A Computational Model of the Emergence of Analogical Reasoning in Category Learning

by

Cesare Bianchi

A Thesis submitted to the
National University of Ireland, Dublin
for the degree of Ph. D.
in the
College of Engineering, Physical and Mathematical Sciences

August 2010

The School of Computer Science and Informatics
Dr. J. Carthy (Head of School)
Under the supervision of
Dr. Fintan Costello

Contents

List of Tables

List of Figures

Abstract

Analogical reasoning is usually seen as involving the transfer of knowledge from a well-known source domain to a less-known target domain. The most successful theory explaining this transfer of knowledge is the structure-mapping theory. According to this theory, the relational structures of the two domains are aligned, and information is transferred from the source to the target domain on the basis of this alignment. This alignment process is computationally expensive and depends on the availability of a well-known source domain. Various theories of category learning based on structure mapping have been developed. These theories assume that learning takes place through the comparison and alignment of members of the same category: this alignment allows the identification of structural commonalities shared by category members.

This thesis investigates the use of similarity and analogy between members of different categories during category learning. A series of experiments show that, during learning, people make use of similarities between members of different categories. These experiments also demonstrate that people use analogical reasoning in early phases of learning, when none of the categories being learned are known. This is a problem for theories of analogy which depend on the availability of a well-known source domain or concept. A new theory is proposed, in which learning takes place via iterative modification of category representations. In this process people learn by initially forming general 'partial-categories' covering the items being learned; these partial-categories are subsequently refined to produce full category learning. In this account two categories will be analogous because they originate from the same partial-category; this account thus gives an emergentist explanation for the origin of analogical reasoning in category learning. This account also provides for efficient use of the computational and memory resources available during reasoning. A computational model which implements this theory has been tested using the same learning tasks given to participants in the experiments. The model accurately reproduced people's results, demonstrating that participants tended to reason as the model would predict.

Declaration

I declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at this, or any other, University or institute of tertiary education.

Cesare Bianchi
August 2010

List of Publications

Bianchi C. (2012). *An Alternative Account on the Origin of Analogical Reasoning*. Cognitive Systems 7-3, November 2012, 261 – 273. First published in 2010 as a book chapter in "Practices of Cognition: Recent Researches in Cognitive Sciences", University of Trento, Italy, ISBN 978-88-8443-350-3

Bianchi C., Costello F. (2009). *An Interactive Test to Study the Relevance of Analogical Reasoning and Concept Separability in Category Learning*. Proceedings of the 2nd International Analogy Conference

Costello F., Bianchi C. (2009). *A Two-Process Account of Analogical Category Learning.* Proceedings of the 2nd International Analogy Conference

Bianchi C., Costello F. (2009). *Un Modello dell'Emergenza ed Uso del Ragionamento Analogico nell'Apprendimento di Categorie*. Sistemi Intelligenti 2009 (3), 405-416

Bianchi C., Costello F. (2008). *Analogical Reasoning helps learning of Similar Unknown Concepts: the use of Analogies between Categories in Category Learning*. Proceedings of the 19th Irish Conference on Artificial Intelligence and Cognitive Science

Bianchi C., Costello F. (2008). *The role of analogical reasoning in category learning.* Proceedings of the 29th International Congress of Psychology

Chapter 1

Introduction

Analogical reasoning is the ability to spot similarities between concepts or domains and to use those similarities to transfer knowledge from one concept to another. A common example of analogical reasoning is the completion of quadruplets like:

Boat : Sea :: Aircraft : ?

Analogy plays a central role in learning, in language, and in everyday reasoning. Some remarkable cases of analogical reasoning can also be found in the history of science (Del Re, 2000; Gentner, 1993; Hoffman, 1980), where it has been used to discover new knowledge in a partially unknown domain being investigated. Theories of analogy in scientific discovery suggest that an already-known 'source' domain similar to the problem domain is first found and then knowledge is transferred from that source domain to the problem or 'target' domain (Gentner, 1983). A typical example of analogy being used in this way is the analogy between structure of the solar system and structure of the atom. This case is also representative of a third common use of analogy: the teaching of a novel concept using another well-known concept as a source of knowledge and basis for understanding.

While the standard view of analogy focuses on the transfer of knowledge from well-known source domains to novel target domains, in real life most concepts and most domains are rarely well known; people, especially children, must often learn domains and concepts for which no already-known source is available. Can analogy play a role in these early stages of learning? Can analogies between partially learned target concepts be used to facilitate learning? How would analogical reasoning operate in this type of early learning situation? In this thesis I aim to address these questions.

One possible advantage of the early use of analogical reasoning is clear: instead of learning several concepts separately, a learning process that uses analogies between partially learned concepts can minimize the cost of learning, both in terms of memory and time. If domains are learned separately with no identification of between-domain analogies, then any common elements or structures shared across domains will be represented one time for each domain. If analogies between domains are identified during learning, however, those common elements need be represented once only, decreasing memory load. Similarly, if analogies between partially learned domains are identified during learning, then understanding gained about one domain can be applied to the other domain, decreasing learning time. My first hypothesis, therefore, is that

1. analogy will be used to facilitate learning between simultaneously-learned categories.

There are at least two possible ways in which analogy can be used when simultaneously learning categories with similar structures. One is by using mutual alignment (Kurtz & Loewenstein, 2007; Gentner, Loewenstein & Thompson, 2003; Kurtz, Miao, & Gentner, 2001) and thus structure mapping to transfer knowledge between the similar categories been learned. The other way does not require structure mapping and consists of two phases where partial categories are first formed and then refined to create the final categories. Therefore there are two alternative hypotheses:

2a. Categories are first learned and then compared and mutually aligned using structure mapping; or

2b. Learning consists of two phases where one or more partial categories are first formed and then refined, so from each partial category stem some final categories.

If the hypothesis 2a is found to be correct, the existing theories and models of analogical reasoning in category learning are already sufficient to explain the results, and need only to be extended to account for similarities between categories. Progressive alignment (Kuehne, Forbus, & Gentner, 2000) will be used not only to compare the exemplars of the first category learned, but after it is learned, also to compare the exemplars from the other category. If this is the case, then one category is first learned and then continuous comparison takes place and mutual alignment and structure mapping are used to learn the other similar categories. Therefore a direct comparison of items of similar categories should help learning, and the experiments will test this prediction.

Hypothesis 2b on the opposite expects that in order to unify the learning of multiple categories, the learning process occurs in two phases, with a first phase of partial learning followed by a phase of refinement. In the first phase some general "partial categories" are formed, representing sets of similar categories; in the second phase these partial categories are modified or refined in order to arrive at the final categories or concepts being learned.

In this thesis "partial categories" refer to drafts of categorization criteria, which can confuse between categories and which consequently will sometimes classify items in the wrong category. According to hypothesis 2b, partial categories can be further modified and refined to become "final categories", i.e. categorization criteria which always classify items in the right category.

This process of partial learning minimizes the memory and time effort while maximizing the amount of information extracted in each stage. Because of this advantage, I expect that people tend to find similar classification criteria even when easily discriminable criteria are available. This contrasts with standard views of categorization, and with machine learning models of learning, which predict that easily-discriminable categories are more easily learned (Doumas & Hummel, 2005; Kuehne, Forbus, & Gentner, 2000; Muggleton, 1991; Nosofsky, Palmeri, & McKinley, 1994).

This two-step process gives an emergent explanation for the origin of analogical reasoning in category learning: in this explanation, categories are seen as similar, and so analogous to each other, because they stem from a common partial category, and were learned together as part of that category.

A series of specific predictions come from these hypotheses. The following predictions will refer to an hypothetical experiment with four categories: two defined by numerical relations between their compositing elements (e.g. in one category "same number of squares and triangles", in the other "twice as many"), the other two by some other non-relational criteria (e.g. presence of some distinctive element: a circle or a star).

The predictions are that, when people are learning multiple categories simultaneously

A) According to hypothesis 1, there is a connection between the learning of structurally similar categories. According to the example, time between learning of the two "numerical" categories will be less than time between learning of one "numerical" category and another non-relational category.

B) According to hypothesis 2a, the simultaneous presentation of items from similar categories should facilitate learning. So for example if we present simultaneously exemplars from the "numerical" categories to a group of participants, they will take less to solve the test than another group presented with one exemplar at a time.

C) According to hypothesis 2b, before learning is complete, any errors in categorisation are not random but they are more frequent across similar categories. So for example it is more probable to answer that an exemplar from one "numerical" category pertains to the other "numerical" category, than it pertains to one of the other two remaining categories.

D) According to hypothesis 1, even if given alternative solutions, people find solutions which have similar structures rather than different structures, when learning multiple concepts.

E) According to hypothesis 1, relational similarity helps learning, in respect to categories defined simply by features. In the example, the "numerical" categories would be learned first, although they may be more complex than the other non-relational categories.

To test these hypotheses and predictions we need a task in which novel similar concepts are learned simultaneously. One candidate is a category-learning task in which some of the categories are structurally similar. Current theories of analogical category learning assume that learning takes place through the comparison and alignment of members of the same category: this alignment allows the identification of structural commonalities shared by category members (Kuehne et al., 2000). The role of analogies between categories has not been investigated in the literature to date.

In order to have categories with analogical similarities between them (similarities that can be exploited using analogical reasoning), those categories must be defined not just by features, but by relations. This is the case with categories such as "pilot", "captain", and "driver"; these categories are defined by the relation of their *instances* with vehicles, the relations with the other people, the environment, and so on. These relations are similar, thus one category can be easily mapped onto another using analogical reasoning. In testing the role of analogical reasoning in the simultaneous learning of multiple novel categories, I will design a number of novel artificial categories with this type of relational structure. A series of experiments will examine how people learn these categories.

1.1. Organization of the Thesis

A review of the current literature is presented in Chapter 2. This highlights the lack of a theory and a model for the early use of analogy during learning of novel similar concepts.

Experiment 1 is described in Chapter 3. This experiment investigates whether or not the simultaneous presentation of instances of categories with similar structures facilitates learning, as the structure mapping theory predicts. The experiment also investigates if learning of similar categories is related, and if partial categories are formed in the early stages of learning. The aim is to determine if the learning of one category is independent from the learning of another similar category or if instead the learning of the two are connected. The results of this experiment will help us to determine if analogical reasoning starts to operate even before complete learning of any category has taken place.

Experiment 2 and experiment 3 are presented in Chapter 4 and Chapter 5 respectively. These experiments will confirm the results from the first experiment. In addition, they will also test the prediction that even if given alternative solutions, people find solutions which have similar structures rather than different structures.

It will be shown that the current models of category learning and analogy could be adapted to explain the results of the experiments, but the existence of partial learning opens the way to a new theory and a new model, based on the formation of partial hypotheses and their subsequent refinement. A new theory of analogical reasoning in category learning will be proposed and instantiated as a computational model in Chapter 6. This new theory is based on the modification of concepts and their refinement through subsequent stages. The computational parsimony of this new approach is also shown, and the plausibility of the theory is confirmed by a computational model based on this theory. It will be shown that this computational model is able to reproduce the results from the participants in the experiments, and to predict their learning patterns. This model provides the same answers to the same category instances as given by participants in the experiments.

Chapter 2

Background research

2.1. Introduction

This thesis considers the simultaneous learning of similar concepts in the absence of any background knowledge. The first hypothesis of this thesis is that:

1. Analogy can be established between simultaneously-learned categories with similar structures, to aid the learning of both categories.

As said in Chapter 1, there are at least two possible ways in which analogy can be used when simultaneously learning categories with similar structures. From those derive two alternative hypotheses:

2a. Categories are first learned and then compared and mutually aligned using structure mapping; or

2b. Learning consists of two phases where one or more partial categories are first formed and then refined, so from each partial category stem some final categories.

As said previously, in this thesis "partial categories" refer to drafts of categorization criteria, which can confuse between categories and which consequently will sometimes classify items in the wrong category. According to hypothesis 2b, partial categories can be further modified and refined to become "final categories", i.e. categorization criteria which always classify items in the right category.

This thesis will present three experiments which will test these hypotheses. In order to study these hypotheses and eventually build a computational model, there are various academic fields from which draw useful knowledge.

The first hypothesis requires that analogy can be used to learn structurally similar categories which are both unknown and which are learned simultaneously. Studies about analogical reasoning, relational categories and simultaneous learning of categories will be therefore used. To test the first hypothesis the experiments will need to show that learning of similar categories is related. In other terms, that time between learning of similar categories is less than time between learning of dissimilar ones. To further test the first hypothesis the experiments can also show that people tend to find similar criteria even if they are given the opportunity to find alternative criteria.

Hypothesis 2a expects that a continuous comparison takes place and that mutual alignment and structure mapping are used to progressively align (Kuehne, Forbus, & Gentner, 2000) exemplars[1]. Therefore a direct comparison of items of similar categories should help learning, and the experiments will test this prediction. This chapter will thus include studies about structure mapping, mutual alignment and direct comparison of similar items. Studies about relational categories learning models will be also presented because if the hypothesis is correct these models could be extended to be able to simultaneously learn similar categories using mutual alignment.

[1] It should be noted that the current theories about progressive alignment only predict its use between exemplars of the same category, although these theories can be expanded.

Hypothesis 2b expects that partial classification criteria can be created and then refined. The experiments will therefore test if a first phase of "partial learning" (i.e. confusion between similar categories) will be followed by a second phase of "final learning". Studies about the modification of concepts, the "theory theory" and some other machine learning models (such as the Inductive logic programming - Lavrac & Dzeroski, 1994; Muggleton, 1991; Muggleton & Raedt, 1994) will be consequently presented in this chapter. Given that the partial learning can occur without the person being conscious of that, studies about implicit learning will be also presented.

Since this work is mainly about analogical reasoning, the existing literature will be selected to include only studies pertaining to analogy, structural relations, comparison, similarities, transfer of knowledge, predicate-based machine learning models, etc. Some other studies in related fields (e.g. the "prototype vs. examples vs. rules" debate in category learning, or the "attribute-value learners") will be only briefly presented for sake of completeness, although not pertaining to the core of this work.

2.2. Analogical Reasoning in Category Learning

2.2.1. The Standard View of Analogical Reasoning

Gentner's (1983) paper lays the foundations of studies on analogy, and in fact it has been cited more than 2000 times according to Google scholar. This paper establishes the classical theory of analogy known as “structure-mapping theory”. Based on this theory, analogy is clearly distinguished from other types of comparisons, such as literal similarity and abstraction. This theory is based on some important distinctions between types of predicates. The first distinction made is that between object attributes and relationships; the second one is between first-order predicates (taking objects as arguments) and second- and higher-order predicates (taking propositions as arguments). Starting from that distinction, the structure-mapping theory features two fundamental rules:

1. Relations between objects, rather than attributes of objects, are mapped from the base domain to the target domain, which means that analogy doesn't affect the specific content of the domains.

2. In analogy, not every single predicate is mapped from a domain onto another. The particular relations chosen are determined by "systematicity" (see Gentner, 1983), with higher-order relations that connect the lower-order relations into an interconnected structure that can be described as a system of relations.

Starting from the classical studies by Gentner (1981, 1983) quite an amount of literature has been produced based on the idea of analogical reasoning as a transfer of knowledge from a well-known domain to another less well-known one (e.g. the solar system and the atom). According to this idea, analogy is a higher-order type of reasoning, as can be found in some cases of scientific discovery. The process allows the common structure between two domains to be exploited. One structure is methodically mapped onto the other, in order to find the items in the target which correspond to the roles inferred from the base domain. It seems therefore impossible, according to this theory, that analogy can occur between two partially understood domains, which is instead what is expected according to my first hypothesis.

2.2.2. SME

Gentner’s structure-mapping theory has been implemented in the “structure-mapping engine” (Falkenhainer, Forbus, & Gentner, 1986). The structure-mapping engine (SME) uses an algorithm specified by a set of “constructor rules” and “evidence rules”. It begins with a large set of random matches that are gradually screened into one or few structurally consistent mappings between a given base and a target. It also provides a structural evaluation for each mapping, according to the constraints of systematicity and structural consistency, which define and distinguish analogy from other kinds of inference.

For some cases (such as scientific discovery) this is a useful and maybe common way of making analogies. But it is difficult to see how it can work in the common situation in which there isn't a complete source of knowledge from which carry out the mapping process. Anyway, Thibaut, French & Vezneva (2010) mentioned the likelihood that analogy can happen between concepts, also if the knowledge of the structure of both concepts increases during learning. Also the theorization of mutual alignment (see below) seems to partially overcome this problem.

2.2.3. Simultaneous learning

In recent years it has been suggested (Kurtz & Loewenstein, 2007; Gentner, Loewenstein & Thompson, 2003; Kurtz, Miao, & Gentner, 2001) that, in problem solving, the simultaneous presentation of two partially understood problems can produce a better understanding of both. These studies importantly highlight the role of comparison of similar problems in order to encode a representation of a higher-order abstract problem.

Kurtz, Miao and Gentner (2001) propose an alternative to the standard base-to-target mapping paradigm, which they call mutual alignment. But as they say:

> Mutual alignment can enhance understanding, but there must be at least rudimentary prior knowledge to make two-way information transfer possible. Learning by mutual alignment can succeed without analogical retrieval, but it is not bootstrapping from ground zero.
>
> *(Kurtz, Miao, & Gentner, 2001)*

Besides, in order to foster mutual alignment participants must be actively stimulated to do so by joint presentation, listing similarities, joint interpretation and other techniques.

The study carried out by Gentner, Loewenstein and Thompson (2003) shows that "analogical encoding - comparing two instances of a to-be-learned principle - is a powerful means of promoting rapid learning, even for novices". In particular, this study tested whether analogical encoding could improve novices' ability to transfer principles learned from examples to actual negotiation problems. "In accord with the analogical encoding hypotheses, novices learned and transferred better when they were instructed to compare the study cases", which provided evidence that "analogical encoding fosters the extraction of the common relational schema inherent in the cases and that this in turn promotes the ability to transfer the knowledge to new cases". Therefore it can be "effective even early in learning, when learners may lack knowledge of an appropriate base domain".

This idea hasn't been developed to account for the use of analogy in simultaneously learning of similar novel categories. The experiments in this thesis will address this gap. Hypothesis 2a tests if mutual alignment is used during simultaneous learning of similar categories. In order to foster mutual alignment the experiments will use the direct comparison of items (see below for other studies). According to the above studies, the comparison of items from similar categories should help learning by allowing mutual alignment.

2.2.4. Modification of Concepts

In a small study Clement (1981) hypothesized the modification of concepts in problem solving. Some participants in his experiment mentioned that they modified a known problem in order to adapt it to match a novel one. In that case, analogy proceeded through modification instead of structure-mapping.

Another study which uses the idea of modification of concepts is Hahn (2003). This study aims to measure the similarity between known concepts using the quantity of transformations needed to transform one object into another one. Participants were shown pairs of exemplars composed of shapes which could be transformed one into the other, using one ore more geometric transformations (e.g. stretching, moving, rotating, change the fill, etc.). They were asked to rate how similar were the two exemplars. Although this study was only about measuring the similarity of known objects, its core concept could be used to test the use of transformations in learning.

Kokinov (2007) also proposes the idea of modifying concepts using re-representation, in order to form analogies. Yan, Forbus & Gentner (2003) suggest that re-representation could be done through structure-mapping, while other studies (Keane, 1995, 1997; Keane, Ledgeway, & Duff, 1994) propose that the learning of classification criteria is incremental.

If the hypothesis 2b will be found correct, the computational model will need an algorithm for the modification of its formed partial hypotheses. Although this algorithm could be one of the complex and intelligent algorithms briefly reviewed in this chapter, the model should use a more trivial algorithm of random modification. The reason for this choice is that if a model can reproduce the participants' results with a simple modification algorithm, it would only work better with more intelligent modification algorithms, such those presented in this chapter.

2.2.5. Comparison of similar items

Some studies (Gentner & Medina, 1998; Gentner & Namy, 1999) have investigated the role of similarities within a category. They proposed that items of the same category are compared and mapped onto each other in order to extract the common structure. In particular, Gentner & Medina asked children and adults to learn relational concepts. They presented their subjects with exemplars, and also helped them by labelling the exemplars. They hypothesize that learning happens through the juxtaposition and the comparison of the shown exemplars, allowing to abstract the common structure:

> Comparison is very often the critical path in the development of rules, especially early in development when higher-order knowledge is sparse and requires the support of concrete commonalities.
>
> [...] at any age, structural alignment provides the necessary bridge in applying rules and abstract knowledge to ongoing experience. Comparison is fundamental to the development and use of rules in cognition.
>
> *(Gentner & Medina, 1998)*

On the other hand, Gentner and Namy (1999) examined the categorization behaviours of 4-year-old children "when asked to select a match for a target object (e.g., an apple) between a perceptually similar, out-of-kind object (e.g., a balloon) and a perceptually different category match (e.g., a banana)". They concluded that

> […] children who learn a novel word as a label for multiple instances of the category are more likely to select the category match over the perceptual match. Children who learn a label for only one instance are equally likely to select either alternative. This effect is present even when individual target instances are more perceptually similar to the perceptual choice than to the category choice
>
> *(Gentner & Namy, 1999)*

Which means that a structural alignment process (invited by the common linguistic label) yields a deeper, more conceptual encoding.

Many studies followed (e.g. Boroditsky, 2007; Oakes, Kovack-Lesh, Horst, 2009; Kotovsky & Gentner, 1996), showing that the comparison of similar items helps learning. Namy & Gentner (2002) for example showed that

> [...] comparison facilitates categorization only when the targets to be compared share relational commonalities.
>
> *(Namy & Gentner, 2002)*

An amount of studies (e.g. Waxman & Klibanoff, 2000; Gentner, Loewenstein & Hung, 2007; Gentner & Namy, 2006; Gelman, Raman & Gentner, 2009) investigated the role of language in structural alignment and learning, showing that the comparison of similar items helps a deeper understanding by promoting analogical encoding.

2.2.5.1. Pairwise presentation

Some other studies (Kurtz & Gentner, 1998; Kurtz & Boukrina, 2004) show that the pairwise presentation of similar items fosters the comparison and the mutual alignment using structure mapping, and thus facilitates learning. In particular, Kurtz and Boukrina (2004) conclude that

> […] when task constraints emerge that engage the learner to apply the machinery of comparison, superior performance in learning relational categories is achieved. These findings are most naturally understood in terms of learning to construct richer, more sophisticated encodings of category instances. While this is a difficult process, it is made easier by comparison.
>
> *(Kurtz & Boukrina, 2004)*

The experiments in the present work will therefore use direct comparison of items of structurally similar categories to test if mutual alignment done by structure mapping facilitates learning also in the case of structurally similar categories simultaneously learned.

