) (Ashby, Ennis, & Spiering, 2007). Besides the cortical-striatal pathway that relies on the same three-factor learning as in the COVIS model, a cortical-cortical pathway is added. This pathway links the sensory association cortex directly to the premotor area, resulting in faster reaction times. In this pathway, two factors are needed to strengthen the synapses. Firstly, there must be pre-synaptic activation and secondly, there must be post-synaptic activation. When the post-synaptic activation passes a certain threshold, the synapses are strengthened. On the other hand, when the post-synaptic activation fails to reach the threshold, the link between the active synapses will be diminished. Consequently, when this cortical-cortical pathway is used, the lack of (positive) feedback in the no-feedback trials should not disturb learning. Nevertheless, there might be one problem for SPEED to explain our results, probably because SPEED did not consider semisupervised learning. In the case when learning depends on the cortical-striatal pathway (three factor learning), the no-feedback trial corresponds to the condition in which striatal activation could pass the threshold but dopamine stays below the baseline. This condition is described as 'long-term depression' (LTD) in SPEED (p.640, Ashby et al., 2007) and results in weaker synapse strength. Since the main goal of SPEED is to simulate categorization automaticity, it is postulated that early in category learning the cortical-striatal pathway is dominant. Hence early in learning, feedback is crucial. Only when the learning process progresses, the role of the cortical-cortical pathway increases and the importance of feedback diminishes (see Ashby et al. (2007) for more details). This implies that in the first blocks, the strengthening of the synapses, and thus learning, is entirely

dependent on the positive feedback-trials. Similarly to the COVIS model, the absence of positive feedback-trials in the first blocks will reduce the active synapses and interfere learning. Although it is unknown at what rate the synapses are enhanced or weakened, these competing mechanisms will complicate the enhancing of the cortical-cortical pathway. Consequently, at least in the first blocks disturbance in learning is expected. Again, no such interference effects could be observed since the learning processes of the 50% condition of Experiment 1 and the 100% condition were similar. Besides, no advantage of the unrelated (and hence no-disturbing) fillers could be found in the first blocks of Experiment 2.

However, it cannot be ruled out that the irrelevance of the no-feedback trials was specific to the used information-integration task in this paper. Note that the stimulus dimensions were highly correlated and that both the feedback-trials and no-feedback trials were randomly chosen from this highly correlated bivariate normal distributions. Consequently, the no-feedback trials did not provide any new information about the category task. This constellation could be masking possible influences of the no-feedback trials on the learning process. In other words, it may be possible that, if the no-feedback trials give extra information about the category task, the no-feedback trials may have an impact on the learning process. To discover the nature of the no-feedback trials entirely, more research is needed.

Footnotes

[1]Unkown during the data collection of this study, the semisupervised classification learning paradigm was simultaneously used in the partial condition by Ashby and O'Brien (2007).

Appendix

Four models were fit to each participant's data: three rule-based decision bound models, i.e. the vertical and the horizontal unidimensional models, the general conjunctive classifier and one information-integration model, i.e. the general linear classifier. More details of the models can be found in Ashby (1992); Ashby and Gott (1988); and Maddox and Ashby (1993).

Rule-Based Models

The one-dimensional classifier (DIM). The DIM assumes that participants set a decision criterion on one of the stimulus dimensions. Because our stimuli have two stimulus dimensions, orientation and frequency, two one-unidimensional classifiers were fit. One for the frequency dimension (DIM-F) and one for the orientation dimension (DIM-O). An example of an one-dimensional rule used for categorization is: 'Respond A if the orientation is steep, otherwise respond B'. These models have two parameters: a decision criterion along the choosen classification dimension and a perceptual noise variance parameter.

The general conjunctive classifier (GCC). The GCC assumes that the rule used by participants is a conjunction, for example: 'Respond A if the orientation is steep and the frequency small, otherwise respond B'. The GCC has three parameters: a criterion on each dimension (i.e., orientation and frequency) and one parameter for the perceptual noise variance.

Information Integration Model.

The general linear classifier (GLC). The GLC assumes that a linear decision bound divides the stimulus space into response regions. Confronted with a stimulus, the perceived location in the stimulus space is determined and the corresponding categorization response elicited. These decision bounds require linear integration of both stimulus dimensions, resulting in a information integration decision strategy. The GLC has three parameters: the slope and the intercept of the linear decision bound and a perceptual noise variance.

Selecting the best-fitting model.

The model parameters were estimated using the method of maximum likelihood. To select the best-fitting model for every participant's set of responses, the model with the smallest Bayesian information criterion (BIC) is choosen. The BIC penalizes for the number of free parameters and is defined as BIC = rlnN − 2lnL, where r is the number of free

parameters, N is the sample size and L is the likelihood of the model given the data (Schwarz, 1978). These BIC scores are shown in Table 2.2, Table 2.3 and Table 2.4. The BIC score corresponding to the best fitting model are underlined.

References

Ashby, F.G. (1992). Multidimensional models of categorization. In F.G. Ashby (Ed), *Multidimensional models of perception and cognition* (pp. 449-483). Hillsdale, NJ: Erlbaum.

Ashby, F. G., Alfonso-Reese, L. A., Turken, A., & Waldron, E. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review, 105*, 442–481.

Ashby, F. G., & Ell, S. (2001). The neurobiology of human category learning. *Trends in Cognitive Science, 5*, 204–210.

Ashby, F. G., Ennis, J., & Spiering, B. (2007). A neurobiological theory of automaticity in perceptual categorization. *Psychological Review, 114*, 632–656.

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

Ashby, F. G., & Maddox, W. T. (1990). Integrating information from separable psychological dimensions. Journal of Experimental Psychology: *Human Perception & Performance, 16*, 598-612.

Ashby, F. G., & Maddox, W. T. (1992). Complex decision rules in categorization: Contrasting novice and experienced performance. *Journal of Experimental Psychology: Human Perception & Performance, 18*, 50–71.

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology, 56*, 149–178.

Ashby, F. G., Maddox, W. T., & Bohil, C. (2002). Observational versus feedback training in rule-based and information-integration category learning. *Memory & Cognition, 30* (5), 666–677.

Ashby, F. G., & O'Brien, J. (2007). The effects of positive versus negative feedback on information-integration category learning. *Perception & Psychophysics, 69* (6), 865–878.

Ashby, F. G., Queller, S., & Berretty, P. M. (1999). On the dominance of unidimensional rules in unsupervised categorization. *Perception & Psychophysics, 61* (6), 1178–1119.

Ashby, F. G., & Valentin, V. (2005). Multiple systems of perceptual category learning: Theory and cognitive tests. In I. H. Cohen & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science* (pp. 547–572). New York: Elsevier.

Ashby, F. G., & Waldron, E. M. (1999). On the nature of implicit categorization. *Psychonomic Bulletin & Review, 6*, 363–378.

Billman, D., & Davies, J. (2005). Consistent contrast and correlation in free sorting. *American Journal of Psychology, 118*, 353–383.

Cincotta, C., & Seger, C. (2007). Dissociation between striatial regions while learning to categorize via feedback and via observation. *Journal of Cognitive Neuroscience, 19*, 249–265.

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

Kruschke, J. K. (1992). Alcove: An exemplar-based connectionist model of category learning. *Psychological Review, 99*, 22–44.

Lassaline, M., & Murphy, G. (1996). Induction and category coherence. *Psychonomic Bulletin & Review, 3*, 95–99.

Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin & Review, 9*, 829–835.

Love, B. C. (2003). The multifaceted nature of unsupervised category learning. *Psychonomic Bulletin & Review, 10*, 190–197.

Maddox, W. T., & Ashby, F. G. (1993). Comparing decision bound and exemplar models of categorization. *Perception & Psychophysics, 53*, 49–70.

Maddox, W. T., & Ashby, F. G. (2004). Dissociating explicit and procedural-learning based systems of perceptual category learning. *Behavioural Processes, 66*, 309–332.

Maddox, W. T., Ashby, F. G., & Bohil, C. (2003). Delayed feedback effects on rule-based and information-integration category learning. *Journal of Experimental Psychology: Learning, Memory & Cognition, 29*, 650–662.

Maddox, W. T., Ashby, F. G., Ing, A., & Pickering, A. (2004). Disrupting feedback processing interferes with rule-based but not information-integration category learning. *Memory & Cognition, 32* (4), 582–591.

Maddox, W. T., & Filoteo, J. (2005). The neuropsychology of perceptual category learning. In I. H. Cohen & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science* (pp. 573–599). New York: Elsevier.

Medin, D., & Schwanenflugel, P. (1981). Linear separability in classification learning. *Journal of Experimental Psychology: Human learning and memory, 7*, 355–368.

Medin, D., & Wattenmaker, W. (1987). Family resemblance, conceptual cohesiveness and category construction. *Cognitive Psychology, 19*, 242–279.

Schwarz, G. (1978). Estimating dimension of a model. *Annals of Statistics, 6*, 461–464.

Shepard, R., Hovland, C., & Jenkins, H. (1961). Learning and memorization of classifications. *Psychological Monographs, 75* (13), 1–42.

Spalding, T., & Murphy, G. (1996). Effects of background knowledge on category construction. *Journal of Experimental Psychology: Learning, Memory & Cognition, 22*, 525–538.

Chapter 3

Semisupervised category learning facilitates the development of automaticity

Vandist, K., Storms, G. & Van den Bussche, E. (2019). *Attention, Perception & Psychophysics, 81* (1), 137-157.

Electronic supplementary material: the online version of this article (https://doi.org/10.3758/s13414-018-1595-7) contains supplementary material.

Abstract

In human category learning, learning is studied in a supervised, an unsupervised or a semisupervised way. The scarce human semisupervised category learning studies all focus on early learning. However, the impact of semisupervised category learning late in learning, when automaticity develops, is unknown. Therefore, in Experiment 1, all participants were first trained on the information-integration category structure for two days until they reached an expert level. Afterwards, half of the participants learned in a supervised way and the other half in a semisupervised way during two succeeding days. Both groups received an equal amount of feedback trials. Finally, all participants took part in a test day where they were asked to respond as quickly as possible. Participants were significantly faster on this test in the semisupervised group than in the supervised group. This difference was not found on day 2, implying that the no-feedback trials in the semisupervised condition facilitated automaticity. Experiment 2 was designed to test whether the higher amount of trials in the semisupervised condition of Experiment 1 caused the faster response times. We therefore created an almost supervised condition where participants almost always received feedback (95%) and an almost unsupervised condition where participants almost never received feedback (5%). All conditions now contained an equal amount of trials as the semisupervised condition of Experiment 1. The results showed that receiving feedback almost always or almost never led to slower response times than the semisupervised condition of Experiment 1. This confirms the advantage of semisupervised learning late in learning.

Introduction

Throughout the ages correct categorization has remained essential for human survival. In prehistory, classifying a predator as a harmless animal was mortal. Nowadays, classifying traffic signs correctly is important to avoid accidents. These examples explain why categorization has received continuous attention in the field of cognitive science (e.g., Ashby & Maddox, 2005; Ashby & Maddox, 2010; Medin & Schaffer, 1978; Nosofsky, 1987; Pothos & Chater, 2002). Although many categories that people use are acquired during childhood (French, Mareschal, Mermillod & Quinn, 2004), adults also learn new categories. In human category learning research the focus lies on the learning process itself. In order to understand this learning process, exemplars and non-exemplars of unfamiliar categories are typically presented (Ashby & Maddox, 2005). The behavior of participants is observed during the period when their ability to assign stimuli to these categories increases from chance level to a certain stable above-chance level (Ashby & Maddox, 2005).

Supervised and unsupervised learning

In the past, human category learning was studied using supervised or unsupervised learning paradigms. In *supervised learning paradigms*, the participant is presented with a stimulus which has to be classified in two or more contrasting categories. Immediately after this response, feedback is always provided about the correct category label. Mostly, the participant knows the number of contrasting categories in advance (see

Shepard, Hovland, & Jenkins, 1961 for a description of a basic experiment). Numerous studies have demonstrated that, based on this paradigm, participants can learn very complex categories if a sufficient amount of trials is provided (e.g., Ashby, Queller & Berretty, 1999; Ashby, Maddox & Bohil, 2002; Maddox, Filoteo, Hejl & Ing, 2004; McKinley & Nosofsky, 1995; Medin & Schwanenflugel, 1981; Maddox, Ashby, & Gottlob, 1998). In *unsupervised learning paradigms*, the participant never receives feedback or information about the category to which the presented stimulus belongs. The goal is to identify an intuitive or natural classification for a set of objects (Clapper & Bower, 1994; Love, 2002; Medin, Wattenmaker & Hampson, 1987; Milton, Longmore & Wills, 2008; Pothos & Chater, 2002, 2005; Pothos et al., 2011). Depending on the paradigm, the number of contrasting categories may or may not be known in advance. Findings based on such unsupervised paradigms reveal that performance is dominated by the use of unidimensional rules, regardless of the complexity of the underlying category structure or the number of training trials (Ashby et al., 1999). These unidimensional rules (e.g., “small stimuli belong to category A and large stimuli to category B”) are easy to verbalize and to apply, whereas complex categorization rules are mostly hard to express. In conclusion, in unsupervised learning people have the tendency to use very simple categorization rules, whereas in supervised learning participants are able to learn very complex categorization structures.

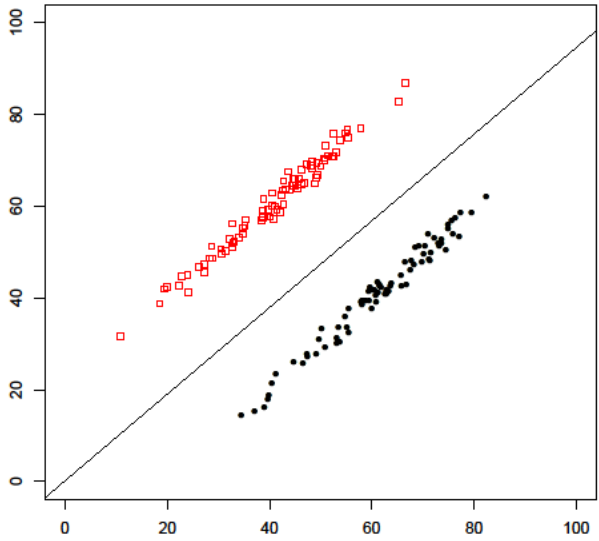
Semisupervised learning

Vandist, De Schryver and Rosseel (2009) argued that both supervised and unsupervised learning are ecologically rare. Translated to daily life, supervised learning entails that for every object that we observe we immediately receive correct information about its category label. In most category learning situations, it seems very unlikely that this occurs after each single encounter of a category member. Strictly speaking, supervised learning would imply that, when we walk in the woods, a label 'tree' is attached to every single tree. Moreover, this information is unambiguous, implying that the information provider and receiver always mean the same object. However, ambiguity occurs often in very rich environments. For example, when walking in the woods a parent might point to a bird and call it "bird" while the child may be watching a nest and hence learns the wrong label. Unsupervised learning on the other hand entails that we never receive any information about objects' categories. This implies that during our entire lives, nobody ever informs us about the name of an object or about which objects belong together. Both types of category learning therefore do not represent our daily reality. In a previous study Vandist et al. (2009) argued that people instead learn in a *semisupervised way*: when confronted with (new) objects (e.g., a dog), sometimes category information will be provided ("look, a dog") and sometimes not. This idea is supported by Gibson, Rogers and Zhu (2013). In the *semisupervised category learning paradigm* this realistic scenario is incorporated. In a block of trials a predetermined percentage of category responses is followed by feedback (i.e., feedback or labelled trials). The remaining trials do not receive feedback (i.e., no-feedback or unlabelled trials). For example, a block in

a 25% semisupervised classification learning paradigm consists of 25% feedback trials and 75% no-feedback trials.

Vandist et al. (2009) compared the effects of this semisupervised learning process to supervised and unsupervised learning processes by using the information-integration structure. This category structure is frequently used in category learning research (e.g. Ashby et al., 2002; Ashby & Ell, 2001; Ell & Ashby, 2006; Maddox, Ashby, Ing & Pickering, 2004; Maddox & Filoteo, 2011; Maddox & Ing, 2005; Maddox, Pacheco, Reeves, Zhu & Schnyer, 2010; Paul, Boomer, Smith & Ashby, 2011; Spiering & Ashby, 2008; Vermaercke, Cop, Willems, D'Hooge & Op de Beeck, 2014). Figure 3.1 shows an example of the information-integration category structure used in the experiments reported in the current study. Although the within category correlation is very high, this structure is difficult to learn. To obtain high-level performance, participants have to combine the perceptual information of the underlying stimulus dimensions simultaneously at some predecisional stage (Ashby & Gott, 1988). This perceptual integration could take many forms – in this case, by calculating a weighted linear combination of the dimensional values. The optimal decision bound is almost impossible to describe verbally (Ashby, Alfonso-Reese, Turken & Waldron, 1998) and it can hardly be discovered via an explicit reasoning process (Ashby & O'Brien, 2007), which makes the category structure difficult to master for humans (Vermaercke et al., 2014). Feedback is essential to learn the structure successfully (Ashby et al., 1999).

Figure 3.1



An example of the information-integration category structure used in the learning task in Experiment 1.The "A" stimuli are printed in squares, the "B" stimuli in solid circles. The decision bound that divides the two categories perfectly is shown in black. The Y-axis is the Orientation dimension, the X-axis the Frequency dimension.

The results of the study of Vandist et al. (2009) indicated that, as expected, learning the information-integration structure was successful in the supervised condition but not in the unsupervised condition. In a 50% semisupervised condition, participants managed to learn the structure, suggesting that feedback after every trial is not necessary to learn the complex structure.

The impact of no-feedback trials in semisupervised learning

An additional goal of the study of Vandist et al. (2009) was to understand the contribution of the no-feedback trials on the learning process. In the no-feedback trials, participants were shown a stimulus, processed it and categorized it. It was only clear after the categorization that no feedback followed. Crucially, it was investigated whether the processing of the stimulus in the no-feedback trials had an impact on the learning process or whether the experience was simply neglected. To achieve this, the no-feedback trials were replaced by irrelevant fillers, where no categorization whatsoever takes place. The number of feedback trials was identical in the condition with no-feedback trials and the condition with filler trials. The results indicated that the learning process in both conditions was similar. Hence, the no-feedback trials neither harmed nor helped learning. Apparently, when we encounter an object early in the learning process, we classify it and when no feedback follows, this has no effect on our category learning. The semisupervised learning was also studied giving feedback after 25% of the trials. In this 25% semisupervised condition, participants failed to learn the task. This failure was not due to the low relative percentage of feedback trials in a block, because when the absolute number of trials was doubled (i.e., 25% feedback was maintained, but twice as many feedback trials were received), almost all participants were able to master the category structure. Thus, when given enough trials, even 25% feedback sufficed to learn the category structure. Again, this result suggests that the no-feedback trials have little impact on the initial learning process, but that

learning is rather determined by the absolute number of feedback trials one receives.

An important question is whether these findings imply that people encounter objects, classify them and then simply delete this experience because no confirmation or correction is provided. If so, this would be in contrast to findings from machine learning where machines do use no-feedback trials to extend the knowledge gained from feedback trials. Remarkably, when supervised and semisupervised machine learning are compared, semisupervised machine learning can even achieve faster optimal performance (Chapelle, Scholkopf & Zien, 2006; Zhu & Goldberg, 2009). In machine learning semisupervised learning is therefore the most often used method, also due to practical implications: semisupervised learning requires fewer feedback items which must be annotated one by one by humans and therefore reduces time investment (Zhu & Goldberg, 2009).

Since Vandist et al. (2009) several human semisupervised category learning studies were conducted and the findings are not always consistent. In the study of Mcdonnell, Jew and Gureckis (2012) no impact of the no-feedback trials was found. In this study the category label was shown only on some trials, but not on others. In the labelled trials all stimuli originated from one subset of the full category. In the unlabelled trials, the presented stimuli covered the full category. After the training phase, the category presentation of the participant was tested. Mcdonnell et al. (2012) found that a large weight was given to the labelled stimuli, making the unlabelled trials irrelevant.

In other studies, the impact of the no-feedback trials depended on the circumstances. First, semisupervised learning was observed in a speeded

classification task but not if the responses were self-paced (Rogers, Kalish, Gibson, Harrison & Zhu, 2010). Second, participants did use the no-feedback trials when the underlying categories were distinct and the gap between the categories was big. However, if the underlying categories were more ambiguous and the space between the categories was small but still existing, no effect of the no-feedback trials was found (Vong, Perfors, & Navarro, 2014). Third, Kalish, Zhu and Rogers (2015) showed that the effect of the no-feedback trials depends on the age of the participants: young children (between 4 and 6 years old) were influenced by the no-feedback trials whereas no effects were found for older children (between 7 and 8 years old).

Finally, some studies did show that the no-feedback trials aided learning (Gibson, Rogers, Kalish & Zhu, 2015; Kalish, Zhu & Rogers, 2011; Lake & McClelland, 2011; Zhu, Gibson, Jun, Rogers, Harrison & Kalish, 2010). However, all of these studies used unidimensional stimuli and a simple underlying category structure. Feedback was always given after a specific subset of stimuli. Based on these feedback trials only, a certain decision bound that splits the two categories can be expected. The stimuli of the no-feedback trials had a different mean and distribution than the stimuli of the feedback trials because the latter were extremes of the category. If participants take these no-feedback trials into account, the decision bound will be shifted. These studies showed that participants indeed use a shifted decision bound, implying that the no-feedback trials do have an impact on the learning process (Gibson et al., 2015; Kalish et al., 2011; Lake & McClelland, 2011; Zhu et al, 2010). Still, it is unlikely that in our daily life feedback is always provided after the same subset of examples and other examples of the category are never followed by

feedback. Contrarily, we believe that every example of a category can be followed by feedback.

Automaticity

Given the inconsistent research results on human semisupervised learning and the advantages of semisupervised learning in machines, we aim to further investigate the role of no-feedback trials in human semisupervised category learning. In the current study, we specifically investigate the role of no-feedback trials in developing *automaticity*. Once a learner reaches automaticity, cognitive or motor skills are executed faster, more accurately and require less attention in comparison to initial learners (Ashby & Crossley, 2012; Ashby, Turner, & Horvitz, 2010). Although various definitions and criteria of automaticity exist researchers agree that automaticity is the result of extensive overtraining after the skilled behavior is well learned (Ashby et al., 2010; Hélie, Waldschmidt, & Ashby, 2010; Moors & De Houwer, 2006; Schneider & Chein, 2003; Nosofsky & Palmeri, 1997; Shiffrin & Schneider, 1977). Especially in categorization this is the main consensus since several studies showed that the criteria for automaticity proposed by Schneider and Shiffrin (1977) as no interference of dual task performance (Waldron & Ashby, 2001; Zeithamova & Maddox, 2006, 2007) and decrease in performance after switching keys (Ashby, Ell & Waldron, 2003; Maddox, Bohil, & Ing, 2004; Maddox, Glass, O'Brien, Filoteo & Ashby, 2010; Spiering & Ashby, 2008) already apply for initial information-integration category learning. Consequently, in this article automaticity

will be defined as the result of overtraining after good performance was obtained.

In cognitive science, two influential models of expertise presented in the literature are Logan's (1988) instance theory of automaticity and Rickard's (1997) component power laws theory. Both models assume that feedback remains essential through the development of automaticity and hence only make predictions about supervised learning, not about semi-supervised learning. In category learning, two important models explicitly deal with automaticity: the Exemplar-based random walk model (EBRW-model) of Nosofsky and Palmeri (1997) and the Subcortical Pathways Enable Expertise Development (SPEED model) of Ashby, Ennis and Spiering (2007). The EBRW-model assumes that expertise develops as the number of stored exemplars increases. The more stored exemplars, the faster the response will be elicited. Since only exemplars followed by feedback will be stored (and activated as belonging to the category), supervised learning is essential and no clear predictions can be made about semisupervised learning based on this model. For semisupervised learning, the most relevant model about the development of automaticity in the information-integration structure is the SPEED-model. In category learning, an important model about the development of automaticity in the information-integration structure is the Subcortical Pathways Enable Expertise Development or SPEED model (Ashby, Ennis, & Spiering, 2007). The SPEED-model assumes that categorization is regulated by two different pathways, a slow and a fast one. The *slow pathway* is supposed to originate in the visual cortex, passes by the basal ganglia and the thalamus and ends in the premotor cortex. This is an indirect and subcortical pathway which includes at least four synapses. When positive feedback is given (after a correct

categorization), dopamine in the striatum will be released and the active synapses will be strengthened. When negative feedback is given (after an incorrect answer) or no feedback at all, the strength of the synapses will be weakened. On the contrary, Ashby et al. (2007) state that the *fast pathway* only involves one synapse. This is a direct route from the visual association areas to the premotor cortex. In this cortical-cortical pathway, synapses are strengthened when there is both pre- and postsynaptic activation (i.e., Hebbian learning). This occurs independent of feedback.

In the SPEED-model, the development of categorization automaticity is defined as a gradual process. Early in learning, the main pathway is the slow subcortical pathway. As learning progresses, the fast cortical-cortical pathway becomes more salient and the subcortical pathway becomes less important. Eventually, experts only rely on the cortical-cortical pathway for their categorization (Ashby et al., 2007). Because this pathway is independent of feedback and the strength of the connections increases with the number of categorization responses, SPEED predicts that late in learning every type of trial will strengthen the connections, regardless of whether the categorization response is followed by feedback. Hence, late in learning, adding extra no-feedback trials to the training would have an impact on the development of automaticity and faster response times can be expected.

To test this hypothesis in Experiment 1, participants were trained in a supervised way on the information-integration structure for two days. After reaching an expert level with regards to the trained category structure, half of the participants continued to practice supervisedly on days 3 and 4. The other half practiced according to a 25% semisupervised

scheme. Both groups of participants received an equal amount of feedback trials, implying that the semisupervised group received four times as many trials as the supervised group. Based on the premises of the SPEED-model, we hypothesize that the categorization in the semisupervised condition will be more automatic, as indexed by faster response times. If the no-feedback trials in semisupervised learning have no impact on the automaticity process, a similar level of automaticity (and thus equal response times) should be observed in both conditions.

Experiment 1

Method

Participants. In total 34 participants (22 women, average age 21.4 years, *SD*=1.97, range=18-26 years) took part in the experiment in return for payment. If participants participated for two days, they received 20 euro; if they participated for five days, the payment was 35 to 40 euro.