2.2.5.2. Within vs. Between categories similarities

All of the studies presented above investigate the similarities between items of the same category, which can be called “within category” similarities. But also the classification criteria of distinct categories can be structurally similar. This is the case of a more general case of similarity, which can be called "between category" similarity.

The "between category" case of similarity has been usually neglected in the literature, apart for a few cases related to problem solving. As mentioned above, it has been proposed (Kurtz & Loewenstein, 2007; Kurtz et al., 2001) that solving two similar problems can facilitate learning, and it has been recently studied (Hammer, Diesendruck et al, 2009) the case of the simultaneous learning of categories with similar and dissimilar features, but it has never been generalized to the learning of categories with similar structures, which can be thus learned also using analogy.

In order to fill this gap, this thesis will investigate the case of similarities between different categories, all learned simultaneously.

2.3. Rules and Relations in Categories

In order to investigate my hypotheses, categories with similar relational structure will be needed. It is thus interesting to review previous works about relational categories, and how they can be learned.

2.3.1. Relational Categories

Relational categories have been defined as categories "whose membership is determined by a common relational structure rather than by common properties" (Gentner & Kurtz, 2005). "Passenger", "bridge" or "barrier" are examples of relational categories. An amount of literature has been already produced on relational categories. The experiments will use this type of category because it is easier to create relational classification criteria with structural similarities between different categories.

2.3.1.1. The Learning of Relational Categories

There have been various approaches to categorization, based on prototypes, examples or rules. Since I will use relational categories, the approach that is most suitable for my research is the rule-based approach, with rules defined by predicates. A brief review of the various approaches will show why the alternatives are not suitable for relational categories.

The debate on whether categories are defined by prototypes (Rosch, 1978), stored examples (Nosofsky, 1988) or rules (Nosofsky et al., 1994) is still active (Feldman, 2003), but in recent years it has been proposed that distinct neural systems are responsible for the different strategies (Ashby et al., 1998; Ashby & Maddox, 2005; Smith, Patalano, & Jonides, 1998). It is therefore possible that all of these approaches are generally used by people, and the best suited for the specific problem encountered is chosen, as also suggested by Greg Murphy (2002).

For the specific kind of problems studied in this work, which have categories defined by relations, it is clear that only a representation based on predicates can be used. In fact, the prototype approach cannot work with categories defined just by relations, since different members can share no features. No prototype can thus exist for these kind of categories. The example-based approach would require a huge amount of memory, since many distinct members of the category must be learned (each very different from the others). It has been proposed both that rules are defined just as boundaries in the representational space, and that they are defined as more complex predicates. The boundaries-based approach cannot explain the learning of relational categories, since for many such categories there aren't clear boundaries in the representational space. In contrast the predicate-based approach, as used for example by Inductive Logic Programming (see below) is more general and can model every kind of category, including relational categories.

If hypothesis 2b will be found correct, the computational model which will be built will need a flexible representation of categories, and the predicate-based approach is the perfect candidate.

It is therefore important to contrast the *classic* "rules as boundaries" approach (Nosofsky et al., 1989) and the "rules as predicates" approach (Muggleton, 1991), as explained above. The "rules" that the computational model will use are not boundaries but predicates (see the Inductive Logic Programming section).

2.3.2. The "theory-theory"

The so-called "theory-theory" started as an explanation of children's early conceptions of the mind (Gopnik, 1984, 1988; Butterworth, Harris, Leslie & Wellman, 1991; Astington, Harris & Olson, 1988; Frye & Moore, 1991), and was in fact also referred to as "children's theory of mind". In summary, it stated that children form and change theories about other people's mental states.

But the idea that the formation and modification of theories (like in science) is central in human cognition, was also extended from the theorization on people's minds to the theorization on the rest of the world, and from children to all human beings:

> […] knowledge is structured in a theory-like way, and […] knowledge changes in a way that is analogous to theory change in science.
>
> *(Gopnik, 2000)*

According to this idea, categories are defined by theories, and as theories can undergo revising, so do the classification criteria, by a scientific process:

> […] A theory postulates a complex but coherent set of causal entities, the theoretical entities, and specifies causal relations among them, the laws. Just as a spatial map allows for new predictions and interventions in the world, so a causal map allows for a wide range of causal predictions and interventions, including experiments. And just as theories are revisable in the light of new experience, rather than hard-wired, so causal maps, like spatial ones, can be updated and revised .
>
> […] it is part of the very nature of theory formation systems that they are perpetually in flux.
>
> *(ibidem)*

The hypothesis 2b is coherent with these ideas, and in fact it predicts that hypotheses are created and then tested and refined. The model which could be built based on hypothesis 2b, should therefore take into account the "theory theory". The classification criteria implemented by the model should have a theory-like structure, which can be tested and revised and if needed discarded.

2.3.3. Implicit learning

Learning is considered to be implicit when people acquire new information without intending to do so, and in such a way that the resulting knowledge is difficult to express (Berry & Dienes, 1993). However, the debate on what exactly implicit learning is and how exactly it works is going on since thirty years. It is largely debated, for example, the extent to which Implicit learning produces unconscious and/or abstract knowledge, or the extent to which Implicit learning uses independent memory and processing systems (Cleeremans, Destrebecqz & Boyer, 1998).

Implicit learning experiments have usually three components: 1. exposure to some complex rule-governed environment under incidental learning conditions; 2. a measure that tracks, through performance on the same or on a different task, how well subjects can express their newly acquired knowledge about this environment; and 3. a measure of the degree to which subjects are conscious of the knowledge they have learned. The paradigms, built on these components, which have been largely explored are: artificial grammar learning (Reber, 1989, 1993), sequence learning (Reed & Johnson, 1994; Lewicki, Hill & Bizot, 1988), and dynamic system control (Berry & Broadbent, 1984).

A large debate exists also about the neurological bases of implicit vs. explicit learning (Ashby & Casale, 2003). Amnesic patients are for example widely used to test implicit learning (Knowlton, Ramus & Squire, 1992; Reber & Squire, 1994), and neuroimaging is used to understand which brain areas are specifically involved under different tasks or instructions (Raush et al. 1995; Hazeltine, Grafton & Ivry, 1997; Berns, Cohen & Mintun, 1997).

In case hypothesis 2b is proved correct, the two-phase account of learning opens the possibility that implicit learning occurs. In fact, while subjects are explicitly instructed to learn the "final" categories, the resulting middle phase of partial learning is something that can occur without the subjects being consciously aware.

I am not interested in knowing if the learning of partial categories is explicit or implicit. For some people it might be explicit while for others it might be implicit; this is not the key point of this study.

It was nevertheless interesting to point out, for future research, the fact that if people learn what we propose as "partial categories", they could learn them in an implicit way, without being conscious of that (partial) learning.

2.3.4. Existing models

It has already been suggested that analogical reasoning can be used during the learning of categories. The two most interesting models of category-learning using analogical reasoning are SEQL (Kuehne et al., 2000) and DORA (Doumas & Hummel, 2005).

If hypothesis 2a will be found correct, these models could be extended to be able to compare items of different categories in order to mutually align them and extract all the available information to help learning.

2.3.4.1. SEQL

Kuehne et al. (2000) propose to extend the existing SEQL model (which is a model of abstraction-making using structural alignment - Skorstad, Gentner, & Medin, 1988) with a new algorithm, called GEL (*Generalization and Exemplar Learning*). The extended model is able to learn relational categories from examples, through an iterative process of abstraction. The GEL algorithm is the core of this model; it has a memory for examples and generalizations. When a new example arrives, it tries to compare it, using the SME engine, to the stored examples and generalizations. If the new example is similar enough (compared to a pre-set threshold) to a stored example or generalization, GEL creates a new generalization with the structural overlap. In this way it is able to progressively abstract categories from examples. Learning is not supervised; this model doesn't use labels to check if its deductions are correct. It just produces its best guesses on how the presented examples could be categorized.

This implies that, in order to be found, the categories must have definitions whose structures are different enough to be discerned. If two categories are too similar, the model would find only one general category. This is the most important limitation, in the light of the present study. In fact, the present work investigates analogy between similar categories. SEQL would be mislead by its use of analogy, instead of being helped. But with some extensions and tweaking the model could be adapted to be able to use mutual alignment also between categories.

2.3.4.2. DORA

In 2005 Doumas and Hummel proposed extending the LISA model (Hummel & Holyoak, 2003) to make it able to discover new relations. DORA (Discovery Of Relations by Analogy) is the resulting model (Doumas & Hummel, 2005). It is not a model of category learning, since it can learn only one concept at a time. It is however interesting to briefly present it (due to its complexity, an exhaustive presentation would be too long), since it is the most interesting alternative explanation of the use of analogical reasoning in the learning of new abstract relational concepts.

DORA (and LISA) represents propositions in a hierarchy of units. At the bottom there are semantic units, which represent concepts in a distributed manner. At the next level there are token units which represent specific relational roles and objects. So, for example, in a proposition like "the cat chases the mouse", the roles "chaser" and "chased" are represented at this level, as well as the specific objects "the cat" and "the mouse". The object "cat" would be connected to a set of semantic units representing its features (e.g. "has fur", "four-legged"), while the role "chaser" would be connected to other semantic units. At the third level there are role-binding units, which encode the bindings of specific roles ("chaser") to specific fillers ("cat"). At the top of the hierarchy there are proposition units which bind sets of role-bindings ("cat-chaser", "mouse-prey") into complete relational structures.

The novelty of DORA, with respect to LISA, is that it is able to compare distinct propositions (i.e. distinct structures of this hierarchy) in order to create new knowledge. If some semantic units "fire" in both the propositions, a new role unit is created and bound to the common units. From these new units, new role-binding units and even new abstract proposition units can be created.

The main limitation of DORA is that it can only learn one concept at a time, and only by comparing two examples. It could be possible to extend DORA to make it able to learn more than one category. Even doing so, when presented with examples from categories with similar structures (as in this present work), DORA would probably learn, like SEQL, only one general category. Due to its functioning, it would be confused by the similarities between the categories, instead of being helped by these similarities. As for SEQL, with the right extensions DORA could do mutual alignment and simultaneously learn similar categories.

2.3.4.3. Summary

Some theories and models have been already proposed dealing with the learning of relational categories using analogy. The common characteristic of the proposed models is that they are based on intersection discovery: a schema is learned from examples by keeping what the examples have in common and discarding details on which they differ. The result is some kind of predicate that describes the common relations of the examples.

The common limitation of the existing models is that, although some of them can exploit structural similarities within categories, when faced with similarities between concepts they would be confused by them. If hypothesis 2a will be found correct, these models could be extended so to not be confused by those similarities and instead use mutual alignment to help learning.

2.4. Machine Learning

The idea that underlies all of this work is that exploiting the similarities between novel concepts can facilitate learning, by minimizing the required time and memory and maximizing the amount of used information. This attempt at optimization is common in many satisficing machine learning models. Many models of analogical reasoning (e.g. the "structure-mapping engine" - Falkenhainer, Forbus & Gentner, 1986) avoid the problem of memory and time constraints. Because some ideas will be borrowed from the domain of machine learning, a brief review is given in the following.

2.4.1. Attribute-Value Learners

A field very close to the kind of reasoning investigated in this work, is Inductive Learning. With this description are grouped all the models that learn general concepts by inductive inference, from (relatively) few examples. There are many such models. This section is about the models which classify items according to the values of their attributes, for example Star (Michalski, 1983) or ID3 (Quinlan, 1986). The next section is dedicated to models which classify items according to relations.

2.4.1.1. Star

Michalski's model is based on a process of analysis of the presented items' attributes, generation of generalizations and subsequent restructuring of those generalizations. Although it uses predicates to represent the concept generalizations, the actual implementation includes only attributes and logical operation (e.g. colour = black AND shape != triangle). Given this limitation, Star isn't able to learn relational concepts, unless relations are represented as attributes. More recent models (discussed below) can use more complex predicates, which include relations.

Because the method of partial learning and restructuring can be used to optimize the used memory, in case a model based on hypothesis 2b will be built, the model will borrow this method.

2.4.1.2. ID3

Quinlan's model generates a decision tree, based on the presented items. Since with big training sets the quantity of all the possible decision trees is impossible to handle, ID3 starts with a random subset, and chooses the simplest tree to correctly classify the items in the subset. Then it tries to classify some other items and iteratively includes in the subset the misclassified items, which allow it to correct the original tree. Therefore, also this model uses some kind of partial learning and refinement, as expected by hypothesis 2b. However, decision trees are based on attributes, and are therefore not suitable for relational learning, unless relations are represented as attributes.

2.4.2. Relational Learners

Another class of learning algorithms, called "relational learners", in contrast create descriptions of relations. These descriptions are generally represented as predicates, which, unlike the ones used by Star, can express relations.

An important example is the class of "inductive logic programming" learners (Lavrac & Dzeroski, 1994; Muggleton, 1991; Muggleton & Raedt, 1994). These systems are based on logic predicates, which are generated (through quite complex and very optimized systems) by induction, starting from examples. Many models have been developed, but they all share the same macro-algorithm, as stated by Muggleton and Raedt (1994). It is an iteration of creation of hypotheses (in the form of predicates) and pruning of the existing hypotheses, until the remaining predicates are able to correctly classify all the items.

Some models follow a "specific-to-general" pattern. They start from the examples and background knowledge, and repeatedly generalize their hypotheses by applying inductive inference rules. One limit of these models is that these rules are very complex inference logic rules, which typical people rarely use. Another limit is that they follow only one direction of reasoning.

Another class of models follow the inverse pattern: "general-to-specific". They start with the most general hypothesis (i.e. the inconsistent clause) and repeatedly specialize the hypothesis by applying deductive inference rules in order to remove inconsistencies with the negative examples. These models have the same limits as the "specific-to-general" ones: the inference rules they use are too complex, and they can only reason in one direction.

There are also some models which are able to reason in both directions, which in my opinion is the most sensible strategy. Nevertheless, they still have the problems of using very complex inference rules, which aren't used by typical people.

As already hypothesized above (paragraph on the modification of concepts), if hypothesis 2b will be found correct, the model will use a more trivial algorithm for the modification of hypotheses, that is random modification. If the model will be able to reproduce the participants' results with a simple modification algorithm, it would only work better with more intelligent modification algorithms like inductive or deductive inference rules.

2.4.3. Summary

Given the relational nature of my experiments, inductive logic programming would be a good candidate to build a model based on hypothesis 2b. Unfortunately, all the algorithms proposed in the past years are too perfect learners and use too complex logic, compared to human beings. They have been developed to optimize the performance on computers and to exploit machines' innate logic, not to be psychologically plausible. Moreover, the inductive logic programming algorithms aren't able to exploit the similarities between categories: each category is learned independently.

Nevertheless, the computational model could borrow many concepts from machine learning and from inductive logic programming. For example, the proposed model could use intensive descriptions (i.e. predicates) to represent hypotheses. The concept of "heuristic" can also be borrowed from machine learning, although the heuristics used are very simple compared to the heuristics of the other common algorithms. Apart from some basic heuristic for hypothesis testing, the heuristic at the core of the model could be phrased as "find partial hypotheses and reuse/refine them as soon as possible". As shown in Chapter 6, this heuristic can give birth to analogical reasoning in an emergentist way, and can exploit the simultaneous-learning of similar categories.

Some other basic aspects of machine learning can be also borrowed by the proposed model. It could learn partial rules and iteratively refine them. In contrast to many machine learning algorithms, the modification of hypotheses should go in any direction (while usually it is either "general-to-specific"or "specific-to-general", rarely both).

2.5. Conclusions

In this chapter previous works in fields related to this theses have been reviewed to show how they are related, what knowledge can be used, and where are the gaps of knowledge that can be filled. This chapter presented studies about analogical reasoning, relational categories and simultaneous learning of categories, which are related to our first hypothesis, which says that analogy can be established between simultaneously-learned categories with similar structures, to aid the learning of both categories. Included are also studies about structure mapping, mutual alignment, direct comparison of similar items and relational categories learning models, which are related to hypothesis 2a, which says that categories are continuously compared and mutually aligned using structure mapping. Studies about the modification of concepts, the "theory theory", implicit learning and some other machine learning models (such as the Inductive logic programming) were also presented, which are related to hypothesis 2b, which says that learning consists of two phases where partial categories are first formed and then refined to create the final categories.

From this review, it is clear that the case of the simultaneous learning of categories with similar structures was never explicitly studied before. This thesis will address what happens in such a case, and to do so will use the knowledge drawn from the studies presented in this chapter.

Chapter 3

Experiment 1

3.1. Introduction

Given the breadth of the field investigated, this first experiment verifies only some of the predictions made in the second chapter. The focus for this experiment was mainly on the invention of a paradigm with novel similar relational categories, and the simultaneous presentation of similar exemplars.

The hypotheses tested in this experiment are those already stated in the previous chapters:

1. Analogy can be established between simultaneously-learned categories with similar structures, to aid the learning of both categories. For example, a category defined by an equal number of elements of the same kind is structurally similar to a category defined by a numerical ratio between elements of the same kind, and analogy can be established between both categories when simultaneously learned;

2a. Categories are first learned and then compared and mutually aligned using structure mapping; or

2b. Learning consists of two phases where one or more partial categories are first formed and then refined, so from each partial category stem some final categories.

This experiment tests only a subset of the predictions of my hypotheses. The predictions tested in this experiment are:

1. according to hypothesis 1, learning of similar categories is related, that is, time between learning of similar categories is less than time between learning of dissimilar ones. For example, the time elapsed between the learning of two categories both defined by a numerical relation (e. g. equal number and 1:2) between elements of the same kind is less than the time elapsed between the learning of one category defined by a numerical relation and another category defined by the presence of a distinctive element,
2. according to hypothesis 2a, the simultaneous presentation of exemplars from similar categories should facilitate learning,
3. according to hypothesis 2b, before learning is complete, any errors in categorisation are not random but they are more frequent across similar categories (e. g., errors across two categories defined by a numerical relation between elements of the same kind are more frequent than errors across one category defined by a numerical relation and another category defined by the presence of a distinctive element),
4. according to hypothesis 1, relational similarity (e. g., based on the numerical relation between elements) helps learning, in respect to categories defined simply by features (for instance, defined by the presence of a distinctive element).

The design of the experiment will follow the need to test all these predictions.

In order to test the first prediction (learning of similar categories is related), the experiment must compare how long it takes to learn each category. The null hypothesis is that the learning of each category is independent.

To test the second prediction, there must be some way in which we can compare how difficult the participants found it to learn simultaneously-presented similar categories compared to simultaneously-presented dissimilar categories. The null hypothesis is that it is equally difficult.

To test the third prediction, the pattern of answers and errors before learning is considered to have occurred, must be recorded and analysed. The null hypothesis is that those answers are randomly distributed.

Finally, to test the last prediction, an analysis of which category is learned first must be performed. The null hypothesis is that there is no difference between the number of times each category is learned first.

In order to perform all these analyses, various constraints were considered which served to guide me in the design of the experiment. The following subsections illustrate these constraints and the solutions proposed. The resulting experiment is then described in the "Method" section.

3.1.1. No previous knowledge

Analogical reasoning has been extensively studied in its function of transferring knowledge from a well known domain to a less known one. The focus in this experiment is instead on its use during the learning of completely unknown concepts. Therefore any domain in which previous knowledge is used (like for example the comparison or completion of words, stories, pictures or situations) must be excluded. Otherwise it can always be argued that analogical reasoning occurs because one concept was already (better) known initially, laying the foundation for a *standard* analogical transfer of knowledge.

Obviously it is impossible to invent a completely unknown domain, since the stimuli must be understandable in some way. Yet it is possible to define artificial categories which are rarely (if never) seen in everyday life. What is most important is to include enough complexity to be able to define categories based on relations.

The solution found is to use categories of a graphical abstract nature. So, I have designed exemplars containing a number of coloured shapes of different types (circles, squares, triangles, crosses, stars) and colours (blue, red, yellow, green and pink). In this way I can build my own categories, ensuring that all participants have no prior knowledge of any of the categories.

3.1.2. Definition by similar relations

In many Category Learning tests the categories are defined by the presence of one or more features. In those cases it is difficult to determine if Analogical Reasoning can have a role, even if categories have similarities between them. Therefore there is need of categories defined by relations (e. g., categories defined by a numerical relation between elements of the same kind), a concept already proposed by Gentner and Kurtz (2005).

In order to create a case in which similarities can be detected only using Analogical Reasoning, in this experiment the *analogical* categories must be defined by relations between features. This means that in different exemplars of the same category the features change, but the relation between them is always the same within the category: e. g., the relation between the exemplars is an equal number of elements of the same kind, with the same shape but different colour (so, within a category we shall have 2 red circles and 2 green circles in an exemplar, 3 red circles and 3 green circles in another one, and so on). Those relations must then be similar between (at least) two categories, so Analogical Reasoning can clearly be used.

Moreover, in order to avoid an early "gestalt" processing of the information, graphical relations (such as "left-right", "inside-outside", etc.) must not be used. To be sure that Analogical Reasoning, and not other abilities, is used to detect similarities, only relations that involve higher logical reasoning can be used.

3.1.3. Simultaneous presentation

A factor that can facilitate or hinder learning is the simultaneous presentation of exemplars. It has already been shown (Kurtz & Gentner, 1998; Kurtz & Boukrina, 2004; Kurtz & Loewenstein, 2007; Kurtz et al., 2001) that the simultaneous presentation of exemplars of the same category can facilitate learning. If Analogical Reasoning is involved in the learning of similar categories, according to the structure mapping theory also the simultaneous presentation of exemplars of similar categories should facilitate learning.

In fact the direct comparison between the two (or more) exemplars should help finding the alignable and non-alignable differences (Gentner & Markman, 1994; Markman & Gentner, 1993, 1996) using the structure mapping.

In order to test if the simultaneous presentation of similar (or dissimilar) exemplars helps learning, two groups of participants are needed. One group is presented more often with paired exemplars of similar categories, another group is instead presented more often with paired exemplars of dissimilar categories. This introduces also the need of two distinct kinds of categories: a first kind with similar structures, and a second kind with structures dissimilar from the first kind (i.e. relational categories, defined by a numerical relation between elements of the same kind *versus* features categories, defined by the presence of a distinctive element).

Finally, a third group of participants is needed, which is presented with only one exemplar at a time, in random order. This third group is introduced to check if the presentation of two exemplars at a time makes the test more difficult. It can be argued that the task is different if one or two exemplars are presented each time, and it is exactly this problem that is tested. Is the task easier or more difficult, if two exemplars are presented, given that in other experiments only one exemplar at each time is presented? The null hypothesis is that the difficulty does not change between the single presentation of exemplars, the simultaneous presentation of similar exemplars, and the simultaneous presentation of dissimilar exemplars. If this is the case, it can be inferred that exemplars are not continuously compared and mutually aligned.

In other terms, we know that people normally see only one exemplar at a time. We want to know what happens if they are shown instead two exemplars simultaneously. In particular, the case in which they are shown two exemplars simultaneously may fall into one of two sub-cases: they are shown two *similar* exemplars simultaneously or they are shown two *dissimilar* exemplars simultaneously. Therefore, the hypothesis to test in general is:

> The simultaneous presentation of the exemplars does not change anything in respect to the presentation of only one exemplar at a time.

But here it is necessary to make a distinction: in fact, as noted above, the simultaneous presentation can be of *similar* exemplars or of *dissimilar* ones. So that hypothesis should be split into three hypotheses:

1) The simultaneous presentation of similar exemplars does not change anything in respect to the presentation of only one exemplar at a time;

2. The simultaneous presentation of dissimilar exemplars does not change anything in respect to the presentation of only one exemplar at a time;

3. The simultaneous presentation of similar exemplars does not change anything in respect to the simultaneous presentation of dissimilar exemplars.

3.1.4. Relations vs. Features

Another way to assess if structural, relational similarities can give an advantage over the definition by simple features, is to have one kind of categories defined by similar relations, and another kind of categories defined just by features. For example, a category defined by a numerical relation between elements of the same kind is a *relational* category, while a category defined by the presence of a distinctive element is a *features* category. If the *relational* categories are found before the *features* ones, it would mean that Analogical Reasoning provides enough help to learning that it not only overcomes the additional complexity of the relational categories, but it allows the exploitation of their similarities to make their learning easier.

It can be argued that all the categories defined by features can also be defined by relations, although simple ones ("is", "has", "is present", etc.). Therefore there would be no difference between the two kinds of categories, because Analogical Reasoning would be used also for the *features* categories to exploit their similar structures. Nevertheless, this argument only strengthens the theory that Analogical Reasoning is widely used, and does not change the validity of other constraints and the other results.

It has also been proposed (Goswami, personal communication) that, given the complexity of the task, people will assume that the categories are defined by complex rules, and ignore simple ones. This would make the discovery of the simple rules at least as difficult as the discovery of the complex ones.

Therefore, a null result wouldn't be particularly informative. Yet, given that the experiment allows for such an analysis, it is worth to mention it and to perform this analysis too.

3.2. Method

3.2.1. Participants

The actual experiment was preceded by a preliminary test on 20 volunteers from outside the university, to check if it was solvable and to assess the expected effect sizes for each test and thus the needed number of participants. Based on the estimated effect sizes, the total number of participants was estimated as 30.

Participants were 30 volunteers from the School of Computer Science and Informatics of the University College Dublin: 20 Males and 10 Females, average age 26.2. They were randomly assigned to the three groups (Paired, Unpaired and Single - see below), 10 for each group, and rewarded with a small sum of money.

3.2.2. Materials

The experiment was carried out on normal personal computers in a controlled environment. The test was developed in Flash and was presented in full-screen mode.

According to the group (see below) the participant was randomly assigned to, the screen presented one or two exemplars each time, one on the left and the other on the right, or just one centred. Below each exemplar there were four buttons, labelled "A", "B", "C", "D". The participant had to click on one of the buttons (for each exemplar) to tell which category (A, B, C or D) they thought the exemplar was.

On the far right of the screen there was a text area labelled "Notepad", in which the participant could write a limited amount of text (500 characters). The entered text, each time the shown exemplars changed, was recorded on a server, in order to provide insight into the learning processes.

Another function of the notepad, given that the task involves much reasoning and is therefore memory-consuming, is to help people to write down ideas. It is worth to remember that this test is only about what kind of reasoning people use, and it was known since the beginning that the tasks would be difficult and memory-consuming.

The exemplars of the four categories looked like grey circles on a white background. Inside the circles there was a number of coloured shapes (in order to avoid spatial hints, the shapes were arranged in random order). The number of the coloured shapes in each exemplar could vary from 1 to 12. The shapes could be of 5 different types (circles, squares, triangles, crosses and stars) and 5 different colours (blue, red, yellow, green and pink). Consequently, there were 25 different kinds of elements in total. Each exemplar contained from 1 to 3 different kinds of elements.

Each category was defined by a different rule, and the exemplars were randomly generated by the computer according to those rules. 3.1 summarizes the kinds of classification criteria used in this experiment.

Table 3.1 Simplified examples of the classification criteria.

Category	Criterion
Relational 1 ("A")	Contains the *same quantity* of red circles and green circles. E.g. 2 red circles and 2 green circles, 3 green circles and 3 red circles, etc.
Relational 2 ("B")	Contains *twice the quantity* of red circles and green circles (or the

	opposite). E.g. 2 red circles and 4 green circles, 3 green circles and 6 red circles, etc.
Feature 1 ("C")	Contains a pink square.
Feature 2 ("D")	Contains a blue square.

Rules for categories A and B (*Relational* categories, in the analysis referred to as R1 and R2) were based on the numeric relation between two kinds of elements inside the exemplar, that had the same shape but different colour (for each test, the shape and the 2 colours were chosen randomly right at the beginning and remained fixed throughout the test).

In category A (Figure 3.1), the number of elements with the given shape had to be the same for the two different colours. For instance, an exemplar of category A could be defined by 2 red circles and 2 green circles (or 3 red circles and 3 green circles, and so on).

In category B (Figure 3.1), the number of elements with the given shape had to be twice the number for one colour than it was for the other. For instance, an exemplar of category B could be defined by 2 red circles and 4 green circles, or just the opposite (or 3 red circles and 6 green circles, or the opposite, and so on). The remaining elements were randomly added and had no role: they were just distractors.

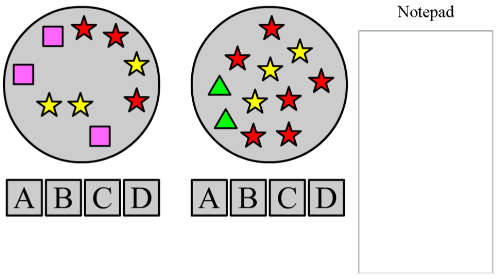


Figure 3.1 Two exemplars are simultaneously shown, one of category A (left), the other of category B (right). In this particular test the categories A and B were defined as same/twice number of red and yellow stars. The other shapes are distractors.

Rules for categories C and D (*Features* categories, in the analysis referred to as F1 and F2) were instead based on the presence of a single distinctive element, of the same shape but different colours for the two categories (Figure 3.2). The remaining elements were randomly added and were just distractors. Obviously, for each test the shape and the 2 colours were chosen randomly at the beginning and remained the same throughout the test. For example, category C could be defined by the presence of a pink square, while category D by a blue square.

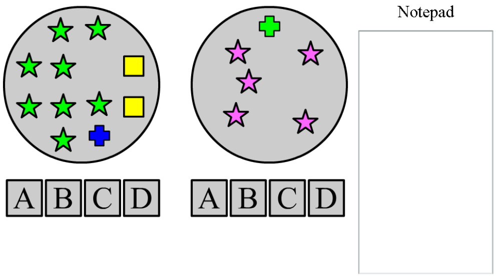


Figure 3.2 Two exemplars are simultaneously shown, one of category C (left), the other of category D (right). In this particular test the category C was defined by the presence of a blue cross, the category D by the presence of a green cross. The other shapes are distractors.

The exemplars were randomly generated by the program, according to the above mentioned rules. In general, each exemplar could contain up to a maximum of 11 “distractor” elements. These distractors contained no category-membership information and were randomly chosen. This design ensured that it was impossible for participants to guess the correct answer by exclusion; instead all the four categories had to be correctly identified.

3.2.3. Design

The participants were divided into three groups: one *paired* group, one *unpaired* group, and one *single* group. The *paired* group (Figure 3.3) saw two exemplars simultaneously on a computer screen, one on the left and one on the right: 5 times out of 6 the two exemplars shown on the screen belonged to similar categories (for instance, A and B, or D and C, or B and B); once out of 6, instead, the two exemplars shown were of dissimilar categories (for instance, A and C, or D and B). The *unpaired* group (Figure 3.4) also saw two exemplars simultaneously on each screen, but only 2 times out of 6 the two exemplars shown belonged to similar categories (A and B, C and D, and so on); more often, i.e. 4 times out of 6, the *unpaired* group saw two exemplars of dissimilar categories (A and C, and so on). Of all the possible ratios between exemplars of similar categories and exemplars of dissimilar categories, this choice of presentation (the two exemplars shown belonged to similar categories 5 times out of 6 for the *paired* group and 2 times out of 6 for the *unpaired* group) was the most arithmetically balanced, although it didn't present only pairings of the same kind, which in any case is not ecologically plausible. With other ratios it would have been impossible to properly balance the different categories in all the possible pairings and, however, it would not have been ecologically plausible. Finally, the *single* group (Figure 3.5) saw only one exemplar on each screen, shown in random order in the centre of the computer screen.

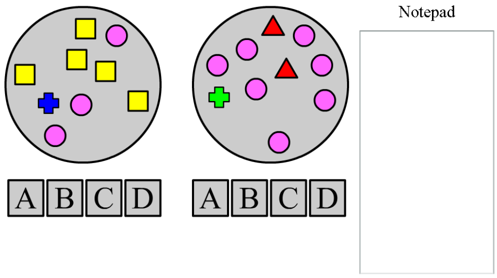


Figure 3.3 Two exemplars are simultaneously shown, of similar categories: one of category C (left), the other of category D (right). They are similar because both present a distinctive element, which is a blue cross for the exemplar on the left and a green cross for the exemplar on the right. This kind of screen is presented more often to participants of the paired group.

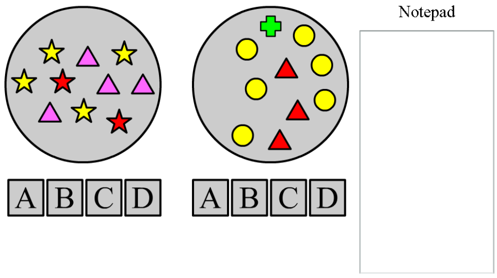


Figure 3.4 Two exemplars are simultaneously shown, of dissimilar categories: one of category B (left), the other of category D (right). They are dissimilar because the one on the left presents a numerical relation (1:2) between elements of the same kind but different colour (there are two red stars and four yellow stars) while the one on the right presents a distinctive element, which is a green cross. This kind of screen is presented more often to participants of the unpaired group.

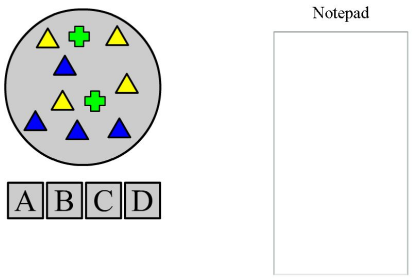


Figure 3.5 Only one figure is shown. This kind of screen is presented to participants of the single group.

The presentation order of the pairs was random, but balanced in cycles of 24 exemplars, so every 12 screens (or 24 in the single group) participants were shown 6 exemplars for each category.

Table 3.2 Possible sequences of screens for the paired, unpaired and single group.

Paired:	Unpaired:	Single:
1. A B	1. A B	1. A
2. B A	2. D B	2. B
3. C D	3. C A	3. C
4. D A	4. B C	4. D
5. A B	5. A A	5. D
6. D C	6. D B	6. A
7. C C	7. C C	7. A
8. A D	8. B D	8. B
9. B B	9. C D	9. B
10. D C	10. D A	10. B
11. C D	11. A D	11. D
12. A B	12. C B	12. C
		13. D
		14. C
		15. C
		16. C
		17. B
		18. A
		19. C
		20. D
		21. A
		22. D
		23. A
		24. B

3.2 gives some examples of possible sequences of screens for a participant of the *paired, unpaired* and *single* group.

The aim of this division into groups was to test if the simultaneous presentation of exemplars of similar categories helps learning. To test this prediction, the performance of the participants had to be compared across the three groups. The number of exemplars and the amount of time needed to finish the test were the crucial factors to evaluate and compare the performances. Furthermore, all the answers given during the test were recorded, in order to analyze all the mistakes made as well as the order in the process of category learning. As already said, the content of the notepad and the answers to the debriefing questions were recorded, too, in order to have a better insight in the reasoning process.

3.2.4. Procedure

1. After reading a screen with the instructions, the participant clicks a button to start the test.

2. For each screen the participant has to choose an answer clicking on one of four buttons (“A”, “B”, “C” or “D”) shown below the exemplar(s) presented: this will answer the question “which category is this exemplar?”

3. The chosen button becomes yellow and a button labelled “Ok” appears below (Figure 3.6). In the case of the *paired* and *unpaired* groups, the “Ok” button will appear only when both of the questions presented in the screen are answered. The participant has to click on “Ok” in order to confirm the answer(s). Before clicking on the "Ok" button, the participant can change the answers an unlimited number of times.

4. After clicking on the "Ok" button, the “Ok” button disappears, and feedback is given (Figure 3.7). If the answer is right, the chosen button becomes green, with a "tick" inside. If the answer is wrong, the chosen button becomes red, with a "cross" inside, and the correct answer becomes green, without any sign. Beside the feedback, a “Next” button appears, allowing the participant to pass to the following exemplar(s) (point 2). The test is therefore self-paced.

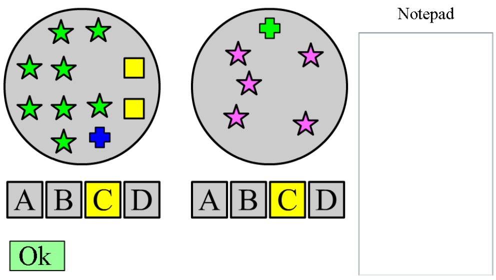


Figure 3.6 Answers are given for both the shown exemplars, and the "Ok" button has appeared. The participant can still change the answers, and must confirm them by clicking the "Ok" button.

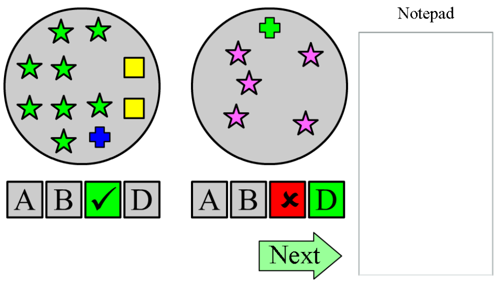


Figure 3.7 After clicking the "Ok" button feedback is given, and the "Next" button appears. The exemplar on the right was correctly of the "C" category, while the exemplar on the left was instead of the "D" category.

Since the exemplars are randomly generated by the program, there isn't a definite, fixed number of exemplars for each test, but it can virtually go on *ad libitum*. Actually the test ends when in each category the participant reaches an accuracy of 83%, calculated on a moving average of 12 exemplars per category (that is 10 correct answers). This criterion is chosen to minimize the probability of passing the test just by chance. The 2 incorrect answers are allowed in order to take into account some possible distractions at the end of the test.

To better understand the learning process, when subjects reach an accuracy of 40% and 60% in each category, and at the end of the test, they are asked if they have found a rule for any of the categories and, if they have, what those rules are. At the end of the test a debriefing question asks what method they have used to solve the test. The test is implemented in Macromedia Flash and all answers are timed with an accuracy of 1 millisecond and are recorded locally and then sent to a server.

3.2.5. A typical test session

A typical test session would work as follows. At the very beginning, the participant had to give answers randomly, not having any previous knowledge. They could take notes (although the available space was intentionally limited), and usually they started writing descriptions of the seen exemplars. Everything the participants wrote on the notepad was recorded. Although no formal analysis has been conducted, an informal review of the notes showed that participants just wrote about the colour, shape and number of elements, ignoring the spatial disposition, which must have already been eliminated as a source of information for categorisation.

After a while, the participant would notice some patterns, and would start to formulate hypotheses about the classification criteria. This could also be seen from reading the notepad recordings. After a while, the participants started to write some kind of rules instead of just descriptions of exemplars. This formulation of rules has also been clearly stated by some participants in their debriefings.

During this particular stage, the answers the participant would give were sometimes correct and sometimes wrong, but they were no longer random, instead they followed their hypothesized rules. It is therefore interesting to analyse the errors made during this period, in order to better understand the type of confusion encountered.

At some point, the participant started to give systematically the right answer to some category, and that is the point we call "learning point" for that category (see below for the algorithm).

The learning processes for the different categories can be unrelated (null hypothesis) or related. In particular, learning of the two relational categories can be related (learning one can help learning the other), as well as learning of the two features categories (since they also are similar to each other). In contrast learning of the two distinct groups of categories (i.e. features vs. relations) is expected to be unrelated.

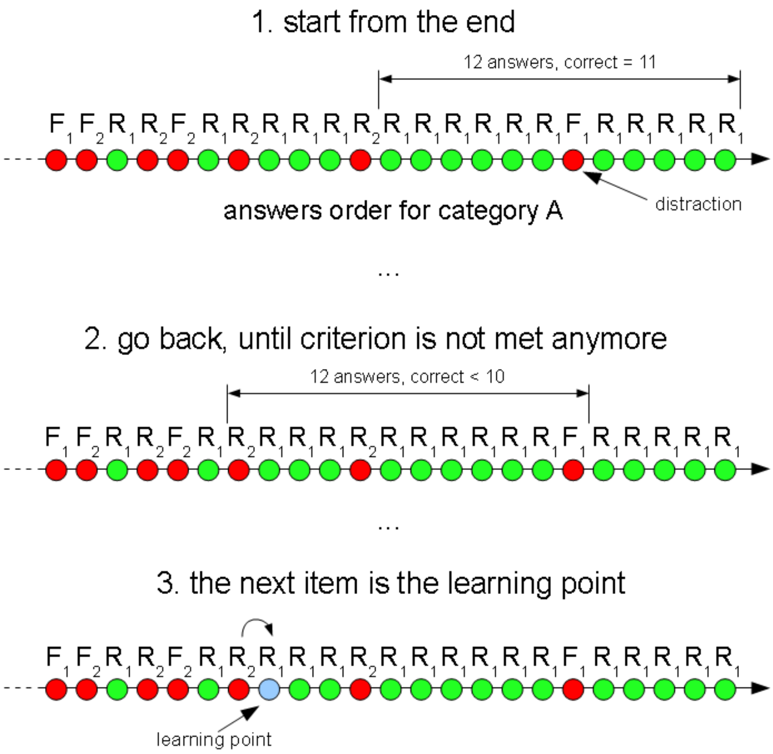
When the participant reached our learning criterion for each category, the test ended. At this stage all four learning points are defined, and it is possible to look at their order and at the intervals between them, to determine if learning of a category is related or not to learning of the other ones.

3.2.6. Estimation of learning points

Since many of the results are based on the estimation of learning points, it can be useful to briefly present the algorithm used for their estimation (Figure 3.8).

Learning points are estimated independently for each category, using a moving average of the correct answers to the shown exemplars of that category. As said above, participants are considered to have learned a category when they reach an accuracy of 83%, calculated on a moving average of 12 exemplars per category (that is 10 correct answers). The learning point for a category is therefore the first correct answer after which the participant only makes two mistakes in the next 11 answers (for that category).

```
for each answer for the given category, starting from the end {
        calculate a moving sum of length 12 of the correct answers for the
        given category;
        when the moving sum > 10 {
                go backward until it is < 10 {
                        the answer immediately following is the learning
                        point;
                }
        }
}
```



1. start from the end
12 answers, correct = 11
F1 F2 R1 R2 F2 R1 R2 R1 R1 R1 R2 R1 R1 R1 R1 R1 R1 F1 R1 R1 R1 R1 R1
answers order for category A
distraction
...
2. go back, until criterion is not met anymore
12 answers, correct < 10
F1 F2 R1 R2 F2 R1 R2 R1 R1 R1 R2 R1 R1 R1 R1 R1 R1 F1 R1 R1 R1 R1 R1
...
3. the next item is the learning point
F1 F2 R1 R2 F2 R1 R2 R1 R1 R1 R2 R1 R1 R1 R1 R1 R1 F1 R1 R1 R1 R1 R1
learning point

Figure 3.8: Algorithm for the estimation of learning points. Green circles are correct answers, red circles wrong answers. The example represents learning of Category R_1.

3.3. Results

3.3.1. Simultaneous presentation factor

To test the prediction that the simultaneous presentation of similar exemplars helps learning, the performance of the participants has to be compared across the three groups. This prediction is related to the hypothesis 2a, that categories are continuously compared and mutually aligned using structure mapping.

When all the categories reached the learning criterion, the test ended. Time and number of exemplars elapsed until that point can both be measures of the difficulty to finish the test. Both were analysed and compared across the groups. No significant difference was found (see Figure 3.9, Table 3.3 and Table 3.4).

Both time and number of shown exemplars were analysed and compared across the groups. No significant difference was found (Time: F(2,27)=.41, *p>.60*; Shown exemplars: F(2,27)=.48, *p>.60*; see Figure 3.9, Table 3.3 and Table 3.4).

Table 3.3 Descriptive statistics for number of shown exemplars to end

Group	Mean	Std. Deviation	N
Paired	138.8	92.87	10
Unpaired	193.8	235.41	10
Single	133.2	74.76	10
Total	155.27	149.61	30

Table 3.4 Descriptive statistics for time to end

Group	Mean	Std. Deviation	N
Paired	22.99	14.45	10
Unpaired	28.72	17.52	10
Single	25.79	8.56	10
Total	25.83	14.13	30

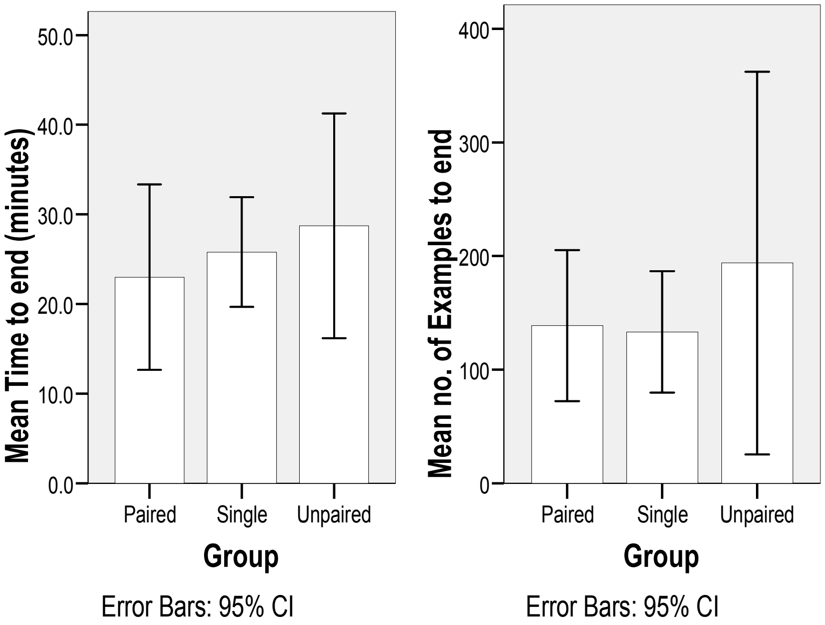


Figure 3.9 Time and number of exemplars required to finish the test

Although the trends seem to favour the hypothesis that direct comparison can help learning (the paired group needs less time and exemplars than the unpaired group), the size of this effect is very small. Therefore direct comparison is not a strong factor to help the learning of similar categories, and it does not even differentiate the performance compared to the single group.