Design. The experiment was organized on five consecutive days. In this way learning could benefit from between-session consolidation due to sleep (Censor, Karni, & Sagi, 2006; Stickgold, James, & Hobson, 2000; Stickgold, Whidbee, Schirmer, Patel, & Hobson, 2000; Stickgold & Walker, 2005). Participants were randomly divided into two conditions: the semisupervised condition (N = 19) and the supervised condition (N = 15). The first two days were equal for both conditions: on each of these days, 400 training trials were presented, divided into five blocks of 80 trials. Each trial was followed by feedback. The goal of this training phase was to master the category structure. Participants who

achieved an average accuracy rate of 90% or more on the last two blocks of the second day were invited to the following phase. Participants who did not reach this expert level were excluded from the remainder of the experiment. On the third and fourth day, participants in the semisupervised condition were presented with 640 trials (8 blocks of 80 trials) and 25% of these trials were randomly followed by feedback. Participants in the supervised condition were shown 160 trials (2 blocks of 80 trials) that where all followed by feedback. Consequently the number of feedback trials was equal in both conditions. On the fifth and final day the test phase took place where all participants received 134 trials in one block. None of these trials were followed by feedback. It was decided to organize this 'test' on a new day, ensuring that participants in both conditions were equally fit. Table 3.1 summarizes the differences between the two conditions.

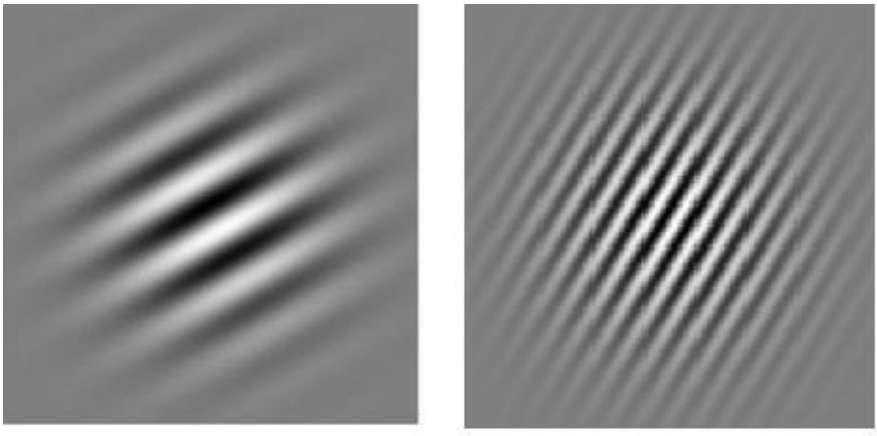
Table 3.1

Condition	Type of trial	Day 1	Day 2	Day 3	Day 4	Test
Semisupervised	Feedback	400	400	160	160	0
	No-feedback	0	0	480	480	134
Supervised	Feedback	400	400	160	160	0
	No-feedback	0	0	0	0	134

Number of trials in each condition of Experiment 1

Stimuli and apparatus. The experiment was conducted using Tscope (Stevens, Lammertyn, Verbruggen & Vandierendonck, 2006). Participants viewed the stimuli on a 17 inch LCD monitor with an 800x600 resolution at a distance of approximately one arm's length. The stimuli were gray 300x300 square-pixel Gabor patches, presented on a black screen. Two examples of Gabor patches can be seen in Figure 3.2. In this study the 'gratings' varied continuously on two dimensions: the spatial orientation and the spatial frequency. These dimensions are perceptually separable. The arbitrary stimulus coordinates were converted to physical units using the following transformations: spatial orientation was referred to in x degrees, with x varying between 0 to 100 degrees. Spatial frequency, expressed in cycles/pixel, was converted using $f(y)=0.01+(y/1500)$, with y varying between 0 and 100. These coordinates originated from the information-integration category structure, an example of which is displayed in Figure 3.1.

Figure 3.2



Two examples of Gabor patches.

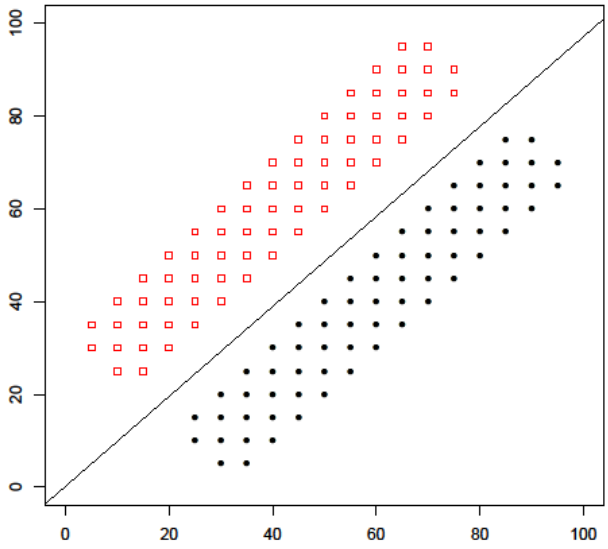
The optimal decision bound, which classifies the stimuli perfectly in two categories, is diagonal. In the semisupervised condition, participants viewed 2080 stimuli in the first four days. These stimuli were generated by randomly sampling from two bivariate normal distributions, leading to 1040 'A' stimuli and 1040 'B' stimuli. Category A had a different mean than category B, but the variance and the covariance of both categories were the same. Due to random sampling, the optimal decision bound varied slightly from block to block, although the mean optimal decision bound in one day was $y=x$. The exact parameter values are shown in Table 3.2. As in the Ashby et al. (1999) study, the mean, the variance and the covariance values were chosen in such a way that a linear decision bound based on one dimension would account for an accuracy of maximum 80%. The stimuli in the supervised condition were constructed in the same way as in the semisupervised condition, the only difference is that in this condition participants viewed 1120 stimuli in the first four days, 560 'A' stimuli and 560 'B' stimuli. On the last day, all participants viewed 134 fixed stimuli, half of which were depicted from the stimuli A range; the other half originated from the stimuli B range as can be seen on Figure 3.3. Again, the optimal decision bound was $y=x$ and the category mean was identical to that of the previous days.

Table 3.2

Category	A	B
Mean (frequency)	40	60
Mean (orientation)	60	40
SD (frequency & orientation)	11.88	11.88
Correlation	0.99	0.99
Average optimal slope	1	
Average optimal intercept	0	

Parameter values that define the categories of Experiment 1.

Figure 3.3.



An example of the information-integration category structure used in the test (= day 5) in Experiment 1.The "A" stimuli are shown in squares, the "B" stimuli in solid circles. The decision bound that divides the two categories perfectly is shown in black. The Y-axis is the Orientation dimension, the X-axis the Frequency dimension.

Procedure. All participants were tested individually in a dimly lit room. On the first two days, participants were informed that they would see stimuli that would appear one by one and that originated from two categories A and B. They were asked to respond by pressing A on the keyboard if they believed that the stimulus was an A and to press B when they believed the stimulus was a B. Participants were informed that they would receive feedback (i.e., the true category label) after each category response. They were also informed that it was possible to do the task without errors. Participants were told that at the end of day 2 the accuracy

would be calculated and only the participants who achieved an average accuracy of 90% or more would be allowed to continue to the next days. At the end of each block, the percentage of the correct responses was printed on the screen. This percentage additionally encouraged them to do better in the next block. A trial started when a stimulus was projected in the middle of the screen until the participant responded. Immediately after the response the stimulus disappeared. The response time was self-paced. After the response, the feedback, consisting of correct/incorrect and the right category label, became visible at the bottom of the screen for 1500 milliseconds. After that a new trial started. The procedure on day 3 and 4 was similar, except that in the semisupervised condition participants were informed that some trials would be followed by feedback and other trials not. In the no-feedback trials the stimulus disappeared immediately after the response and the screen remained blank for 1500 milliseconds. Afterwards a new trial started. Hence, there is no difference in post-response events on feedback trials and no-feedback trials except on the appearance of the feedback on the screen. Again, in both conditions participants were informed that it was possible to obtain maximum accuracy and they were encouraged to achieve this. On day 5 participants were informed that they would see similar stimuli as in the preceding days and that feedback would no longer be given. In contrast to the first four days, participants were urged to respond as quickly as possible. After a response was given, the stimulus disappeared and the screen remained blank during 1500 milliseconds, after which a new trial started. To encourage the participants to respond as fast as possible, the stimulus disappeared when a response was not given within a time limit of 1800 milliseconds. In this case, the message “Respond faster” was shown during the subsequent intertrial interval of 1500

milliseconds. The trials where the participant responded too slowly, were presented again at the end of the block. Thus, for each participant we collected 134 valid categorization responses on day 5.

Results

Selection of participants. Before analyzing the response time patterns in both conditions, it was essential to ensure that the participants mastered the category structure at the end of day 2. Therefore accuracy and model-based analyses were performed. High accuracy rates indicate that the participant made few errors. Nevertheless, it is still unclear whether these errors were just random mistakes or systematic faults. The model-based analyses are a necessary complement to the accuracy. These models were calculated based upon the responses of the participant on the last two blocks of day 2. For each model the corresponding BIC-score were calculated. The best fitting model was the model with the lowest BIC-score. This model is supposed to correspond to the strategy that the participant used to solve the categorization task. The strategy that matched perfect performance in this experiment is called the optimal decision bound. Combining both criteria rules out the possibility that a participant increased the accuracy during the experiment but this improvement was not reflected in the model-based analyses. This can be the case if the errors are systematic and another strategy than the optimal decision bound was preferred. Therefore, only participants that passed both criteria were retained for further analyses.

Criterion 1: High accuracy. Participants could achieve perfect accuracy when using the optimal decision bound (i.e., by integrating the information from the two stimulus dimensions at some predecisional stage). On the other hand, using a unidimensional decision rule, the accuracy could never exceed 80%. This implies that participants with an achieved accuracy of more than 80% probably adopted a (suboptimal) information-integration decision rule. However, since our study aimed at studying automaticity after becoming an expert learner, category learning was considered successful when an average performance of at least 90% was obtained. As can be seen in Table 3.3, 11 participants did not pass this criterion and were therefore excluded from further analyses.

Criterion 2: Optimal decision bound. Figures 3.1 and 3.2 in the supplementary materials show the actual responses during the last two blocks of day 2 for each of the participants retained after applying Criterion 1. These responses (i.e., whether a stimulus belongs to category A or category B) form the basis on which the individual decision bounds were calculated. Four different types of models were fit to each participant's response (see the Appendix for details). These models were introduced by Ashby and Gott (1988) and Ashby and Maddox (1993). Three models, namely the horizontal unidimensional model (DIM-O), the vertical unidimensional (DIM-V) and the general conjunctive classifier (GCC), are rule-based. If participants adopted one of these category decision strategies, category learning failed. The last model, the general linear classifier (GLC), is an information-integration model. Only with this category decision strategy perfect accuracy could be obtained. Consequently, if the general linear classifier was used and this decision bound fell in between the two categories, learning was successful. The model parameters were estimated using the method of

maximum likelihood. To select the best-fitting model to the data, the model with the smallest Bayesian Information Criterion (BIC) was selected. The BIC penalizes for the number of free parameters. BIC is defined as BIC = $r\ln N - 2\ln L$, where r is the number of free parameters, N is the sample size and L is the likelihood of the model given the data (Schwarz, 1978). The BIC values of the four models for each participant are presented in Table 3.3. In the semisupervised condition, all participants except participant 4 favoured the general linear classifier. Hence, participant 4 was excluded from further analyses. As can be seen in Figure 3.1, all optimal decision bounds fell between the two categories. In the supervised condition, all participants favoured a strategy based on the general linear classifier. As can be seen in Figure 3.2, all optimal decision bounds fell between the two categories except for participant 15. Hence, participant 15 was excluded from further analyses. As a result, the final sample used in the subsequent analyses consisted of 21 participants (N=11 semisupervised and N=10 supervised), the average age was 21.3 years (*SD*=1.88, range 18 to 24 years), and 15 of them were women. In the following sections, four types of analyses are described: accuracy and model-based analyses to define the strategy used by the participant, response time (RT) analyses and the speed-accuracy trade-off analyses. The response time analyses were studied to compare the semisupervised learning process to the supervised learning process.

Accuracy analysis. Figure 3.4 shows the average percentage of correct responses and the 95% confidence intervals on each block of trials received during the first four days for the supervised and

semisupervised condition separately. In the semisupervised condition, the accuracy was based on the feedback trials only. Eighty feedback trials were grouped into a block to facilitate comparison to the supervised condition. As expected, the learning process was similar in both conditions. The mean accuracy increased from an average of 73% (*SD*=11.06) in the first block for the semisupervised condition and an average of 75% (*SD*=11.72) for the supervised condition to almost perfect accuracy in the last block of day two (97%, *SD*=2.76 and 95%, *SD*=3.62 respectively). During the blocks on day 3 and 4, the mean accuracy remained high in both conditions. In the semisupervised condition, mean accuracy on the last response block on day 3 was 98% (*SD*=1.80) and 98% on day 4 (*SD*=1.89). Similarly, in the supervised condition mean accuracy on the last block was 96% (*SD*=2.95) on day 3 and 96% (*SD*=3.41) on Day 4. A repeated measures ANOVA was conducted to determine whether the mean accuracy on the last two blocks differed depending on the day (4 levels: day 1, 2, 3 and 4) and condition (2 levels: supervised and semisupervised). Not surprisingly, there was a main effect of day, $F(3,17)=13.97$, $p<.001$, $\eta_p^2=.71$, indicating that the accuracy increased during the succeeding days.

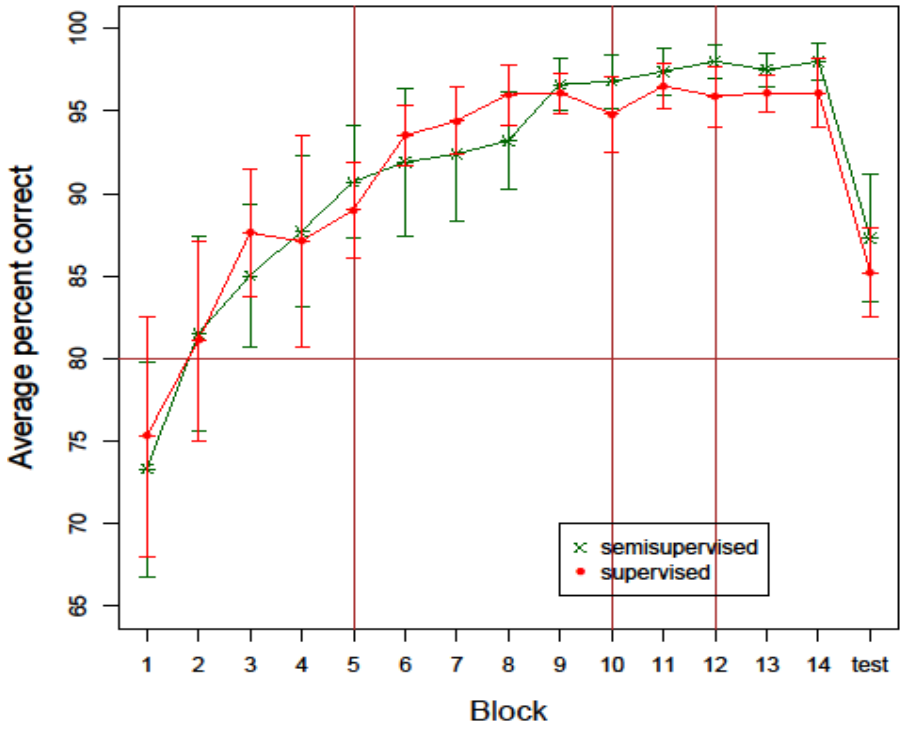
Table 3.3

Condition	Participant	GLC	DIM-O	DIM-F	GCC	Mean accuracy (%)	Continued to day 3
Semisupervised	1	120,76	210,91	151,31	113,88	86,88	no
	2	36,84	176,66	118,31	61,79	98,13	yes
	3	18,44	148,80	167,54	52,90	99,38	yes
	4	103,40	184,60	179,89	75,66	91,88	no
	5	163,81	173,48	225,16	164,46	78,75	no
	6	81,82	148,98	190,57	105,74	93,75	yes
	7	115,00	215,45	143,54	126,24	86,88	no
	8	15,23	146,18	150,06	67,04	100,00	yes
	10	106,76	240,21	104,52	109,57	78,13	no
	11	67,49	202,81	121,77	78,83	94,38	yes
	12	75,97	160,99	170,85	78,36	95,00	yes
	13	45,92	144,50	171,39	85,92	97,50	yes
	15	113,02	233,09	113,54	118,61	81,88	no
	16	45,21	182,43	146,23	95,33	96,88	yes
	17	72,62	166,22	173,66	110,02	95,63	yes
	18	187,60	230,57	191,50	187,31	73,75	no
	19	47,01	117,47	202,99	66,36	95,38	yes
	20	33,64	143,68	163,89	74,42	98,75	yes
	26	154,79	241,89	152,85	154,58	74,38	no

Supervised	1	<u>96,75</u>	176,61	173,71	97,82	93,13	yes
	2	<u>51,72</u>	164,90	156,70	79,18	97,50	yes
	3	<u>36,10</u>	156,05	165,01	89,58	98,13	yes
	4	<u>58,12</u>	165,86	161,56	83,26	96,88	yes
	5	<u>115,82</u>	160,44	199,81	117,14	89,38	no
	6	<u>108,42</u>	215,85	128,97	127,23	88,13	no
	7	<u>90,37</u>	185,83	160,68	106,73	92,50	yes
	8	<u>116,05</u>	204,53	159,70	124,89	89,38	no
	9	<u>49,23</u>	171,32	134,58	86,52	96,25	yes
	10	<u>137,92</u>	227,78	148,88	142,17	83,13	no
	11	<u>63,21</u>	159,75	180,47	103,53	96,25	yes
	12	<u>49,95</u>	150,89	159,37	62,27	96,88	yes
	14	<u>80,05</u>	168,53	176,18	90,21	95,00	yes
	15	<u>89,64</u>	144,01	206,52	102,48	91,88	no
	18	<u>98,25</u>	149,18	194,89	107,51	91,88	yes

Mean accuracy (%) and model-based analyses (BIC scores) of the last 2 blocks of day two (i.e., blocks 9 and 10) for every participant of Experiment 1. The best fitting model is underlined. Only participants who reached an accuracy of minimal 90% and had the best fitting decision bound model based on the GLC, continued on to day 3 of the experiment.
Note: DIM-O: unidimensional classifier based on the orientation; DIM-V: unidimensional classifier based on the frequency; GCC: General Conjunctive Classifier; GLC: General Linear Classifier

Figure 3.4



Mean accuracy (%) along with the 95% confidence intervals by block for all participants in the Semisupervised and Supervised Condition of Experiment 1 from day 1 (blocks 1 to 5), day 2 (blocks 6 to 10), day 3 (blocks 11 to 12), day 4 (block 13 to 14) and day 5 (test). Only the participants who reached a performance of minimum 90% at the end of day 2 and revealed a decision bound based on the optimal decision bound were included in the analyses. In the Semisupervised condition, the accuracy for every 80 feedback trials is used.

Paired sample *t*-tests using the Bonferroni correction for multiple comparisons showed that in comparison to day 1, mean accuracy was significantly higher on days 2, 3 and 4 (resp. $t(20)=6.41$, $p<.001$; $t(20)=6.80$, $p<.001$; $t(20)=7.02$, $p<.001$). Participants also made less errors on day 3 compared to day 2, $t(20)=2.18$, $p<.042$ There was no

main effect of condition, $F(1,19)=1.71$, $p=.21$, $\eta_p^2=.08$, nor an interaction between day and condition ($F<1$, $p=.99$, $\eta_p^2=.007$), suggesting that the accuracy in both conditions increased similarly across days. The accuracy on the test (day 5), where the speed of responding was stressed, was lower in both conditions compared to the accuracy reached at the end of day 4: 87% (SD=6.53) in the semisupervised condition and 85% (SD=4.37) in the supervised condition. This difference was significant for the semisupervised condition, $t(10)=5.42$, $p<.001$, $d=1.64$, and for the supervised condition $t(9)=7.46$, $p<.001$, $d=2.36$. Finally, an independent sample t-test revealed that there was no difference in accuracy between the two conditions on the fifth day, $t(19)=0.91$, $p=.37$, $d=0.40$.

Model-based analysis. Table 3.4 shows the four model fits on day 4. In both conditions, all participants preferred a decision bound based on the general linear classifier, indicating successful learning. Figures 3.3 and 3.4 in the supplementary materials show the actual responses during the last two blocks of day 4 for each participant. Table 3.5 presents the model fits on the test day (i.e., day 5): in the semisupervised condition, most participants chose a strategy based on the optimal decision bound, except for participants 6, 12 and 19. For these participants a strategy based on the general conjunctive classifier fitted slightly better than the optimal decision bound. Figures 3.5 and 3.6 in the supplementary materials present the responses and the best fitting decision bound for every participant during the test day. In the supervised condition, most participants preferred the optimal decision bound, except for participants 1 and 18. For participant 1, the BIC-score of the general conjunctive

classifier was slightly lower than the general linear classifier. Participant 18 clearly preferred a strategy based on the general conjunctive classifier.

Table 3.4

Condition	Participant	GLC	DIM-O	DIM-F	GCC	Mean accuracy (%)
Semi-supervised	2	37,21	152,60	156,77	83,20	98,75
	3	40,21	142,25	169,50	68,89	100,00
	6	67,27	157,05	184,21	80,58	96,88
	8	15,23	151,83	140,60	64,52	97,50
	11	15,23	150,92	162,51	33,45	98,75
	12	41,56	161,91	156,19	82,15	95,00
	13	28,17	153,32	155,14	58,71	98,75
	16	75,25	162,52	157,88	81,30	95,63
	17	25,30	159,73	127,14	62,00	98,13
	19	57,56	138,92	183,19	68,14	96,25
	20	15,23	140,19	145,79	61,59	99,40
Supervised	1	44,37	270,08	164,14	82,63	97,50
	2	47,95	258,59	160,92	80,58	97,50
	3	80,20	256,55	166,29	99,31	94,38
	4	64,59	254,25	137,36	96,71	95,00
	7	91,93	262,33	160,30	108,01	93,13
	9	58,14	258,02	158,47	73,30	96,88
	11	35,34	259,01	143,44	80,13	98,75
	12	40,75	265,04	166,17	85,23	98,13
	14	64,73	258,52	164,25	89,22	96,25
	18	81,47	262,97	189,35	102,19	93,75

Mean accuracy (%) and model-based analyses (BIC scores) on the last two blocks of day 4 (i.e., blocks 7 to 8 semisupervised condition; blocks 1 to 2 supervised condition) for every participant of Experiment 1. The best fitting model is underlined.
Note: DIM-O: unidimensional classifier based on the orientation; DIM-V: unidimensional classifier based on the frequency; GCC: General Conjunctive Classifier; GLC: General Linear Classifier

Table 3.5

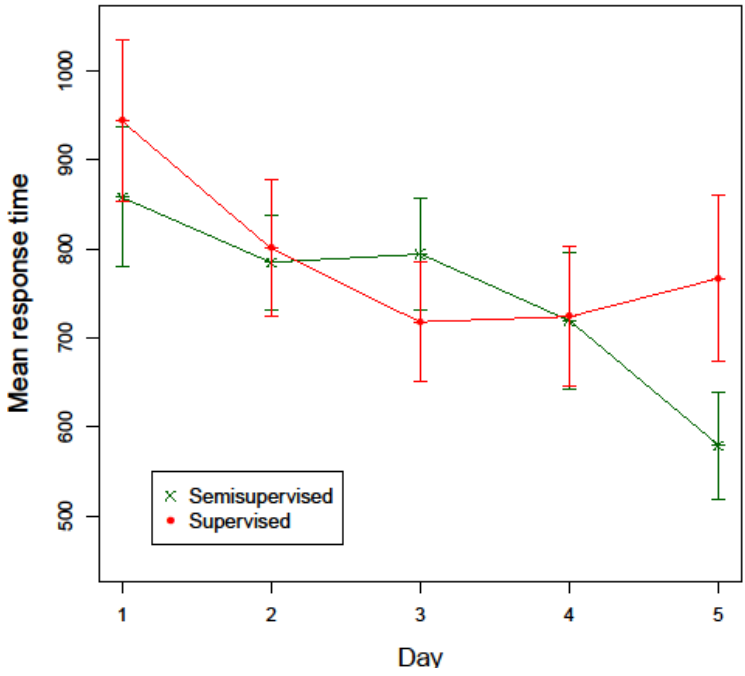
Condition	Participant	GLC	DIM-O	DIM-F	GCC	Mean accuracy (%)
Semi-supervised	2	66,35	166,04	173,69	113,15	94,78
	3	57,34	161,63	175,74	99,05	95,52
	6	146,23	178,73	186,08	141,14	80,60
	8	111,43	154,48	193,88	122,96	84,33
	11	69,64	179,10	160,71	93,81	92,54
	12	118,17	171,33	183,52	111,44	85,07
	13	148,39	159,83	191,96	148,86	73,88
	16	98,29	162,05	185,89	121,93	88,81
	17	91,97	179,02	168,41	116,15	89,55
	19	116,41	169,27	187,80	111,35	84,33
	20	58,78	149,03	184,81	92,53	91,79
Supervised	1	130,18	179,16	182,94	127,17	84,33
	2	103,71	202,83	154,30	120,48	85,83
	3	125,65	168,79	189,29	139,90	83,58
	4	127,66	140,63	197,70	139,62	76,87
	7	69,25	174,38	166,44	109,05	92,54
	9	100,43	171,17	171,63	103,40	85,82
	11	64,10	145,70	181,16	91,09	89,55
	12	95,96	157,49	188,50	123,50	87,31
	14	104,43	153,38	174,07	130,30	80,60
	18	102,03	149,89	197,79	74,00	85,07

Mean accuracy (%) and model-based analyses (BIC scores) on all trials of day 5 for every participant of Experiment 1. The best fitting model is underlined.
Note: DIM-O: unidimensional classifier based on the orientation; DIM-V: unidimensional classifier based on the frequency; GCC: General Conjunctive Classifier; GLC: General Linear Classifier

Analysis of the response times. The mean response times (RTs) along with the 95% confidence intervals for the semisupervised and

supervised condition from day 1 to 4 are presented in Figure 3.5. These RTs were calculated on the last two blocks of each day. For day 1, the mean RTs in the supervised condition was 944 ms (*SD* =145.48) whereas the mean RTs in the semisupervised condition was 858 ms (*SD*=131.22). Importantly for this investigation is that participants were equally fast in the last two blocks of day 2 (semisupervised mean RTs=785 ms, *SD*=90.10 and supervised mean RTs=801 ms, *SD*=122.22). This was confirmed by an independent sample *t*-test, t (19)=0.35, p=.73, d=*0.15*. On days 3 and 4, the mean RTs slowly decreased in the semisupervised condition to, respectively, 794 ms (*SD*=106.53) and 719 ms (*SD*=128.59). This decrease in mean RTs was also observed in the supervised condition: 719 ms (*SD*=108.98) on day 3 and 724 ms (*SD*=126.85) on day 4. A repeated measures ANOVA was conducted to determine whether the mean RTs on the last two blocks differed depending on the day (4 levels: day 1, 2, 3 and 4) and condition (2 levels: supervised and semisupervised). Not surprisingly, there was a main effect of day, $F(3,17)=7.12$, p=.003, η_p^2=*.56*, indicating that the RTs decreased during the succeeding days. There was no main effect of condition, $F<1$, p=.84, η_p^2=.002, but the interaction between day and condition reached significance $F(3,17)=3.92$, p=.027, η_p^2=.41. Post-hoc paired-sample *t*-tests, adjusted by the Bonferroni correction for multiple comparisons, indicated that in the semisupervised condition participants did not speed up by day, none of the paired sample *t*-tests was significant, all $p>.09$. In the supervised condition, responses on later days were all faster in comparison to day 1 (all $p < .02$). None of the other comparisons were significant (all p>.06). The decrease in RTs is thus only present in the supervised condition.

Figure 3.5



Mean response times (ms) and 95% confidence intervals for the semisupervised and the supervised condition of Experiment 1 calculated on the last two blocks of each day.

Decisive to test our hypotheses was the difference in RTs on day 5. In the semisupervised condition the mean RT was 579 ms (*SD*=104.08) whereas the mean RT in the supervised condition was 767 ms (*SD* =153.20). An independent sample *t*-test confirmed that this difference in RTs on the test day was significant, $t(19)=3.32$, $p=.004$, $d=1.45$. Participants in the semisupervised condition responded significantly faster than participants in the supervised condition. Paired sample *t*-tests showed that in the semisupervised condition, participants responded significantly faster on day 5 compared to day 4, $t(10)=4.71$, $p=.001$,

d=1.42 whereas participants in the supervised condition responded equally fast on days 4 and 5, $t(9)=1.39$, $p=.20$, *d=0.44.*

Speed-accuracy trade-off. On the fifth day, participants were instructed to respond as fast as possible. This might result in a speed-accuracy trade-off (SAT): participants gave up decision accuracy in favour of decision speed (see Heitz, 2014).The speed-accuracy trade-off was calculated for each condition. Since data points are limited and errors are rare, the SAT is calculated by the Pearson correlation between the mean RT and the mean accuracy rate (Heitz, 2014). In the supervised condition, there was no SAT-effect, $r=-.21$, $p=.57$. In the semisupervised condition, there was a SAT-effect, $r=.61$, $p=.046$, implying that the faster participants responded, the more errors they made.

Discussion

The objective of Experiment 1 was to test the hypothesis based on the SPEED-model that late in learning, when automaticity develops, participants benefit from semisupervised learning, resulting in faster RTs. Therefore participants were trained during two days until a certain expertise was gained and then either received feedback on all trials (supervised condition), or on 25% of the trials (semisupervised condition) for the next two days. On day 3 to 4 both conditions received an equal amount of feedback trials but the total number of trials differs: in the semisupervised condition, participants categorized a quadruple of trials compared to the supervised condition. The results clearly showed

that the mean RTs on the test (day 5) were significantly faster in the semisupervised condition than in the supervised condition. The participants in the semisupervised condition revealed more automatic behavior than the participants in the supervised condition. Importantly, this difference in mean RTs on day 5 was not observed on day 2, ruling out the possibility that participants in the semisupervised condition were always faster. On day 5 a SAT-effect occurred in the semisupervised condition: participants who tended to respond fast, also made more errors. This was not the case in the supervised condition. Note that the mean accuracy was similar on day 5 in both conditions: even though semisupervised participants sacrificed accuracy for response times, they still performed at the same level with regards to accuracy as supervised participants.

There are three possible explanations for these findings. The first is that semisupervised learning does have an impact late in learning when automaticity develops and that it leads to faster RTs. Second, confirming the SPEED-model, it is possible that the higher number of trials in the semisupervised condition is accountable for the faster RTs. On day 3 and 4 participants in the semisupervised condition responded to a fourfold amount of trials. The SPEED-model postulates that late in learning the fast pathway is dominant. This pathway is assumed to be independent of feedback. According to the SPEED-model, it is the multiple repetitions in the semisupervised condition that lead to faster response times, because regardless whether or not these trials are followed by feedback. Third, the results may have been influenced by a confound. The participants in the supervised condition never experienced no-feedback trials before day 5 and this inexperience might have slowed their performance. In Experiment 2 these explanations were addressed.

Experiment 2

Experiment 2 again examines whether late in learning the nature of feedback (i.e., feedback on every trial, occasional feedback or no-feedback) has an impact on the development of automaticity, indicated by faster RTs. In order to do this without the confounds present in Experiment 1, two control conditions are run and compared to the semisupervised condition of Experiment 1: an almost supervised condition, in which 95% of the trials are followed by feedback and an almost unsupervised condition, in which 5% of the trials are followed by feedback[1]. Instead of using fully supervised and unsupervised control conditions, 'almost' supervised and unsupervised conditions are purposefully chosen so that all participants have experienced no-feedback trials prior to day 5, ruling out the possibility that the novelty of no-feedback trials on day 5 influences the RTs of the supervised condition. The total amount of trials in these two new conditions corresponds to the semisupervised condition of Experiment 1, to test the hypothesis that just the higher amount of trials in the semisupervised condition of Experiment 1 led to faster RTs on day 5.

[1] Technically, the almost supervised condition and the almost unsupervised condition are semisupervised conditions as well, since these conditions consist of supervised and unsupervised trials. However, for clarity, when we refer to the "semisupervised condition" we intend the 25% semisupervised condition throughout the manuscript.

If the total amount of trials is indicative for the development of automaticity, similar RTs are expected in these two new conditions as in the semisupervised condition of Experiment 1. Remark that this is the outcome predicted by the SPEED-model (Ashby et al., 2007). Contrarily,

when the RTs in the two new conditions differ from the semisupervised condition of Experiment 1, this effect will be due to the different amount of feedback trials. If RTs in the *almost supervised condition* are faster than the semisupervised condition, either the higher amount of feedback trials in the supervised condition aids automaticity, or the higher amount of no-feedback trials in the semisupervised condition slows down learning. On the other hand, if RTs in the almost supervised condition are slower than the semisupervised condition, semisupervised learning aids the development of automaticity, despite the lower amount of feedback trials.

For the *almost unsupervised condition*, the SPEED-model predicts similar RTs as in the semisupervised condition of Experiment 1, since only expert participants are selected. When RTs do differ from the semisupervised condition, this will provide us with insight into the minimum amount of feedback trials needed to successfully develop automaticity.

Method

Participants, design, stimuli, apparatus and procedure. In total 38 participants (28 women, average age 20.74 years, *SD*=3.18, range=18 – 30 years) took part in the experiment in return for payment. The background of the participants was similar to the participants of Experiment 1. Also, the time of testing in the academic year was comparable. If participants participated for two days, they received 10 euro; if they participated for five days, the payment was 30 euro. Participants were randomly divided into two conditions: the almost

supervised condition (N=19) and the almost unsupervised condition (N=19). The organization of Experiment 2 was identical to the semisupervised condition of Experiment 1 except on the third and fourth day. In the almost supervised condition, participants were presented with 640 trials (8 blocks of 80 trials) and 95% of these trials were randomly followed by feedback, resulting into 608 feedback trials on days 3 and 4. In the almost unsupervised condition participants also received 640 trials (8 blocks of 80 trials) but only 5% of these trials were randomly followed by feedback, resulting into 32 feedback trials on days 3 and 4. Note that the total amount of trials in the semisupervised condition of Experiment 1 is the same as the total amount of trials in the two conditions of Experiment 2. Table 3.6 presents the conditions schematically.

Results

Selection of participants. As in Experiment 1, accuracy and model-based analyses were performed to ensure that the participants mastered the category structure at the end of day 2.

Criterion 1: High accuracy. Recall that participants could achieve perfect accuracy in this task. As in Experiment 1, the criterion was an average performance of at least 90% on the last two blocks of day 2. As can be seen in Table 3.7, 10 participants did not pass this criterion and were therefore excluded from further analyses.

Table 3.6

Condition	Type of trial	Day 1	Day 2	Day 3	Day 4	Test
Semisupervised	Feedback	400	400	160	160	0
Experiment 1	No-feedback	0	0	480	480	134
	Total	400	400	640	640	134
Almost supervised	Feedback	400	400	608	608	0
	No-feedback	0	0	32	32	134
	Total	400	400	640	640	134
Almost unsupervised	Feedback	400	400	32	32	0
	No-feedback	0	0	608	608	134
	Total	400	400	640	640	134

Number of trials in each condition of Experiment 2 and the semisupervised condition of Experiment 1

Criterion 2: Optimal decision bound. Figures 3.7 and 3.8 in the supplementary materials show the actual responses during the last two blocks of day 2 for each participant who passed the accuracy criterion. As in Experiment 1, these responses were used to calculate the individual decision bounds of four different models and the corresponding BIC-scores. The model with the lowest BIC-score is presumably the strategy that the participant adopted in the last two blocks of day 2.

Table 3.7

Condition	Participant	GLC	DIM-O	DIM-F	GCC	Mean accuracy (%)	Continued to day 2
95%	6	43,92	278,61	158,77	83,84	98,13	yes
	8	34,52	262,25	164,44	88,15	98,75	yes
	10	120,39	259,54	132,87	127,56	83,13	no
	11	90,23	274,58	164,66	92,92	93,75	yes
	12	101,52	258,13	187,90	125,69	91,88	yes
	13	143,06	242,65	144,00	143,52	78,75	no
	14	68,35	264,54	187,13	85,75	95,63	yes
	15	64,38	258,93	169,46	98,06	96,25	yes
	16	104,53	236,14	134,35	84,65	86,88	no
	17	28,63	252,74	151,33	64,86	98,75	yes
	18	15,23	272,62	125,92	79,57	100,00	yes
	19	59,41	264,04	157,14	95,63	96,88	yes
	20	33,80	261,05	153,41	157,96	98,75	yes
	21	39,00	257,39	155,33	91,06	98,13	yes
	22	55,66	274,62	147,26	89,00	96,88	yes
	30	87,82	274,10	174,76	105,20	93,75	yes
	31	61,31	273,55	167,50	66,92	96,25	yes
	33	66,43	260,26	164,10	100,90	96,25	yes
	34	112,86	260,82	135,03	132,94	85,63	no
5%	1	122,08	256,69	157,20	133,90	87,50	no
	2	52,42	258,24	171,21	96,19	97,50	yes
	3	15,23	259,17	133,39	139,17	100,00	yes

4	167,87	170,41	168,98	172,67	54,38	no
5	15,23	262,82	134,55	87,14	99,38	yes
9	47,62	255,85	134,61	98,70	96,88	yes
23	101,30	255,41	161,08	118,23	91,88	yes
24	47,76	280,27	139,44	93,26	96,88	yes
25	86,81	270,69	192,59	101,06	92,50	yes
26	120,97	267,43	176,05	130,84	88,75	no
27	151,78	253,21	154,40	157,07	80,63	no
28	77,48	279,11	174,92	97,19	95,00	yes
32	83,84	268,02	180,18	120,27	94,38	yes
35	136,31	277,26	203,32	128,85	86,88	no
36	73,27	259,55	170,94	103,65	95,00	yes
37	50,27	278,40	181,38	64,42	95,63	yes
38	158,81	250,97	205,01	156,82	82,50	no
39	36,39	253,73	135,00	69,26	97,50	yes
40	52,58	258,91	154,59	74,19	97,50	yes

Mean accuracy (%) and model-based analyses (BIC scores) of the last two blocks of day 2 (i.e., blocks 9 and 10) for every participant of Experiment 2. The best fitting model is underlined. Only participants who reached an accuracy of minimal 90% and had the best fitting decision bound model based on the GLC, continued on to day 3 of the experiment.
Note: DIM-O: unidimensional classifier based on the orientation; DIM-F: unidimensional classifier based on the frequency; GCC: General Conjunctive Classifier; GLC: General Linear Classifier

Only participants who favoured the general linear classifier model with a decision bound falling between the two categories were retained. In both conditions all remaining participants favoured a strategy based on the general linear classifier. As can be seen in Figures 3.7 and 3.8 in the supplementary materials, all optimal decision bounds fell between the two categories. As a result, the final sample used in the subsequent analyses consisted of 28 participants (19 women; average age of 20.9 years, *SD*=3.24 and range 18 to 30 years; N= 15 in the almost supervised condition and N=13 in the almost unsupervised condition).

As in Experiment 1, four types of analyses are reported: accuracy, model-based, response time analyses and SAT analyses. The response time analyses were used to compare the learning process of the almost supervised and almost unsupervised conditions of Experiment 2 to the semisupervised condition of Experiment 1.

Accuracy analysis. Figure 3.6 shows the average percentage of correct responses and the 95% confidence intervals on each block of trials received during the first four days for each of the conditions (almost supervised, almost unsupervised and the semisupervised condition of Experiment 1) separately. In all conditions the accuracy was based on all trials (feedback and no-feedback trials). During the first two days (block 1 to 8) the learning process was similar in the three conditions. The mean accuracy increased from an average of 72% (*SD*=9.93) in the first block for the almost supervised condition, an average of 75% (*SD*=11.76) for the almost unsupervised condition and an average of 73% (*SD*=11.06) for the semisupervised condition to almost perfect accuracy in the last block of day two (97%, *SD*=3.11, 96%, *SD*=2.35 and 97%, *SD*=2.76

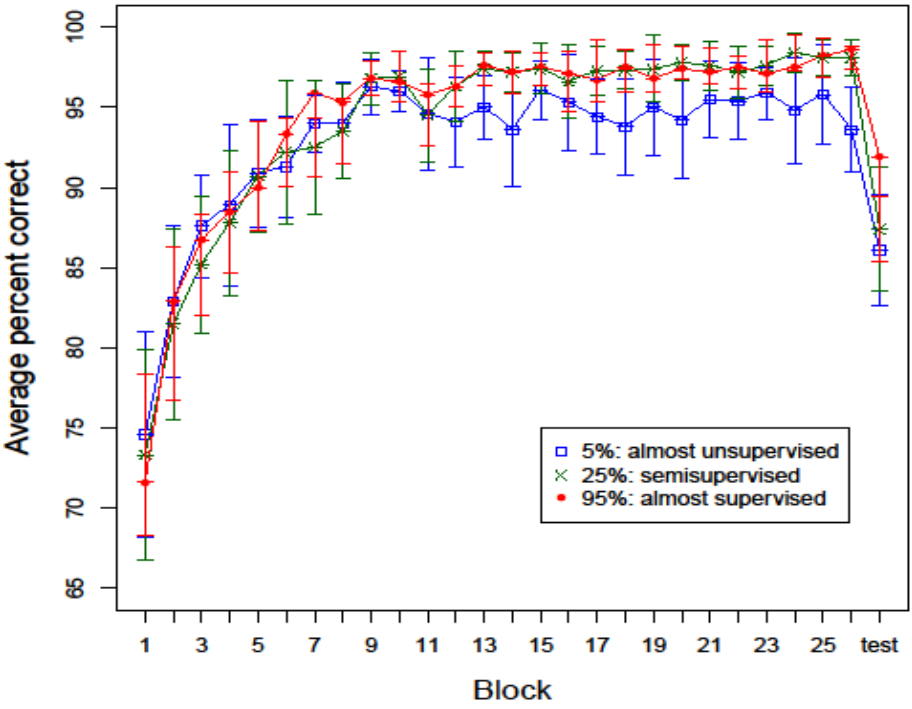
respectively). During the blocks on day 3 and 4, the mean accuracy was almost perfect in the almost supervised condition: the mean accuracy on the last block on day 3 was 98% (*SD*=2.59) and 99% (*SD*=1.48) on day 4. In the almost unsupervised condition the mean accuracy was also high: the mean accuracy on the last block on day 3 was 94% (*SD*=5.54) and 94% (*SD*=4.92) on day 4. For the semisupervised condition of Experiment 1, the mean accuracy on the last block of day 3 was 97% (*SD*=2.00) and 97% (*SD*=2.00) on day 4. A repeated measures ANOVA was conducted to determine whether the mean accuracy on the last two blocks differed depending on the day (4 levels: day 1, 2, 3 and 4) and condition (3 levels: almost supervised, almost unsupervised and semisupervised). Not surprisingly, there was a main effect of day, $F(3,34)=22.67$, $p<.001$, $\eta_p^2=.67$, indicating that the accuracy significantly increased during the succeeding days. Post-hoc paired-sample *t*-tests, adjusted by the Bonferroni correction for multiple comparisons, indicated that in comparison to day 1, mean accuracy was significantly higher on days 2, 3 and 4 (resp. $t(38)=7.79$, $p<.001$; $t(38)=5.60$, $p<.001$; $t(38)=7.18$, $p<.001$). There was no main effect of condition, $F(2,36)=1.60$, $p=.22$, $\eta_p^2=.08$ nor an interaction between day and condition ($F(6,68)=1.60$, $p=.16$, $\eta_p^2=.12$, suggesting that the increase in accuracy across days was similar in the three conditions.

The accuracy on the test (day 5), where the speed of responding was stressed, was lower in all conditions compared to the accuracy reached at the end of day 4, the average difference was -6.7% (*SD*=3.35) in the almost supervised condition, -7.5% (*SD*=7.47) in the almost unsupervised condition and -10.69% (*SD*=6.97) in the semisupervised condition. Paired sample *t*-tests showed that these differences were significant: accuracy significantly dropped between the last block of day

4 and day 5, $t(14)=7.69$, $p<.001$, $d=2.06$, for the almost supervised condition, $t(10)=5.08$, $p<.001$, $d=1.53$ for the semisupervised condition and $t(12)=3.62$, $p=.003$, $d=1.00$ for the almost unsupervised condition.

.

Figure 3.6



Mean accuracy (%) and 95% confidence intervals by block for all participants in the Almost supervised and Almost Unsupervised Condition of Experiment 2 and Semisupervised Condition of Experiment 1 from day 1 (blocks 1 to 5), day 2 (blocks 6 to 10), day 3 (blocks 11 to 18), day 4 (blocks 19 to 26) and day 5 ("t"). Only the participants who reached a performance of minimum 90% at the end of day 2 and revealed a decision bound based on the optimal decision bound were included in the analyses. All trials (feedback and no-feedback trials) were adopted in the analyses.

Finally and crucially, a one-way ANOVA was conducted to determine whether the mean accuracy on day 5 differed between the almost supervised, almost unsupervised and the semisupervised condition of Experiment 1. This was the case: $F(2,36)=4.18$, $p=.02$, $\eta_p^2=.19$. Independent sample *t*-tests corrected by the Bonferroni correction for multiple comparisons showed that this effect is due to the significant difference in accuracy between the almost supervised condition (92%, *SD*=4.08) compared to the semisupervised condition of Experiment 1 (87%, *SD*=6.53), $t(24)=2.19$, $p=.04$ and the almost unsupervised condition (86%, *SD*=6.40), $t(26)=2.94$, $p=.02$. The accuracy in the semisupervised condition (87%, *SD*=6.53) did not differ significantly from the almost unsupervised condition ($t(22)=.50$, $p=1$) and, most crucially, did not differ significantly from the almost supervised condition ($t(24)=2.19$, $p=.11$).

Model-based analysis. Table 3.8 shows the four model fits on day 4. In the almost supervised condition, all participants preferred a decision bound based on the general linear classifier, indicating successful learning. In the almost unsupervised condition, 11 participants of the 13 learned successfully: they revealed a decision bound based on the general linear classifier. Participant 25 and 37 preferred a decision bound based on the general conjunctive classifier, indicating that they switched to another strategy in comparison to day 2.

Figures 3.9 and 3.10 in the supplementary materials show the actual responses during the last two blocks of day 4 for each participant. Table 3.9 presents the model fits on the test day (i.e., day 5).

Table 3.8

Condition	Participant	GLC	DIM-O	DIM-F	GCC	Mean accuracy (%)
Almost Supervised	6	15,23	140,25	158,76	53,95	100,00
	8	53,02	162,44	145,39	88,09	97,50
	11	36,08	159,57	164,48	86,37	98,75
	12	24,06	140,65	160,16	66,73	99,38
	14	21,01	158,51	162,55	71,50	99,38
	15	82,40	198,44	132,80	107,23	93,13
	17	44,84	156,04	169,13	75,68	97,50
	18	23,55	133,00	141,50	57,27	99,38
	19	15,23	128,30	169,71	70,16	99,38
	20	15,23	148,01	134,59	63,61	100,00
	21	27,92	144,63	158,16	66,75	98,75
	22	32,74	122,05	177,36	60,28	98,75
	30	25,60	146,63	171,97	70,64	99,38
	31	43,20	175,01	124,65	71,36	97,50
	33	52,18	140,59	179,36	78,88	96,88
Almost unsupervised	2	52,19	176,38	153,37	91,42	97,50
	3	15,23	155,32	145,85	63,20	99,38
	5	34,97	126,98	149,77	64,91	98,75
	9	15,23	168,19	147,59	66,24	99,38
	23	97,97	168,38	176,41	125,60	92,50
	24	105,09	146,89	205,22	123,40	90,00
	25	150,37	169,19	216,97	150,16	82,50
	28	69,82	166,42	164,11	93,33	95,63
	32	53,37	142,10	185,44	99,94	96,88
	36	81,49	154,82	193,75	116,22	94,38
	37	104,82	160,20	200,55	92,31	90,00
	39	51,68	162,09	176,85	87,80	96,88
	40	59,34	166,72	174,26	102,01	96,88

Mean accuracy (%) and model-based analyses (BIC scores) on the last two blocks of day 4 for every participant of Experiment 2. The best fitting model is underlined.
Note: DIM-O: unidimensional classifier based on the orientation; DIM-F: unidimensional classifier based on the frequency; GCC: General Conjunctive Classifier; GLC: General Linear Classifier

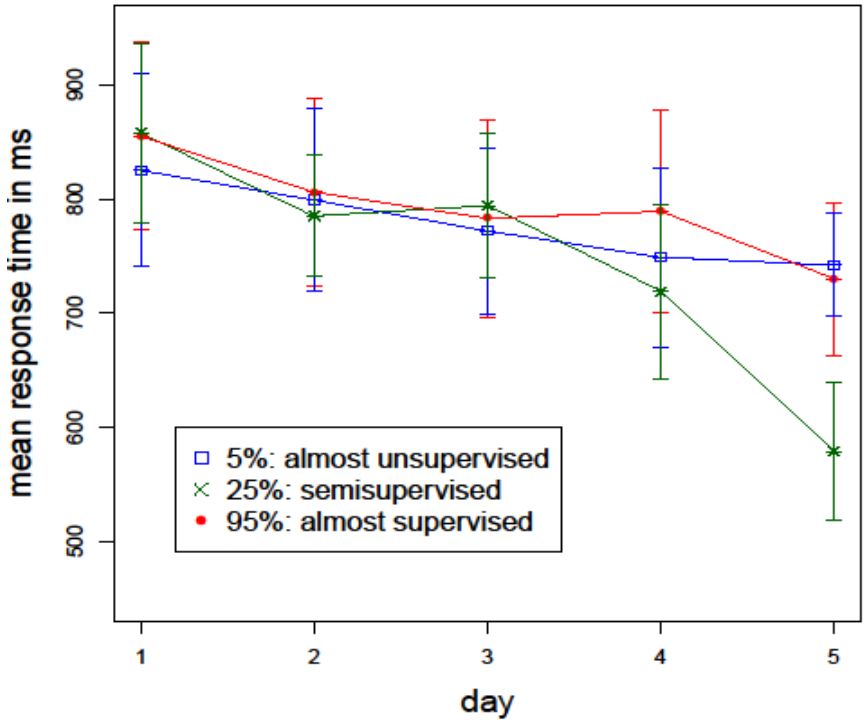
Again, for the almost supervised condition all participants used a decision bound based on the general linear classifier. In the almost unsupervised condition, 11 participants preferred a decision bound based on the general linear classifier whereas participant 23 and 37 adopted a strategy based on the general conjunctive classifier. Figures 3.11 and 3.12 in the supplementary materials present the responses and the best fitting decision bounds for every participant during the test day.

Analysis of the response times. The mean response times (RTs) along with the 95% confidence intervals from day1 to 4 for the almost supervised, almost unsupervised and the semisupervised condition of Experiment 1 from day1 to 4 are presented in Figure 3.7. These RTs were calculated on the last two blocks of each day. For day 1, the mean RTs in the almost supervised condition was 855 ms (*SD* =162.41), the mean RTs in the almost unsupervised condition was 825 ms (*SD*=156.31) and the mean RTs in the semisupervised condition of Experiment was 858 ms (*SD*=131.22). Importantly for this investigation is that participants were equally fast in the last two blocks of day 2 (almost supervised mean RTs=806 ms, *SD*=161.29; almost unsupervised mean RTs=800 ms, *SD*=148.49 and semisupervised mean RT=785 ms, *SD*=90.10). This was confirmed by a one-way ANOVA comparing the RTs of the last two blocks of day two to the three conditions, almost supervised condition, the almost unsupervised condition and the semisupervised condition of Experiment 1, F <1, p=.93, η_p^2=.004. On days 3 and 4, the mean RTs dropped in the almost supervised condition to, respectively, 783 ms (*SD*=170.67) and 772 ms (*SD*=134.68). This decrease in mean RTs was also observed in the almost unsupervised condition: 772 ms (*SD*=134.68)

on day 3 and 749 ms (*SD*=144.54) on day 4. A repeated measures ANOVA was conducted to determine whether the mean RTs on the last two blocks differed depending on the day (4 levels: day 1, 2, 3 and 4) and condition (3 levels: almost supervised, almost unsupervised and semisupervised of Experiment 1). Not surprisingly, there was main effect of day, $F(3,34)=6.81$, $p=.001$, $\eta_p^2=.38$, indicating that the RTs decreased across the succeeding days. Post-hoc paired-sample *t*-tests, adjusted by the Bonferroni correction for multiple comparisons, indicated that participants were significantly faster on day 2, 3 and 4 compared to day 1, $t(38)=2.92$, $p=.03$; $t(38)=2.95$, $p=.05$ and $t(38)=4.03$, $p=.001$ respectively. None of the other comparisons were significant (all p>.42). There was no main effect of condition, $F<1$, $p=.89$, $\eta_p^2=.01$, neither was there a significant interaction between day and condition, $F(6,68)=1.85$, $p=.10$, $\eta_p^2=.14$.

Decisive to test our hypotheses was the difference in RTs on day 5. In the almost supervised condition the mean RTs was 730 ms (*SD=132.08*), the mean RTs in the almost unsupervised condition was 742 ms (*SD* =84.59) and the mean RTs in the semisupervised condition of Experiment 1 was 579 ms (*SD*=104.08). A one-way ANOVA indicated significant differences between the RTs of the three groups on day 5, $F(2,36)=8.07$, $p=.001$, $\eta_p^2=.31$. Post-hoc paired-sample *t*-tests, adjusted by the Bonferroni correction for multiple comparisons, showed that RTs were faster in the semisupervised condition of Experiment 1, compared to the almost supervised condition, $t(24)=3.15$, $p=.004$, and to the almost unsupervised condition, $t(22)=4.26$, $p=.003$. The difference between the almost supervised condition and the almost unsupervised condition was not significant, $t(26)=0.28$, $p>.99$.