Nevertheless, as the following analyses will show, learning of similar categories is related, thus some kind of analogy is used during the test. Given that the direct comparison is not the crucial factor, hypothesis 2a (categories are first learned and then compared and mutually aligned using structure mapping) should be discarded if the prediction based on hypothesis 2b (learning consists of two phases where partial categories are first formed and then refined to create the final categories) is confirmed.

3.3.2. Learning intervals

A trivial way to test the prediction that learning of similar categories is related would be looking at the order in which different categories were learned. But to do so would mean to discard a lot of useful information, and could result in misclassification in some cases. Take, for example, the cases shown in Figure 3.10. They have the same order of learning, but clearly in the first case the two relational categories (R_1 and R_2) are as related (or unrelated) as the two features ones (F_1 and F_2). In fact, the distance, in time, between the learning of R_1 and R_2 is the same than the distance between F_1 and F_2. In the second case, in contrast, the two relational categories are much more related than the other two, although the learning order is the same. From the figure it is clear that the distance, in time, between the learning of R_1 and R_2 is less than the distance between F_1 and F_2. Because of the brief time elapsed between the learning of R_1 and R_2, the two events can be considered related.

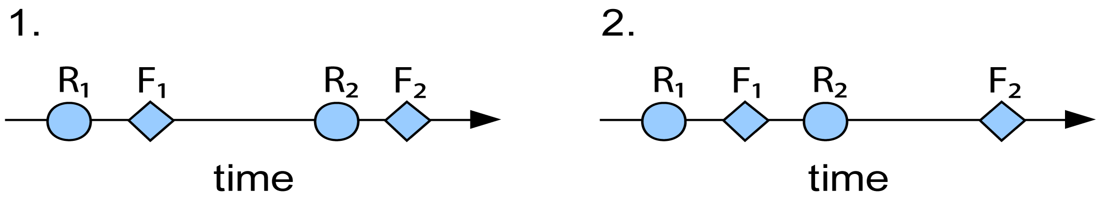


Figure 3.10 Points of learning in time, equivalent for order, but different for intervals

A computation of the average time needed to learn each category is also useless, for two reasons. The first reason is that it is not important if category R1 is learned before or after R2 (or F1 before or after F2), because the attention is on which *kind* of category is learned first. The second and most important reason is that time and number of exemplars needed to finish the test vary a lot between participants, and therefore also the time and number of exemplars elapsed before the learning of each category. Thus, another analysis of learning points must be used, to compare learning points independently for each participant.

As a first step, for each participant, all the intervals between the learning points are calculated (i.e. R_1R_2, R_1F_1, R_2F_1, etc.), in terms of number of exemplars shown until the points of learning. For example, the interval R_1R_2 is computed as the number of exemplars elapsed between the learning of R_1 and R_2 (for each participant).

The important relation is of the learning points of R_1 and R_2, on one hand, and of the points of F_1 and F_2 on the other hand. Those two intervals are therefore of particular interest, and are compared to see which pair is more related.

In a second step, for each participant the average of all the intervals between all the categories (AllIntAvg) is computed:

$$AllIntAvg = \frac{\overline{R_1 R_2} + \overline{F_1 F_2} + \overline{R_1 F_1} + \overline{R_2 F_1} + \overline{R_1 F_2} + \overline{R_2 F_2}}{6} \qquad (3.1)$$

It is used as a baseline for comparison, for each participant. In fact, if the four learning points are all independent from each other, a randomly chosen learning interval (e.g. R_1R_2) should be approximately equal to this average.

If, for a participant, the number of exemplars elapsed between the learning of R_1 and R_2 (and/or the learning of F_1 and F_2) is less than what randomly expected (i.e. *AllIntAvg*), it can be inferred that learning of similar categories is related.

In a third step, it is counted for how many participants AllIntAvg > R_1R_2, for how many AllIntAvg < R_1R_2, AllIntAvg > F_1F_2, etc.

A preliminary MonteCarlo was performed to compute the expected number of times AllIntAvg > R_1R_2 (or F_1F_2). This simulation showed that this comparison is positive roughly 50% of times and negative the remaining 50%, and that the distribution is not normal (Appendix C). Therefore it was not possible to use the *t*-test or other parametric tests, and a binomial test has been performed.

Although the group factor should have no effect on how much learning of similar categories is related, a preliminary analysis was done on each group separately. Given the small size of the groups, no statistically significant result was expected, but the trends can always be studied. As shown in the Table 3.5, 3.6 and 3.7, for each group the interval between relational categories (R_1R_2) is less than the average of all the intervals

(AllIntAvg) and the interval between features categories (F_1F_2) is also less than the average of all the intervals (AllIntAvg). Given that there are no differences between the groups, a joint analysis was performed.

In the joint analysis the interval between relational categories (R_1R_2) is less than the average of all the intervals (AllIntAvg) in 87% of the cases (much more than the expected 50%, binomial test $p < 0.001$, see Table 3.5, 3.6 and 3.7), meaning that the relational categories are more related to each other than expected. Also the interval between features categories (F_1F_2) is less than the average of all the intervals (AllIntAvg), in 82% of the cases (much more than the expected 50%, binomial test $p < 0.001$, see Table 3.5, 3.6 and 3.7), meaning that also the features categories are more related to each other than expected.

These results support the prediction that learning of similar categories is related, and thus the hypothesis that analogy can be used early in learning, between simultaneously learned categories. But given the small effect size of the simultaneous presentation factor (see previous section), it is plausible that some process other than structure mapping is involved when learning novel similar categories. This will be further discussed below and in detail in the model chapter.

Table 3.5 Frequencies of the comparisons of the interval between relational categories (R_1R_2) Vs the average of all the intervals (AllIntAvg) and significance of the binomial tests.

Group	AllIntAvg > R_1R_2	AllIntAvg < R_1R_2	AllIntAvg = R_1R_2	N	Proportion	P
Paired	9	1		10	.9	.021
Single	8	2		10	.8	.109
Unpaired	9	1		10	.9	.021
Total	26	4		30	.87	.001

Table 3.6 Frequencies of the comparisons of the interval between features categories (F_1F_2) Vs the average of all the intervals (AllIntAvg) and significance of the binomial tests.

Group	AllIntAvg > F_1F_2	AllIntAvg < F_1F_2	AllIntAvg = F_1F_2	N	Proportion	P
Paired	9	1		10	.9	.021
Single	8	1	1	10	.89	.039
Unpaired	6	3	1	10	.67	.508
Total	23	5	2	30	.82	.001

Table 3.7 Frequencies of the comparisons of the interval between features categories (F_1F_2) Vs the interval between relational categories (R_1R_2) and significance of the binomial tests.

Group	F_1F_2 > R_1R_2	F_1F_2 < R_1R_2	F_1F_2 = R_1R_2	N	Proportion	P
Paired	5	2	3	10	.71	.453
Single	2	8		10	.2	.109
Unpaired	6	3	1	10	.67	.508
Total	13	13	4	30	.5	1

3.3.3. Analysis of Errors

From the hypothesis 2b, that learning consists of two phases where partial categories are first formed and then refined to create the final categories, stems the prediction that before learning is complete, any errors in categorisation are not random but they are more frequent across similar categories. In other terms, it is more probable to say that a R_1 exemplars belongs to the R_2 category, than it belongs to F_1 or F_2.

Therefore, the incorrect answers given by a participant before learning of any category has occurred can be a good indicator of how the learning process happens for that participant. In particular, my interest is whether they are randomly distributed, or if the mistakes (one could say the "confusion") follow particular patterns.

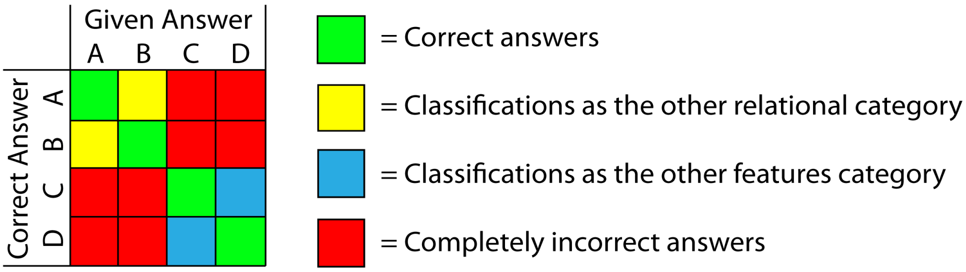
This analysis must be performed on the answers given before any category has been learned, therefore the final point of the data to be analysed is the learning point (see above for the estimation) of the first learned category. Clearly, for each participant the number of answers until that point is different.

It would therefore be pointless to consider the first (or the last, or the middle) *n* answers, since they would be for each participant a different fraction of the given answers. Yet some criterion must be defined to classify the answers in order to analyse the progression of learning. In fact the first answers are expected to be random, while the last answers are very near the correct identification of the first classification criterion. But how many are the "first answers" and how many the "last answers"? Given that every criterion would be arbitrary, the one that seemed most sensible was chosen. The data to be analysed (i.e. the answers from the beginning until the first learning point) was divided into three equal intervals.

Given the value 1 to the first learning point, and the value 0 to the start of the experiment, three intervals have been considered: from 0 to 0.333 (interval 1), from 0.333 to 0.666 (interval 2) and from 0.666 to 1 (interval 3). They represent three possible stages of learning: initial, intermediate and almost complete.

For each of these intervals a contingency table has been collated, of the given answers vs. the correct answers. Four regions can be marked on this contingency table. On the diagonal are the correct answers. The remaining incorrect answers can be further divided into three regions: the mistakes classifying an exemplar in the other relational category, the mistakes classifying an exemplar in the other features category, and the completely incorrect answers (see Figure 3.11).

The mistakes across relational categories and across features categories are of particular interest, because they represent cases of partial learning. Given that the analysis is performed before complete learning has occurred, if mistakes across similar categories are more than randomly expected, it can be an indication that a partial draft of a common classification criterion was already learned.



		Given Answer			
		A	B	C	D
Correct Answer	A	🟩	🟨	🟥	🟥
	B	🟨	🟩	🟥	🟥
	C	🟥	🟥	🟩	🟦
	D	🟥	🟥	🟦	🟩

🟩 = Correct answers

🟨 = Classifications as the other relational category

🟦 = Classifications as the other features category

🟥 = Completely incorrect answers

Figure 3.11 The contingency table of Given VS Expected number of answers. Four regions can be marked: the correct answers, the mistakes classifying an exemplar in the other relational category, the mistakes classifying an exemplar in the other features category, and the completely incorrect answers.

For each of these regions the random expected values have been also computed using the marginal means (as in the C*hi Square* method), and then the given and expected values has been compared for each region for each interval.

Although the group factor should have no effect on the distribution of errors, a preliminary analysis was done on each group separately. Given the small size of the groups, no statistically significant result was expected, but the trends can always be studied and compared.

Table 3.8 Frequencies and statistics of the sign comparisons of given VS expected number of mistakes classifying an exemplar in the other relational category, for each interval, for each group. In bold the significant results.

Count		Other Relational Category			Binomial P
Interval	Group	Giv<Exp	Giv=Exp	Giv>Exp	
1	Paired	2	5	3	1
	Unpaired	2	5	3	1
	Single	2	5	3	1
1 Total		6	15	9	.607
2	**Paired**	**0**	**4**	**6**	**.031**
	Unpaired	**0**	**3**	**7**	**.016**
	Single	1	4	5	.219
2 Total		**1**	**11**	**18**	**.001**
3	Paired	2	3	5	.453
	Unpaired	3	3	4	1
	Single	1	4	5	.219
3 Total		6	10	14	.115

Table 3.9 Frequencies and statistics of the sign comparisons of given VS expected number of mistakes classifying an exemplar in the other features category, for each interval, for each group. In bold the significant results.

Count		Other Features Category			Binomial P
Interval	Group	Giv<Exp	Giv=Exp	Giv>Exp	
1	Paired	2	7	1	1
	Unpaired	2	6	2	1
	Single	2	7	1	1
1 Total		6	20	4	.754
2	Paired	1	4	5	.219
	Unpaired	1	6	3	.625
	Single	1	2	7	.070
2 Total		**3**	**12**	**15**	**.008**
3	Paired	0	6	4	.125
	Unpaired	1	7	2	1
	Single	1	3	6	.125
3 Total		**2**	**16**	**12**	**.013**

Table 3.10 Frequencies and statistics of the sign comparisons of given VS expected number of completely incorrect answers, for each interval, for each group. In bold the significant results.

Count		Completely Incorrect			Binomial P
Interval	Group	Giv<Exp	Giv=Exp	Giv>Exp	
1	Paired	1	5	4	.375
	Unpaired	1	5	4	.375
	Single	1	5	4	.375
1 Total		**3**	**15**	**12**	**.035**
2	Paired	3	3	4	1
	Unpaired	2	6	2	1
	Single	3	1	6	.508
2 Total		8	10	12	.503
3	Paired	3	6	1	.625
	Unpaired	3	6	1	.625
	Single	4	2	4	1
3 Total		10	14	6	.454

Table 3.11 Frequencies and statistics of the sign comparisons of given VS expected number of correct answers, for each interval, for each group.

Count		Correct			Binomial P
Interval	Group	Giv<Exp	Giv=Exp	Giv>Exp	
1	Paired	1	5	4	.375
	Unpaired	0	7	3	.250
	Single	1	5	4	.375
1 Total		**2**	**17**	**11**	**.022**
2	Paired	4	1	5	1
	Unpaired	3	4	3	1
	Single	2	4	4	.687
2 Total		9	9	12	.664
3	Paired	2	3	5	.453
	Unpaired	2	4	4	.687
	Single	3	2	5	.727
3 Total		7	9	14	.189

As shown in the 3.8, 3.9, 3.10 and 3.11, the comparisons between given and expected number of answers are similar for each group. Because the size of each group was too small to give significant results separately, and there was no reason to keep them separate, a joint analysis was performed (the results are the totals in the tables).

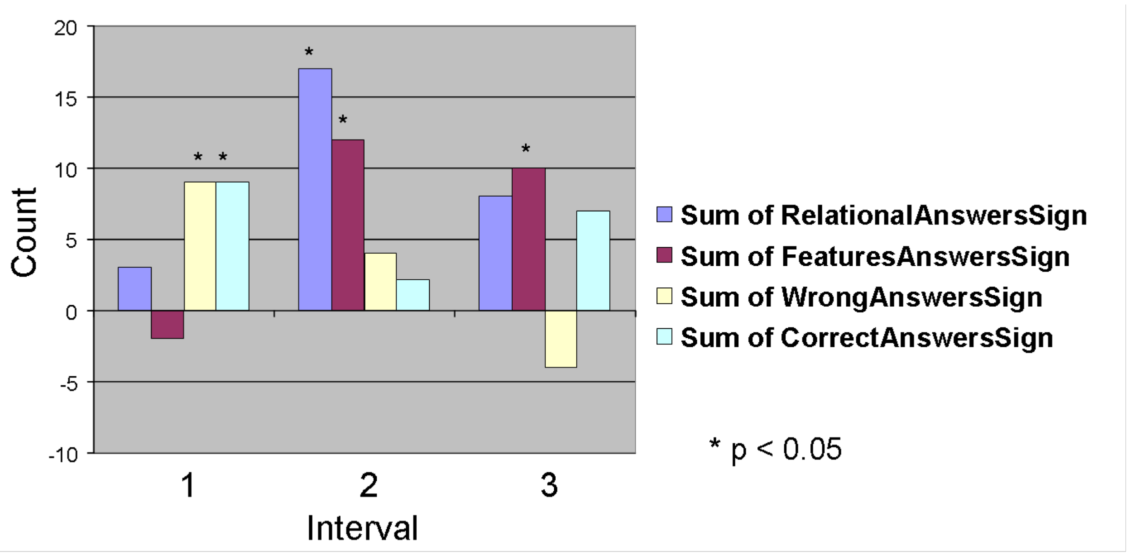


Figure 3.12 Sums of the sign comparisons of given VS expected number of answers for each region of the contingency table, for each interval, joint analysis.

Sum = count (given > expected) - count (given < expected)

As shown by 3.8, 3.9, 3.10, 3.11 and Figure 3.12, the sign tests between given and expected number of answers give these results:

- In the first interval (0 - 0.333) both the completely wrong answers and the correct answers occurred more often than expected. This is an unexpected and unexplainable result. The number of given answers in the remaining regions of the contingency table (Figure 3.11) are not statistically different from what randomly expected.

- In the second interval (0.333 - 0.666) the incorrect relational answers occurred more often than expected and so did the incorrect features answers. Both the completely wrong answers and the correct answers are not different from what randomly expected.

- In the third interval (0.666 - 1) only the incorrect features answers occurred more often than expected.

The pattern that emerges is of particular interest, depicting learning dynamics that have a middle period of "partial learning" for both the relational and features categories. In fact, the mistakes across similar categories occur more frequently than would be expected if these mistakes were random, while the mistakes across dissimilar categories (i.e. between different kinds of categories) are not different from what would be expected if these errors were random.

This result supports the prediction that before learning is complete, any errors in categorisation are not random but they are more frequent across similar categories. In turn, this supports the hypothesis 2b, that learning consists of two phases where partial categories are first formed and then refined to create the final categories.

Of course, structure mapping could provide an alternative explanation for the fact that mistakes across similar categories occur more frequently than expected if these mistakes were random, while mistakes across dissimilar categories (i.e. between different kinds of categories) are not different from what expected if these errors were random. Although it is possible that simultaneously learned categories are aligned and compared using structure mapping, we saw that direct comparison was not a strong factor to help the learning of similar categories, because the paired group didn't need significantly less time and exemplars than the unpaired group, and it did not even differentiate the performance compared to the single group. Therefore, the small effect size of the simultaneous presentation factor reduces the likelihood of the hypothesis that categories are continuously compared and mutually aligned using structure mapping.

According to these results, it is more likely that a partial "draft" of a common classification criterion is first found and then refined to produce the final classification criteria. The learning process would therefore be a continuum of subsequent refinements and adjustments, from some very rough drafts, to a better defined one (able to discern between one kind of category and the other kind) to the final criteria. As will be explained in detail in the model chapter (Chapter 6), from this continuous process of refinement and modification could emerge a form of analogical reasoning different from structure mapping.

3.3.4. Learning order

To test if learning of the relational categories was simpler or more difficult than learning of the features categories, a simple method was used. For each participant, the kind of the first learned category was recorded.

Of the 30 participants in the actual test, 14 learned first a features category, 16 a relational category; no significant difference was found. Also a Chi Square analysis of category kind vs. group wasn't significant (χ^2 (2) = 2.079, $p > 0.05$).

As proposed above, there are various possible explanation for this null result. For example, the features categories could also be learned based on a predicate. Or the complexity of the task could make the participants assume that the categories are defined by complex rules. Or that for the relational categories, the advantage given by the similarities compensates the disadvantage of the more complex criteria to be found.

3.3.5. Participants' debriefings

A last but not less important source of information are the debriefings written by the participants after solving the test. They were asked which method they used to solve the task, and many of them indicated the following aspects[2]:

- Using rules (27 participants)
- Testing and rejecting hypotheses (23 participants)
- Writing previous exemplars on notepad (18 participants)
- Writing hypothesized rules on notepad (7 participants)
- Once a criterion was found, look for similar criteria (5 participants)

Although this is an informal and qualitative analysis, some important knowledge can be deduced:

[2] Some examples of participants' debriefings:

- "First write down colours and shapes, then try to find a rule by looking for similarities. As one rule was violated incorporate number of shapes as well. Find a common rule and test it."
- "Began by looking for family resemblances, eg. tending to have blue crosses, tending to have purple circles, etc. Noticed then that one green square perfectly identified a category and began to look for similar rules."
- "I wrote down attributes of the cases, and tried to find common factors. After I postulated a hypothesis I was able to test it. I assumed my hypothesis to be correct until proven wrong."
- "Initially counted distinct shapes in each diagram, which yielded clues to groups A and B. After I was sure of these but could not see a numerical pattern to C and D I started to examine the particular colours present in the diagrams for each group and noticed the presence of the purple or yellow circle in both."

- The type of categorization used in this task is rule-based and not exemplar-based (or prototype-based).

- A scientific method of trial and error is used.

- It was a memory-intensive task, and the notepad was used to remember the previously seen exemplars.

- The formulation of hypotheses reduces the memory load. In fact, the notepad is used much more to write previous exemplars than to write rules. A rule is a compact way to store information, and needs to be written down less than a quantity of exemplars.

- Very rarely analogical reasoning was consciously and openly used (see point 5, above), to transfer knowledge between categories. Nevertheless, the other results show that structural similarities are of help. Therefore analogical reasoning is used in some way different from the "canonical" structure mapping.

3.4. Summary

Of the four initial questions, only three can be clearly answered by the results of this experiment.

Two results proved that analogical reasoning is used during learning of novel categories with similarities between them. In fact, the learning of one category is quickly followed by the learning of the other similar one, both for the relational and the features categories.

The fact that they are related means that the two learning processes help each other, or even that they are in fact just a single process.

This last hypothesis is further supported by the presence of a middle phase of "partial learning" of the categories. In this phase the actual categories are not clear yet, but the mistakes are made more often with the other similar category than with the other two dissimilar ones. The confusion between similar categories could be explained by the formation of a partial draft of a common classification criterion. This would allow to distinguish one kind of categories from the other kind, without knowing the differences between the individual categories.

If the structure mapping theory could possibly explain these two results, the small effect size of the group factor (i.e. the direct comparison of exemplars) is on the contrary an indication that structure mapping is probably used little if not at all. Otherwise the group that had the opportunity to directly compare exemplars of similar categories should find easier to solve the test than the other group(s). Therefore an explanation that does not require structure mapping is preferable, and will be proposed in the model chapter (Chapter 6).

The presence of relational similarity between two categories did not appear to make their learning easier nor more difficult than the other two. But as already stated in section 3.1.4, there are a number of reasons why this could happen. The most important of these being the fact that the features categories could also be learned based on a predicate. In this case, they would have relational similarities as well.

One limitation of the present experiment is that it could have forced people to notice analogies and use relational similarities. In fact the participants were forced to learn categories with similarities between them; they had no alternative. Therefore they were strongly influenced to use analogies, even if in normal circumstances they wouldn't have used them. This "forced choice", needed in the design, could have biased the experiment in the direction of the use of analogies. In short, we could have discovered the use of analogies just because it was an "analogical experiment".

For these reasons, another experiment with a different design is required. This new experiment must give the participants the opportunity whether or not to use analogies.

Chapter 4

Experiment 2

4.1. Introduction

As in the previous experiment (Chapter 3), this experiment tests the hypothesis that analogy can be established between simultaneously-learned categories with similar structures, to aid the learning of both categories.

Given that in the previous experiment the direct comparison of items of similar categories didn't help (nor hinder) learning, and that there were results showing that learning is split in two phases, this second experiment will not take into account the hypothesis 2a, that categories are first learned and then compared and mutually aligned using structure mapping, and will on the opposite further investigate the hypothesis 2b, that learning consists of two phases where one or more partial categories are first formed and then refined, so from each partial category stem some final categories.

From these hypotheses a series of predictions can be made. The predictions tested in this experiment are:

1. Even if given alternative solutions, people find solutions which have similar structures rather than different structures. To give an example, let's consider two categories each defined by two alternative criteria. For each category only one of the two alternative criteria needs to be discovered in order to solve the test. Of this

two criteria, one (the "analogical criterion") is structurally similar to the “analogical criteria” of the other categories. The other (the "unrelated criterion") is instead unique to that category. Let's give some instances. All the exemplars of a category (“Dual 1”, defined by two alternative criteria) feature (at least) two groups of elements. When we click on each element of the first group, the same piece of music (Music 1) is played, while when we click on each element of the second group, all the elements of the group (included itself) will rotate at the same time. All the exemplars of another category (“Dual 2”, defined by two alternative criteria) also feature (at least) two groups of elements. When we click on each element of the first group, a certain piece of music (Music 2) is played, always the same, but different from the piece of music featured on category “Dual 1”. When we click on each element of the second group, all the elements of the group (included itself) will jump in turn. So we can infer that the first criterion for category “Dual 1”, defined from a certain piece of music being played, is structurally similar to the first criterion for category “Dual 2”, defined by a another piece of music being played, although the music is different across the two categories. The other two criteria (the rotating at the same time for “Dual 1” and the jumping in turn for “Dual 2”) are structurally less similar both to each other and to the music criterion. In the experiment we need to see which criterion (analogical or unrelated) is found for each category.

2. Learning of similar categories is related. For each dual category, people will find just one criterion out of the two, which could be the structurally similar criterion (analogical, i. e. the piece of music) or the structurally dissimilar criterion

(unrelated, i. e. the others). People who find the structurally dissimilar criteria cannot have any facilitation due to analogy, therefore will not be included in this analysis. We thus consider only those who find the structurally similar criteria (in the previous example, music). For these people, we expect that time between learning of similar categories is less than time between learning of dissimilar ones. That is, time elapsed between the learning of categories “Dual 1” and “Dual 2” is less than the time elapsed between, for example, the learning of category “Dual 1” and a third category defined by a single “unrelated” criterion, which consists, for example, in all the elements doing different things, all together.

3. Before learning is complete, considering only people who find the structurally similar criteria (in the previous example, music), any errors in categorisation are not random but they are more frequent across similar categories. For example, errors across categories “Dual 1” and “Dual 2” are more frequent than errors across, for example, category “Dual 1” and the third category (as in the example given in the previous paragraph).

The first prediction tests if analogy really helps learning, and therefore categorization criteria with similar structures are found more easily. The third prediction is directly related to the hypothesis that learning is split in a first phase of partial learning, followed by a phase of refinement. All of the predictions are based on the hypothesis of the early use of analogical reasoning.

In particular, in order to test the first prediction, some categories in the experiment must be defined by alternative criteria, one of which is similar between some categories. The null hypothesis is that people usually find all dissimilar criteria, even when given the alternative to find similar criteria.

To test the second prediction, the experiment must compare how long it takes to learn each category. The null hypothesis is that the learning of each category is independent.

To test the third prediction, the pattern of answers and errors before learning is considered to have occurred, must be recorded and analysed. The null hypothesis is that those answers are randomly distributed.

In order to perform all these analyses, various constraints were considered which served to guide me in the design of the experiment. The following subsections illustrate these constraints and the solutions proposed.

4.1.1. Experience from previous experiment

This experiment is based on the experiences of Experiment 1, and takes into account some issues discovered in the previous experiment. In particular, I wanted to give the participants the freedom to use or not to use analogies. In fact, in the previous experiment the participants were forced to learn categories with similarities between them; they had no alternative. Therefore they were strongly influenced to use analogies, even if in normal circumstances they wouldn't have used them. If the categories were defined by two alternative criterion, one "analogical" and the other not, the participants would have a choice. It would be interesting, then, to see which criterion they discover.

Given the null result from the previous experiment about direct comparison, the simultaneous presentation factor is abandoned. Since there is no longer any need of a particular presentation order, in this experiment the presentation of exemplars is just randomized.

Finally, many participants of the previous experiment reported that after a while they had problems to concentrate, because it was very boring. Therefore this new experiment needed to be more entertaining, and possibly to resemble a game.

4.1.2. Alternative solutions

As already stated above, two alternative solutions must be available to the participants. One solution involves the use of criteria with similarities between them, across different categories. For instance, given two exemplars of two different categories, when we click on some elements of the first one, the same piece of music (Music 1) is played, and when we click on some elements of the second one, another piece of music (Music 2) is played. The alternative solution, in contrast, has different criteria for each different category. For instance, given two exemplars of two different categories, clicking on some elements of the first one will elicit a reaction (the elements rotate at the same time), while clicking on some elements of the second one will elicit a completely different reaction (the elements jump in turn). Therefore, some categories (which will be called "dual categories") must be defined by two alternative criteria, only one of which needs to be discovered in order to solve the test. Of this two criteria, one, which will be called the "analogical criterion", is structurally similar to the analogical criteria of the other categories. The other "unrelated criterion" is instead unique to that category.

4.1.3. Entertainment

A number of analogical reasoning experiments use static figures (e.g. Goswami & Brown, 1990a, 1990b; Kokinov, Bliznashki, Kosev, & Hristova, 2007; Lipkens & Hayes, 2009; Rattermann, Gentner, & DeLoache, 1990; Thibaut, French, & Vezneva, 2008), yet the limits of such approach are evident, considering both the little ecological plausibility they have, and the little interest and poor attention they cause in participants.

To overcome these issues, interactivity is introduced, in order to make the test more entertaining and to make it resemble a game. As a result, it is also more ecologically plausible and probably easier to solve. It allows also for more complex definitions of categories and similarities between them. A task with interactive exemplars, in fact, introduces, among other things, cause and synchronicity relations (for example, when we click on some elements a piece of music is played, or the elements react with an action all together, or a tone is played in synchronicity with a movement of the elements), which are essential in our everyday life.

Also for the sake of entertainment and to be more game-like, a cover story must be used to explain the task (the full text is available in Appendix D). The story I chose involves a toy factory that was experimenting a new machine to produce interactive toys. In the cover story, this machine unfortunately didn't work well. It didn't label the toys, and some of them had their circuits broken (see below about this last statement). Thus the factory hires an expert (the participant) to discover how to classify the toys. The participant has to study the toys by clicking on their composing elements to elicit some reactions, which provide the key to classifying the toys. In contrast to the previous experiment, in this experiment shapes and colours are meaningless. The participant can try different labels until the correct label is found for each toy. When the participant correctly classifies at first try enough toys for each kind, the test ends.

4.1.4. Avoid elimination

In order to be sure that all the categories are learned and participants don't answer by elimination, a last "residual" category must be introduced. This category doesn't have a definition of its own, since its exemplars will be composed only by distractor elements. It is exactly this lack of categorization criterion that defines this category, which will be chosen when all the criteria for the other categories fail. In order to make the task clear and simple, the name for this category must be exemplary of its definition. The name "Wrong" is chosen.

4.1.5. Insight of learning process and learned rules

Having some categories defined by two alternative criteria, the problem arises of how to discover which criterion is found. Fortunately, having exemplars with interactive elements, it is possible to record all the interactions (namely: clicks) of the user with the elements. If some elements pertain to one criterion and some other to the alternative criterion, it is easy to discern with which criterion the participants are interacting. The "analogical group of elements" is the set of all the elements of that exemplar pertaining to the "analogical criterion". In the same way, the "unrelated group of elements" is the set of all the elements of that exemplar pertaining to the "unrelated criterion".

Two arrow buttons are provided, to navigate back and forth through the previous exemplars. This new method provides also the opportunity to see how often the participants go back to check the previously answered exemplars, and thus how much use is made of deductive reasoning.

4.2. Method

The experiment described in this chapter is a rather peculiar category learning test: its peculiarity lies in the fact that during the test the participants are shown the exemplars and immediately after they are told if their answer is right or wrong, which makes this experiment different from the usual category learning tests.

4.2.1. Participants

The test was administered to an heterogeneous set of participants. They were both from inside and outside the university, of various ages and educations, and the test was performed in different environments with different levels of supervision. All these factors were recorded and analysed to test for differences in difficulty, but no significant difference was found. The choice to extend the set of participants was done to be able to generalize the results, because in the first experiment the participants were only from Computer Science.

Participants were 28 volunteers randomly chosen: 11 Males and 17 Females, average age 25.7 (SD 9.8). They were randomly assigned to the three groups (U_1, U_2, U_3 - see below), 9 for U_1 and U_2, 10 for U_3.

4.2.2. Materials

The current experiment was carried out on various personal computers equipped with headphones, colour screen and mouse. With the exception of a factor described below, the three groups of participants had the same test. During the test, each participant was shown 4 exemplars per screen. On the left of each screen they could see 5 stacks of draggable labels (corresponding to the 5 categories) with invented names: Babed, Didom, Golev, Liset (these names have been invented to be as much equivalent as possible, in terms of phonetic complexity) and Wrong, which is indeed a non-category, as explained later in this chapter. Inside each exemplar they could find an area where to stick the labels. In the screen there was also a timer which showed the time elapsed from the beginning of the experiment.

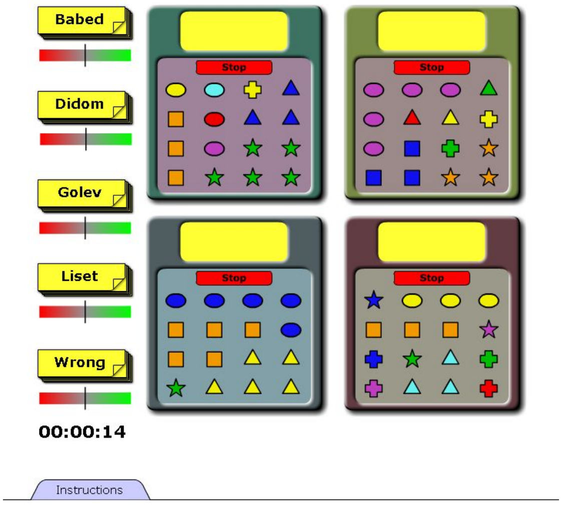


Figure 4.1 A screen from a test just started. Four exemplars are shown, with on the left five "stacks" of labels which can be dragged on the exemplars.

The participants' task was to learn to correctly classify the exemplars shown: to classify an exemplar, they had to click on the labels on the left and then drag them onto the exemplar. Once each label was dropped onto each exemplar, a feedback was immediately given on the correctness of the answer. If the answer was wrong, a red cross would appear on the label; if it was right, a green tick would appear on a side of the label. In case the answer was wrong, the participants had to keep trying with other labels, until they found the correct one. When all the 4 exemplars were correctly answered, the participants could proceed to the next screen (as explained in the procedure section) with an arrow button at the bottom of the screen. Another arrow button could be used to go the previous screen(s). Below each label there was a coloured bar with an indicator for the level of learning achieved for the corresponding category (i.e. the moving average of correct answers given for that category). The exemplars looked like rectangles of a neutral colour, which contained 16 coloured elements arranged in a 4 by 4 grid and area for the labels to stick onto. As in the previous experiment, the elements could be of 5 different shapes (cross, star, ellipse, square and triangle) and 7 different colours (yellow, orange, red, light blue, violet, blue and green), with a total of 35 possible combinations. 4.1 summarizes the kinds of classification criteria used in this experiment.

Table 4.1 Simplified examples of the classification criteria for a participant assigned to group 1 (see below for the explanation of the groups).

Category	Criterion 1	Criterion 2
Dual 1 (e.g. "Babed")	Click on any of the four blue triangles -> Music *#1* is played	Click on any of the three green circles -> All green circles do the same (randomly chosen) movement (e.g. jump) *at the same time*
Dual 2 (e.g. "Golev")	Click on any of the five red circles -> Music *#2* is played	Click on any of the four yellow squares -> All yellow squares do the same (randomly chosen) movement (e.g. rotate) *in turn*
Single analogical (e.g. "Liset")	Click on any of the three blue squares -> Music *#3* is played	
Single unrelated	Click on any of the three yellow	

(e.g. "Didom")	circles -> All yellow circles do *different randomly chosen* movements (e.g. one fades, one changes color, one rotates) *at the same time*	
Residual ("Wrong")	Click on some of the four red squares -> *Some randomly chosen* elements do some random movement (e.g. a red square jumps, a blue circle rotates, a green triangle jumps, etc.)	

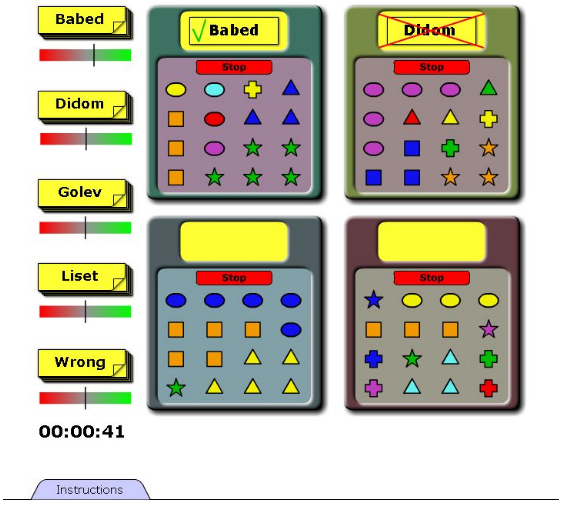


Figure 4.2 Two exemplars are labelled, and feedback (correct answer and wrong answer) is given.

In order to learn to correctly classify the exemplars, the participants had to click on the elements inside them. When clicked, the elements could elicit a reaction, which could be of three different kinds: 1. a music or a tone could be played with a particular instrument; 2. the elements could move; 3. both. The movements that the elements could make include jumping, rotating, zooming, blurring, a change of shape or colour, etc. For some actions (like zooming or jumping, ...) the volume of the played tone could be also synchronized with the movement. On the one hand, this helped make the test more entertaining for participants, because it resembled a game; on the other hand, that allowed an insight on the reasoning processes of participants. In fact every single click was recorded in order to be later analysed.

Each exemplar contained 3 groups of elements: each group was formed of elements of the same shape and colour, adjacent to each other. The number of elements for each group varied between 3 and 5. The quantity was randomly chosen, as were the shape, colour and position of each group inside the exemplar, and they varied even between exemplars of the same category during each test. Therefore to sum up to 16 elements for each exemplar a certain number of distractors, of unrelated shapes and colours, were added. For instance, an exemplar of Golev category could be composed of 5 blue crosses, 5 orange triangles, 5 green stars and 1 blue square. Another exemplar of the same category (Golev) could be composed of 5 yellow triangles, 4 blue squares, 5 red stars, 1 orange cross and 1 violet cross. Clearly, neither the shapes nor the colours nor the number of the elements could help to classify the exemplars, which makes this experiment different from Experiment 1. In fact shapes, colours and quantities are randomly chosen, so the key resides just in the reactions.

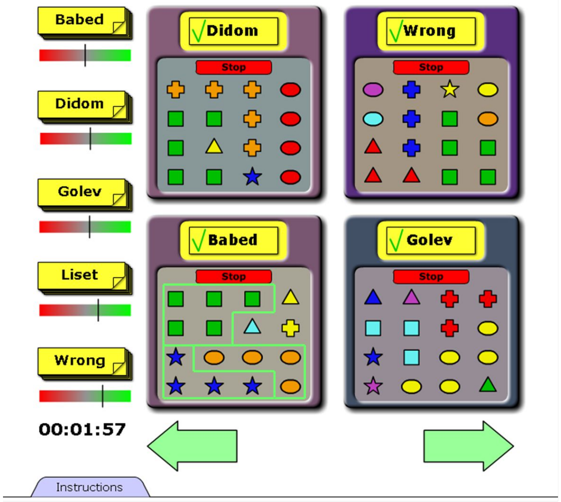


Figure 4.3 Some of the 16 elements form 3 groups of same shape and colour, the remaining elements are distractors. The lines show the three groups of elements in one exemplar of the Babed category.

In this experiment there are 4 different categories. Two of them are "dual categories" because they have 2 defining criteria, the other two are instead "single categories" because they have just 1 defining criterion. We have a total of 6 classification criteria: of these criteria, 3 have a similar structure (we shall call them "analogical"), whilst the other 3 have structures dissimilar from the former ones and also between them (we shall call them "unrelated"). All of the 3 analogical criteria are defined by this rule: when participants click on each element of the corresponding groups a music is played. For each criterion it is always the same piece of music, but it's different for criteria A_1, A_2 and A_3 (i.e. criterion A_1 has piece #1, A_2 has #2, A_3 has #3). There are also other pieces of music present in the test, which are sometimes played when clicking on distractors. So the structural commonality between the three analogical criteria lies in the fact that a music is played, whilst the difference is in the piece of music which is played. The 3 distinct pieces of music are chosen among 12 different ones and assigned randomly to each participant at the beginning of the test. The remaining 9 pieces of music are instead used by the distractors. To give some instances, let's reconsider the example which was made before in this chapter. An exemplar of Golev category is composed of 5 blue crosses, 5 orange triangles, 5 green stars and 1 blue square; another exemplar of the same category (Golev) is instead composed of 5 yellow triangles, 4 blue squares, 5 red stars, 1 orange cross and 1 violet cross. When participants click on each blue cross of the former exemplar, the same piece of music is played as when each yellow triangle of the latter is clicked. So we can infer that these exemplars belong to the same category (Golev, in this case): the structural commonality between the two exemplars lies in the fact that, when the elements of one of the three groups composing each exemplar are clicked, they elicit the same reaction (i.e. play the same piece of music). The elements belonging to the other groups (i.e. the 5 orange triangles and the 5 green stars for exemplar 1; the 4 blue squares and the 5 red stars for exemplar 2) show instead distinct reactions, which are different from each other and from the one used for the classification criterion. Also some of the distractors (in the example, the blue square, the orange cross and the violet cross) elicit distinct reactions, but their behaviours are different from the one used for the classification criterion.

Each of the unrelated criteria (U_1, U_2 and U_3) is instead defined by a distinct kind of synchronicity between the reactions of the elements of the group, as follows:

U_1. When participants click on each element of the group for the U_1 criterion the entire group react with the same action (randomly chosen) at the same time. For instance, when an element of the group is clicked, all the elements of the same group (included itself) will rotate at the same time; when another element of the same group is clicked, all the elements of the same group (included itself) will blink at the same time, and so on.

U_2. When participants click on each element of the group for the U_2 criterion the entire group react with the same action (randomly chosen) in turn. For instance, when an element of the group is clicked it will zoom, then another element of the same group will zoom at its turn, then another element will zoom and so on until they start back again. Even in this case, when a different element of the same group is clicked, the elicited reaction will be different: for instance, the element will blink and after it all of the elements of the group will blink in turn.

U_3. Finally, when participants click on each element of the group for the U_3 criterion all the elements of the group perform different actions (randomly chosen) at the same time. For instance, when an element of the group is clicked, all the elements of the group (included itself) will react: one element will change its colour, another one will fade, another one will jump, and so on. As in the previous cases, for each element of the same group which is clicked, the actions performed will be yet different for every element of the group.

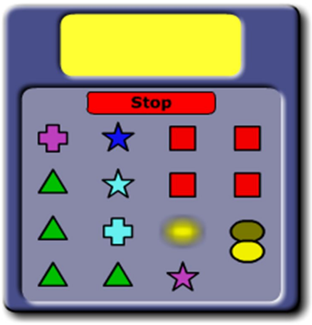


Figure 4.4 One element of the group with the U_3 criterion (in this case, the yellow ellipses) was clicked, and all the elements of the group perform different actions at the same time.

Resuming, we have one category which can be classified using one of the analogical criteria (A_1) and/or one of the unrelated criteria (U_1); another category can be classified using another analogical criterion (A_2) and/or another unrelated criterion (U_2). A third category instead must be classified using the third unrelated criterion (U_3), and the last category must be classified using the third analogical criterion (A_3). For each participant, the assignment of a criterion to a category is random and just obeys one rule: for dual categories (which have 2 defining criteria), one of the criteria has to be of the analogical type, whilst the other criterion must be of the unrelated type. So participants can get to classify the dual categories by learning the analogical criterion or the unrelated criterion (or by learning both). Whilst as for single categories (which have just 1 defining criterion), participants are obliged to classify them by learning their single criterion.

As we have already seen in the example above discussed, in each exemplar shown in the test there are 3 groups of elements. Of these, 1 or 2 groups of elements are related to some classification criterion (we shall call them "active groups"), while the remaining group(s) of elements are not related to any classification criteria (we shall call them "non-active groups"). The groups of elements in the exemplars are always independent from each other. To each "active" group of elements is randomly assigned one classification criterion, so there is a one-to-one correspondence between groups and criteria. Thus, in a dual category two groups are "active", one for each criterion, and one group is "non-active"; in a single category only one group is "active" (and corresponds to one criterion) and two groups are "non-active". In addition, as stated before, there is also a variable number of distractors, in order to sum up to 16 elements for each exemplar. Some of the remaining non-assigned elements, which are randomly chosen, are given random behaviours. However, these behaviours are always different from the ones used for the classification criteria. For instance, an exemplar of Liset category is composed of 5 blue stars, 5 orange crosses, 5 green triangles (3 groups) and 1 violet square (distractor). In this case, for example, the blue stars are the “active” group, because when participants click on each blue star the same piece of music is played, which is the reaction related to the classification criterion (analogical); the orange crosses and the green triangles are instead the “non-active” groups, because the elements belonging to these groups show distinct reactions, which are different from each other and from the one used for the classification criterion. Finally, the violet square is just a distractor element and also its behaviour differs from the one used for the classification criterion.

In order to be sure that all the categories are learned and participants don't answer by elimination (in fact, once participants have learned 3 of the 4 categories, they could correctly classify the remaining category by elimination), a last category has been introduced, named “Wrong”. This category doesn't have a definition of its own, since its exemplars will be composed only by distractors. It is exactly this lack of classification criterion that defines this category, which will be chosen when all the criteria for the other categories fail. In order to make the task clear and simple, the name for this category must be exemplary of its definition, which explains the name “Wrong”.