Figure 3.7



Mean response times (in ms) along with the 95% confidence intervals for the almost supervised and the almost unsupervised condition of Experiment 2 along with the semisupervised condition of Experiment 1. The mean response times are calculated on the last two blocks of each day.

Speed-accuracy trade-off. On day 5 participants were deliberately asked to respond as fast as possible. The speed-accuracy trade-off was again calculated for each condition. In the almost supervised condition, there was no SAT-effect, r=-.08, p=.79. In contrast, there was a SAT-effect in the almost unsupervised condition, r=.85, p<.001: faster responses were correlated with more errors. Recall that in the semisupervised condition of Experiment 1 we also observed a SAT-effect.

Table 3.9

Condition	Participant	GLC	DIM-O	DIM-F	GCC	Mean accuracy (%)
Almost Supervised	6	63,81	181,73	155,45	104,78	92,53
	8	57,49	171,93	165,23	109,59	95,53
	11	98,93	170,44	178,41	121,78	89,56
	12	93,64	176,87	171,30	115,55	90,29
	14	59,40	178,45	159,58	95,90	94,77
	15	137,33	191,39	174,85	145,64	81,35
	17	85,57	166,05	179,45	88,06	90,29
	18	65,02	154,51	184,79	111,94	93,29
	19	58,48	171,18	165,67	109,62	93,29
	20	25,86	165,86	165,83	101,39	98,50
	21	71,43	162,14	179,59	122,45	94,02
	22	58,08	173,09	164,69	109,56	94,77
	30	69,64	164,96	175,69	115,15	94,02
	31	78,10	136,74	193,03	99,51	88,80
	33	88,22	151,80	193,36	125,25	88,05
Almost Unsupervised	2	74,23	222,84	170,29	120,24	94,77
	3	87,93	233,26	184,67	126,47	89,55
	5	102,91	226,44	196,58	127,34	82,83
	9	112,09	250,90	198,29	119,71	76,11
	23	146,69	216,83	180,00	130,53	79,86
	24	66,29	225,13	191,32	123,43	91,80
	25	102,65	229,95	211,97	119,91	81,35
	28	81,91	226,28	201,71	102,30	85,83
	32	103,47	222,37	193,09	123,79	87,32
	36	116,88	222,70	187,54	125,02	85,83
	37	112,96	234,52	216,60	79,68	76,86
	39	71,56	219,76	151,07	97,00	91,05
	40	49,29	223,11	166,02	106,68	95,53

Mean accuracy (%) and model-based analyses (BIC scores) on all trials of day 5 for every participant of Experiment 2. The best fitting model is underlined.
Note: DIM-O: unidimensional classifier based on the orientation; DIM-F: unidimensional classifier based on the frequency; GCC: General Conjunctive Classifier; GLC: General Linear Classifier

Discussion

In Experiment 2 the total amount of trials was equal in each condition to investigate whether the nature of feedback (almost supervised, semisupervised or almost unsupervised) had an impact on the development of automaticity. Two control conditions were run to compare to the semisupervised condition of Experiment 1: the almost supervised condition (in which mainly feedback trials were given) and the almost unsupervised condition (in which mainly no-feedback trials were given). The results indicated that the mean RTs on day 5 in the semisupervised condition of Experiment 1 were significantly faster than in the two control conditions of Experiment 2, indicating that a combination of feedback and no-feedback trials boosted the automaticity process. This result cannot be due to general faster RTs in the semisupervised condition, as RTs were equal for all conditions on day 2. This result can also not be affected by the novelty of the no-feedback trials on the test day, since all participants already experienced feedback and no-feedback trials on days 3 and 4. Remarkably, in the almost unsupervised condition, two participants did not reveal a strategy based upon the optimal decision bound anymore on day 4. Apparently, they changed their categorization strategy during day 3 and day 4. This was not the case in the almost supervised condition nor in the semisupervised condition. This might imply that for some participants a minimal percentage of feedback is still needed late in learning – even though almost perfect accuracy was obtained long before. On the test day, the mean accuracy was higher in the almost supervised condition compared to the almost unsupervised condition. There was no significant difference

in mean accuracy in the semisupervised condition compared to the almost supervised condition and the almost unsupervised condition. In the semisupervised and the almost unsupervised condition, a few participants showed a drop in accuracy on day 5. In both latter conditions, this drop in accuracy was due to a few participants. This drop in performance was also reflected in the model-based analyses: in the almost supervised condition all participants adhered to a strategy based upon the optimal decision bound whereas in the semisupervised and the almost unsupervised conditions three and two participants respectively switched. When these participants who switched strategy were omitted from the analyses, a one-way ANOVA revealed that the difference in accuracy disappeared, $F(2,31)=2.67$, $p=.12$ $\eta_p^2=.13$. The mean accuracy was 87% *(SD=5.90)* for the almost unsupervised, 89% *(SD*=7.04*)* for the semisupervised and 92% *(SD=4.08)* for the almost supervised condition. Still, the effect of faster RTs in the semisupervised condition remained, one-way ANOVA $F(2,31)=6.54$, $p=.004$, $\eta_p^2=.30$. Furthermore, this effect was due to the significant difference between the semisupervised (577ms, *SD*=123.14) and the almost supervised (730 ms, *SD*=132.08), $t(21)=2.71$, $p=.015$ and the difference between the semisupervised and the almost unsupervised (760 ms, *SD*=79.20), $t(17)=3.96$, $p=.005$ while the difference between the almost supervised and the almost unsupervised condition was not significant, $t(24)=0.67$, $p>.99$. These post-hoc *t*-tests were corrected for multiple comparisons by Bonferroni. These results show that the faster response times in the semisupervised condition were not due to quick random guessing by some participants, resulting in fast RTs and low accuracy, since the effect remains when the switch participants were omitted.

On day 5 a SAT-effect occurred in the almost unsupervised condition of Experiment 2: participants who tended to respond fast, also made more errors. This was not the case in the almost supervised condition. This could explain the significantly lower accuracy in the almost unsupervised condition on day 5. Note that we also observed a SAT-effect on day 5 for the semisupervised condition. Even though this condition showed a SAT, they show a comparable accuracy as the almost supervised condition.

As a conclusion, Experiment 2 shows that, even when the total amount of trials is the same, the development of automaticity is enhanced by 25% semisupervised learning. Only in this feedback scheme the accuracy remained high and response times were strikingly faster. When feedback was almost always provided, participants maintained a high accuracy but were not able to accelerate their responses. In the almost unsupervised condition, there is a drop in accuracy – some participants even unlearned the category structure – and response times remained high. These results are in contrast to the SPEED-model which predicts a similar development of automaticity between the three conditions as the number of trials, regardless of feedback, is identical in all conditions.

General Discussion

This study investigated the impact of semisupervised category learning late in the learning process when automaticity develops. Participants were first trained in a supervised way during two days on the information-integration category structure and only participants who performed at least 90% accurately and used a decision bound similar to

the optimal decision bound were included in the actual experiments. In Experiment 1, half of the participants were trained in a 25% semisupervised way on days 3 and 4: only a quarter of the trials were followed by feedback. The other half were trained in a supervised way, implying that feedback was given after every categorization response. Both conditions received an equal amount of feedback trials. On the fifth day, differences in performance between the semisupervised and supervised learners were studied. Participants were urged to respond as fast as possible on this test day. Accuracy was similar in both groups on day 5, which is to be expected, as accuracy was already above 90% at the end of day 2. However, the results clearly showed that participants in the semisupervised condition responded significantly faster than the participants in the supervised condition on day 5. This effect cannot be due to general faster RTs and/or higher accuracy levels in the semisupervised condition, as evidenced by similar RTs and accuracies for both conditions at the end of day 2. Thus, the findings of Experiment 1 imply that late in learning the no-feedback trials in the semisupervised condition aided the development of automaticity.

However, two confounds hampered a clear conclusion that late in learning semisupervised learning is superior. First, even though the amount of feedback trials was equal in the semisupervised and the supervised conditions, the total number of trials differed. Participants in the semisupervised condition received four times as many trials on day 3 and day 4 as participants in the supervised condition and hence had more practice. It is therefore possible that the larger total number of trials caused the faster RTs in the semisupervised condition. Second, participants in the supervised condition never received no-feedback trials on days 3 and 4. Perhaps the sudden encounter of no-feedback trials on

day 5 might have slowed them down on this test day. To exclude these alternative explanations, two control conditions were administered in Experiment 2 which were compared to the semisupervised condition of Experiment 1. The percentage of feedback trials on day 3 and day 4 was manipulated. In the almost supervised condition 95% of the trials were randomly followed by feedback. In the almost unsupervised condition 5% of the trials were randomly followed by feedback. This implies that in all conditions (almost supervised, semisupervised and almost unsupervised), participants encountered both feedback and no-feedback trials on days 3 and 4. Crucially, the total amount of trials was identical in all three conditions. Despite these alterations, the results of Experiment 2 again showed that the RTs on the fifth day were significantly slower for the almost supervised and the almost unsupervised conditions compared to the semisupervised condition, whereas RTs did not differ at the end of day 2. Since the total number of trials was now identical in all three conditions, we can conclude that the semisupervised learners achieved automaticity faster. Hence, late in learning semisupervised learning should be preferred. These results are not in line with the predictions of the SPEED-model since it stipulates that the type of trial (feedback or no feedback) should not have an impact on the development of automaticity late in learning, and would therefore expect similar RTs in all the conditions as they all contain the same amount of trials. Why semisupervised category learning is superior late in learning requires further investigation. Perhaps semisupervised learning increased participants' motivation and attention compared to a condition where feedback is almost always offered. Contrarily, participants in the almost unsupervised learning condition often reported

frustration and they perhaps gave up to perform to their maximum ability.

In the almost unsupervised condition there was a Speed accuracy trade-off (SAT-)effect: participants who tended to respond fast also made more errors. This can explain why the mean accuracy in the almost supervised condition was significantly higher than in the almost unsupervised condition. Although we also observed a SAT-effect in the semisupervised condition, there was no difference in accuracy between the almost supervised and the semisupervised conditions, suggesting that, even though participants in the semisupervised condition sacrificed accuracy for speed, they still remained at a similar accuracy level as participants in the almost supervised condition. The model-based analyses on day 5 showed that a few participants in the semisupervised and the almost unsupervised condition switched in strategy. This was not the case in the almost supervised condition where all participants adhered to the optimal decision bound. Apparently, some participants in the almost unsupervised and in the semisupervised condition unlearned the category structure. Note that, when these participants were omitted from analyses, the faster RTs in the semisupervised condition remained, whereas the difference in accuracy between almost supervised and almost unsupervised conditions disappeared. These findings suggest that the faster RTs in the semisupervised condition were not due to quick random guessing. It rather seems that individual differences are decisive. In this study, the learning scheme in a condition was fixed and individual differences in learning were not taken into account. It is possible that some participants need a higher percentage of feedback-trials on day 3 and 4 even though they successfully learned the structure at the end of day 2. Another explanation could be that the switch point, that is, the

point in the learning process on which supervised learning becomes less effective in favour of semisupervised learning, is later for some participants than the fixed point of 800 trials (end of day 2) used in this study. Thus, even though semisupervised learning appears to be the best learning mode late in learning, there may be exceptions for some participants. These individual differences in learning are interesting directions for further research.

The current study supports the idea that no-feedback trials aid the learning process, as shown in machine learning (Chapelle et al., 2006; Zhu et al, 2009) and in a few human semisupervised studies (Kalish et al., 2011; Lake & McClelland, 2011; Zhu et al., 2010). Contrary to the studies of Kalish et al. (2011), Lake and McClelland (2011) and Zhu et al. (2010), in our study all the stimuli (as compared to a fixed subset) could be followed by feedback, mimicking real-life category learning. As in the study of Rogers et al. (2010), semisupervised learning was found when the participants were urged to respond as fast as possible. It is possible that, in order to observe semisupervised learning, speed of responding is essential. Again, this can be an interesting direction for future research. Our study also proves that not only young children (Kalish et al., 2014), but young adults too are able to learn in a semisupervised way.

The results of our study may seem to be in contrast to the study of Vandist et al. (2009) where no effect of the no-feedback trials was found in learning the information-integration category structure. However, there are important differences between both studies. First and most importantly, Vandist et al. (2009) focused on the effects early in the learning process whereas the current study dealt with the effects late in

learning, when automaticity develops. Second, the impact of the no-feedback trials was studied on the accuracy levels in the study of Vandist et al. (2009). In this study response time was the dependent variable. Third, in the Vandist et al. (2009) study, participants in the semisupervised condition learned from the start in a semisupervised way whereas in the present study semisupervised learning was only introduced after expert performance was obtained.

Combining the results of both studies suggests that early in learning the information-integration structure, the no-feedback trials do not have an impact, but that late in learning the no-feedback trials facilitate automaticity. Indeed, the effects of semisupervised learning, at least in the information-integration structure, might be especially apparent late in the learning process. These results can also explain why former semisupervised category studies failed to find convincing effects, as they all focused on initial learning processes. This also makes sense if we relate this to category learning in children. When a child is first confronted with items of an unknown category, parents label most of the presented items. As the child gets more and more familiar with the category, the parent still labels information but less often. When the parent has the idea that the child has acquired the category, label information diminishes but still takes place from time to time. In fact, it is ecologically plausible that semisupervised learning takes place when a solid basis of category expertise has first been acquired and from that point on, it aids learning. Experiment 2 even suggests that from a certain expert level on, semisupervised category learning might be essential to the development of automaticity, since semisupervised category learning fastens the development of automaticity whereas (almost) supervised learning seems to slow it down. Although speculative, it could for

example be that the continuous feedback in the almost supervised condition makes the expert learner less attentive or less motivated, leading to decreased performance. Nevertheless, our results also indicated that a certain percentage of feedback trials is still needed to develop automaticity successfully: when feedback was rare in the almost unsupervised condition, the mean RTs remained at the same level and at the end of day 4 a few participants even unlearned the category structure.

In conclusion, this is the first study that examines the effect of 25% semisupervised learning late in learning. In Experiment 1, faster RTs were observed in a 25% semisupervised condition in comparison to a supervised condition when the total amount of feedback trials in both conditions was the same. In Experiment 2 the total amount of trials was kept identical in all conditions, but still the 25% semisupervised condition of Experiment 1 showed faster response times in comparison to the almost supervised and almost unsupervised conditions of Experiment 2. Hence, late in learning, (25%) semisupervised learning seemed to have a beneficial effect as the no-feedback trials facilitated automaticity. A learning condition containing a certain amount of no-feedback trials even seems to outperform a condition where (almost) all trials are followed by feedback, as long as a minimal percentage of the trials is followed by feedback.

Appendix

Four models were fit to each participant's data: two rule-based decision bound models, i.e. the vertical and the horizontal unidimensional models, and one information-integration model, i.e. the general linear classifier. More details of the models can be found in (Ashby & Gott, 1988; Ashby, 1992; Maddox & Ashby, 1993).

Rule-based Models

The unidimensional classifier (DIM). The DIM assumes that participants set a decision criterion on one of the stimulus dimensions. Because our stimuli have two stimulus dimensions, orientation and frequency, two unidimensional classifiers were fit: one for the frequency dimension and one for the orientation dimension. An example of an unidimensional rule used for categorization is: "Respond A if the orientation is steep, otherwise respond B". In this case the frequency is irrelevant. These models have two parameters: a decision criterion along the chosen classification dimension and a perceptual noise variance parameter.

The general conjunctive classifier (GCC). assumes that the rule used by participants is a conjunction, for example: "Respond A if the orientation is steep and the frequency is small, otherwise respond B". The GCC has three parameters: one for each for the single-decision criterion placed along each dimension (i.e., orientation and frequency) and finally one for the perceptual noise variance.

Information-Integration Model

The general linear classifier (GLC). The GLC assumes that a linear decision bound divided the stimulus space into response regions. Confronted with a stimulus, the perceived place in the stimulus space is determined and the contributed categorization response elicited. These decision bounds require linear integration of both stimulus dimensions, resulting in an information-integration decision strategy. The GLC has three parameters: the slope and the intercept of the linear decision bound and finally the perceptual noise variance.

References

Ashby, F.G. (1992). Multidimensional models of categorization. In F.G. Ashby (Ed), *Multidimensional models of perception and cognition* (pp. 449-483). Hillsdale, NJ: Erlbaum.

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*, 442–481.

Ashby, F. G., & Ell, S. W. (2001). The neurobiology of human category learning. *Trends in Cognitive Sciences,* 5,204-210.

Ashby, F. G., Ell, S. W., & Waldron, E. M. (2003). Procedural learning in perceptual categorization. *Memory & Cognition*, *31*(7), 1114–1125.

Ashby, F. G., Ennis, J. M., & Spiering, B. J. (2007). A neurobiological theory of automaticity in perceptual categorization. *Psychological Review*, *114*, 632–656.

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

Ashby, F. G., & Maddox, W. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, *37* (3), 372-400.

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology, 56*, 149–178.

Ashby, F. G., & Maddox, W. T. (2010). Human category learning 2.0. *Annals of the New York Academy of Sciences*, 1224, 147-161.

Ashby, F. G., Maddox, W. T., & Bohil, C. J. (2002). Observational versus feedback training in rule-based and information-integration category learning. *Memory & Cognition, 30*, 666–677.

Ashby, F. G., & O'Brien, J. B. (2007). The effects of positive versus negative feedback on information-integration category learning. *Perception & Psychophysics, 69*, 865–878.

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

Ashby, F. G., Turner, B. O., & Horvitz, J. C. (2010). Cortical and basal ganglia contributions to habit learning and automaticity. *Trends in Cognitive Sciences, 14*(5), 191-232.

Ashby, F. G., & Crossley, M. (2012). Automaticity and multiple memory systems. Wiley Interdisciplinary Reviews: Cognitive Science, 3, 353–376.

Censor, N., Karni, A., & Sagi, D. (2006). A link between perceptual learning, adaptation and sleep. *Vision Research, 46*, 4071–4074.

Chapelle, O., Schölkopf, B., Zien, A (2006). Semi-Supervised Learning. MIT Press, Cambridge, MA, USA.

Clapper, J. P., & Bower, G. H. (1994). Category invention in unsupervised learning. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 20*, 443–460.

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

French, R. M., Mareschal, D., Mermillod, M., & Quinn, P. C. (2004). The role of bottom-up processing in perceptual categorization by 3- to 4-month-old infants: simulations and data. *Journal of experimental psychology: General, 133* (3), 382-397.

Gibson, B. R., Rogers, T. T., & Zhu, X. (2013). Human Semi-Supervised Learning. *Topics in Cognitive Science, 5*, 132–172.

Gibson, B.R., Rogers, T.T., Kalish, C.W., & Zhu, X (2015). What causes categoryshifting in human semi-supervised learning? In Proceedings of the 37th Annual Conference of the Cognitive Science Society (CogSci).

Heitz, R. P. (2014). The speed-accuracy trade-off: history, physiology, methodology and behavior. *Frontiers in Neuroscience, 8* (150), 1-15.

Hélie, S., Waldschmidt, J. G., & Ashby, F. G. (2010). Automaticity in rule-based and information-integration categorization. *Attention, Perception & Psychophysics, 72*, 1013–1031.

Kalish, C. W., Rogers, T. T., Lang, J., & Zhu, X. (2011). Can semi-supervised learning explain incorrect beliefs about categories? *Cognition, 120*, 106–118.

Kalish, C.W., Zhu, X., & Rogers, T.T. (2015). Drift in children's categories: when experienced distributions conflict with prior learning. *Developmental Science. 18*(6), 940-956.

Lake, B. M., & McClelland, J. L. (2011). Estimating the strength of unlabeled information during semi-supervised learning. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, 1400–1405.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review, 95*, 492-527.

Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin & Review, 9*, 829–835.

Maddox, W. T., & Ashby, F. G. (1993). Comparing decision bound and exemplar models of categorization. *Perception & Psychophysics, 53*, 49–70.

Maddox, W. T., Ashby, F. G., & Gottlob, L. R. (1998). Response time distributions in multidimensional perceptual categorization. *Perception & Psychophysics, 60*(4), 620–637.

Maddox, W. T., Ashby, F. G., Ing, A. D., & Pickering, A. D. (2004). Disrupting feedback processing interferes with rule-based but not information-integration category learning. *Memory & cognition, 32* (4), 582-591.

Maddox, W. T., Bohil, C. J., & Ing, A. D. (2004). Evidence for a procedural-learning-based system in perceptual category learning. P*sychonomic Bulletin & Review, 11*, 945–952.

Maddox, W. T., & Filoteo, J. V. (2011). Stimulus range and discontinuity effects on information-integration category learning and generalization. *Attention, Perception & Psychophysics, 73*, 1279–1295.

Maddox, W. T., Filoteo, J. V., Hejl, K. D., & Ing, A. D. (2004). Category number impacts rule-based but not information-integration category learning: further evidence for dissociable category-learning systems. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 30*, 227–245.

Maddox, W. T., Glass, B. D., O'Brien, J. B., Filoteo, J. V., & Ashby, F. G. (2010). Category label and response location shifts in category learning. *Psychological Research, 74*, 219–236.

Maddox, W. T., & Ing, A. D. (2005). Delayed feedback disrupts the procedural-learning system but not the hypothesis-testing system in perceptual category learning. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 31*, 100–107.

Maddox, W. T., Pacheco, J., Reeves, M., Zhu, B., & Schnyer, D. M. (2010). Rule-based and information-integration category learning in normal aging. *Neuropsychologia, 48*, 2998–3008.

McDonnell, J.V.**,** Jew, C.J., and Gureckis, T.M. (2012). "Sparse category labels obstruct generalization of category membership." In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), Proceedings of the 34th Annual Conference of the Cognitive Science Society**.**

McKinley, S. C., & Nosofsky, R. M. (1995). Investigations of exemplar and decision bound models in large, ill-defined category structures. *Journal of Experimental Psychology. Human Perception and Performance*, *21*, 128–148.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85* (3), 207-238.

Medin, D. L., & Schwanenflugel, P. J. (1981). Linear separability in classification learning. *Journal of Experimental Psychology: Human Learning & Memory*, *7*, 355-368.

Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive Psychology*, *19*, 242–279.

Milton, F., Longmore, C. A., & Wills, A. J. (2008). Processes of overall similarity sorting in free classification. *Journal of Experimental Psychology. Human Perception and Performance*, *34*, 676–692.

Moors, A., & De Houwer, J. (2006). Automaticity: a theoretical and conceptual analysis. *Psychological Bulletin*, *132*, 297–326.

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *13*, 87–108.

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review, 104,* 266-300.

Paul, E. J., Boomer, J., Smith, J. D., & Ashby, F. G. (2011). Information-integration category learning and the human uncertainty response. *Memory & Cognition*, *39*, 536–554.

Pothos, E. M., & Chater, N. (2002). A simplicity principle in unsupervised human categorization. *Cognitive Science, 26 (3),* 303-343.

Pothos, E. M., & Chater, N. (2005). Unsupervised categorization and category learning. *The Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology*, *58*, 733–752.

Pothos, E. M., Perlman, A., Bailey, T. M., Kurtz, K., Edwards, D. J., Hines, P., & McDonnell, J. V. (2011). Measuring category intuitiveness in unconstrained categorization tasks. *Cognition*, *121*, 83–100.

Rickard, T. C. (1997). Bending the power law: A CMPL theory of strategy shifts and the automatization of cognitive skilss. *Journal of Experimental Psychology: General, 126,* 288-311.

Rogers, T. T., Kalish, C., Gibson, B. R., Harrison, J., & Zhu, X. (2010). Semi-supervised learning is observed in a speeded but not an unspeeded 2D categorization task. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 2320–2325).

Schneider, W., & Chein, J. M. (2003). Controlled & automatic processing: Behavior, theory, and biological mechanisms. *Cognitive Science, 27* (3), 525-559.

Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics, 6,* 461-464.

Shepard, R., Hovland, C., & Jenkins, H. (1961). Learning and memorization of classifications. *Psychological Monographs*, *75*, 1–42.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review, 84* (2), 127-190.

Spiering, B. J., & Ashby, F. G. (2008). Response processes in information-integration category learning. *Neurobiology of Learning and Memory*, *90*, 330–338.

Spiering, B. J., & Ashby, F. G. (2008). Initial training with difficult items facilitates information-integration, but not rule-based category learning: Research article. *Psychological Science*, *19*, 1169–1177.

Stevens, M., Lammertyn, J., Verbruggen, F., & Vandierendonck, A. (2006). Tscope: A C library for programming cognitive experiments on the MS windows platform. *Behavior Research Methods*, *38*, 280–286.

Stickgold, R., & Walker, M. P. (2005). Memory consolidation and reconsolidation: What is the role of sleep? *Trends in Neuroscience, 28*, 408–415

Stickgold, R., James, L., & Hobson, J. A. (2000). Visual discrimination learning requires sleep after training. *Nature Neuroscience*, *3*, 1237–1238.

Stickgold, R., Whidbee, D., Schirmer, B., Patel, V., & Hobson, J. A. (2000). Visual discrimination task improvement: A multi-step process occurring during sleep. *Journal of Cognitive Neuroscience*, *12*, 246–254.

Vandist, K., De Schryver, M., & Rosseel, Y. (2009). Semisupervised category learning: the impact of feedback in learning the information-integration task. *Attention, Perception & Psychophysics*, *71*(2), 328–341.

Vermaercke, B., Cop, E., Willems, S., D'Hooge, R., & Op de Beeck, H. P. (2014). More complex brains are not always better: rats outperform humans in implicit category-based generalization by implementing a similarity-based strategy. *Psychonomic Bulletin & Review*, *21*, 1080–6.

Vong, W. K., Perfors, A., & Navarro, D. J. (2014). The Relevance of Labels in Semi-Supervised Learning Depends on Category Structure. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society,* 1718–1723.

Waldron, E. M., & Ashby, F. G. (2001). The effects of concurrent task interference on category learning: evidence for multiple category learning systems. *Psychonomic Bulletin & Review*, *8*, 168–176.

Zeithamova, D., & Maddox, W. T. (2006). Dual-task interference in perceptual category learning. *Memory & Cognition*, *34*, 387–398.

Zeithamova, D., & Maddox, W. T. (2007). The role of visuospatial and verbal working memory in perceptual category learning. *Memory & Cognition*, *35*, 1380–1398.

Zhu, X., Gibson, B.R., Jun, K.-S., Rogers, T.T., Harrison, J. & Kalish, C. (2010). Cognitive models of test-item effects in human category learning. In *Proceedings of the 27th International Conference on Machine Learning,* 1247–1254.

Zhu, X., & Goldberg, A. B. (2009). Introduction to Semi-Supervised Learning. *Morgan and Claypool Publishers.*

Chapter 4

Semisupervised category learning: the impact of feedback investigated in the prototype distortion category structure

Vandist, K., Storms, G. & Van den Bussche, E.

Abstract

The scarce human semisupervised category learning studies report mixed results with regards to the role of no-feedback trials in the category learning process. It remains unclear whether instances where a categorization response is not followed by feedback aids the categorization learning, hampers it or has no impact at all. A possible explanation of these inconsistent findings is that the effect of no-feedback trials depends on the underlying category structure being investigated. Here, we focused on the prototype distortion category structure to examine the impact of no-feedback trials in semisupervised category learning. In the "A, not A" task participants indicate whether a stimulus belongs to the category or not. Experiment 1 demonstrated that feedback is not needed after each trial, since learning was also successful in semisupervised conditions with only 50% and 25% feedback. However, when no feedback at all was given, no successful learning occurred. Experiment 2 investigated the contribution of no-feedback trials in semisupervised learning by replacing the no-feedback trials by irrelevant fillers. The results showed that no-feedback trials had no impact on learning. Experiment 3 addressed the hypothesis that positive effects of no-feedback trials, resulting in higher accuracy rates, are expected when feedback is provided only sporadically. However, half of the participants failed to learn in a 10% semisupervised condition. Experiment 4 confirmed that also in this 10% condition no-feedback trials had no impact on learning. Together, these results show that even though the prototype distortion category structure can be learned in a semisupervised way, no-feedback trials do not aid the learning process.

Introduction

Categorization is a skill that we perform countless times a day. When a child reaches adulthood, thousands of categories are learned (Ashby & O'Brien, 2005). These categories vary from simple (is an animal big or small?) to very complex (is a bat a bird or a mammal?). When needed, adults are still capable to learn new categories (Ashby & Maddox, 2005, 2011). In the field of human category learning participants have to classify unfamiliar stimuli into categories in order to understand this learning process. The behavior of the participants is observed during the period where their ability to assign stimuli to these categories increases from chance level to a certain stable above-chance level (Ashby & Maddox, 2005, 2011; Ashby & O'Brien, 2005).

Until recently, human category learning has been studied using supervised or unsupervised paradigms. In *supervised learning paradigms*, the participant receives a stimulus which has to be classified in two or more contrasting categories. Immediately after this response, feedback is provided about the correct category label (see Shepard, Hovland, & Jenkins, 1961 for a description of a basic experiment). Numerous studies have demonstrated that, based on this paradigm, participants can learn very complex categories if a sufficient amount of trials is provided (e.g., Ashby, Maddox, & Bohil, 2002; Ashby, Queller, & Berretty, 1999; Maddox, Ashby, & Gottlob, 1998; Maddox, Filoteo, Hejl, & Ing, 2004; McKinley & Nosofsky, 1995; Medin & Schwanenflugel, 1981). In *unsupervised learning paradigms*, the participant never receives feedback or information about the category label. The goal is to identify an intuitive or natural classification for a set of objects (Clapper & Bower, 1994; Love, 2002;

Medin & Wattenmaker, 1987; Milton, Longmore, & Wills, 2008; Pothos & Chater, 2002, 2005; Pothos et al., 2011). Findings based on these unsupervised paradigms reveal that participants especially use unidimensional rules, regardless of the amount of trials or the underlying category structure (Ashby et al., 1999). For example, they classify all the small animals in one category and all the large animals in another category. In conclusion, in supervised learning participants are able to learn very complex categorization structures, whereas in unsupervised learning people have the tendency to use very simple categorization rules.

The study of Vandist, De Schryver and Rosseel (2009) argued that both supervised and unsupervised learning occur very rarely. Translated to daily life, supervised learning entails that for every object that we observe, we immediately receive correct information about its category label, which renders it low ecological validity. Contrarily, unsupervised learning implies that during our entire lives, nobody ever tells us which objects belong together or informs us about the category label of an object. Both types of category learning therefore do not represent daily reality. Vandist et al. (2009) argued that people instead learn in a *semisupervised way*: when confronted with (new) objects (e.g., a dog), sometimes category information will be provided ("look, a dog") and sometimes not. This realistic scenario was incorporated in the *semisupervised category learning paradigm*. In a block of trials a predetermined percentage of category responses is followed by feedback (i.e., feedback or labeled trials). The remaining trials do not receive feedback (i.e., no-feedback or unlabeled trials). Vandist et al. (2009) tested the semisupervised category learning using the information-integration category structure. This is a complex category

structure which can be learned successfully based on supervised, but not unsupervised learning (Ashby et al., 1999). Their results showed that semisupervised learning can be as successful as supervised learning whereas unsupervised learning fails (Vandist et al., 2009). Thus, feedback is not needed after every single trial to master a complex category structure.

An important question in semisupervised learning is the effect of the *no-feedback trials* on the learning process. Do these no-feedback trials speed up the learning process (which would, for example, be indexed by faster response times, higher accuracy rates or a more adequate decision bound), slow it down (which could be indexed by slower response times, lower accuracy rates or an inadequate decision bound) or have no impact at all (which could be indexed by similar response times, similar accuracy rates or a decision bound solely based on the stimuli of the feedback trials)? The human semisupervised studies show mixed results. On the one hand, several studies (Kalish, Zhu & Rogers, 2015; Lake & McClelland, 2011; Zhu, Gibson, Jun, Rogers, Harrison, & Kalish, 2010; Zhu, Rogers, Qian, & Kalish, 2007) reported an effect of the no-feedback trials. In these studies the effect of the no-feedback trials was investigated on the decision bound that the participants used to classify an item into one of two categories. The stimuli presented in the no-feedback trials were selected from a different distribution than the stimuli in the feedback trials, resulting in a shifted decision bound. In a test after the training phase, participants preferred the shifted decision bound to classify. Further, in the study of Rogers, Kalish, Gibson, Harrison and Zhu (2010) participants showed a higher accuracy in the semisupervised compared to the supervised condition when they were forced to answer as quickly as possible.

On the other hand, other studies found no effect of the no-feedback trials (McDonnell, Jew & Gureckis, 2012; Kalish et al., 2015; Rogers et al., 2010; Vandist et al., 2009). For example, in Vandist et al. (2009) the accuracy rates in the semisupervised condition were investigated and compared to a condition where the no-feedback trials were replaced by irrelevant fillers. Although participants in the semisupervised condition encountered twice as many stimuli during the experiment, the stimuli in the no-feedback trials seemed to be ignored by the participants since the accuracy rates were similar in both conditions. Finally, no studies reported a detrimental effect of no-feedback trials on the learning process.

Thus, whether and under which conditions no-feedback trials play an enhancing role in semisupervised learning, seems to be unclear. One factor that might explain the contradictory results, is the fact that the aforementioned studies used different *underlying category structures* to assess category learning. Ashby and O'Brien (2005) distinguish four different types of category learning structures, namely rule-based, information-integration, unstructured and prototype distortion learning structures (see also Poldrack & Foerde, 2008 and Smith & Grossman, 2008). They postulate that depending on these underlying category structures, a different category learning system is involved and this results in other learning characteristics. For example, several studies have indicated that in supervised learning, depending on the underlying category structure, manipulating feedback has a differential impact on learning. When feedback is delayed within a trial, no effect on learning rule-based category structures was observed whereas learning the information-integration category structure was impaired (Maddox, Ashby & Bohil, 2003; Maddox & Ing, 2005). Similarly, when feedback

is deferred to the end of a block and rearranged (positive outcomes first), no effect on rule-based category structures was observed whereas no successful learning occurred using an information-integration category structure (Smith et al., 2014). Considering that the effects of manipulating when to give feedback in supervised learning differ depending on the category structure, it seems plausible that this also applies for semisupervised learning. It has for example already been shown that the underlying category structure can determine the impact of feedback trials when comparing semisupervised to unsupervised learning. For example, Vong, Navarro and Perfors (2016) used a free sorting task where participants could sort the stimuli into as many categories as they like. When the underlying category structure was distinct, implying that the categories are clearly separated, the feedback trials had no influence. The classification was so obvious that participants did not need feedback. In contrast, when the category structure was ambiguous, the no-feedback trials were not sufficient to learn the classification. In this case, providing feedback, in the form of adding label information, had a large impact on how participants classified: even when only one exemplar of every category was labeled, performance improved significantly.

Based on this we can approach the existing literature on semisupervised category learning and the role of no-feedback trials therein, from a new angle, taking into account the different category structures that are used.

First, in *rule-based category structures*, the category membership can be determined by means of some explicit reasoning process. Frequently, the rule that maximizes accuracy (i.e., the optimal rule) can

be verbalized (Ashby, Alfonso-Reese, Turken & Waldron, 1998). An example is: “Respond A when the stimulus is small; respond B when the stimulus is big”. In many cases, only a single stimulus dimension is relevant, although structures where the optimal strategy is a logical conjunction are rule-based too (Ashby et al., 1998). All semisupervised studies that used a simple rule-based category structure and unidimensional stimuli, found an effect of no-feedback trials on the preferred decision bound (Gibson, Zhu, Rogers, Kalish & Harrison, 2010; Kalish, Rogers, Lang & Zhu, 2011; Lake & McClelland, 2011; Zhu et al., 2007). Using a simple rule-based category structure with two dimensional stimuli also led to superior semisupervised learning, albeit dependent on the age of the participants and the percentage of feedback trials (Kalish et al., 2015). Specifically, young children always seemed to use the no-feedback trials to learn whereas older children only used the no-feedback trials when feedback is scarce. Thus, in general, for rule-based category structures, no-feedback trials seem to aid category learning.

In *information-integration category structures*, it is impossible to describe the optimal decision rule verbally (Ashby et al., 1998). Here, information about at least two stimulus dimensions must be predecisionally integrated to respond optimally (Ashby & Gott, 1988). The cognitive system accomplishes dimensional integration, but it does so implicitly and procedurally. It is closely associated with motor or skill learning. An example can be choosing the appropriate category of motor programs during a tennis match (e.g., backhand, topspin,…) (Ashby & O’Brien, 2005). It is difficult to explain in words how the solution is reached. Although this category structure takes more time to learn and seems arbitrary at the start, it is structured because the within-

category correlation of the stimulus dimensions is high. Due to this high within-category correlation, pigeons can learn information-integration category structures as fast as rule-based structures (Smith et al., 2011). The semisupervised study of Vandist et al. (2009) is the only reported study using this information-integration category structure and showed no effect of the no-feedback trials on the learning process.

Third, *unstructured categories* lack any coherent structure and are the most difficult type to learn (Shepard et al., 1961). An example is the category 'personal numbers', including phone numbers, zip codes, social security numbers, etc. (Ashby & O'Brien, 2005). There are no known semisupervised category learning studies using unstructured categories. However, we expect that, given the high difficulty to learn these stimuli even in a fully supervised way and the impossibility to learn these stimuli unsupervised, semisupervised learning will be difficult to obtain with this category structure and it seems unlikely that no-feedback trials will have an impact.

Finally, in *prototype distortion category structures* a category exists of a prototype and many highly similar exemplars. These exemplars are created by randomly distorting the prototype (Casale & Ashby, 2008). This type of learning is often referred to as repetition priming and is supposed to rely on the perceptual representation system (Schacter, 1990; Zeithamova, Maddox & Schnyer, 2008). This perceptual representation system causes a feeling of familiarity every time an A stimulus is presented. The not A stimuli are randomly made, are perceptually dissimilar and do not elicit a familiarity feeling. In the 'A not A task', participants have to classify the stimulus as being a member of the category or not (Kéri, Kelemen, Benedek, & Janka,

2001; Knowlton & Squire, 1993; Zeithamova et al., 2008). To be successful, participants can use the decision rule 'if the stimulus feels familiar, respond A; if the stimulus feels unfamiliar, respond not A' (Casale & Ashby, 2008). The perceptual representation system does not depend on feedback for learning, the simple repetition is sufficient (Schacter, 1990; Wiggs & Martin, 1998). Casale and Ashby (2008) showed that it is possible to learn the A not A task supervised and unsupervised. When the distortion between the prototype and the other category exemplars was low, performance was even better unsupervised than supervised. There are no known semisupervised category learning studies using the A not A category structure. However, we expect that participants will benefit from semisupervised learning since feedback is not crucial to learn this category structure successfully. We therefore predict that the no-feedback trials will facilitate the learning process.

Thus, looking at the semisupervised learning literature from a different angle, no-feedback trials seem to enhance learning when using simple rule-based category structures, but seem to have no impact on learning complex information integration category structures. In the current paper semisupervised human category learning was investigated using the prototype distortion category structure. This structure is interesting from the point of view of semisupervised category learning, as this structure can also be learnt unsupervised, sometimes even outperforming supervised learning. This makes this structure highly relevant to study the effect of no-feedback trials on learning.

In Experiment 1, we examined whether it is possible to learn the 'A, not A' task semisupervised as compared to supervised and

unsupervised learning. Therefore, there were two control conditions, the supervised condition (100% feedback trials) and the unsupervised condition (0% feedback trials). Based on the study of Casale and Ashby (2008), we expected successful learning in both the supervised and unsupervised conditions. Additionally, we implemented two semisupervised conditions: a 50% condition, in which half of the trials in a block were randomly followed by feedback and a 25%, in which a quarter of the trials in a block were randomly followed by feedback. Since former studies reported that learning can be successful unsupervised and supervised, successful learning is also expected in the semisupervised conditions. The performance, measured in accuracy, was compared between these 4 conditions (0%, 25%, 50% and 100% feedback).

Experiment 2 further addressed the impact of the no-feedback trials on the learning process. Therefore, in two new conditions the no-feedback trials were replaced by irrelevant fillers and these conditions were compared to the semisupervised conditions of Experiment 1. If the no-feedback trials are helpful for learning, improved accuracy performance should be observed in the semisupervised conditions compared to the filler conditions. If the no-feedback trials harm the learning process or are ignored by the participants, worse or similar performance should be obtained respectively, compared to the filler conditions.

In Experiment 3 we explored the minimum percentage of feedback needed in order to learn the ‘A, not A’ task successfully. A 10% semisupervised condition was used and was compared to the 100% and 0% conditions of Experiment 1.

Finally, in Experiment 4, we tested the hypothesis, based on the study of Kalish et al. (2015) in older children, that, when feedback is given only sporadically, the no-feedback trials have a stronger impact. We did this by comparing the 10% semisupervised condition of Experiment 3 to a 10% filler condition in which the no-feedback trials were replaced by irrelevant fillers.

Experiment 1

Method

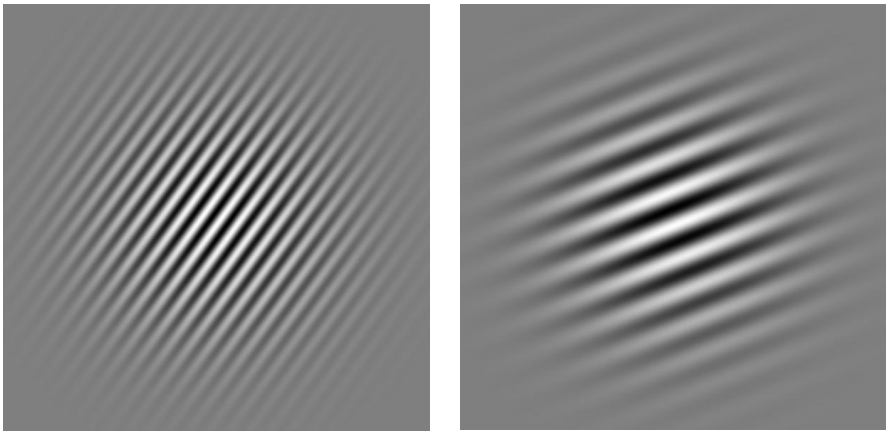
Participants. In total 41 participants (34 women, average age 22.6 years, *SD*=2.42, range=18-30) took part in the experiment in return for a payment of 15 euro. All participants reported normal vision.

Design. Experiment 1 consisted of two parts: a learning phase and a test. During the learning phase, the amount of feedback received was manipulated, creating four between-subject conditions. In the 100% condition (n=10), participants always received feedback, resulting in a supervised classification learning condition. In the 0% condition (n=9), participants never received feedback, leading to the unsupervised classification learning condition. In the 50% (n=12) and the 25% (n=10) conditions, participants received feedback in exactly 50% or 25% of the trials, respectively. In this way, two semisupervised classification learning conditions were created. The percentage of feedback trials a participant received, remained the same in all blocks of the learning phase. The presentation of feedback and no-feedback trials was randomized within blocks. After this learning phase, all

participants received the same test. In the test no feedback was provided.

Stimuli and apparatus. Participants viewed the stimuli on a 17 inch LCD monitor with a 800 x 600 resolution at a distance of approximately one arm's length. The experiment was conducted using Tscope (Stevens, Lammertyn, Verbruggen, & Vandierendonck, 2006). In all conditions gray 300 x 300 square-pixel Gabor patches were shown on a black screen (two examples can be seen in Figure 4.1).

Figure 4.1



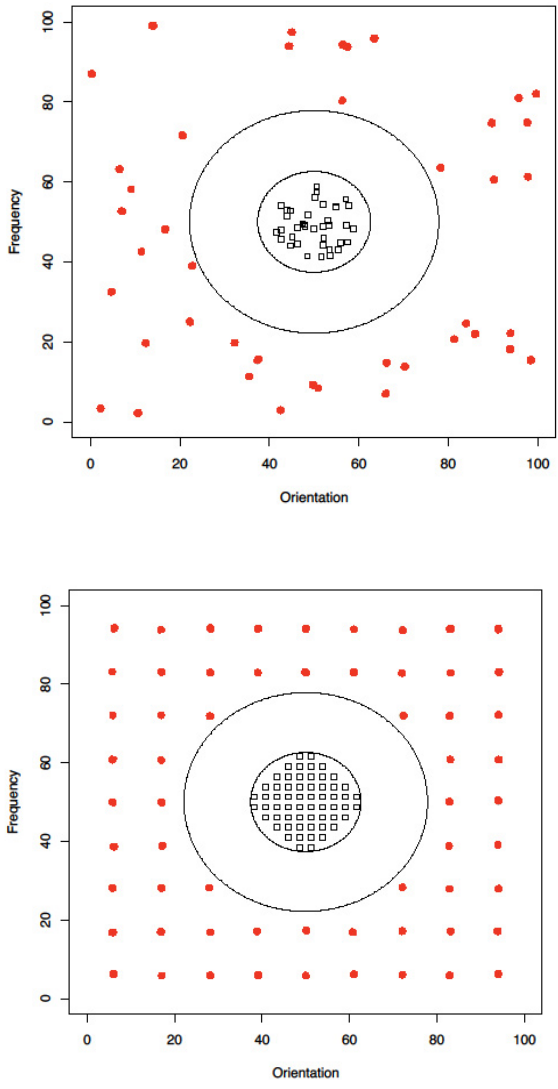
Two examples of Gabor patches. Left panel: an exemple of category A. Right panel: an example of not category A.

It was our goal to simulate natural learning as close as possible. Therefore, we selected stimuli with continuous stimulus dimensions since natural objects often consist of continuous dimensions (for example, the thickness of the trunk of trees). Furthermore, the stimuli in our experiments consisted of 2 stimulus dimensions, namely spatial

frequency and spatial orientation, to approach the complexity of stimuli in real life. The arbitrary stimulus coordinates were converted to physical units using the following transformations: spatial frequency (in cycles/pixel) was converted using f(y) = 0.01 + (y/1500), with y varying between 0 and 100. Spatial orientation is referred to in x degrees, varying between 0 and 100 degrees (Vandist et al., 2009). The coordinates originate in the A, not A category structure, as displayed, for example, in Figure 4.2.

For the learning phase, 640 stimuli were generated by random sampling for every participant. The 320 ‘A’ stimuli originated from a bivariate normal distribution with mean (50 50) and standard deviation (5 5). All A stimuli fell within the radius of sqrt(500/pi) of the mean. This category A spot corresponded to 5% of the total area. Consequently, the within-category similarity was high, which is a necessary condition to elicit the perceptual representation system. The 320 ‘not A’ stimuli were equally distributed from the area outside the radius of sqrt(1000/pi) + 10 with mean 50 50, implying that the distance from a not A stimulus to the A prototype was at least 27.84 for each dimension. The not A stimuli were visually distinct and their stimulus space corresponded to 75.66% of the total area. Consequently, almost 20% of the total area was not used to generate stimuli and this area fell in between the A and not A stimulus areas. The optimal decision bound, which classifies the stimuli perfectly in two categories, was circular.

Figure 4.2



The upper panel shows an example of the underlying category structure in the learning phase. The lower panel shows the underlying category structure of the test. Stimuli of category A are always depicted as squares, stimuli of not category A as filled red circles.

For the test, 60 ‘A’ stimuli and 60 ‘not A’ stimuli were generated. These test stimuli were the same for all participants in all conditions. As in the learning task, the ‘A’ stimuli fell within the radius of sqrt(500/*pi*) of the mean (50 50). Within this radius, the stimuli were equally divided over the space. The not ‘A stimuli’ again fell outside the radius of sqrt(1000/*pi*) + 10 with mean (50 50) and they were equally distributed within this space.

We deliberately used randomized feedback since we believe that in real life every stimulus can be followed by feedback or not. This is in contrast to the previous studies (Kalish et al., 2015; Lake & McClelland, 2011; Zhu et al., 2010; Zhu et al., 2007) in which a vast subset of the category was always followed by feedback and another subset never. We also believe that in real life feedback is given irregularly within a block and not grouped over blocks. The performance of the participant was measured during the learning phase (training) and on a test. Helie, Shamloo and Ell (2017) advised to use tests because they provide insight in what participants learned exactly. The test was the same for all participants in all conditions: the same stimuli were given and no feedback was provided. As a result, the test could be used to compare performance and categorization representations over conditions.

Procedure. All participants were tested individually in a dimly lit room. Every participant started with the learning phase, consisting of 640 trials. These trials were distributed over 8 blocks. In each block, an equal number of ‘A’ and ‘not A’ stimuli was guaranteed. Participants were told that they would receive stimuli one by one, belonging to category A or absolutely not to category A. They were asked to respond

by pressing A on the keyboard if they judged the stimulus to be an 'A' and to press 'N' when they judged the stimulus not to be an A. A correct response, in all conditions, is then defined as an 'A' response when the stimulus indeed belonged to category A or an 'N' response when the stimulus did not belong to category A. They were told that at the beginning of the experiment their responses would be guesses, but that they could potentially reach 100% accuracy as the experiment continued. Depending on the condition, the procedure differed. In the 100% condition, participants were informed that they would receive feedback (i.e., the true category label) after each category response. In the 50% and 25% conditions, participants were told that feedback would follow sometimes after a response. In the 50% condition, feedback followed randomly after half of the trials, whereas in the 25% condition feedback followed randomly after a quarter of the trials only. In the 0% condition, participants were informed that they would never receive feedback.

A feedback trial started when the stimulus was projected in the middle of the screen. After the response of the participant, the stimulus disappeared and feedback became visible for 1500 msec. A no-feedback trials had the same timing as the feedback trials except that the response was followed by a blank screen for 1500 msec. In all conditions, except for the 0% condition, the mean accuracy of the feedback trials was displayed after each block. Response speed was not stressed during the learning phase.

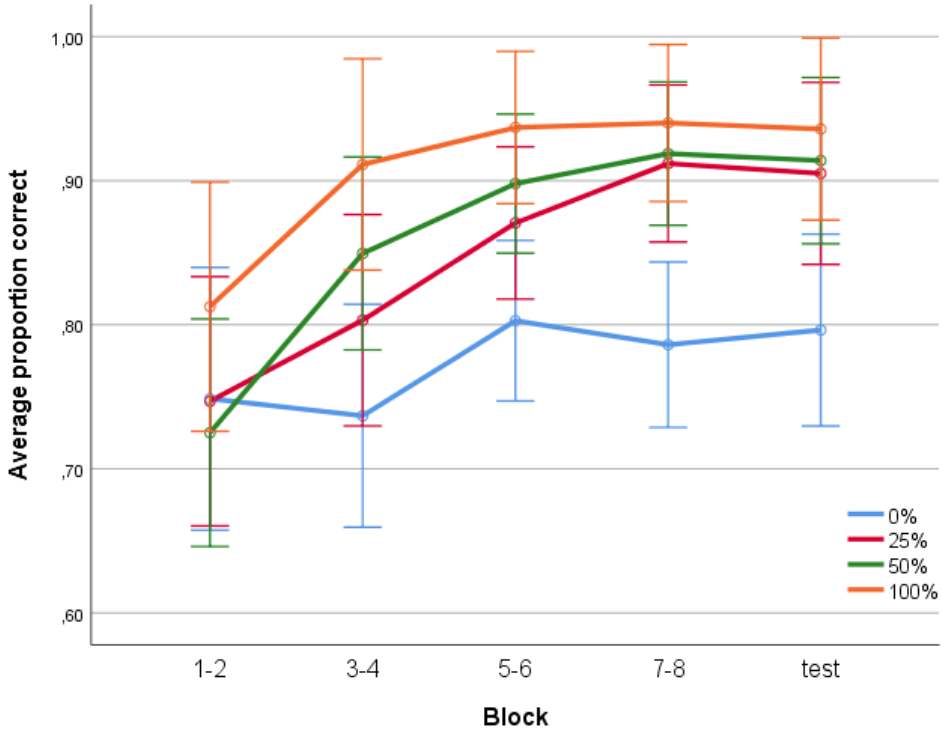
After the learning phase, participants received the test. During the test 60 A stimuli and 60 not A stimuli were shown. These 120 stimuli were randomized within one block. Participants were asked to perform

similarly as they did during the last block of the learning phase. They were informed that no feedback would be provided during the test. It was stressed to respond as quickly as possible. The course of a trial was the same as a no-feedback trial in the learning phase.

Results

Learning phase. Participants could achieve perfect accuracy when they categorized based on the perceptual similarity (of the A category). The accuracy analyses were based on all trials, namely both the feedback and no-feedback trials. For all analyses, the eight blocks were grouped and averaged per two blocks (i.e., blocks 1+2, 3+4, 5+6, 7+8). This was done in order to be able to compare all conditions reported in this paper on a sufficient amount of trials (in Experiment 2 and 4 filler conditions were used and the averages were calculated on the trials with stimuli only). This way we retained enough response trials per analysis unit. Learning was considered successful when participants obtained an accuracy of minimum 85%. This criterion was chosen in order of Experiment 4 and allowed participants, if they learned successfully, to make one accidental mistake in each block. Figure 4.3 shows the mean accuracy in proportion per two blocks for every condition.

Figure 4.3



Mean accuracy in proportion by every two response blocks and the test of Experiment 1. The error bars reflect the 95% confidence intervals.

In Table 4.1 the mean accuracy for every participant in the last two blocks of the learning phase is displayed. In the 100%, 50% and 25% conditions, learning was successful: the mean accuracy across all participants increased from an average of 81% (*SD*=14.41) in the 100% condition, 73% (*SD*=12.25) in the 50% condition and 75% (*SD*=12.90) in the 25% condition in the first two blocks to almost perfect accuracy in the last two blocks, respectively 94% (*SD*=4.86), 92% (*SD*=9.87) and 91% (*SD*=7.39).