So the introduction of this “non-category” obliges participants to learn the classification criterion of each category (or, for dual categories, at least one of the 2 possible classification criteria). This non-category serves also another purpose. Given the complexity of the criteria, it would be almost impossible to be sure that the task isn't solved using some "shortcut", instead of finding the full criterion for each category. But this non-category has elements with behaviours similar to the ones defining the actual categories. Therefore, without learning the complete criteria for the other categories, a "Wrong" exemplar could be mistaken for an exemplar of another category.

For each participant, the assignment of a category to a label (Babed, Didom, Golev or Liset) is random, except for the non-category, which is always "Wrong".

The participants can solve the task in different ways. If, for example, they find rules A_1, A_2, A_3 and U_3, it can be inferred that these criteria are more easily learned because the structural similarities between A_1, A_2 and A_3 help learning through analogy. Actually, A_3 and U_3 have to be necessarily learned because each of them is the only defining criterion for each of the "single" categories. On the other hand, U_1 and/or U_2 could be discovered instead of A_1 and/or A_2, showing that similarities can create confusion.

Table 4.2 Example of one of the three possible allocation of criteria to categories. For dual categories, participants can find the analogical criteria and/or the unrelated criteria.

Dual category 1	Dual category 2	Single analogical category	Single unrelated category	Residual category
(e.g. Babed)	(e.g. Golev)	(e.g. Liset)	(e.g. Didom)	(Wrong)
$A_1 + U_1$	$A_2 + U_2$	A_3	U_3	Only distractors

4.2.3. Design

Three distinct groups of participants are created and the assignment of participants to each group is random. The test is the same for each of the three groups, except for the criterion used for one of the two single categories. One of the single categories, indeed, is always defined by an analogical criterion (A_1, A_2 or A_3). The distinction between A_1, A_2

and A_3 is merely casual, therefore these three criteria are interchangeable. On the contrary, the other single category is defined by an unrelated criterion (U_1, U_2 or U_3). Since the criteria U_1, U_2 and U_3 are much different between them, the difficulty of the test could vary according to which criterion is assigned to the single category. That is why three groups of participants are created, one group with the U_1 criterion defining one of the single categories, the second group with the U_2 criterion defining one of the single categories, and finally the third group with the U_3 criterion defining one of the single categories. So the presence of all the possible equivalent dispositions ensures that, even if there are differences due to the different criteria, they are balanced by the design. This design also allows estimate the resulting difficulty of the test for each unrelated criterion. Besides, as stated above, at the very beginning of the test, for each participant is randomly assigned the correspondence between name of categories ("Liset", "Golev", "Babed" and "Didom") and kind of categories (dual or single, defined by A_1, A_2, A_3, U_1, U_2, U_3), except for the non-category which is always "Wrong". For instance, a participant from group 1 will have A_1 and U_3 criteria assigned to the category named "Babed", U_1 criterion assigned to the category named "Didom", A_3 assigned to "Golev", A_2 and U_2 assigned to "Liset". Another participant from group 1 will have U_1 criterion assigned to the category named "Babed", A_2 and U_3 assigned to "Didom", A_1 and U_2 assigned to "Golev" and A_3 assigned to "Liset".

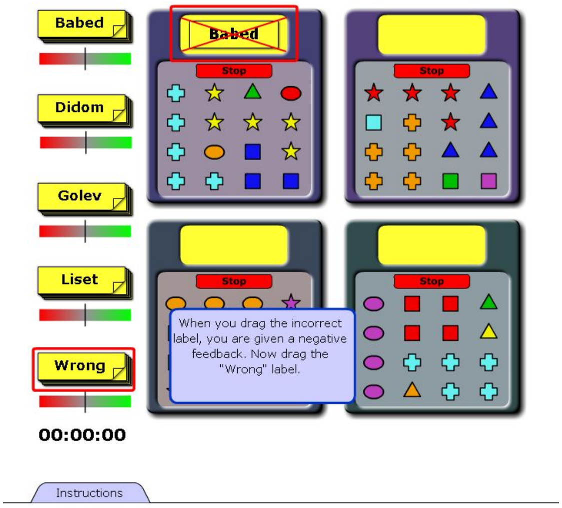
Table 4.3 Summary of the allocation of criteria to categories, for each group.

Group	Dual category 1	Dual category 2	Single analogical category	Single unrelated category	Residual category
1	$A_1 + U_2$	$A_2 + U_3$	A_3	U_1	Only distractors
2	$A_1 + U_1$	$A_2 + U_3$	A_3	U_2	Only distractors
3	$A_1 + U_1$	$A_2 + U_2$	A_3	U_3	Only distractors

4.2.4. Procedure

Before starting the very test, participants are shown a screen where they are told to put their headphones on and adjust the volume until they can clearly hear a music played by the program. This ensures that participants have access to all of the multimedia features. When participants click on the “Next” button, they are shown the instructions to the test in the form of a cover story, which explains the task in a more entertaining, game-like way (the full text is available in Appendix D). The story I made up involves a toy factory that was experimenting a new machine to produce interactive toys. In the cover story, this machine unfortunately didn't work well. It didn't label the toys, and some of them had their circuits broken (see below about this last statement). Thus the factory hires an expert (the participant) to discover how to classify the toys. The participant has to study the toys by clicking on their composing elements to elicit some reactions, which provide the key to classifying the toys. As stated above, in contrast to the previous experiment, in this experiment shapes and colours are meaningless. The participant can try different labels until the correct label is found for each toy. When the participant correctly classifies at first try enough toys for each kind, the test ends.

After given these instructions, participants are shown a tutorial in which they are told how to use the test interface. Popup instructions guide them to actions like closing or re-opening the instructions tab, clicking on the interactive elements, dragging the labels on the exemplars, recognizing and distinguishing between positive and negative feedbacks, understanding that all of the 4 exemplars shown have to be answered, using the “Next” and “Back” arrow buttons (which will be explained below), etc.



When you drag the incorrect label, you are given a negative feedback. Now drag the "Wrong" label.

Figure 4.5 Popup instructions are given during the tutorial.

Once the tutorial is over, the very test begins. Four exemplars at a time are shown on the computer screen, in random order. Five stacks of labels with the names of the 4 categories ("Babed", "Golev", "Liset", "Didom") and the name "Wrong" (for the non-category) are available on the left of the screen, one stack for each name. On the very first screen are presented all of the 4 categories (i.e. no wrong exemplars) in random order. In the subsequent 2 screens (i.e. 8 exemplars) are presented, on the whole (i.e. randomly arranged), one exemplar for each category, and 4 wrong exemplars. Then the presentation of categories is balanced, in order to present the same number of exemplars for each category, plus a random number of wrong exemplars, every 4 screens.

To give an answer, the participant must drag a label over an exemplar. If the given answer is correct, positive feedback is given. Otherwise negative feedback is given, and the participant must try with another label, until the correct answer is found. When all of the currently shown exemplars are correctly answered, a "Next" arrow button appears, to proceed to the next four exemplars. As stated above, a "Back" arrow button is always available (except for the very first screen) to go back to the previously seen (and correctly answered) exemplars.

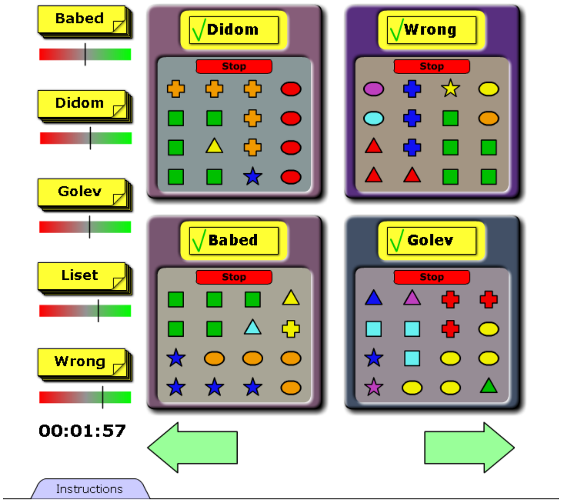


Figure 4.6 When all the exemplars are correctly answered, a "Next" arrow button appears. The "Back" arrow button is present from the second screen, to go back to the previously answered exemplars.

The test ends when the participant gives 3 correct answers (at the first attempt, i.e. without trying again) in a row for each category. This criterion is different from the one used in the previous experiment, but ensures the same low probability of random solution. It is chosen to minimize the length of the test, and therefore the boredom and probability of distraction mistakes.

At the end of the test a debriefing question asks participants to write a report about the classification criteria (to teach a worker to continue the job, according to the cover story). The test was implemented in Macromedia Flash and all interactions were timed with an accuracy of 1 millisecond and were recorded on a server. Recorded clicks and answers were recorded locally and sent to the server in batches, in order to avoid the network latency.

4.2.5. A typical test session

A typical test session works very similarly to the first experiment. In the beginning the participant is forced to answer randomly, and probably gives the wrong answers. So they have to keep answering until they find the correct answers. Then they can go on to try to answer the next exemplars, but they can also go back to compare the new exemplars to the previous ones.

Some people extensively used these comparisons with previous exemplars, to the point that they could extract all the needed information just seeing a few exemplars for each category. Many people in contrast just ignored this opportunity and just kept going on, relying on their memory. Two distinct strategies can be therefore imagined for these two kinds of people (although it is a gradient, not a sharp division). The former kind uses a more "scientific" strategy based on the falsification of hypotheses, while the latter uses instead a strategy based on reinforcement.

As in the previous experiment, after a while the participant noticed some patterns, and started to formulate hypotheses about the classification criteria. During this particular stage, the answers the participant gave were sometimes correct and sometimes incorrect, but they were no longer random, instead they followed their hypothesized rules. It is therefore interesting to analyse the errors made during this period, in order to better understand the type of confusion encountered.

At some point, the participant started to give systematically the right answer to some category, and that is the point we call "learning point" for that category (see below for the algorithm).

The learning processes for the different categories can be unrelated (null hypothesis) or related. In particular, the learning of the three analogical categories (i.e. A_1, A_2 and A_3) can be related (learning one can help learning the other), when the found criteria are the analogical ones.

After a criterion is found for a category, for all the subsequent exemplars the participant starts to click randomly until he finds the group of elements corresponding to that criterion, then he gives the correct answer. Therefore the last click before answering is with high probability on an element having the found criterion. The analysis of these last clicks can reveal which criterion (for the dual categories) the participants find.

After learning the classification criteria for all the four categories which have classification criteria, using exclusion the participant can correctly classify also the "Wrong" category. At this point the participant quickly reaches the learning criterion for each category, and the test ends. At this stage all four learning points are defined, and it is possible to look at the intervals between them, to determine if the learning of a category is related or not to the learning of the other ones.

4.2.6. Estimation of learning points

As in the previous experiment, many of the results are based on the estimation of learning points. So it can be useful to briefly present the algorithm used for their estimation (Figure 4.7). It is very similar to the one used in the previous experiment.

Learning points are estimated independently for each category, using a moving average of the correct answers to the shown exemplars of that category. As said above, participants are considered to have learned a category when they give at least 3 correct answers in a row at the first attempt for that category. The learning point for a category is therefore the first correct answer after which the participant doesn't make mistakes anymore (for that category).

```
for each answer for the given category, starting from the end {
        calculate a moving sum of length 3 of the correct answers (at first
        attempt) for the given category;
        when the moving sum = 3 {
                go backward until it is < 3 {
                        the answer immediately following is the learning point;
                }
        }
}
```

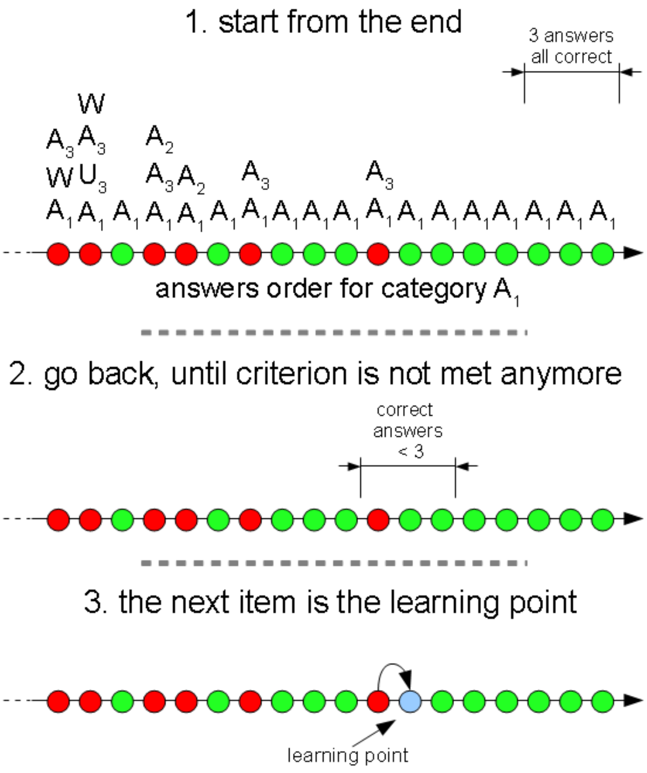

1. start from the end
3 answers all correct
W
A3 A3 A2
W U3 A3 A2 A3 A3
A1 A1 A1 A1 A1 A1 A1 A1 A1 A1 A1 A1 A1 A1 A1 A1 A1 A1
answers order for category A1
2. go back, until criterion is not met anymore
correct answers < 3
3. the next item is the learning point
learning point

Figure 4.7 Algorithm for the estimation of learning points.

Green circles are correct answers at first attempt, red circles wrong answers. The example represents learning of Category A_1.

4.3. Results

4.3.1. Difficulty

Two different measures of performance can be extracted from the test, as in the previous experiment: the number of exemplars and the amount of time needed to finish.

4.3.1.1. Differences between groups

An analysis of variance showed no significant difference of difficulty between the groups of participants.

Both time and number of shown exemplars were analysed and compared across the groups. No significant difference was found (Time: F(2,25)=1.19, *p>.30*; Shown exemplars: F(2,25)=.516, *p>.60*; see Figure 4.8, 4.4 and 4.5).

Table 4.4 Descriptive statistics for number of shown exemplars to end

Group	Mean	Std. Deviation	N
1	42.11	29.801	9
2	53.22	19.766	9
3	45.30	21.370	10
Total	46.82	23.517	28

Table 4.5 Descriptive statistics for time to end

Group	Mean	Std. Deviation	N
1	29.2611	22.58297	9
2	39.4870	16.29889	9
3	27.8367	13.55484	10
Total	32.0393	17.85130	28

It is reasonable from these results to think that the 3 unrelated criteria (U_1, U_2, U_3) are of equal difficulty and that they do not introduce any bias. In any case, all the remaining analyses will be done also separately to check that across the groups there are not different results, and only in that case a joint analysis will be performed.

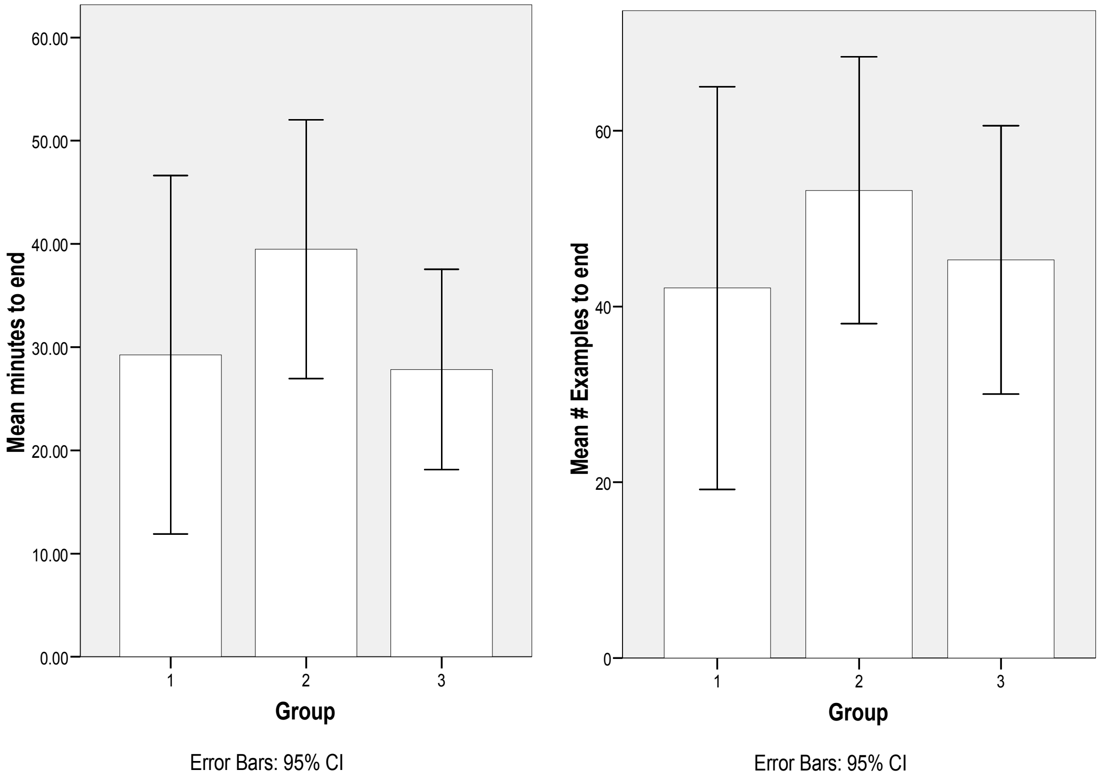


Figure 4.8 Time and number of exemplars required to finish the test

4.3.1.2. Other variables

As in the previous experiment, differences of difficulty were tested also for variables such age, sex, education and discipline of studies, in addition to the environment variable. None of these factors was significant, assuring that no bias was introduced, and that the results can be generalized to a population broader than that of Computer Science.

4.3.2. Analysis of Clicks

The novel factor introduced in this experiment is the presence of alternative solutions. This can test the prediction that when given alternative solutions, people find similar criteria instead of dissimilar ones. If they do so, it can be inferred that analogy helps finding similar criteria.

Because the exemplars are interactive, it is possible to find out which solution was discovered by each participant simply by analysing their clicks. As explained above, it is sufficient to see on which group of elements the participant clicked just before answering (obviously after they learned that category).

In fact the first clicks of a participant on a novel exemplar are random, until they find elements corresponding to the learned criterion, at which point they can give their answer. Therefore the total number of clicks on other elements is greater than the number of clicks on significant elements, but the very last click is probably on a significant element.

For the dual categories (that is, the categories defined by two alternative criteria), after their learning points, the last clicks before answering are counted both for the elements pertaining to the analogical criteria and to the unrelated criteria. A sign test was done to confront the last clicks on analogical and unrelated elements.

Although the group factor should have no effect on how much learning of similar categories is related, a preliminary analysis was done on each group separately. Given the small size of the group, no statistically significant result was expected.

Table 4.6 Frequencies and significance of the sign comparisons of last clicks on analogical elements Vs unrelated elements.

Group	Analogical > Unrelated	Analogical < Unrelated	P	N
1	8	1	.039	9
2	7	2	.180	9
3	8	2	.109	10
Total	23	5	.001	28

In all the groups the clicks on the analogical elements were more than the clicks on unrelated elements. Given that there are no differences between the groups, a joint analysis was performed. The clicks on the analogical elements were significantly more than on unrelated elements ($p < 0.001$), showing that the analogical criteria were found more often than the unrelated criteria (see Table 4.6).

The new and most important result of this experiment is that: *even if given an alternative*, the participants found the analogical criteria. This is another confirmation of our hypothesis that analogy can be used between simultaneously-learned categories. The similarities between the analogical criteria help to find those criteria instead of the unrelated criteria. In this way it is possible to minimize the memory and time efforts.

4.3.3. Learning intervals

In order to test the prediction that learning of similar categories is related, the same method was used as in the previous experiment. Given the different design of this experiment, some changes had to be made. In fact, in this experiment there are only two kinds of categories: the "analogical categories" (i.e. the ones that contain one of the three analogical criteria: A_1, A_2 and A_3) and the "unrelated category" (i.e. the category containing only an unrelated criterion - $U_{1,}$ $U_{2,}$ U_3, according to the group).

Since the "Wrong" category is derived by exclusion, and is consequently "learned" together with the last learned category, it is not considered in the analysis of the learning intervals. It would be a mistake to consider the "Wrong" category a learned category, because its exemplars can be correctly identified only by exclusion, since they do not have any real classification criterion (and therefore after all the other classification criteria are learned and excluded).

The analysis was performed only for the 23 participants who found the analogical criteria, since for the other cases it is meaningless. In fact, if the criteria found by a participant are all dissimilar, it is pointless to check if the learning times of two categories are more related than the others. Only if a participant found the analogical criterion for the dual category, there could be a structural similarity between analogical criteria which could help learning.

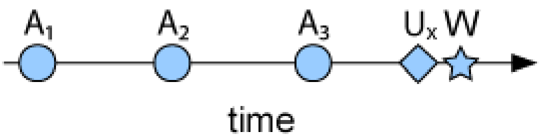


Figure 4.9 Points of learning

As in the previous experiment, all the intervals between the learning points (Figure 4.9) are calculated (i.e. A_1A_2, A_1A_3, A_1U_x, etc.), in terms of the number of exemplars shown. Then an average of the intervals between analogical categories (AnalogIntAvg) is computed:

$$AnalogIntAvg = \frac{\overline{A_1 A_2} + \overline{A_1 A_3} + \overline{A_2 A_3}}{3} \tag{4.1}$$

As in the previous experiment, the average of all the intervals between all the categories (AllIntAvg) is used as a baseline for comparison. In fact, if the four learning points are all independent from each other, a randomly chosen learning interval should be approximately equal to this average.

If this latter average (AllIntAvg) is greater than the former (AnalogIntAvg), it can be inferred that learning of similar categories is related.

Like in the previous chapter, a preliminary MonteCarlo was performed to compute the expected number of times AllIntAvg > AnalogIntAvg. This simulation showed that this comparison is positive roughly 38% of times and negative the remaining 62%, and that the distribution is not normal (Appendix C). Therefore it was not possible to use the *t*-test or other parametric tests, and a binomial test has been instead performed.

Although the group factor should have no effect on how much learning of similar categories is related, a preliminary analysis was done on each group separately. Given the small size of the group, no statistically significant result was expected, but the trends can always be studied. As shown in the 4.7, for each group the average of the intervals between analogical categories (AnalogIntAvg) is significantly less than the average of all the intervals (AllIntAvg). Given that there are no differences between the groups, a joint analysis was performed.