Table 4.1.

		Condition			
		100%	**50%**	**25%**	**0%**
Participant	1	98	93	98	92
	2	97	96	80	81
	3	93	94	86	79
	4	99	88	97	85
	5	96	99	90	81
	6	90	91	93	86
	7	98	100	97	58
	8	83	99	92	81
	9	94	98	79	66
	10	91	64	100	
	11		97		
	12		84		

Mean accuracy (%) on the aggregated last two blocks of the learning phase for every participant of Experiment 1.
Note: each condition contained different participants.

The individual accuracy rates showed that all participants learned successfully, except participant 8 in the 100% condition, participants 10 and 12 in the 50% condition and participants 2 and 9 in the 25% condition. Contrarily, there is no convincing indication that the learning of participants in the 0% condition increased any further after the first two blocks: the mean accuracy stayed similar throughout all blocks, ranging from 75% (*SD*=14.67) in the first two response blocks to 79% (*SD*=10.59) in the last two blocks. However, three participants (1, 4 and 6) did manage to reach an accuracy of more than 85%. Mean accuracies were submitted to a repeated measures analysis of variance with block (i.e., 4 levels: aggregated blocks 1-2, 3-4, 5-6, 7-8) as within-subjects factor and condition (4 levels: 0%, 25%, 50% or 100% feedback) as between-subjects factor. Results show that there was a main effect of block, $F(3,35)=13.76$, $p=.001$, $\eta_p^2=.54$, indicating that across the blocks

the accuracy increased. Paired sample *t*-tests using the Bonferroni correction for multiple comparisons showed that in comparison to block 1-2, mean accuracy was significantly higher on block 3-4, 5-6 and 7-8 (resp. $t(40)=3.93$, $p<.01$; $t(40)=6.38$, $p<.001$; $t(40)=6.35$, $p<.001$). There was also a significant increase in accuracy between block 3-4 and blocks 5-6 and 7-8 (resp. $t(40)=4.17$, $p<.001$; $t(40)=3.80$, $p<.01$). There was no significant difference between block 5-6 and block 7-8, $t(40)=1.44$, $p=.90$. There was a main effect of condition as well, $F(3,37)=3.81$, $p=.02$, $\eta_p^2=.24$. Independent sample *t*-tests, corrected for multiple comparisons by Bonferroni, revealed that this effect is only due to the lower mean accuracy in the unsupervised condition (77%, $SD=10.41$) in comparison to the supervised condition (90%, $SD=5.84$), $t(12.30)=3.35$, $p=.04$. Finally, there was no interaction effect, $F(9,86)=1.77$, $p=.09$, $\eta_p^2=.13$.

Test. Even in this part it was possible for participants to achieve maximum accuracy. The mean accuracy rates for every participant are printed in Table 4.2. Again, in the 100%, the 50% and the 25% learning was successful, with respective mean accuracies of 94% ($SD=4.46$), 91% ($SD=7.88$) and 91% ($SD=5.53$). In the 0% condition the mean accuracy was only 80% ($SD=17.53$). A one-way ANOVA confirms that these differences in accuracy between the conditions on the test were significant, $F(3,37)=3.74$, $p=.02$, $\eta_p^2=.23$.

Table 4.2.

		Condition			
		100%	**50%**	**25%**	**0%**
Participant	1	94	98	93	93
	2	96	97	83	87
	3	99	89	88	90
	4	88	91	98	93
	5	98	98	88	74
	6	85	94	93	88
	7	91	89	99	67
	8	95	95	85	86
	9	96	94	84	39
	10	95	68	93	
	11		94	98	
	12		90		

Mean accuracy (%) on the test for every participant of Experiment 1.
Note: each condition contained different participants.

Discussion

The objective of Experiment 1 was to test semisupervised category learning using the prototype distortion category structure. Therefore the 'A, not A' task was used and the obtained results were compared to 100% and 0% learning conditions. The results showed that participants were in general successful in the 100% and in both semisupervised conditions (50% and 25%): the mean accuracy increased over time and exceeded 85% at the end of the learning phase and in the test. By contrast, in the 0% condition most participants failed to learn. The mean accuracy stayed below 85% during the last two blocks of the learning phase and in the test.

The results of the semisupervised conditions indicated that participants were able to master the prototype distortion category

structure in a semisupervised way, implying that feedback is not needed after every trial to learn successfully. Nevertheless, it is still unclear what the role is of the no-feedback trials in the learning process. This was addressed in Experiment 2. In Experiment 2, two follow-up conditions were run to explore the role of the no-feedback trials during learning. Therefore, in the 50% and the 25% condition, the no-feedback trials were replaced by irrelevant filler trials.

Experiment 2

Method

Participants. Twenty-one participants (17 women, mean age 23.5 years, *SD*=5.12, range 18-37 years) participated in the experiment in return for payment of 15 euro. None of them had participated in Experiment 1.

Procedure, stimuli and apparatus. The procedure, stimuli, apparatus and category task were identical to the 25% and 50% semisupervised conditions as described in Experiment 1. The difference between the 25% and the 50% conditions of Experiment 1 and the 25% (n=10) and 50% (n=11) filler conditions of Experiment 2 was that the stimuli in the no-feedback trials were replaced by unrelated filler-trials: when 'X' appeared on the screen, the participant had to press X on the keyboard. When 'Not X' was presented on the screen, the participant had to press the N key. These fillers did not disclose anything about the

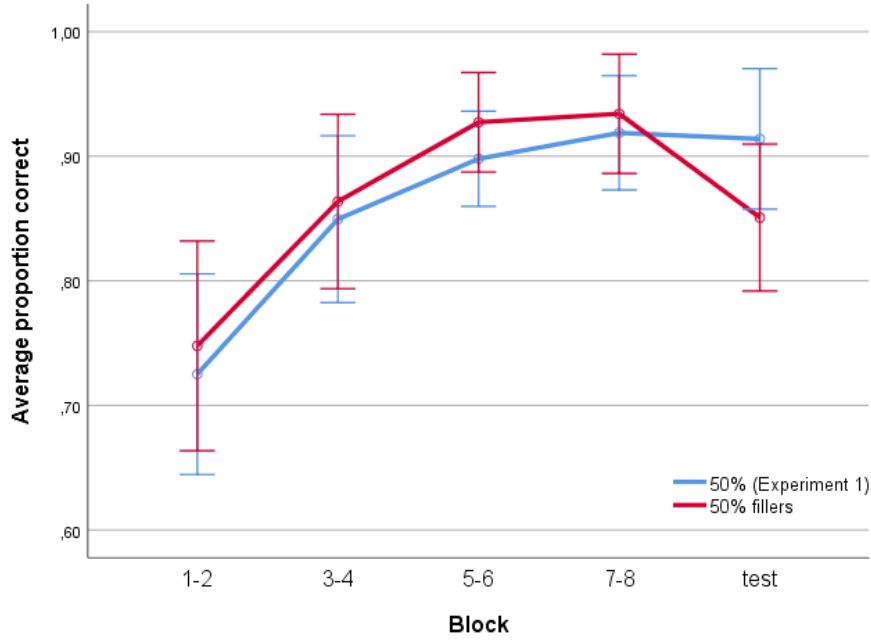
underlying category task nor did they cue the outcome of the following feedback trial. This was clearly stressed in the instructions to the participant. These two filler conditions were compared to the two semisupervised conditions of Experiment 1.

Results

Learning phase. In the filler conditions, the accuracy was calculated on the feedback trials only since the filler trials did not carry any information. Figures 4.4 and 4.5 show the mean accuracy in proportion per two blocks for the 25% and 50% semisupervised conditions of Experiment 1 and the 25% and 50% filler conditions of Experiment 2. In Table 4.3 the mean accuracy for every participant in the last two blocks of the learning phase is displayed. In the 50% filler condition, learning was successful: the mean accuracy across all participants increased from an average of 75% (*SD*=14.59) in the first two blocks to an almost perfect accuracy of 93% (*SD*=3.92) in the last two blocks. In the 25% filler condition, learning was successful as well: the mean accuracy rose from 73% (*SD*=18.34) in the first two blocks to 95% (*SD*=2.36) in the last two blocks. The individual accuracy rates on the last two blocks showed that in both conditions all participants learned successfully. To compare the performance between the 50% semisupervised condition of Experiment 1 and the 50% filler condition, the mean accuracies were submitted to a repeated measures analysis with block (i.e., 4 levels: aggregated blocks 1-2, 3-4, 5-6, 7-8) as within-subjects factor and condition (2 levels: 50% semisupervised and 50% filler) as between-subjects factor. Results showed that there was a

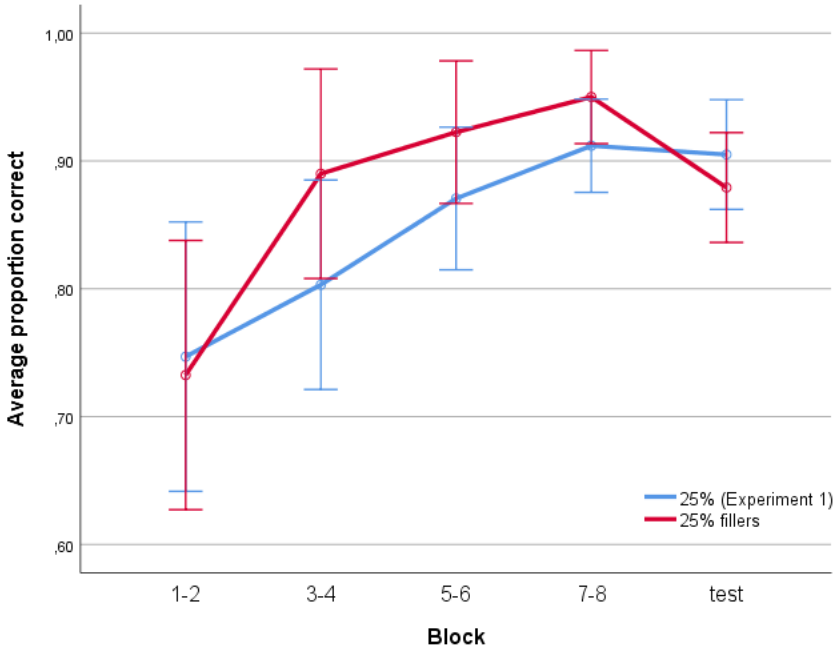
main effect of block, $F(3,19)= 17.62$, $p<.001$, $\eta_p^2=.74$, indicating that across the blocks the accuracy increased. There was no main effect of condition, $F<1$, $p=.55$, $\eta_p^2=.02$ and the interaction effect was also not significant, $F<1$, $p=.77$, $\eta_p^2=.06$. Similarly, the performance between the 25% semisupervised condition of Experiment 1 and the 25% filler condition was compared in a repeated measures analysis with block (i.e., 4 levels: aggregated blocks 1-2, 3-4, 5-6, 7-8) as within-subjects factor and condition (2 levels: 25% semisupervised and 25% filler) as between-subjects factor.

Figure 4.4.



Mean accuracy in proportion by every two response blocks and the test for the 50% semisupervised condition of Experiment 1 and the 50% filler condition of Experiment 2. The error bars reflect the 95% confidence intervals.

Figure 4.5



Mean accuracy in proportion by every two response blocks and the test for the 25% semisupervised condition of Experiment 1 and the 25% filler condition of Experiment 2. The error bars reflect the 95% confidence intervals.

The results were similar: there was a main effect of block, $F(3,16)=10.66$, $p<.001$, $\eta_p^2=.67$, due to increasing accuracy across the blocks. There was again no main effect of condition, $F(1,18)=1.08$, $p=.31$, $\eta_p^2=.06$ and no interaction effect, $F(3,16)=1.39$, $p=.28$, $\eta_p^2=.21$.

Table 4.3.

		Learning phase		Test	
		50% fillers	**25% fillers**	**50% fillers**	**25% fillers**
Participant	1	94	90	87	88
	2	90	95	89	89
	3	99	95	88	91
	4	94	95	93	88
	5	86	98	93	98
	6	90	98	88	95
	7	91	93	70	86
	8	96	95	98	92
	9	96	95	61	73
	10	93	98	81	79
	11	99		89	

Mean accuracy (%) on the aggregated last two blocks of the learning phase and the test for every participant of Experiment 2.
Note: the 50% and 25% filler conditions contained different participants.

Test. In the test of Experiment 2 learning seemed successful in the 50% and 25% filler conditions with mean accuracies of 85% (*SD*=10.82) and 88% (*SD*=7,28) respectively. In the 50% filler condition, three participants (7, 9 and 10) failed to obtain 85% or higher. In the 25% filler condition, two participants (9 and 10) scored below 85%. All other participants of Experiment 2 learned successfully. The mean accuracy rates for every participant are shown in Table 4.3. To compare the accuracy in the test phase of the 50% semisupervised condition of Experiment 1 to the 50% filler condition, a one-way ANOVA was conducted with condition as factor. It revealed that there was no significant difference between the conditions, $F(1,21)=2.59$, $p=.12$, $\eta_p^2=.11$. Similarly, there was also no significant difference between the 25% semisupervised condition of Experiment 1 and the 25% filler condition, $F<1$, $p=.38$, $\eta_p^2=.04$.

Discussion

To understand the impact of the no-feedback trials, the 25% and 50% semisupervised conditions of Experiment 1 were compared to 25% and 50% filler conditions of Experiment 2 respectively, in which the feedback trials were replaced by irrelevant fillers. The results suggested no impact of the no-feedback trials on the learning process, as no differences between the semisupervised and the filler conditions were observed with regards to accuracy performance.

Experiment 1 already indicated that the prototype distortion category structure can be learned in a semisupervised way. In Experiment 3 the minimum amount of feedback needed to achieve learning of this category structure is investigated. In the 10% semisupervised condition, feedback was provided after only 10% of the trials. This semisupervised condition was compared to the 100% and the 0% conditions of Experiment 1 in order to examine the semisupervised learning effects.

Experiment 3

Method

Participants. Eleven participants (11 women, mean age 20.3 years, *SD*=2.90, range=18-27) participated in the experiment in return for payment of 15 euro. None of them had participated in Experiments 1 or 2.

Procedure, design, stimuli and apparatus. The procedure, design, stimuli, apparatus and category task of the 10% condition were identical to the semisupervised conditions described in Experiment 1, except that only 10% of the trials were randomly followed by feedback.

Results

Learning phase. In the 10% semisupervised condition, all trials were used to calculate the accuracy. In Table 4.4 the mean accuracy for every participant in the last two blocks of the learning phase is displayed. Figure 4.6 presents the mean accuracy in proportion per two blocks for the 0% and 100% conditions of Experiment 1 and the 10% semisupervised condition of Experiment 3. In the 10% condition the learning of participants did not strongly increase after the first two blocks: the mean accuracy across all participants was 71% (*SD*=8.18) in the first two blocks and 80% (*SD* = 16.99) in the last two blocks. However, the individual accuracy rates indicated that six participants did learn successfully, namely, participants 1, 3, 6, 7, 8 and 11. To compare the performance between the 0% and 100% conditions of Experiment 1 and the 10% semisupervised condition of Experiment 3, mean accuracies were submitted to a repeated measures analysis of variance with block (i.e., 4 levels: aggregated blocks 1-2, 3-4, 5-6, 7-8) as within-subjects factor and condition (3 levels: 100%, 0% and 10% semisupervised) as between-subjects factor. Results showed a main effect of block, $F(3,25)=7.85$, $p<.01$, $\eta_p^2=.49$, indicating that during the experiment the accuracy increased. There was also a main effect of condition, $F(2,27)=6.52$, $p<.01$, $\eta_p^2=.33$.

Table 4.4.

		Learning phase	Test
Participant	1	94	93
	2	46	51
	3	99	97
	4	62	90
	5	67	64
	6	97	91
	7	89	81
	8	86	89
	9	77	73
	10	72	84
	11	94	92

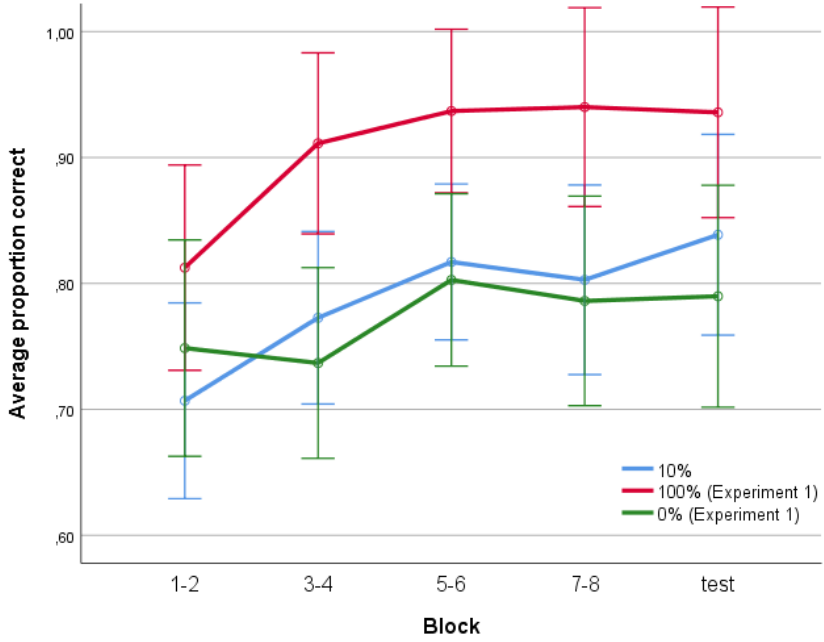
Mean accuracy (%) on the aggregated last two blocks of the learning phase and the test for every participant of Experiment 3.

Independent sample t-tests corrected with the Bonferroni correction for multiple comparisons revealed that there was a significant difference between the accuracy in the 100% condition (90%, *SD*=5.84) and the 10% condition (77%, *SD*=10.48), $t(19)=3.34$, $p<.01$. The 100% condition of Experiment 1 reached a significantly higher accuracy. There was no significant difference between the 0% condition of Experiment 1 and the 10% condition, $t(18)=.13$, $p>.99$. The interaction between block and condition was not significant, $F<1$, $p=.47$, $\eta_p^2=.06$.

Test. Table 4.4 presents the mean accuracy for every participant on the test. The mean accuracy was 82% (*SD*=14.14) in the 10% semisupervised condition. Five out of 11 participants obtained an accuracy score below 85% and failed to learn. A one-way ANOVA with condition (0%, 10% and 100%) revealed that there was no effect

of condition, $F(2,27)=3.13$, $p=.06$, $\eta_p^2=.19$, suggesting that the accuracy in the test was similar in all conditions.

Figure 4.6.



Mean accuracy in proportion by every two response blocks and the test for the 100% and 0% conditions of Experiment 1 and the 10% semisupervised condition of Experiment 3. The error bars reflect the 95% confidence intervals.

Discussion

The objective of Experiment 3 was to expose what the minimum percentage of feedback trials is to learn the prototype distortion category structure successfully. Therefore, a 10% semisupervised condition was administered and compared to the 100% and 0% conditions of Experiment 1. The results showed that, although the

accuracy increased during the learning process in the 10% condition, the accuracy remained significantly lower than in the 100% condition of Experiment 1 in the learning phase. Interestingly, about half of the participants did learn successfully in the 10% feedback condition whereas the other half failed. Although the test revealed no significant difference between the 10% condition and the 100% and 0% conditions of Experiment 1, the individual accuracy rates showed that half of the participants failed to learn. In conclusion, 10% semisupervised learning seemed to be too low to learn successfully for at least half of the participants.

Finally, in Experiment 4 the 10% semisupervised condition of Experiment 3 was compared to a 10% filler condition, in which the no-feedback trials were replaced by unrelated fillers. In this condition we addressed the hypothesis that if feedback is scarce (i.e., 10%), the no-feedback trials have a stronger impact on learning, as observed by Kalish et al. (2015) and Vong et al. (2016). If this hypothesis is true, better performance (i.e. higher accuracy) should be observed in the 10% semisupervised condition compared to the 10% filler condition.

Experiment 4

Method

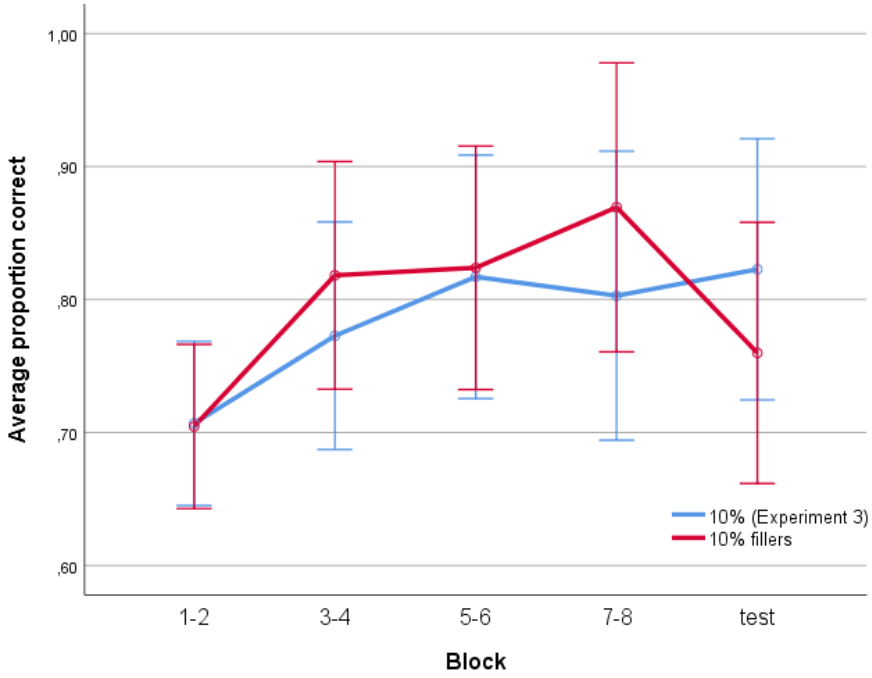
Participants. Eleven participants (11 women, mean age 20.1 years, *SD*=1.76, range=19-24) participated in the experiment in return for payment of 15 euro. None of them had participated in Experiments 1, 2 or 3.

Procedure, design, stimuli and apparatus. The procedure, design, stimuli, apparatus and category task of the 10% filler condition were identical to the filler conditions described in Experiment 2, except that 90% of the trials were irrelevant fillers now.

Results

Learning phase. The accuracy was calculated on the feedback trials only. Table 4.5 presents the mean accuracy for every participant in the last two blocks of the training phase. In Figure 4.7 the mean accuracy in proportion by every two response blocks is displayed for the 10% of Experiment 3 and the 10% filler condition. In the 10% filler condition successful learning seemed to occur: the mean accuracy across all participants was 70% (*SD* = 11.21) in the first two blocks and 87% (*SD* = 17.56) in the last two blocks. However, three participants (1, 4 and 8) failed to reach an accuracy over 85% on the last two blocks, suggesting no successful learning. To compare the performance between the 10% semisupervised condition of Experiment 3 and the 10% filler condition of Experiment 4, mean accuracies were submitted to a repeated measures analysis with block (i.e., 4 levels: aggregated blocks 1-2, 3-4, 5-6, 7-8) as within-subjects factor and condition (2 levels: 10% semisupervised and 10% filler) as between-subjects factor. Results showed a main effect of block, $F(3,18)=4.73$, $p<.05$, $\eta_p^2=.44$, indicating that the mean accuracy increased across blocks. There was no main effect of condition, $F<1$, $p=.56$, $\eta_p^2=.02$, nor a significant interaction effect, $F<1$, $p=.42$, $\eta_p^2=.14$.

Figure 4.7



Mean accuracy in proportion by every 2 response blocks and the test for 10% semisupervised condition of Experiment 3 and the 10% filler condition of Experiment 4. The error bars reflect the 95% confidence intervals.

Test. In Table 4.5 the mean accuracy for every participant on the test is displayed. The mean accuracy on the test was 76%, (*SD*=16.96). Eight out of eleven participants scored below 85% and thus failed to learn. An independent samples *t*-test showed that there was no significant difference between the 10% semisupervised condition of Experiment 3 and the 10% filler condition, $t(20)=0.95$, $p=.36$, $d=.40$.

Table 4.5

		Learning phase	Test
Participant	1	38	63
	2	94	79
	3	100	83
	4	81	93
	5	88	52
	6	94	40
	7	94	88
	8	81	83
	9	94	88
	10	94	84
	11	100	84

Mean accuracy (%) on the aggregated last two blocks of the learning phase and the test for every participant of Experiment 4

Discussion

In Experiment 4, the impact of the no-feedback trials in the 10% semisupervised condition of Experiment 3 was investigated. Therefore a 10% filler condition was designed and performance on the 10% semisupervised condition of Experiment 3 was compared to this 10% filler condition. The hypothesis was tested that, when feedback is scarce, the no-feedback trials become more important to the participants and can have an impact as shown by Kalish et al. (2015). Results revealed no difference in accuracy between the 10% semisupervised and the 10% filler conditions during the learning phase and the test. Hence, the no-feedback trials in the 10% semisupervised condition did not seem to have an impact: even when feedback was scarce, participants ignored the no-feedback trials.

Chapter 4

General Discussion

In this study, we started from the observation that semisupervised learning effects in the literature are mixed and that these mixed results might be due to the used underlying category structure. The semisupervised studies were reviewed, based on several category structures proposed by Ashby and O'Brien (2005). In simple rule-based category structures successful semisupervised learning effects have been reported. Under such circumstances, the participants use the stimulus information in the no-feedback trials, resulting in a shifted decision bound (Gibson et al., 2010; Kalish et al., 2011; Lake & McClelland, 2011; Zhu et al., 2007) or higher accuracy when participants are forced to answer as quick as possible (Rogers et al., 2010). In contrast, in information-integration category structures, participants seem to ignore the stimulus information in the no-feedback trials (Vandist et al., 2009). There are no known semisupervised category learning studies using the unstructured category structure or the prototype distortion category structure. The aim of the present study was to test semisupervised category learning using the prototype distortion category structure and the role of no-feedback trials therein, as the properties of this specific structure might be especially beneficial to promote the use of no-feedback trials.

In four experiments, an 'A not A task' was used and feedback was given randomly within blocks. Based on the characteristics of the perceptual representation system formulated by Schacter (1990) and Wiggs and Martin (1998), on which prototype distortion category structures rely, we expected that the no-feedback trials would contribute to the learning process, resulting in higher accuracy rates.

The stimuli, like many natural objects, were constructed with continuous stimulus dimensions. Experiment 1 showed that participants were able to learn the prototype distortion category structure supervised and semisupervised (for both 50% and 25% feedback conditions), whereas participants failed to learn unsupervised. Our results indicated that participants do not require feedback after every single trial to learn this category structure successfully, but that a minimum of feedback is needed. Experiment 2 addressed the impact of the no-feedback trials on the learning process. Two filler conditions were administered to compare to the 25% and 50% semisupervised conditions of Experiment 1. In the filler conditions the stimuli in the no-feedback trials were replaced by irrelevant stimuli, keeping the number of feedback trials the same. Our results showed that the no-feedback trials did not have an impact on the learning process, as results were similar when these no-feedback trials were replaced by irrelevant fillers. In Experiment 3, a 10% semisupervised condition was added to reveal the minimum relative percentage of feedback needed to learn successfully. The results showed that 6 out of 11 participants learned the category structure successfully and obtained a high accuracy, whereas the other half failed to learn. As a group, performance in the learning phase was significantly worse compared to the 100% condition of Experiment 1. In Experiment 4 we tested the hypothesis that, when feedback is scarce, no-feedback trials become more important and do have an impact on the learning process. However, again no difference was found between the 10% semisupervised condition of Experiment 3 and a 10% filler condition. This result indicated that even when feedback is rare, the no-feedback trials did not have an impact on the learning process and participants still seemed to neglect the no-feedback trials.

Interestingly, in the 10% semisupervised condition of Experiment 3, half of the participants failed to learn whereas the other half of the participants succeeded in mastering the categorization task. This result might be due to individual differences in category learning. The research on individual differences in category learning, however, is being developed fairly recent (for a review see Shen & Palmeri, 2016) and our findings stress the necessity to further investigate what determines how much feedback an individual participant needs to successfully learn the category structure.

In this study we wanted to investigate semisupervised learning in the perceptual representation system. We deliberately chose to keep the stimulus space of the A stimulus small (in this study it was 5% of the overall stimulus space), ensuring that the category A stimuli were highly similar. The higher the within-category similarity of the A stimulus, the more likely the perceptual representation system is active and will be used to make categorization decisions instead of other systems (Casale & Ashby, 2008). Unlike in the study of Casale and Ashby (2008), we did not replicate the finding that, when the distortion of the A stimuli is low, better performance is obtained in the 0% than in the 100% condition. It is possible that we were not able to elicit the perceptual representation system in this study for three reasons. First, the used distortion of the A stimuli in this study, which was 5% of the stimulus space, might have been too large. It is known that the perceptual representation system is extremely sensitive to distortion in comparison to other category memory systems. As a result, when dissimilarity increases, the probability of responding ‘A’ decreases quickly (Casale & Ashby, 2008). Second, the relevant stimulus dimensions in our study were distorted. Several studies have shown that

the activation of the perceptual representation system is more impaired when the distortion is applied to relevant stimulus dimensions than to irrelevant features (e.g., Roediger & Srinivas, 1993; Wiggs & Martin, 1998). Third, we used artificial Gabor patches with two continuous stimulus dimensions. When two dimensions are used, only five exemplars can be the nearest neighbors of the prototype (Odlyzko & Sloane, 1979). The study of Casale and Ashby (2008), for example, used dot patterns, where each of the nine dots could move horizontally and vertically, resulting in 18 dimensions and a minimum of 7398 nearest neighbors (calculations based on the study of Odlyzko & Sloane, 1979). As a result, the more stimulus dimensions that are used, the more stimuli that are similar to the distorted prototype can be produced and the more likely it is that the perceptual representation system will be activated. As a conclusion, it is possible that, although we used the 'A not A task' and we distorted the prototype of category A in this study, the perceptual representation system was not activated enough during learning. Therefore, future research should replicate the current semisupervised learning study using the stimuli and the low distortion stimulus space of the study of Casale and Ashby (2008).

In conclusion, this study shows that early in the learning process participants are able to master the prototype distortion category structure semisupervised when feedback is given regularly (25% or 50%). It seems that during the learning process the no-feedback trials are ignored, since replacing the no-feedback trials by task-uninformative fillers does not have an effect on the accuracy performance. Even when feedback is given only sporadically (e.g., after 10% of the trials), participants ignore the stimulus information given in the no-feedback trials. This result is surprising because in the

perceptual representational system, which is activated by the prototype distortion category structure, learning does not depend on feedback. However, due to procedural choices, it is possible that the distortion of the ‘A stimuli’ was too high and that the perceptual representation system was not activated enough during learning. Finally, the observed individual differences in (semisupervised) learning deserve further investigation.

References

Ashby, F.G., Alfonso-Reese, L.A., Turken, A.U., & Waldron, E.M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*, 442–481.

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

Ashby, F.G., & Maddox, W.T. (2005). Human category learning. *Annual Review of Psychology*, *56*, 149–178.

Ashby, F.G., & Maddox, W.T. (2011). Human category learning 2.0. *Annals of the New York Academy of Sciences*, 1224, 147-161.

Ashby, F.G., Maddox, W.T., & Bohil, C.J. (2002). Observational versus feedback training in rule-based and information-integration category learning. *Memory & Cognition*, *30* (5), 666–677.

Ashby, F.G., & O'Brien, J.B. (2005). Category learning and multiple memory systems. *Trends in Cognitive Sciences, 9* (2), 83-89.

Ashby, F.G., Queller, S., & Berretty, P.M. (1999). On the dominance of unidimensional rules in unsupervised categorization. *Perception & Psychophysics*, *61* (6), 1178–1199.

Casale, M.B., & Ashby, F.G. (2008). A role for the perceptual representation memory system in category learning. *Perception & Psychophysics*, *70* (6), 983–999.

Clapper, J.P., & Bower, G.H. (1994). Category invention in unsupervised learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20* (2), 443–460.

Gibson, B.R., Zhu, X., Rogers, T.T., Kalish, C.W., & Harrison, J. (2010). Humans learn using manifolds reluctantly. In J. Lafferty (Ed), *Advances in neural information processing systems,* vol 24, pp 730-738. Red Hook, NY: Curran Associates, Inc.

Helie, S., Shamloo, F., & Ell, S.W. (2017). The effect of training methodology on knowledge representation in categorization. *Plos One, 12* (8), e0183904.

Kalish, C.W., Rogers, T.T., Lang, J., & Zhu, X. (2011). Can semi-supervised learning explain incorrect beliefs about categories? *Cognition, 120* (1), 106–118.

Kalish, C.W., Zhu, X., & Rogers, T.T. (2015). Drift in children's categories: when experienced distributions conflict with prior learning. *Developmental Science*, 18 *(6), 940-956*.

Kéri, S., Kelemen, O., Benedek, G., & Janka, Z. (2001). Intact prototype learning in schizophrenia. *Schizophrenia Research*, *52* (3), 261–264.

Knowlton, B.J., & Squire, L.R. (1993). The learning of categories: parallel brain systems for item memory and category knowledge. *Science*, *262*, 1747–1749.

Lake, B.M., & McClelland, J.L. (2011). Estimating the strength of unlabeled information during semi-supervised learning. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, 1400–1405.

Love, B.C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin & Review*, *9* (4), 829–835.

Maddox, W.T., Ashby, F.G., & Bohil, C.J. (2003). Delayed feedback effects on rule-based and information-integration category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29* (4), 650–662.

Maddox, W.T., Ashby, F.G., & Gottlob, L.R. (1998). Response time distributions in multidimensional perceptual categorization. *Perception & Psychophysics*, *60* (4), 620–637.

Maddox, W.T., Filoteo, J.V., Hejl, K.D., & Ing, A.D. (2004). Category number impacts rule-based but not information-integration category learning: further evidence for dissociable category-learning systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30* (1), 227–245.

Maddox, W.T., & Ing, A.D. (2005). Delayed feedback disrupts the procedural-learning system but not the hypothesis-testing system in perceptual category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31* (1), 100–107.

McDonnell, J.V., Jew, C.A., Gureckis, T.M. (2012). Sparse category labels obstruct generalization of category membership. *Proceedings of the Annual Meeting of the Cognitive Science Society, 34,* 749-754.

McKinley, S.C., & Nosofsky, R.M. (1995). Investigations of exemplar and decision bound models in large, ill-defined category structures. *Journal of Experimental Psychology: Human Perception and Performance*, *21* (1), 128–148.

Medin, D.L., & Schwanenflugel, P.J. (1981). Linear separability in classification learning. *Journal of Experimental Psychology: Human Learning & Memory, 7* (5), 355-368.

Medin, D.L., Wattenmaker, W.D. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive Psychology*, *19* (2), 242–279.

Milton, F., Longmore, C.A., & Wills, A. J. (2008). Processes of overall similarity sorting in free classification. *Journal of Experimental Psychology: Human Perception and Performance*, *34* (3), 676–692.

Odlyzko, A.M., & Sloane, N.J.A. (1979). New bounds on the number of unit spheres that can touch a unit sphere in n-dimensions. *Journal of Combinatorial Theory Series A, 26* (2)*,* 210-214.

Poldrack, R.A., & Foerde, K. (2008). Category learning and the memory systems debate. *Neuroscience and Biobehavioral Reviews*, *32* (2), 197-205.

Pothos, E.M., & Chater, N. (2002). A simplicity principle in unsupervised human categorization. *Cognitive Science*, *26* (3), 303-343.

Pothos, E.M., & Chater, N. (2005). Unsupervised categorization and category learning. *Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, *58* (4), 733–752.

Pothos, E.M., Perlman, A., Bailey, T.M., Kurtz, K., Edwards, D. J., Hines, P., & McDonnell, J.V. (2011). Measuring category intuitiveness in unconstrained categorization tasks. *Cognition*, *121* (1), 83–100.

Roediger, H.L., & Srinivas, K. (1993). Specificity of operations in perceptual priming. In P. Graf & M.E.J. Masson (Eds), *Implicit memory: New directions in cognition, development and neuropsychology* (pp.17-48). Hillsdale, NJ: Lawrence Erlbaum Associates.

Rogers, T.T., Kalish, C., Gibson, B.R., Harrison, J., & Zhu, X. (2010). Semi-supervised learning is observed in a speeded but not an unspeeded 2D categorization task. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society,* p. 2320–2325.

Schacter, D.L. (1990). Perceptual representation systems and implicit memory: Toward a resolution of the multiple memory systems debate. *Annals of the New York Academy of Sciences, 608,* 543-571.

Shen, J. & Palmeri, T. J. (2016). Modelling individual difference in visual categorization. *Visual cognition, 24* (3), 260-283.

Shepard, R.N., Hovland, C.I., & Jenkins, H.M. (1961). Learning and memorization of classifications. *Psychological Monographs*, *75* (13), 1–42.

Smith, E.E., & Grossman, M. (2008). Multiple systems of category learning. *Neuroscience and Biobehavioral Reviews, 32* (2), 249-264.

Smith, J.D., Ashby, F.G., Berg, M.E., Murphy, M.S., Spiering, B., Cook, R.G., & Grace, R.C. (2011). Pigeons' categorization may be exclusively nonanalytic. *Psychonomic Bulletin & Review*, *18* (2), 414–421.

Smith, J.D., Boomer, J., Zakrzewski, A.C., Roeder, J.L., Church, B.A., & Ashby, F.G. (2014). Deferred feedback sharply dissociates implicit and explicit category learning. *Psychological Science*, *25* (2), 447–57.

Stevens, M., Lammertyn, J., Verbruggen, F., & Vandierendonck, A. (2006). Tscope: A C library for programming cognitive experiments on the MS windows platform. *Behavior Research Methods*, *38*, 280–286.

Vandist, K., De Schryver, M., & Rosseel, Y. (2009). Semisupervised category learning: the impact of feedback in learning the information-integration task. *Attention, Perception & Psychophysics*, *71* (2), 328–341.

Vong, W.K., Navarro, D. J., & Perfors, A. (2016). The helpfulness of category labels in semi-supervised learning depends on category structure. *Psychonomic Bulletin & Review*, *23* (1), 230–238.

Wiggs, C.L., & Martin, A. (1998). Properties and mechanisms of perceptual priming. *Current Opinion in Neurobiology, 8* (2)*,* 227-233.

Zeithamova, D., Maddox, W.T., & Schnyer, D.M. (2008). Dissociable prototype learning systems: evidence from brain imaging and behavior. *The Journal*

of Neuroscience : The Official Journal of the Society for Neuroscience, *28* (49), 13194–13201.

Zhu, X., Gibson, B.R., Jun, K.S., Rogers, T.T., Harrison, J. & Kalish, C. (2010). Cognitive models of test-item effects in human category learning. In *Proceedings of the 27th International Conference on Machine Learning,* 1247–1254.

Zhu, X., Rogers, T., Qian, R., & Kalish, C. (2007). Humans Perform Semi-supervised Classification Too. In *Proceedings of the 22Nd National Conference on Artificial Intelligence (AAAI'07),* 864–869.

Chapter 5

General Discussion

Goal of this dissertation

This dissertation departed from the observation that in human category learning, learning is investigated using supervised or unsupervised paradigms. In supervised learning, every single stimulus is followed by the true category label, whereas in unsupervised learning, no information about the category label is ever provided. We argued that in real life human category learning is more likely to occur in a semisupervised way: some stimuli receive feedback (i.e., feedback trials) and others do not (i.e., no-feedback trials). The goal of this dissertation was to investigate the possibilities and limitations of human semisupervised category learning. In addition, we examined whether the no-feedback trials that do not receive feedback in a stream of feedback trials speed up the learning process, slow it down or have no impact at all.

Overview of the empirical results

Chapter 2 started with the creation of a semisupervised category learning paradigm. This paradigm was tested in two semisupervised conditions: a 50% condition, in which half of the trials was followed by feedback and a 25% condition, in which a quarter of the trials was followed by feedback The feedback trials and the no-feedback trials in these conditions were presented intermixed. The obtained results in these semisupervised conditions were compared to a supervised and an unsupervised condition. The information-integration category structure was used, which is a complex category structure, impossible to learn without feedback. This is due to the impossibility to verbally describe the optimal decision bound, which reflects the best possible division of the stimuli into two (or more) categories (Ashby, Queller & Berrety,

1999). To obtain high-level performance in the information-integration category structure, participants have to combine the perceptual information of the underlying stimulus dimensions simultaneously at some predecisional stage (Ashby & Gott, 1988). In Experiment 1, participants learned successfully when feedback was always provided (supervised condition) whereas no learning occurred when no feedback was provided (unsupervised condition). In the semisupervised conditions, the participants who received 50% feedback trials learned successfully, whereas learning failed for almost all participants who only received 25% feedback trials. To study whether this failure was due to the relative low percentage of feedback trials or to the low total number of feedback trials, Experiment 2 was designed. In this second experiment, the 25% condition was extended so that this condition contained the same total number of feedback trials (i.e., 400) compared to the 50% condition of Experiment 1, randomly mixed with 1200 no-feedback trials. Now participants did learn successfully. The results clearly showed that it is the total number of feedback trials (and not the percentage of feedback) that is decisive for successful semisupervised learning. Additionally in Experiment 2, the impact of the no-feedback trials on the learning process was investigated in the 50% semisupervised condition. To achieve this, a new condition was created where the 50% no-feedback trials were replaced by irrelevant filler trials. In an irrelevant filler trial participants saw the letter 'Y' or 'N' on the screen and they had to press the corresponding key on the keyboard. Participants were informed that these letters did not give any information about the previous or the following trial. The results revealed that participants in this filler condition learned as quickly as in the 50% semisupervised condition of Experiment 1. Consequently, the no-feedback trials in the semisupervised condition did not have an impact on the learning process.

Chapter 3 built further on the results and discussion of Chapter 2. Therefore the same information-integration category structure was used. In contrast to Chapter 2, in which semisupervised learning was introduced from the start of the experiment and was thus used early in the learning process, we now focused on the effects of semisupervised category learning late in the learning process. In Experiment 1 all participants were first trained fully supervisedly during two succeeding days. Only participants who reached a certain expert level were selected for the actual experiment. On the next two days, half of the participants further trained supervisedly (i.e, supervised condition), while the other half received a 25% semisupervised training (i.e., semisupervised condition). The absolute number of feedback trials was the same in both conditions. On the final day, all participants took part in the same speeded test where no feedback was provided. Importantly, the response times of both groups was similar at the end of the first two training days. Interestingly, although the accuracy on the test day was equal for both groups, the participants of the semisupervised condition responded significantly faster on the test than the participants of the supervised condition. This result suggested that the categorization performance of the semisupervised learners was more automatic than the supervised learners. Still, alternative explanations could not be excluded. First, although both groups received an equal amount of feedback trials, this implied that the semisupervised learners received a higher total amount of trials and thus more (unsupervised) training. Second, until the test day, the supervised learners had no experience with receiving no feedback after a classification. This might have caused a 'surprise' effect on the no-feedback test trials, which perhaps slowed down the supervised learners. In Experiment 2, these alternative explanations were addressed. Again, participants were trained supervisedly during two succeeding

days and only experts were allowed to the actual experiment. On the next two days, half of the participants were trained in an *almost* supervised way (95% of the trials were randomly followed by feedback) and the other half in an *almost* unsupervised way (5% of the trials were randomly followed by feedback). As such, all participants had experience with no-feedback trials prior to the test day. The total number of trials of these two conditions was kept identical to the total number of trials of the semisupervised condition of Experiment 1. Hence, the relative number of feedback trials now differed between the conditions and the semisupervised condition contained less feedback trials than the almost supervised condition. Nevertheless, the results again indicated that the semisupervised learners responded significantly faster on the test, despite similar response times of all groups at the end of the second training day. As a conclusion, late in learning, when automaticity develops, participants seem to benefit from semisupervised learning as the no-feedback trials facilitated automaticity.

Chapter 4 was inspired by the suggestions of the majority of category learning researchers that multiple category-learning systems exist. Depending on the category structure of the given task, a specific category-learning system becomes active. Several studies report different effects of variables such as multitasking or delaying feedback, depending on the category learning system that is used. It is hence possible that semisupervised category learning results differ according to the used learning system. Since in both Chapters 2 and 3 the information-integration category structure was tested (tapping into the procedural learning system), the main goal of Chapter 4 was to test semisupervised category learning in another learning system. Specifically, the perceptual representation system was investigated. This

system can be activated by using the prototype distortion category structure. This category structure implies that all A category stimuli are slight variations of the prototype (which is the mean of the category). These stimuli have to be recognized in a stream of 'not A' stimuli, which are random stimuli, unrelated to category A. In general, better performance is obtained when this category structure is learned unsupervised (compared to supervised) on condition that the distortion is low and that there is no overlap between the A and not A stimuli (Casale & Ashby, 2008). In Experiment 1 two semisupervised conditions (50% and 25%) were compared to a supervised and an unsupervised condition. The semisupervised learning scheme was introduced from the start of the experiment. Results showed that participants failed to learn in the unsupervised condition. However, participants did learn successfully in the supervised and both semisupervised conditions. Experiment 2 focused on the effect of the no-feedback trials on the learning process. The 50% and the 25% conditions of Experiment 1 were compared to filler conditions. In the 50% filler and the 25% filler condition the no-feedback trials were replaced by irrelevant fillers. Participants were informed that these fillers did not carry any information about the categorization task. Again, as in Chapter 2, the no-feedback trials did not have an impact early in this learning process. Experiment 3 addressed the minimum relative percentage of feedback needed to master the prototype distortion category structure. When a 10% semisupervised condition was used, half of the participants failed to learn. Finally, in Experiment 4 the effect of the no-feedback trials was again investigated by replacing the 90% no-feedback trials of Experiment 3 by irrelevant fillers. The results showed that, again, the no-feedback trials did not aid nor hamper the learning process.

Taken together, the obtained results of this dissertation suggest three recurrent conclusions. First, in comparison to supervised learning, semisupervised learning is at least as successful. When participants learn semisupervisedly, high accuracy rates can be achieved, regardless of the category structure of the given task. What is crucial for successful semisupervised learning, is that the total number of feedback trials is sufficiently high (cf. Chapters 2 and 4). Second, the onset of the semisupervised learning determines whether the no-feedback trials have an effect on the learning process. Specifically, when semisupervised learning is implemented from the beginning and hence the no-feedback trials are present early in the learning process, no effects of these no-feedback trials on learning are observed (cf. Chapters 2 and 4). The no-feedback trials neither aid, not hamper the learning process and seem to be simply ignored by the participants. However, when semisupervised learning is only introduced after a certain level of expertise has been acquired, the no-feedback trials do have an effect on learning. More specially, they enhance the development of automaticity: participants classified stimuli significantly faster on the test day and this without decreases in accuracy performance (Chapter 3). Third, in all studies we observed individual differences despite the similar background of our participants (psychology students). In Chapters 2 and 3 not all participants succeeded in mastering the information-integration category structure. In Chapter 2, it was a surprising finding that the mean accuracy of the 25% condition of Experiment 1 (run on a single day; 75% accuracy for the two last blocks) was much lower than the mean accuracy of the 25% condition of Experiment 2 (run on two consecutive days) after the first day (88% accuracy for the two last blocks on day 1). This is striking, as at until that point the experiments were exactly the same in both conditions. In Chapter 3, although participants obtained an accuracy

above 90% after the second training day, some participants unlearned the category structure by the end of the test day. In Chapter 4 half of the participants learned successfully in the 10% semisupervised condition whereas the other half did not. These observed individual differences will be discussed later in this general discussion.

Link to the broader literature

In 2009, we were among the first to study human semisupervised learning in cognitive science. Since then, several human semisupervised studies have been conducted. All researchers emphasize the importance of semisupervised learning research. Not only because it is the most natural way of human learning, but also because investigating semisupervised learning can help us to understand incorrect beliefs about categories (see discussion in Kalish, Rogers, Lang & Zhu, 2011). For example, a common misunderstanding is that bats belong to the category "birds". This is probably due to the fact that a bat can fly, which is a typical feature of birds. Meanwhile, the typical features for mammals that a bat possesses (e.g., the bat has an uterus and nurses her offspring) are less obvious. Usually people learn to divide animals into categories in a semisupervised way, since it is impossible to provide feedback for each encounter with more than a million known different animal species. In most cases semisupervised participants will label new animals correctly. However, when features of the animal belong to several contrasting categories (as is the case for the bat), mistakes are more easily made. Researchers believe that increasing our understanding of these mechanisms will improve category learning (e.g., Kalish, et al., 2011).

All previous semisupervised human category learning studies have focused on the early learning process. Some studies investigated semisupervised learning by itself, others compared it to supervised or to unsupervised learning performance. In the following parts, we will discuss the human semisupervised category learning studies and their relation to the findings of the current dissertation. Due to the sometimes fundamental differences in design and procedures, only cautious comparisons to our studies can be made.

Finding 1: semisupervised category learning is successful

Many studies showed that participants can learn semisupervisedly when the underlying category learning structure is simple (Gibson, Rogers, Kalish & Zhu, 2015; Rogers, Kalish, Gibson, Harrison & Zhu, 2010; Kalish et al., 2011; Kalish, Zhu & Rogers, 2015; Lake & McClelland, 2011; LaTourrette & Waxman, 2019; McDonnell, Jew & Gureckis, 2012; Patterson & Kurtz, 2018; Zhu, et al., 2010; Zhu, Rogers, Qian & Kalish, 2007). These results are in line with Chapter 4: we showed that semisupervised learning was successful in the (presumably) simple 'A not A' category structure. Semisupervised learning is even successful when complex categories have to be learned which have been proven to be impossible to learn without feedback. For example, the study of Gibson, Zhu, Rogers, Kalish and Harrison (2010) showed that people were able to learn manifolds, this is a common category structure in machine learning, semisupervisedly. In Chaper 2 and 3 we found that people are able to learn the complex information-integration category structure semisupervisedly. Similarly to Chapter 2, Gibson et al. (2010)

concluded that a minimal amount of feedback trials is needed to learn complex categories successfully in a semisupervised way.

Some studies compare the effect of semisupervised learning to unsupervised learning performance. The study of Vong, Perfors and Navarro (2016) revealed that semisupervised learners outperformed unsupervised learners when the underlying category structure is ambiguous. This result implies that a little amount of feedback is sufficient to boost performance. Accordingly, in Chapters 2 and 4 we too observed that semisupervised participants performed better than unsupervised participants.

When semisupervised learning is compared to supervised learning, most results indicate that semisupervised learning is at least as effective as supervised learning. Patterson and Kurtz (2018) reported that semisupervised participants obtained a higher mean accuracy in a simple categorization task compared to supervised participants. Rogers et al. (2010) found that, depending on the type of instructions, semisupervised participants obtained a similar or a higher accuracy in comparison to supervised learners. McDonnell et al. (2012) reported no difference between semisupervised and supervised participants. This is in line with our observations of similar (Chapters 2 and 4) or improved (Chapter 3) performance of semisupervised learners compared to supervised learners. Contrarily, Lake and McClelland found that semisupervised learning was successful but less effective than supervised learning. Of course, due to differences in designs and procedures in the experiments, we need to be cautious when comparing these studies.

In summary, participants can learn both simple and complex category structures semisupervisedly and, in most cases, semisupervised learning performance proves to be at least as good as supervised learning performance. An important precondition for this, is that a sufficient amount of feedback trials is provided. Yet, it remains unclear how the no-feedback trials contribute to the semisupervised learning process.

Finding 2: no-feedback trials have no impact on the semisupervised learning process early in learning, but speed up late learning

One of the main questions in human semisupervised category learning involves the impact of the no-feedback trials. Note that the effect of the no-feedback trials on learning can be threefold: they can speed up the learning process, slow it down or have no impact at all. Studies approach the investigation to the impact of the no-feedback trials in different ways.

In a few studies the impact of the no-feedback trials is inferred from a comparison between performance of semisupervised participants and supervised participants. Patterson and Kurtz (2018) showed that semisupervised learning performance (in terms of accuracy) is similar or even better than supervised learning performance. Similarly, LaTourrette and Waxman (2019) showed that even two-year-old infants learn semisupervisedly as well as supervisedly, even when only two feedback trials were presented in the semisupervised condition. Both studies

argued that these results can only be obtained when the no-feedback trials benefit the learning process.

In other studies the impact of the no-feedback trials is investigated by comparing performance in the semisupervised condition and in the unsupervised condition. These studies argue that superior performance in the former cannot be reached by the small amount of feedback trials alone. Indeed, Gibson et al. (2010) showed that semisupervised participants learned a manifold category structure successfully, while receiving only four labeled stimuli. In the study of LaTourrette and Waxman (2019) only two feedback trials were sufficient to outperform unsupervised learning.