In the joint analysis the average of the intervals between analogical categories (AnalogIntAvg) is less than the average of all the intervals (AllIntAvg) in 91% of the cases (much more than the expected 38%, binomial test p < 0.001, see 4.7), meaning that the analogical categories are more related to each other than expected. This result supports the prediction that learning of similar categories is related, and thus the hypothesis that analogy can be used early in learning, between simultaneously learned categories.

Table 4.7 Frequencies of the comparisons of interval averages and significance of the binomial tests.

Group	AllIntAvg > AnalogIntAvg	AllIntAvg < AnalogIntAvg	N	Proportion	P
1	6	2	8	.75	.034
2	7	0	7	1	.001
3	7	0	7	1	.001
Total	20	2	22	.91	.001

4.3.4. Analysis of Errors

From the hypothesis 2b, that learning consists of two phases where partial categories are first formed and then refined to create the final categories, stems the prediction that before learning is complete, any errors in categorisation are not random but they are more frequent across similar categories. In other terms, it is more probable to say that an A_1 exemplar belongs to the A_2 category, than it belongs to U_x. Obviously, this analysis can be performed only for the 23 participants who found the analogical criteria, since for the other cases it is meaningless, as explained above for the learning intervals.

The details of this prediction and subsequent analysis are the same as in the first experiment. The only difference from the first experiment is that only three regions can be marked on this contingency table. On the diagonal are the correct answers. The remaining incorrect answers can be further divided into only two regions: the mistakes classifying an exemplar inside the analogical type of category and the answers completely wrong (see Figure 4.10). Since the "Wrong" category is derived by exclusion, and is consequently "learned" together with the last learned category, it is not considered in the analysis of the errors.

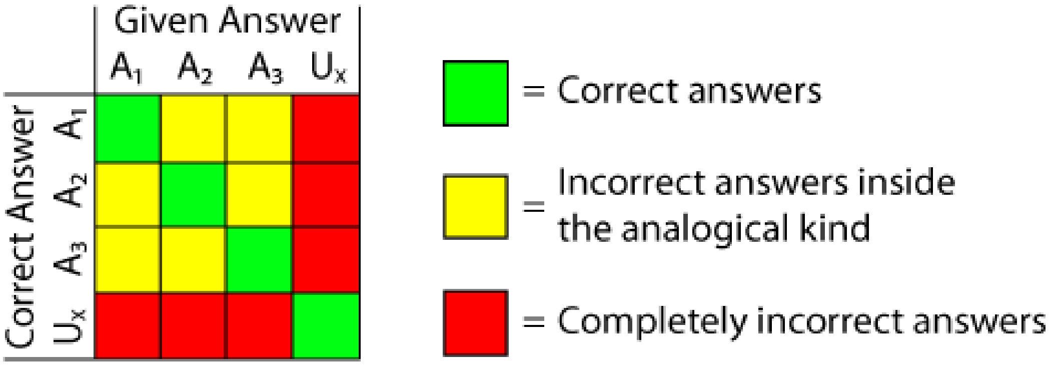


Figure 4.10 The contingency table of Given VS Expected number of answers. Three regions can be marked: the correct answers, the mistakes classifying an exemplar in the other analogical category, and the completely incorrect answers.

Although the group factor should have no effect on the distribution of errors, a preliminary analysis was done on each group separately. Given the small size of the groups, no statistically significant result was expected, but the trends can always be studied and compared.

Table 4.8 Frequencies and statistics of the sign comparisons of given VS expected number of mistakes classifying an exemplar in the other analogical category, for each interval, for each group. In bold the significant results.

Count		Other Analogical Category			Binomial P
Interval	Group	Giv<Exp	Giv=Exp	Giv>Exp	
1	1	3	2	3	1
	2	3	1	3	1
	3	4		4	1
1 Total		10	3	10	1
2	1	2		6	.289
	2	1		6	.125
	3	2		6	.289
2 Total		**5**	**0**	**18**	**.011**
3	1	2		6	.289
	2	1		6	.125
	3	2		6	.289
3 Total		**5**	**0**	**18**	**.011**

Table 4.9 Frequencies and statistics of the sign comparisons of given VS expected number of completely incorrect answers, for each interval, for each group. In bold the significant results.

Count		Completely Incorrect			Binomial P
Interval	Group	Giv<Exp	Giv=Exp	Giv>Exp	
1	1	2	4	2	1
	2	4		3	1
	3	4		4	1
1 Total		10	4	9	1
2	1	3	3	2	1
	2	5		2	.453
	3	5		3	.727
2 Total		13	3	7	.263
3	1	7		1	.700
	2	7			.160
	3	6		2	.289
3 Total		**20**	**0**	**3**	**.001**

Table 4.10 Frequencies and statistics of the sign comparisons of given VS expected number of correct answers, for each interval, for each group.

Count		Correct			Binomial P
Interval	Group	Giv<Exp	Giv=Exp	Giv>Exp	
1	1	4	3	1	.375
	2	3	1	3	1
	3	6	0	2	.289
1 Total		13	4	6	.167
2	1	3	1	4	1
	2	4	0	3	1
	3	5	0	3	.727
2 Total		12	1	10	.832
3	1	3	1	4	1
	2	3	1	3	1
	3	4	0	4	1
3 Total		10	2	11	1

As shown in the 4.8, 4.9 and 4.10, the comparisons between given and expected number of answers are similar for each group. Because the size of each group was too small to give significant results separately, and there was no reason to keep them separate, a joint analysis was performed (the results are the totals in 4.8, 4.9 and 4.10).

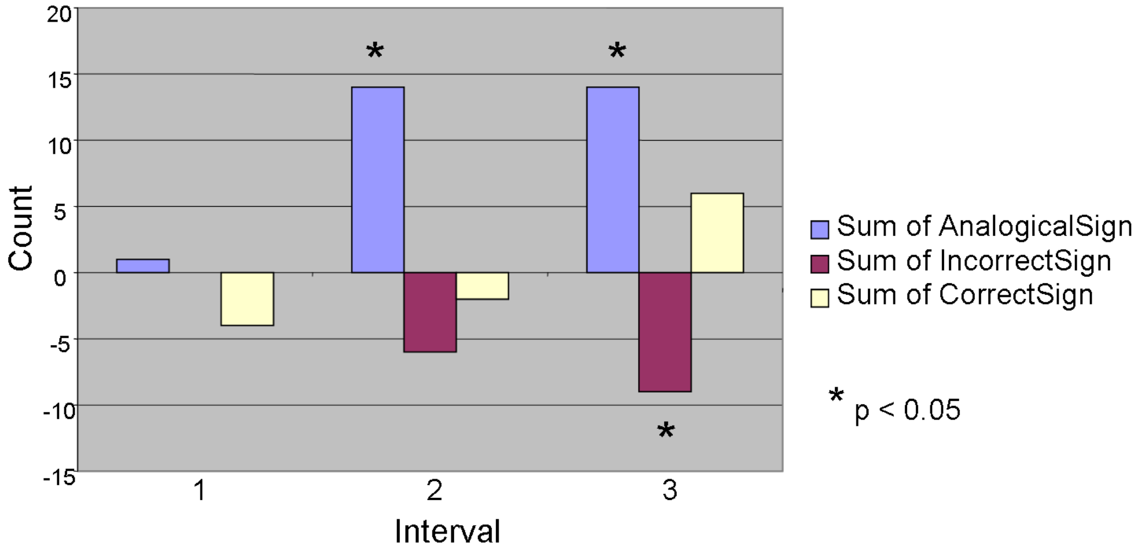


Figure 4.11 Sums of the sign comparisons of given VS expected number of answers for each region of the contingency table, for each interval, joint analysis.

Sum = count (given > expected) - count (given < expected)

As shown by 4.8, 4.9 and 4.10 and Figure 4.11, the sign tests between given and expected number of answers confirm the results of the first experiment:

- In the first interval (0 - 0.333) there are no significant results. The number of given answers in each region of the contingency table (Figure 4.10) are not statistically different from what randomly expected.

- In the second interval (0.333 - 0.666) the incorrect analogical answers occurred more often than expected. Both the completely wrong answers and the correct answers are not different from what randomly expected.

- In the third interval (0.666 - 1) the incorrect analogical answers occurred more often than expected and the completely wrong answers occurred less often than expected.

The pattern that emerges is of particular interest and confirms the learning dynamics already emerged in the first experiment. These learning dynamics have a middle period of "partial learning" for the categories which have similar criteria (it is worth to remember that this analysis was performed only for people who found the analogical criteria). In fact, the mistakes across similar categories occur more frequently than would be expected if these mistakes were random, while the mistakes across dissimilar categories (i.e. between different kinds of categories) are not different from what would be expected if these errors were random.

This result further supports the prediction that before learning is complete, any errors in categorisation are not random but they are more frequent across similar categories. In turn, this confirms the hypothesis that analogical reasoning starts to operate from the very beginning of learning, and helps finding similarities even between categories which are only partially learned.

4.3.5. Use of the previous exemplars

In contrast to the previous experiment, no notepad was available, and the participants weren't allowed to take notes in any way. Instead two arrows allowed them to go back and forth through the previously seen (and correctly answered) exemplars. Thus, before giving any answer, they could go back to compare the present exemplar(s) to the previous ones.

Because all the clicks, before and after answering, were recorded, a measure of the use of the previous exemplars during learning is the number of clicks done after answering, before any learning has occurred.

A Pearson correlation with the difficulty showed that people who go back more often find the task easier (correlation with number of exemplars: r = - .459 p < .02; with time: r = - .260 p > .05). This result is consistent with the theory previously outlined, that some people use a more scientific strategy based on the falsification of hypotheses, and some other use instead a strategy based on reinforcement. Given the complexity of the task, it isn't surprising that the more "scientific" people find it easier to solve it. This result is also a confirmation of the hypothesis that people operate under memory constraints.

4.4. Summary

This second experiment tested whether for people it is easier to find structurally similar categorization criteria or structurally dissimilar ones, when given a choice, and also confirmed the results from the first experiment. The results support the hypothesis that structural similarity helps finding categorization criteria, a fact that is also an indication of the pervasiveness of analogical reasoning. Other results from the previous experiment are also confirmed by the present experiment: the learning of similar categories is related, and a phase of partial learning precedes the discovery of the final categories. This latter result is an indication that analogical reasoning starts to operate from the beginning, when no concept is clearly learned. It also suggests that the learning of similar categories is in fact a single process.

A limitation of this experiment is that the analogical criteria are defined by music. If a participant has a preference for music (e.g. if music were perceptually more salient), this could have biased the results. In order to exclude this possibility and generalize the results, a third experiment had to be performed, with analogical criteria defined by other kinds of actions.

Chapter 5

Experiment 3

5.1. Introduction

The third experiment is very similar to the second, of which it is an extension. The hypotheses tested are the same, as well as the predictions, as are the constraints that arise from them and the analyses performed.

During the execution of the second experiment, the doubt arose whether the analogical criteria were preferred because they were similar, or because music was a more salient feature than movement. This third experiment was performed to generalize the results to the case in which the similar criteria are based on movement.

Some participants in experiment 2 reported that remembering five categories was too difficult, so in this experiment there are only four categories. Apart from these differences, the present experiment is identical to the previous one.

5.2. Method

5.2.1. Participants

As in experiment 2 the test was administered to an heterogeneous set of participants. They were both from inside and outside the university, of various ages and educations, and the test was performed in different environments with different levels of supervision. All these factors were recorded and analysed to test for differences in difficulty, and also in this experiment no significant difference was found. This assures that no bias was introduced, and allows us to generalize the results.

Participants were 41 volunteers randomly chosen: 21 Males and 20 Females, average age 26.1 (SD 9.4). They were randomly assigned to the four groups: 11 for group number 2, 10 each to the other three groups (see below for the experiment design).

5.2.2. Materials

Differently from the second experiment, in this experiment there are only three main categories (Babed, Didom, Golev), plus the residual category Wrong. Therefore on the left of each screen the participants could see only 4 stacks of draggable labels. The task is identical to the previous experiment.

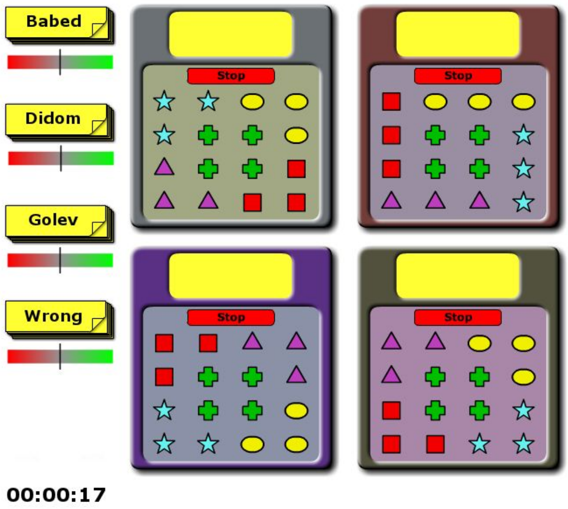


Figure 5.1 A screen from a test just started. Four exemplars are shown, with on the left four "stacks" of labels which can be dragged on the exemplars.

Table 5.1 Simplified examples of the classification criteria for a participant assigned to group 1 (see below for the explanation of the groups).

Category	Criterion 1	Criterion 2
Dual (e.g. "Babed")	Click on any of the three red squares -> All red squares do the *same* movement (e.g. jump) *at the same time*	Click on any of the four green crosses -> Music #1 is played
Single analogical (e.g. "Golev")	Click on any of the three yellow circles -> All yellow circles do the *same* movement (e.g. rotate) *at the same time*	
Single unrelated (e.g. "Didom")	Click on any of the three blue stars -> All blue stars do the same movement (e.g. zoom) *in turn*	
Residual ("Wrong")	Click on some of the four green crosses -> *Some randomly chosen* elements do some random movement (e.g. a red square jumps, a blue star rotates, a green cross jumps, etc.)	

In this experiment there are 3 different categories. One of them is a “dual category” because it has 2 defining criteria, the other two are instead “single categories” because they have just 1 defining criterion. 5.1 summarizes the kinds of classification criteria used in this experiment.

So for example, for one participant the Babed category could have one group of elements which when clicked always jump at the same time, and another group of elements which when clicked always play the same music. The Golev category could have one group of elements which clicked always rotate at the same time, which is a criterion structurally similar to the first criterion of the Babed category. The Didom category could have one group of elements which when clicked always zoom in turn.

For each group of participants there are 4 classification criteria which must be learned in order to solve the test: of these criteria, 2 have a similar structure (we shall call them "analogical" - in the example, jump or rotate at the same time), whilst the other 2 have structures dissimilar from the former ones and also between them (we shall call them "unrelated" - in the example, play a music or zoom in turn).

Differently from the second experiment, the analogical criteria are not based on music but on synchronicity of movement, and are practically identical to the U_1 and U_2 criteria from the second experiment. The only difference is that in this experiment the criteria which were called U_1 and U_2 must each have two instantiations so they can act as analogical criteria (like criteria A_1 and A_2 in the second experiment). Therefore there will be in total four analogical criteria: A_{X1}, A_{X2}, A_{Y1} and A_{Y2}. Two groups of participants will have A_{X1} and A_{X2} for the analogical criteria (in the example, jump or rotate at the same time), while two other groups will have A_{Y1} and A_{Y2} (e.g. zoom or fade in turn).

- A_{X1} and A_{X2}

  When participants click on each element of the group for criteria A_{X1} or A_{X2} the entire group react with the same action at the same time. For instance, when an element of the group is clicked, all the elements of the same group (included itself) will rotate at the same time. For each criterion it is always the same movement, but it is different for criteria A_{X1} and A_{X2} (e.g. criterion A_{X1} jumps, A_{X2} rotates). The two movements are chosen among the 10 available movements and assigned randomly to each participant at the beginning of the test.

So the structural commonality between the two criteria lies in the fact that an action is performed at the same time, but the kind of action is different for the two criteria.

- A_{Y1} and A_{Y2}

When participants click on each element of the group for criteria A_{Y1} or A_{Y2} the entire group react with the same action in turn. For instance, when an element of the group is clicked it will zoom, then another element of the same group will zoom at its turn, then another element will zoom and so on until they start back again. For each criterion it is always the same movement, but it is different for criteria A_{Y1} and A_{Y2}. (e.g. criterion A_{Y1} zooms, A_{Y2} fades). The two movements are chosen among the 10 available movements and assigned randomly to each participant at the beginning of the test.

So the structural commonality between the two criteria lies in the fact that an action is performed in turn, but the kind of action is different for the two criteria.

For each group of participants one of the unrelated criteria (U) is always defined by music, as happened for the A_1, A_2 and A_3 criteria in the second experiment, where a piece of music was played. Since there is only one version of this criterion in this experiment, the piece of music played is always the same for the associated category, and is assigned randomly to each participant at the beginning of the test. There are also other pieces of music in the test, which are sometimes played when a distractor element is clicked.

The other unrelated criteria depends on which analogical criteria are used for the group of participants, because this experiment was designed to counterbalance and test the effect of any possible salience or preference of one kind of criteria over another kind (e.g. music over synchronised movements).

If the group of participants uses criteria A_{X1} and A_{X2} for the analogical criteria (e.g. jump or rotate at the same time), the other unrelated criteria is A_{Y1} (e.g. zoom in turn), if instead the group uses criteria A_{Y1} and A_{Y2} for the analogical criteria (e.g. zoom or fade in turn), the other unrelated criteria is A_{X1} (see 5.3 in the next section).

Resuming, we have one category which can be classified using one of the analogical criteria and/or one of the unrelated criteria. A second category instead must be classified using the second unrelated criterion, and the last category must be classified using the second analogical criterion.

The groups of elements in the exemplars are always independent from each other. To each "active" group of elements is randomly assigned one classification criterion, so there is a one-to-one correspondence between groups and criteria. Thus, in the dual category two groups are "active", one for each criterion, and one group is "non-active"; in a single category only one group is "active" (and corresponds to one criterion) while two groups are "non-active". In addition there is also a variable number of distractors, in order to sum up to 16 elements for each exemplar. Some of the remaining non-assigned elements, which are randomly chosen, are given random behaviours. However, these behaviours are always different from the ones used for the classification criteria.

In order to be sure that all the categories are learned and participants don't answer by elimination (in fact, once participants have learned 2 of the 3 categories, they could correctly classify the remaining category by elimination), as in the second experiment a last category has been introduced, named “Wrong”. This category doesn't have a definition of its own, since its exemplars are composed only by distractors. It is exactly this lack of classification criterion that defines this category, which will be chosen when all the criteria for the other categories fail. In order to make the task clear and simple, the name for this category must be exemplary of its definition, which explains the name “Wrong”.

So the introduction of this “non-category” obliges participants to learn the classification criterion of each category (or, for dual categories, at least one of the 2 possible classification criteria). This non-category serves also another purpose. Given the complexity of the criteria, it would be almost impossible to be sure that the task isn't solved using some "shortcut", instead of finding the full criterion for each category. But this non-category has elements with behaviours similar to the ones defining the actual categories. Therefore, without learning the complete criteria for the other categories, a “Wrong” exemplar could be mistaken for an exemplar of another category.

For each participant, the assignment of a category to a label (“Babed”, “Didom” or “Golev”) is random, except for the non-category, which is always "Wrong". The assignment of a criterion to a category is random and just obeys one rule: for the dual category (which has 2 defining criteria and therefore 2 active groups of elements), one of the criteria has to be of the analogical type, whilst the other criterion must be of the unrelated type. So participants can get to classify the dual category by learning the analogical criterion or the unrelated criterion (or by learning both). Whilst as for single categories (which have just 1 defining criterion and therefore just 1 active group of elements), participants are obliged to classify them by learning their single criterion.

The participants can solve the task in different ways. If, for example, the dual category is defined by both criteria A_{X1} and U_1 (as in the example in 5.2) and the single categories are each defined by criteria A_{X2} and A_{Y1}, the participant can solve the test finding criteria A_{X1}, A_{X2} and A_{Y1} or instead U_1, A_{X2} and A_{Y1}.

If they find criteria A_{X1}, A_{X2} and A_{Y1}, it can be inferred that these criteria are more easily learned because the structural similarities between A_{X1} and A_{X2} help learning through analogy. Actually, A_{X2} and A_{Y1} have to be necessarily learned because each of them is the only defining criterion for each of the "single" categories. On the other hand, U_1 could be discovered instead of A_{X1}, showing that similarities can create confusion.

Table 5.2 Example of one of the four possible allocations of criteria to categories. For the dual category, participants can find the analogical criteria and/or the unrelated criteria.

Dual category	Single analogical category	Single unrelated category	Residual category
(e.g. Babed)	(e.g. Golev)	(e.g. Didom)	(Wrong)
$A_{X1} + U_1$	A_{X2}	A_{Y1}	Only distractors
Jump at the same time + Play a music	Rotate at the same time	Zoom in turn	Random behaviours

5.2.3. Design

Four distinct groups of participants are created and the assignment of participants to each group is random. The test is the same for each of the four groups, except for the criteria used for the definition of categories.

As said above, in this experiment there are two kinds of analogical criteria, based on the unrelated criteria U_1 and U_2 from the second experiment, while the unrelated U criterion in this experiment is based on the criteria A_1, A_2 and A_3 from the second experiment. In this way, combining the two experiments, each of the three kinds of criteria is used both as analogical and unrelated, and when used as unrelated it is used both in a single and in a dual category. In other terms, all kinds of criteria serve all the possible roles in the design of the experiment, ensuring a complete balancing.

Two groups of participants will have the analogical criteria A_{X1} and A_{X2}, while two other groups will have A_{Y1} and A_{Y2}. The dual category is defined by two criteria, one of which is one of these two analogical criterion (as explained above). The other analogical criterion defines one of the single categories. The other criterion defining the dual category and the other criterion defining the remaining single category, are alternated between the two groups. One of these alternative criteria is U, the other, as explained above, depends on the analogical criteria. If the group of participants uses criteria A_{X1} and A_{X2} for the analogical criteria, the other unrelated criteria is A_{Y1}, if instead the group uses criteria A_{Y1} and A_{Y2}, the other unrelated criteria is A_{X1} (see 5.3).

The presence of all the possible equivalent permutations ensures that, even if there are differences due to the different criteria, they are balanced by the design. This design also allows to estimate the resulting difficulty of the test for each permutation of the criteria. Besides, as stated above, at the very beginning of the test, for each participant is randomly assigned the correspondence between name of categories (“Didom”, “Golev” and “Babed”) and kind of categories (dual or single, analogical or unrelated), except for the non-category which is always "Wrong". For instance, a participant from group 1 will have

A_{X1} and U criteria assigned to the category named "Babed", A_{Y1} criterion assigned to the category named "Didom" and A_{X2} assigned to "Golev". Another participant from group 1 will have A_{Y1} criterion assigned to the category named "Babed", A_{X2} assigned to "Didom" and A_{X1} and U assigned to "Golev".