Alternatively, several studies investigated the role of the no-feedback trials based on the preferred decision bound of the participant, which is supposed to reflect the strategy that the participant used (Kalish, et al., 2011; Kalish, et al., 2015; Lake et al., 2011; McDonnell et al., 2012; Rogers et al., 2010; Zhu, et al., 2010; Zhu, et al., 2007). In all these studies the no-feedback trials contained stimuli selected from a different distribution than the stimuli in the feedback trials. The idea is that when semisupervised participants take all trials into account (i.e., the feedback *and* no-feedback trials), the decision bound to classify new items will be shifted in comparison to a decision bound based only on the feedback trials. Hence, the effect of the no-feedback trials is investigated in the preferred decision bound. When comparing the decision bound based on all trials in the semisupervised condition to the decision bound based on the feedback trials only in this condition, the results indeed showed this shift in most studies, suggesting that the no-feedback trials did have an effect on learning (Kalish et al., 2011; Kalish et al., 2015; Lake &

McClelland, 2011; Rogers et al., 2010; Zhu et al., 2010; Zhu et al., 2007). Only the study of McDonnell et al. (2012) observed no effect of the no-feedback trials using this strategy. To understand why the no-feedback trials are (mostly) taken into account during category learning, Gibson et al. (2015) used several models to fit and predict the obtained human data. The results revealed that the best fitting model was a semisupervised learning model, implying that the shifted decision bound is the result of the sensitivity of humans to all observations (i.e., feedback and no-feedback) during their search to find an optimal classification. Note that in our studies the stimuli in the feedback and the no-feedback trials originated from the same distribution. We argued that this is more likely in daily life category learning. Consequently, in our studies it is impossible to observe effects of no-feedback trials on the decision bound.

Finally, the effect of the no-feedback trials during semisupervised learning can also be investigated by an irrelevant filler condition. In this condition the number of feedback trials is the same as in the semisupervised condition. The difference is that the no-feedback trials are replaced by irrelevant fillers that are completely unrelated to the category that has to be learnt. The duration of the two conditions is similar. This approach was first introduced in Chapter 2. Comparing the 50% filler condition to the 50% semisupervised condition, we observed no effect of the no-feedback trials since the mean accuracy and the obtained decision bounds were similar for both conditions. Similarly, in Chapter 4, the effect of the no-feedback trials in the semisupervised conditions was investigated in matching filler conditions. Again no effect of the no-feedback trials was found for the 50%, the 25% and the 10% semisupervised conditions. In the study of Rogers et al. (2010) the effect of the no-feedback trials in the semisupervised condition was also

investigated by comparing to an irrelevant filler condition. The effect of the no-feedback trials depended on the given instructions. When the response of the participants was self-paced, the no-feedback trials did not have an impact. Contrarily, when the participants were forced to respond as quickly as possible, semisupervised participants obtained a much higher accuracy. In this case the no-feedback trials improve the learning process.

Critically, all the aforementioned studies focused on the role of the no-feedback trials early in learning. In Chapter 3, we investigated the impact of the no-feedback trials late in learning by comparing the response times of the semisupervised participants to the (almost) supervised participants. Although the almost supervised participants received an equal total amount of trials, semisupervised participants responded significantly faster, implying an enhanced development of automaticity. We argued that this observation is due to the positive impact of the no-feedback trials.

In summary, the semisupervised literature to date is contradictory with regards to the impact of no-feedback trials on the early learning process. Some studies reported a beneficial impact, implying that the no-feedback trials aid learning. Other studies (including the current Chapters 2 and 4) observe no impact at all. Since several approaches are used to investigate the effect in combination with various experimental designs, it is clear that more research will be needed to obtain a better understanding of the impact of the no-feedback trials. When semisupervised learning is introduced late in the learning process, the no-feedback trials improved the development of automaticity (Chapter

3): in comparison to (almost) supervised participants, semisupervised participants responded significantly faster.

Finding 3: individual differences

In all the empirical Chapters of this dissertation differences between participants within the same condition were observed. This was a striking observation since not only the age range was rather limited (18 until 37 years), but also all participants were psychology students and arguably rather homogeneous in cognitive abilities.

As in most cognitive psychology research, category learning research tries to identify the *general* underlying mechanisms of human behavior. However, people vary in how they behave. These individual differences in performance can occur for many different reasons, ranging from uncontrolled variables in the lab, to differences in motivation or attention, to variations in past experience or to relatively stable individual differences such as intelligence or personality (Shen & Palmeri, 2016). Although more and more studies mention individual learning performances (for example, Ashby, et al., 1999 and all our empirical Chapters), the majority of the category learning research focuses on group averages, treating variability between participants as noise. Nevertheless, averaging results is not without risk. Averaging can produce results that are qualitatively different from the behavior of any single individual within that aggregate. Estes (1956) illustrated this risk in case participants show all-or-none learning but at different time points in the experiment. The individual learning curve follows a step function that rises from chance to perfect performance in an instant. When participants reach this 'aha experience' at different time points, which is likely, averaging all these individual curves will lead to a gradual,

continuously rising curve, which is not an accurate reflection of the individual learning curves. Thus, analyzing the results of an individual participant might sometimes provide more accurate information regarding the category learning process. However, since the number of observations per participant is finite, the cost of individual analyses is significantly less statistical power. To avoid the disadvantages of both approaches, Shen and Palmeri (2016) suggest to provide averaging over group as well as analyses at the individual level, as we did in our studies.

Only relatively recently, researchers in cognition have begun to examine individual differences systematically and tried to link these differences to performance. In category learning, Lewandowsky, Yang, Newell and Kalish (2012) related individual *working memory capacity*, which is the ability to hold and manipulate information in memory for brief periods of time (Oberauer, 2019), to individual learning performances on rule-based and information-integration category structures. They found that category learning performance on both category structures was positively associated with working memory capacity, implying that the higher the working memory capacity of a participant, the better the category performance was. Similarly, Craig and Lewandowsky (2012) reported a positive association between a participant's working memory capacity and his/her speed of category learning. These findings can possibly explain the results obtained in Chapters 2, 3 and 4: the finding in Chapter 3 that not all participants reached the 90% expertise criterion at the end of day 2 can be due to differences in capacity in working memory between participants. The same explanation can hold for Chapter 4 in which we observed successful learning even in the 10% semisupervised condition by half of the participants. Note however, that differences in working memory as a

possible cause for individual differences in category learning performance cannot be validated in our studies, since we did not collect individual working memory data.

Additionally, Craig et al. (2012) and Little and McDaniel (2015) showed that individuals' *strategies* can differ in a single category-learning task. This was also the case in Chapters 2 and 3. Here, we calculated for each participant four different strategies based on their responses of the last two blocks. The strategy resulting in the lowest BIC-score was assumed to be the strategy that the participant preferred to classify. In Chapters 2 and 3 we observed in all conditions, except the 50% filler condition of Experiment 2, that different participants preferred different strategies.

Finally, to our knowledge, there are no studies investigating the effect of individual differences in *motivation or attention* on category learning performance. Nevertheless, this can obviously have an impact on performance. Several participants of the almost unsupervised condition of Chapter 3 admitted to be frustrated by the lack of feedback whereas other participants in this condition could keep up their motivation to perform as well as possible. This difference in frustration tolerance might have affected their performance. However, our current data do not allow to objectively ascertain this. Remark that differences in motivation might also provide a feasible explanation for the better performance at the end of day 1 in the 25% condition of Experiment 2 in Chapter 2. In the 25% condition of Experiment 1 in Chapter 2, participants only came to the lab on one day to perform a category learning task. In Experiment 2, they came to the lab on two consecutive days. It could be that, since the participants in Experiment 2 knew in advance that they had to return the

next day to perform the same task again, this knowledge motivated them during the first day to perform optimally. This in turn could perhaps have led to their increased performance compared to the participants of Experiment 1, who performed the task on a single day. In any case, albeit speculative, the impact of individual motivation on category learning deserves further attention in future research.

In summary, in all empirical Chapters of this dissertation individual differences in performance were observed. Therefore it is argued that future research should provide aggregated data (as most studies do) and individual data. We argued that observed individual differences could be due to differences in working memory capacity, preferred strategy or motivation and attention. It is important that in future research these variables are taken into account and linked to learning performance.

Link to semisupervised category learning models

Gibson, Rogers & Zhu (2013) adapted existing supervised category learning models to incorporate semisupervised learning. We discuss two models here, namely the semisupervised exemplar model and the semisupervised prototype model.

The *semisupervised exemplar model* is based on the Generalized Context Model (GCM) proposed by Nosofsky (1986, 2011). The GCM stipulates that all previously viewed examples (or stimuli) are stored along with their category label. When a new stimulus has to be

categorized, these stored examples are used to estimate the most likely category label. Next, this estimation is compared to the given feedback: when the estimation was correct, the strengths of the parameters of the model are boosted. When the estimation is incorrect, the values of the parameters are adjusted. For a stimulus in a no-feedback trial, in which the category label remains uncertain, this mechanism cannot be used. In this case the likelihood of the stimulus belonging to every category is calculated. Based on this likelihood the 'weight' of the stimulus is divided over the categories. Remark that the semisupervised exemplar model stipulates that, in general, the no-feedback trials facilitate the learning process. This hypothesis is not in line with the results of Chapters 2 and 4: when the no-feedback trials were replaced by irrelevant fillers, no effect of the no-feedback trials was observed. However, the hypothesis on which the semisupervised exemplar model is essentially based, is compatible with the findings of Chapter 3, where the no-feedback trials did improve the learning process, more precisely the development of automaticity.

The *semisupervised prototype model* is based on the prototype model proposed by Minda and Smith (2011). The model assumes that learning results in the formation of a prototype model (for each category). A prototype is a summary of a category and participants only need to retain the distribution parameters of every category. For the feedback trials the parameters of each category are computed by the maximum likelihood estimate. In the no-feedback trials the parameters are estimated by the expectation maximization algorithm (as introduced by Dempster, Laird & Rubin, 1977). This estimation implies that a stimulus unlikely to belong to the category contributes little to its parameter estimates whereas a stimulus very likely to belong to the category acts as a

feedback trial. A stimulus whose category status is uncertain will contribute slightly to the estimate of many category prototypes. Problematic with this approach is that the initial parameters of the algorithm have a major impact on the obtained parameters at the end of the experiment (Wu, 1983).

Applied to the results in this dissertation, the semisupervised prototype model can explain why the no-feedback trials did not have an impact on the learning process in Chapter 2. The explanation derived from the model is that the category status of the stimuli presented in the no-feedback trials were all uncertain. However, this seems unlikely since the stimuli of the no-feedback trials originated from two bivariate normal distributions, implying that the coordinates of most stimuli were close to the mean of the category. Furthermore, the stimuli were chosen in such a way that the mean of the two categories coincides with the parameters of the two prototypes. Taken together, based on the semisupervised prototype model some positive effect of the no-feedback trials should have been expected. In Chapter 3, the number of total trials in the semisupervised condition was identical as in the almost supervised condition. Yet, the semisupervised participants responded significantly faster than the almost supervised participants. This beneficial effect of the no-feedback trials late in learning cannot be explained by the semisupervised prototype model: when the category of the stimulus in a no-feedback trial is clear, this no-feedback trial acts as a feedback trial. Consequently, the semisupervised prototype model predicts that both conditions should be equally fast. Finally, in Chapter 4, no effect of the no-feedback trials was observed using the prototype category structure. This finding is in line with the prediction of the semisupervised prototype model: the mean parameters of the prototypes of category ‘A’ and

category ‘not A’ are the same. In that case the category status of the stimulus presented in the no-feedback trial is unclear and the effect of the corresponding estimates are negligible.

Summarizing, the development of human semisupervised category learning models is still very fresh and more (detailed) models are desirable.

Limitations of the studies and future directions

In the next paragraphs, the most important general limitations of our studies are discussed. Based on these limitations, future research ideas are suggested.

A first limitation concerns the stimuli used in this dissertation, which were Gabor patches. In category learning research stimuli can vary by their modality (visual, auditory, tactile,…), by their real-world versus artificial nature, by their binary (absent or present) versus continuous dimensions, by the number of stimulus dimensions and finally by their separable versus integral dimensions (Ashby & Valentin, 2018). In this dissertation the Gabor patches are visual, artificial and have two continuous dimensions that are separable. There is a preference in the category learning research for the use of artificial stimuli (Ashby & Valentin, 2018). Although real-world stimuli have a greater ecological validity, the (varying) life-time history of participants with the stimuli could affect how they learn the categories constructed for the categorization experiment. For example, when birds are used as stimuli, participants can classify based on perception (e.g., similarity), on

relationships (e.g., use for food) or on variables that are difficult to tap (e.g., typical geographical location). Additionally, little is known about the perceptual representation of most real-world stimuli. On the other hand, when artificial stimuli are used, it can be expected that the categorization judgments are closely tied to perceptual features of the stimuli (Ashby & Valentin, 2018; Folstein, Van Pette & Rose, 2008). Next, the common choice in category learning for continuous dimensions has the advantage that an infinite number of unique stimuli can be created. Instead, using binary stimulus dimensions the number of the maximum possible stimuli is 2^r where r is the number of stimulus dimensions (Ashby & Valentin, 2018). The dimensions of the used Gabor patches in this dissertation are separable: participants can perceive the orientation of the Gabor patch separately from the frequency of the gratings. Separable dimensions increase the risk of attending to only one dimension, while ignoring the other. With integral dimensions, however, it is difficult to attend selectively to a single dimension. A common example of integral dimensions are the saturation and the brightness of a color: when the brightness of a color changes, participants have the impression that the saturation also varies. Nosofsky and Palmeri (1996) showed that, although the underlying category structures were the same, different supervised learning results are obtained using separable versus integral dimension stimuli. Indeed, in the study of Gibson et al. (2013) the use of stimuli with separable dimensions in Chapter 2 was criticized. The authors stipulated that if participants only attend to one dimension, they cannot be influenced by the distribution of the unlabeled items along the unattended stimulus dimension. This might explain why no beneficial effects of the no-feedback trials were found in Chapter 2. The authors predict that participants will be more strongly influenced by the full distribution of no-feedback trials when stimulus dimensions are integral

(Gibson, et al., 2013). This is a feasible hypothesis that should be addressed in future research. Summarizing, it will be interesting to investigate how the obtained results in this dissertation will hold using other types of stimuli such as higher number of dimensions, integral stimulus dimensions or real world stimuli.

A second general limitation is that in every study the number of underlying categories was limited to maximum two. Until now, the maximum number of categories used in human semisupervised category learning is three. Hence an interesting topic in future human semisupervised category learning can be how people learn multiple categories semisupervisedly.

A third general limitation is that the number of stimuli in each category was the same. This is not representative of real-world categories. For example, the category of dogs includes more different species than the category of cats. Consequently, to mimic natural learning more closely, studying the effect of unbalanced categories on semisupervised category learning would be an interesting path.

A fourth limitation is the experimental environment. In our studies, the stimuli were presented on the computer screen. It was clear for participants which stimuli had to be learned; no other (distracting) information was visible on the screen. Contrarily, in real life, the to be learned stimuli are embedded in a rich environment. It may happen that the teacher labels an object but has another object in mind than the child thinks he or she does. Indeed, children have difficulties in identifying the intended referent of the caregiver (e.g., Cartmill et al., 2013; Gillette, Gleitman, Gleitman & Lederer, 1999). To mimic natural category

learning more realistically, the effect of misunderstanding on (semisupervised) category learning should be investigated.

A fifth limitation involves the small sample sizes. It was not easy to obtain a sufficient amount of participants, especially in Chapter 3 where not only a selection of the participants had to be made based on their performance on day 2, but the participants also had to be present in the lab on five consecutive days. Small sample sizes decrease the statistical power and increase the risk of making a type II error. This error implies failing to appreciate a significant effect when there is one to be detected. This limitation can be solved by increasing the number of participants in each condition.

The sixth limitation concerns aggregation of the results in a condition. Aggregations are common in category learning research and are necessary necessarily for the most common used statistical techniques such as *t*-tests, repeated measures,… Nevertheless, these aggregations are in contrast to the recurrent finding in this dissertation of individual differences. A benefit of this dissertation is the reporting of individual performances in all studies and the individual determination of the strategy in Chapter 2 and 3. Consequently, it is essential that future research reports individual performance and uses statistical techniques which take individual performances more into account (e.g., mixed effects models).

The seventh limitation concerns the learning mode. In all our studies, semisupervised learning was investigated in an intentional learning task. This implies that participants were instructed to learn the category structure by labeling every stimulus. Learning can also be investigated

in incidental learning tasks. In that case, learning the category is not the primary objective when one interacts with a stimulus; instead, it is a consequence of completing another task (Love, 2002). It might be interesting to investigate semisupervised learning by combining intentional with incidental learning tasks.

The final limitation concerns the interpretation of the test in Chapters 3 and 4. It is contradictory to investigate semisupervised learning effects (more precisely, the effect of the no-feedback trials) in the training phase and to assume that the test phase, which consists solely of no-feedback trials, has no further impact on the learning process. Indeed, Kalish et al. (2011) observed that participants learn during transfer. Zaki and Nosofsky (2004) called this phenomenon 'learning during transfer' and it cannot be ruled out that this effect might have influenced our data. Nevertheless, it is difficult to avoid this effect when enough data of the participant is needed to make a reliable estimate of the used strategy. As in most human category learning studies, we assume that the strategy of the participant had stabilized after a sufficient amount of trials. In this dissertation the test was given after 640 trials in an easy category structure (Chapter 4) and after at least 1120 trials in a complex category structure (Chapter 3), which is higher than in other studies. A solution can be to use trial-by-trial methods to analyze the data.

In the following paragraphs future ideas for research are discussed. Some ideas are inspired by findings of this dissertation, some by real life category learning and others by other human category learning studies.

In all studies we observed individual differences in learning. It might be interesting in future research to individualize the number of training trials instead of presenting a fixed number of trials to all participants of

a condition. This can be accomplished by setting a predetermined criterion such as 90% accuracy on the last 50 trials. The advantage of this approach is that participants are in a more similar phase of learning when they receive the test afterwards. Further, in this general discussion we suggested that differences in motivation might explain the observed individual differences. Therefore, it might be interesting to manipulate motivation during the experiment. This can for example be done by providing extrinsic rewards if a participant achieves a certain goal.

To mimic real life learning more closely, it will be interesting to manipulate the percentage of feedback trials between blocks during training. In the current studies all blocks consisted of the same percentage of feedback trials. Contrary, in real life participants receive a lot of feedback at the start and this percentage of feedback trials diminishes during the training. Therefore studying the effect of diminishing feedback on learning performance might be informative. Additionally, when children are uncertain about a category label of a stimulus, they ask the teacher what the label is. In our studies (and in other semisupervised studies) this option was not available. Hence, it would be interesting (and perhaps more motivating) if participants could choose when to receive feedback. Additionally, when in education new categories have to be learned, the items are instructed in a fixed manner: first the typical items are learned and only when these items are acquired, the atypical items are introduced. Translated to the human category learning research, this approach implies that first the prototypes must be learned and only if performance is sufficient, the outliers of the category are presented. Perhaps this approach stimulates semisupervised category learning performance as well. This could be explored in future research.

Finally, based on current human category learning findings, it might be interesting to investigate the effect of order on semisupervised learning performance. In unsupervised learning, Zeithamova and Maddox (2009) found that during the training phase, incidental learners benefit from a fixed order. This implies that first all category A stimuli are shown and then all category B stimuli are shown. It would be interesting to investigate whether this also holds for semisupervised learning.

In sum, due to the few existing semisupervised category learning studies, many interesting research questions can be formulated to improve our understanding of the optimal conditions for learning.

Implications of the obtained results for daily life

Over the course of the studies the "how might you apply this work to a real world setting?" question has come up repeatedly. Remark that this is a common question in basic research. The research presented in this dissertation has several implications. Firstly, it stimulates a better understanding of real-life human learning behavior by mimicking more realistic natural learning conditions. The research in this dissertation and in other human semisupervised category learning studies convincingly shows that semisupervised learning can be as successful as supervised learning. This result implies that many (unsupervised) encounters do not hamper learning. On the contrary, these no-feedback trials are enhancing category knowledge under certain conditions. A second implication is related to improving category learning. Chapter 3 for example revealed that late in learning the less expensive semisupervised learning obtained

better results than the more costly supervised learning. Subsequently, the study of Kalish et al. (2011) in which incorrect beliefs of categories are linked to semisupervised learning, is a promising start to gain a better insight into which stimuli need to be learned supervisedly and which stimuli can be learned equally well semi- or unsupervisedly. Additionally, the results of this research can be used in machine learning to develop machines that learn more as humans do. In machine learning several studies revealed not only that semisupervised learning is successful, but also that the information of the no-feedback trials are used to extend the knowledge gained in the feedback trials. A major additional benefit is the lower cost of semisupervised learning in comparison to supervised learning (Chapelle, Zien & Scholkopf, 2006; Zhu & Goldberg, 2009). Remark that supervised learning requires to label all stimuli. This is not only time-consuming, it even often requires an expert in the field (Balcan & Blum, 2010). The clearest application of semisupervised category learning in a real world setting can be found in education. Given that a key goal of education is to provide learners with a set of learning experiences that enable their continued learning (Bransford & Schwartz, 1999), research that elucidates how to integrate feedback and no-feedback trials to maximize learning is directly informative for this goal.

References

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

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

Ashby, F. G., & Valentin V. V. (2018). The categorization experiment: Experimental design and data analysis. In E. J. Wagenmakers & J. T. Wixted (Eds.), *Stevens handbook of experimental psychology and cognitive neuroscience, Fourth Edition, Volume Five: Methodology*. New York: Wiley.

Balcan, M.F., & Blum, A. (2010). A discriminative model for semi-supervised learning. *Journal of the ACM, 57*(3), 19:1-19:46.

Bransford, J.D., & Schwarz, D.L. (1999). Chapter 3 : Rethinking transfer : A simple proposal with multiple implications. *Review of Research in Education, 24*(1), 61-100.

Cartmill, E. A., Armstrong, B. F., Gleitman, L. R., Goldin-Meadow, S., Medina, T. N., & Trueswell, J. C. (2013). Quality of early parent input predicts child vocabulary 3 years later. *Proceedings of the National Academy of Sciences of the USA, 110*(28), 11278–11283.

Casale, M.B., & Ashby, F.G. (2008). A role for the perceptual representation memory system in category learning. *Perception & Psychophysics*, *70* (6), 983–999.

Chapelle, O., Scholkopf, B., & Zien, A. (Eds) (2006). Semi-supervised learning. Cambridge, MA: MIT Press.

Craig, S., & Lewandowsky, S. (2012). Whichever way you choose to categorize, working memory helps you learn. *The Quarterly Journal of Experimental Psychology, 65*(3), 439-464.

Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological), 39*(1), 1-38.

Gibson, B. R., Zhu, X., Rogers, T. T., Kalish, C.W., & Harrison, J. (2010). Humans learn using manifolds,reluctantly. In J. Lafferty (Ed.), *Advances in neural information processing systems 23, vol. 24*, pp. 730–738. Red Hook, NY: Curran Associates, Inc.

Gibson, B.R., Rogers, T.T., Kalish, C.W., & Zhu, X (2015). What causes categoryshifting in human semi-supervised learning*? In Proceedings of the 37th Annual Conference of the Cognitive Science Society (CogSci)*.

Gibson, B.R., Rogers, T.T. & Zhu, X. (2013). Human semi-supervised learning. *Topics in Cognitive Science, 5*, 132-172.

Gillette, J., Gleitman, H., Gleitman, L., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition, 73*(2), 135–176.

Estes, W.K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin, 53*(2), 134-140.

Folstein, J. R., Van Petten, C., & Rose, S. (2008). Novelty and conflict in the categorization of complex stimuli: N2 and P3 correlates. *Psychophysiology,45*, 467–479.

Kalish, C. W., Rogers, T. T., Lang, J., & Zhu, X. (2011). Can semi-supervised learning explain incorrect beliefs about categories?. *Cognition, 120* (1), 106-118.

Kalish, C.W., Zhu, X., & Rogers, T.T. (2015). Drift in children's categories: when experienced distributions conflict with prior learning. *Developmental Science. 18*(6), 940-956.

Lake, B. M., & McClelland, J. L. (2011). Estimating the strength of unlabeled information during semi-supervised learning. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, 1400–1405.

LaTourrette, A., & Waxman, S. R. (2019). A little labeling goes a long way: Semi-supervised learning in infancy. *Developmental Science, 22*(1), e12736.

Lewandowsky, S., Yang, L.X., Newell, B.R., & Kalish, M.L. (2012). Working memory does not dissociate between different perceptual categorization tasks. *Journal of Experimental Psychology: Learning, Memory and Cognition, 38*(4), 881-904.

Little, J.L., & McDaniel, M.A. (2015). Individual differences in category learning: memorization versus rule abstraction. *Memory & Cognition, 43*(2), 283-297.

Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin & Review, 9*, 829–835.

McDonnell, J.V., Jew, C.J., and Gureckis, T.M. (2012). "Sparse category labels obstruct generalization of category membership." In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society*.

Minda, J.P., & Smith, J.D. (2011). Prototype models of categorization: Basic formulation, predictions and limitations. In E.M. Pothos & A.J. Wills (Eds.), Formal approaches in categorization (pp. 40-64). Cambridge, UK: Cambridge University Press.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General, 115*(1), 39.

Nosofsky, R.M. (2011). The generalized context model: An exemplar model of classification. In E.M. Pothos & A.J. Wills (Eds.), Formal approaches in categorization (pp. 18-39). Cambridge, UK: Cambridge University Press.

Nosofsky RM, & Palmeri TJ. (1996). Learning to classify integral-dimension stimuli. *Psychonomic Bulletin and. Review,. 3*, 222–26.

Oberauer, K. (2019). Is rehearsal an effective maintenance strategy for working memory? *Trends in Cognitive Science, 23*(9), 798-809.

Patterson, J.D., & Kurtz, K.J. (2018). Semi-supervised learning: A role for similarity in generalization-based learning of relational categories. *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, p2211-2217.

Rogers, T.T., Kalish, C., Gibson, B.R., Harrison, J., & Zhu, X. (2010). Semi-supervised learning is observed in a speeded but not an unspeeded 2D categorization task. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society,* p. 2320–2325.

Shen, J., & Palmeri, T. J. (2016). Modelling individual difference in visual categorization. *Visual Cognition, 24*(3), 260–283.

Vong, W. K., Perfors, A., & Navarro, D. J. (2014). The Relevance of Labels in Semi-Supervised Learning Depends on Category Structure. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society,* 1718–1723.

Wu, C. F. J.(1983). On the convergence properties of the EM algorithm. *The Annals of Statistics, 11*(1):95–103.

Zaki, S.R., & Nosofsky, R.M. (2004). False prototype enhancement effects in dot pattern categorization. *Memory and Cognition, 32*, 390-398.

Zeithamova, D., & Maddox, W.T. (2009). Learning mode and exemplar sequencing in unsupervised category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 731-741.

Zhu, X., Gibson, B.R., Jun, K.-S., Rogers, T.T., Harrison, J. & Kalish, C. (2010). Cognitive models of test-item effects in human category learning. In *Proceedings of the 27th International Conference on Machine Learning,* 1247–1254.

Zhu, X., & Goldberg, A.B. (2009). Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning, 3*(1), 1-130, Morgan & Claypool.

Zhu, X., Rogers, T., Qian, R., & Kalish, C. (2007). Humans Perform Semi-supervised Classification Too. In *Proceedings of the 22Nd National Conference on Artificial Intelligence (AAAI'07),* 864–869.