Table 5.3 Summary of the allocation of criteria to categories, for each group.

Group	Dual category	Single analogical category	Single unrelated category	Residual category
1	A_{X1} + U	A_{X2}	A_{Y1}	Only distractors
2	A_{X1} + A_{Y1}	A_{X2}	U	Only distractors
3	A_{Y1} + U	A_{Y2}	A_{X1}	Only distractors
4	A_{Y1} + A_{X1}	A_{Y2}	U	Only distractors

5.2.4. Procedure

The procedure is identical to experiment 2, except that it is adapted for only four categories.

5.3. Results

5.3.1. Difficulty

As in the two previous experiments, two different measures of performance are extracted from the test: the number of exemplars needed to finish and the amount of time needed to finish.

Of the 41 volunteers who finished the test, 2 had to be excluded as outliers (they exceeded by more than 3 standard deviations the average time needed to finish).

5.3.1.1. Differences between groups

An analysis of variance showed no significant difference of difficulty between the groups of participants.

Both time and number of shown exemplars were analysed and compared across the groups. No significant difference was found (Time: F(3,35)=.542, *p>.60*; Shown exemplars: F(3,35)=1.687, *p>.15*; see Figure 5.2, 5.4 and 5.5).

Table 5.4 Descriptive statistics for number of shown exemplars to end

Group	Mean	Std. Deviation	N
1	19,50	17,75	10
2	20,20	10,78	10
3	37,11	32,02	9
4	31,90	18,34	10
Total	26,92	21,33	39

Table 5.5 Descriptive statistics for time to end

Group	Mean	Std. Deviation	N
1	33,47	20,48	10
2	26,92	25,24	10
3	35,95	21,41	9
4	39,07	20,82	10
Total	33,80	21,69	39

It is reasonable from these results to think that all the classification criteria are of equal difficulty and that they do not introduce any bias. In any case, all the remaining analyses will be done also separately to check that across the groups there are not different results, and only in that case a joint analysis will be performed.

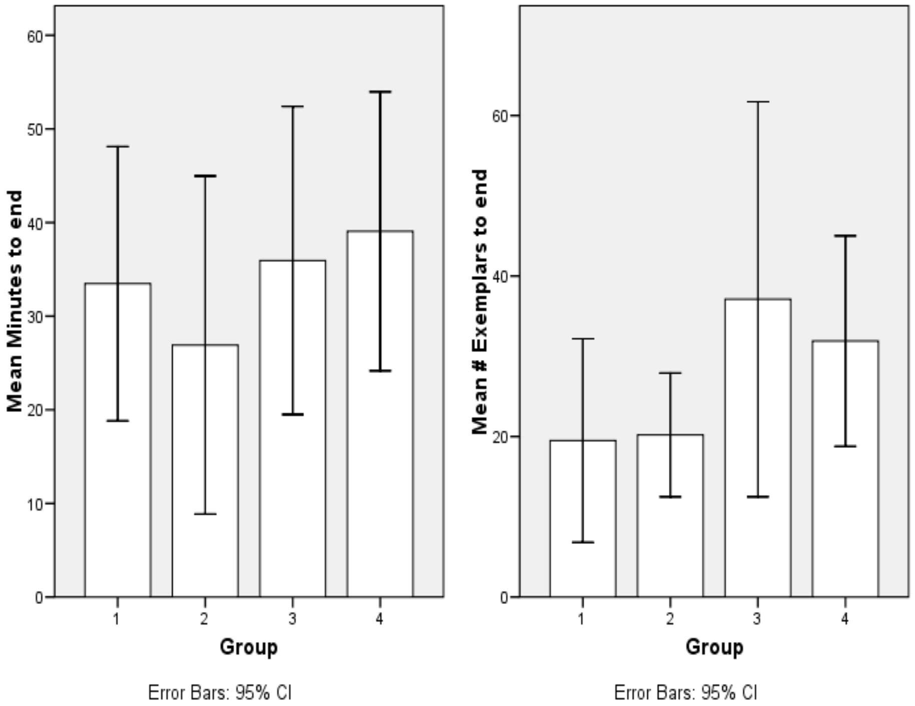


Figure 5.2 Time and number of exemplars required to finish the test

5.3.1.2. Other variables

As in the previous experiment, differences of difficulty were tested also for variables such age, sex, education and discipline of studies, in addition to the environment variable. None of these factors was significant, assuring that no bias was introduced, and that the results can be generalized to a population broader than that of Computer Science.

5.3.2. Analysis of Clicks

As in the previous experiment, this experiment has alternative solutions. This can test the prediction that when given alternative solutions, people find similar criteria instead of dissimilar ones. If they do so, it can be inferred that analogy helps finding similar criteria.

As in the previous experiment, for the dual categories (that is, the categories defined by two alternative criteria), after their learning points, the last clicks before answering are counted both for the elements pertaining to the analogical criteria and to the unrelated criteria. A sign test was done to confront the last clicks on analogical and unrelated elements.

Although the group factor should have no effect on how much learning of similar categories is related, a preliminary analysis was done on each group separately. Given the small size of the group, no statistically significant result was expected, but the trends can always be studied.

Table 5.6 Frequencies and significance of the sign comparisons of last clicks on analogical elements Vs unrelated elements.

Group	Analogical > Unrelated	Analogical < Unrelated	Ties	P	N
1	6	4	0	.377	10
2	6	4	0	.377	10
3	5	3	1	.363	9
4	7	3	0	.172	10
Total	24	14	1	.072	39

In all the groups the clicks on the analogical elements were more than the clicks on unrelated elements. Given that there are no differences between the groups, a joint analysis was performed. The clicks on the analogical elements were not significantly more than on non-analogical elements. Given the low power of the sign test and the closeness to significance, a Wilcoxon signed ranks test was also performed, which was significant ($p < 0.01$). The trend is the same as in the previous experiment, that is, the analogical criteria were found more often than the non-analogical one.

Although the preference for similar criteria over separable ones is weaker than in the previous experiment, it is still present, and is the same in each group. This means that in the previous experiment the preference wasn't caused by a greater salience of music. When the similar criteria are defined by movements they are still preferred to separable criteria. Therefore the similarity between criteria is the crucial factor, and not some bias introduced by a greater salience of some feature.

This is the most important result of this experiment; it provides confirmation of and permits generalization of the results of the previous experiment and eliminates the the hypothesis that they were caused by some bias. The similarities between the analogical criteria help to find those criteria instead of the unrelated criteria.

5.3.3. Learning intervals

In order to test the prediction that learning of similar categories is related, the same method was used as in the previous experiment. Given the different design of this experiment, some changes had to be made. In fact, in this experiment there are only two "analogical categories" (i.e. the ones that contain one of the two analogical criteria) thus there is only one interval, and there is no need to compute an average.

Since the "Wrong" category is derived by exclusion, and is consequently "learned" together with the last learned category, it is not considered in the analysis of the learning intervals. It would be a mistake to consider the "Wrong" category a learned category, because its exemplars can be correctly identified only by exclusion, since they do not have any real classification criterion (and therefore after all the other classification criteria are learned and excluded).

The analysis was performed only for the 24 participants who found the analogical criteria, since for the other cases it is meaningless. In fact, if the criteria found by a participant are all dissimilar, it is pointless to check if the learning times of two categories are more related than the others. Only if a participant found the analogical criterion for the dual category, there could be a structural similarity between analogical criteria which could help learning.

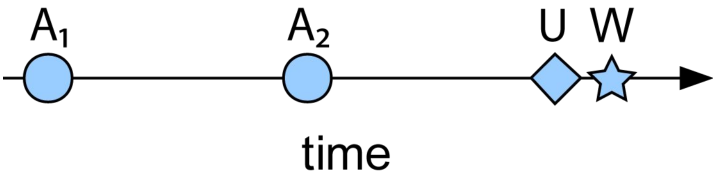


Figure 5.3 Points of learning

As in the previous experiment, all the intervals between the learning points (Figure 5.3) are calculated (A_1A_2, A_1U, A_2U), in terms of the number of exemplars shown.

As in the previous experiment, the average of all the intervals between all the categories (AllIntAvg) is used as a baseline for comparison. In fact, if the three learning points are all independent from each other, a randomly chosen learning interval should be approximately equal to this average.

If this average (AllIntAvg) is greater than the interval between the two analogical categories, it can be inferred that learning of similar categories is related.

Like in the previous chapter, a preliminary MonteCarlo was performed to compute the expected number of times AllIntAvg > A_1A_2. This simulation showed that this comparison is positive roughly 44% of times and negative the remaining 56%, and that the distribution is not normal (Appendix C). Therefore it was not possible to use the *t*-test or other parametric tests, and a binomial test has been instead performed.

Although the group factor should have no effect on how much learning of similar categories is related, a preliminary analysis was done on each group separately. Given the small size of the group, no statistically significant result was expected, but the trends can always be studied. As shown in the 5.7, for two groups the interval between the two analogical categories is significantly less than the average of all the intervals (AllIntAvg), and for the remaining two groups the trend is the same. Given that there are no differences between the groups, a joint analysis was performed.

In the joint analysis the interval between the two analogical categories is less than the average of all the intervals (AllIntAvg) in 88% of the cases (much more than the expected 44%, binomial test $p < 0.001$, see 5.7), meaning that the analogical categories are more related to each other than expected. This result further confirms the prediction that learning of similar categories is related, and thus the hypothesis that analogy can be used early in learning, between simultaneously learned categories.

Table 5.7 Frequencies of the comparisons of interval averages and significance of the binomial tests.

Group	AllIntAvg > AnalogIntAvg	AllIntAvg < AnalogIntAvg	N	Proportion	P
1	5	1	6	.83	.063
2	5	1	6	.83	.063
3	5	0	5	1	.016
4	6	1	7	.86	.032
Total	21	3	24	.88	.001

5.3.4. Analysis of Errors

From the hypothesis that learning is split in a first phase of partial learning, followed by a phase of refinement, stems the prediction that before learning is complete, the errors are not random but they are more frequent across similar categories. In other terms, it is more probable to say that an A_1 exemplars belongs to the A_2 category, than it belongs to U. Obviously, this analysis can be performed only for the 24 participants who found the analogical criteria, since for the other cases it is meaningless, as explained above for the learning intervals.

The details of this prediction and subsequent analysis are the same as in the second experiment. The only difference from the second experiment is that the analogical categories are only two.

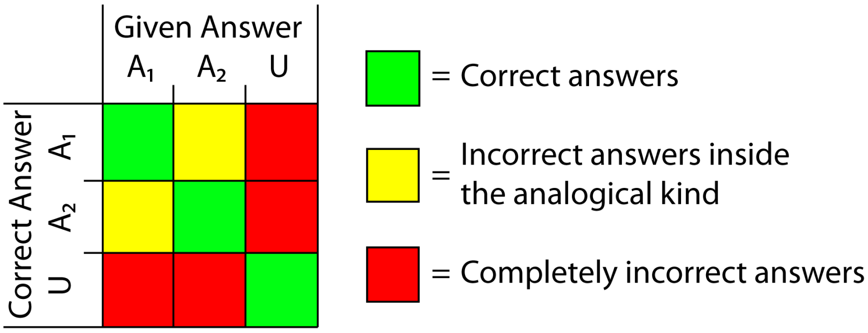

Given Answer
A1 A2 U
Correct Answer
A1 A2 U
= Correct answers
= Incorrect answers inside the analogical kind
= Completely incorrect answers

Figure 5.4 The contingency table of Given VS Expected number of answers. Three regions can be marked: the correct answers, the mistakes classifying an exemplar in the other analogical category, and the completely incorrect answers.

Although the group factor should have no effect on the distribution of errors, a preliminary analysis was done on each group separately. Given the small size of the groups, no statistically significant result was expected, but the trends can always be studied and compared.

Table 5.8 Frequencies and statistics of the sign comparisons of given VS expected number of mistakes classifying an exemplar in the other analogical category, for each interval, for each group. In bold the significant results.

Count		Other Analogical Category			Binomial P
Interval	Group	Giv<Exp	Giv=Exp	Giv>Exp	
1	1	2	2	2	1
	2	3		3	1
	3	2		3	1
	4	3	1	3	1
1 Total		10	3	11	1
2	1	1	1	4	.37
	2	1	1	4	.37
	3	1		4	.37
	4	2		5	.45
2 Total		**5**	**2**	**17**	**.01**
3	1	1	1	4	.37
	2	1	1	4	.37
	3		1	4	.12
	4	1	2	4	.37
3 Total		**3**	**5**	**16**	**.01**

Table 5.9 Frequencies and statistics of the sign comparisons of given VS expected number of completely incorrect answers, for each interval, for each group.

Count		Completely Incorrect			Binomial P
Interval	Group	Giv<Exp	Giv=Exp	Giv>Exp	
1	1	1	3	2	1
	2	2	3	1	1
	3	3		2	1
	4	3	1	3	1
1 Total		9	7	8	1
2	1	2	3	1	1
	2	2	3	1	1
	3	3		2	1
	4	5		2	.45
2 Total		12	6	6	.23
3	1	4	1	1	.37
	2	4		2	.68
	3	4		1	.37
	4	4		3	1
3 Total		16	1	7	.09

Table 5.10 Frequencies and statistics of the sign comparisons of given VS expected number of correct answers, for each interval, for each group.

Count		Correct			Binomial P
Interval	Group	Giv<Exp	Giv=Exp	Giv>Exp	
1	1	2	2	2	1
	2	2	2	2	1
	3	3		2	1
	4	3	2	2	1
1 Total		10	6	8	.81
2	1	1	3	2	1
	2	2	3	1	1
	3	2		3	1
	4	5		2	.45
2 Total		10	6	8	.81
3	1	2		4	.68
	2	2		4	.68
	3	1		4	.37
	4	2	1	4	.68
3 Total		7	1	16	.09

As shown in the 5.8, 5.9 and 5.10, the comparisons between given and expected number of answers are similar for each group. Because the size of each group was too small to give significant results separately, and there was no reason to keep them separate, a joint analysis was performed (the results are the totals in 5.8, 5.9 and 5.10).

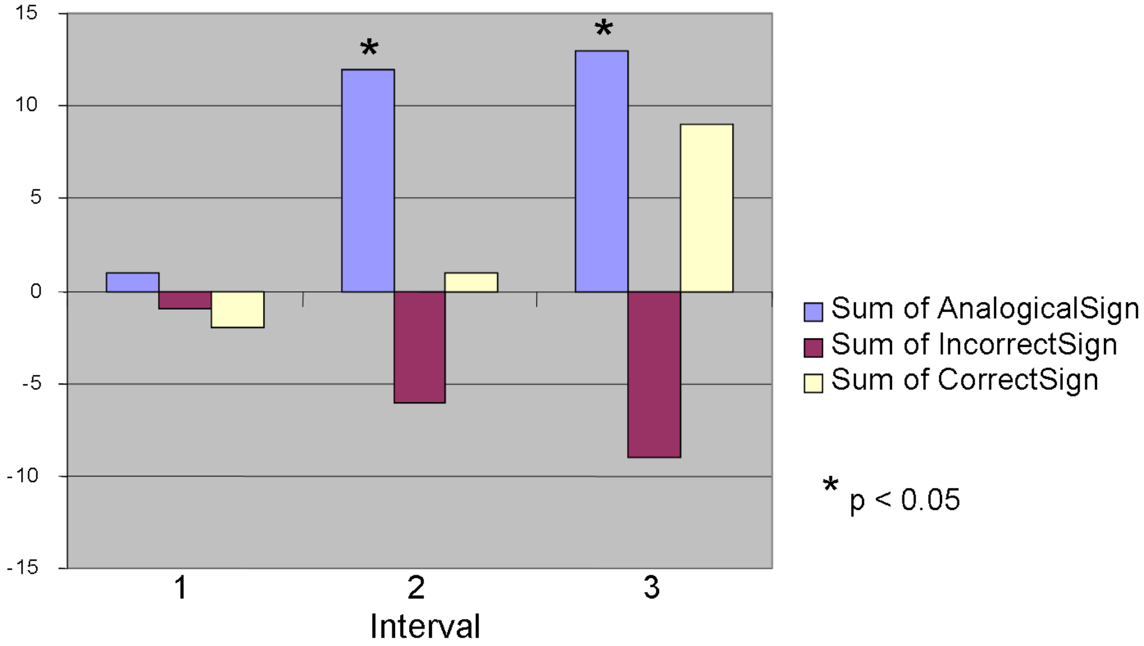


Figure 5.5 Sums of the sign comparisons of given VS expected number of answers for each region of the contingency table, for each interval, joint analysis.

Sum = count (given > expected) - count (given < expected)

As shown by 5.8, 5.9, 5.10 and Figure 5.5, the sign tests between given and expected number of answers confirm the results of the first experiment:

- In the first interval (0 - 0.333) there are no significant results. The number of given answers in each region of the contingency table (Figure 5.4) are not statistically different from what randomly expected.

- In the second interval (0.333 - 0.666) the incorrect analogical answers occurred more often than expected. Both the completely wrong answers and the correct answers are not different from what randomly expected.

- In the third interval (0.666 - 1) the incorrect analogical answers occurred more often than expected. Both the completely wrong answers and the correct answers are not different from what randomly expected.

The pattern that emerges confirms the learning dynamics already emerged in the first and second experiment. These learning dynamics have a middle period of "partial learning" for the categories which have similar criteria (it is worth to remember that this analysis was performed only for people who found the analogical criteria). In fact, the mistakes across similar categories occur more frequently than would be expected if these mistakes were random, while the mistakes across dissimilar categories (i.e. between different kinds of categories) are not different from what would be expected if these errors were random.

This result confirms the prediction that before learning is complete, any errors in categorisation are not random but they are more frequent across similar categories. In turn, this confirms the hypothesis that analogical reasoning starts to operate from the very beginning of learning, and helps finding similarities even between categories which are only partially learned.

5.3.5. Use of the previous exemplars

As in the previous experiment, the participants could go back to check the previous exemplars. So a measure of the use of the previous exemplars during learning is the number of clicks done after answering, before any learning has occurred.

In this experiment, the Pearson correlation with the difficulty wasn't significant (correlation with number of exemplars: $r = - .211$ $p > .05$; with time: $r = .148$ $p > .05$).

One explanation is that, since this task was easier and had only four categories, there was less of a need to check the previous exemplars in order to succeed in the task.

5.4. Summary

This third experiment confirms the results of the previous ones, and allows generalization from the previous results to cases in which the similar criteria are not defined by music. As in the previous experiment, this experiment showed that for people it is easier to find structurally similar categorization criteria or structurally dissimilar ones, when given a choice; the learning of similar categories is related; a phase of partial learning precedes the discovery of the final categories.

5.5. General conclusions about the experiments

A clear and novel view of the use of analogical reasoning in category learning emerges from the results of the three experiments. Not only are the similarities within a category (i.e. between the exemplars of the same category) exploited, as already shown in other works, but the similarities between different categories are exploited too. This confirms the first hypothesis of this work: analogy can be established between simultaneously-learned categories with similar structures, to aid the learning of both categories.

When given the choice to find structurally similar categorization criteria or structurally dissimilar ones, for people it is easier to find similar criteria, confirming a strong bias toward the use of similarities. This is not surprising, considering the economy of such a choice, both in computational and memory terms, and further supports the first hypothesis.

But what is more interesting, is that structural similarities are exploited even before complete learning of one category occurs, and that there is a form of partial learning, hinting that some process different from structure mapping also produces a form of analogical reasoning. This confirms the hypothesis 2b: learning consists of two phases where partial categories are first formed and then refined to create the final categories.

Although the existing theories and models of analogical reasoning in category learning could be adapted to explain these results, the existence of this partial learning opens the way to a new theory and a new model, based on the formation of partial hypotheses and their subsequent refinement. In other terms, this would mean that people don't create solutions which are immediately correct, but instead they reason in steps, generating partial solutions and modifying them, and that this simple mechanism can account for the emergence of analogical reasoning. It is interesting, at this point, to explore this opportunity and build such a model, which is what will be done in the next chapter.

Chapter 6

A computational model

6.1. Introduction

From the presented experiments, some empirical results can be summarized:

- For people, the learning of one category is temporally related to the learning of other categories with similar relational structures;
- Categories with similar relational structures are not learned individually, but in a single process;
- The final learning of similar categories is preceded by a phase of partial learning;
- Direct comparison of exemplars from both similar and/or dissimilar categories does not help or hinder learning;
- When given an alternative, people tend to learn similar criteria (for different categories) than to learn dissimilar criteria;
- In a task with relational categories, participants often form hypotheses for categorization rules and test those hypotheses in order to reject or refine them;

All of these results are in agreement with the initial hypotheses that:

- Analogy can be established between simultaneously-learned categories with similar structures, to aid the learning of both categories.

2b. Learning consists of two phases where partial categories are first formed and then refined to create the final categories.

The results of the experiments showed that people tend to learn various categories simultaneously (without using structural mapping), and that this process is usually split in two phases (partial learning, then refinement). Therefore, a test should be done in order to verify if a model based upon these assumptions can predict participants' results. This is precisely why this proposed new model has been developed. Although it is not based on the existing theories and models of category learning and analogical reasoning, this model is not incompatible with them, but on the contrary it can well complement the existing ones.

The criteria used to develop it, in addition to the need to explain the said results, are to maximize its simplicity. The ideas which form this model can be generalized to other domains of learning and reasoning. In order to do so, more experiments could be needed, and this would be outside the current scope. This model could help expand our knowledge of human reasoning by complementing the explanation of a large class of phenomena in learning.

The idea at the core of the model is that people don't arrive immediately at a final solution in category learning, instead they form partial hypotheses and then refine them in subsequent steps. If two (or more) categories have similarities between them, the final classification criteria can stem from the same partial hypothesis. In other words, learning of similar categories is a single process, in which various partial hypotheses are formed, tested and then refined. This could complement the standard explanations of the kind of analogical reasoning used by people when learning novel categories with structural similarities.

What is tested is therefore the heuristic of "form and test partial hypotheses and refine and re-use them". This chapter will show that the results in the experiments are easily explained by this single heuristic. In future works, this heuristic could also be integrated in other analogical learning models to form a more complete and powerful model of analogical reasoning.

6.2. Standard explanations of the experimental results

According to the *standard* category learning theory (Rosch, 1978), categories are defined by a set of common attributes. On the nature of those attributes, Rosch didn't limit them to the perceptual ones, although she started from them for reasons of simplicity. But since then almost all the categorization theories use the notion of attributes in the restricted and "dimensional" sense of *perceptual* attributes (Kittur, Hummel, & Holyoak, 2004), completely neglecting relations. These perceptual attributes are something that can be present or not (not necessarily in a binary form, although many models assume such further simplification), and the structure of the category definition is generally "flat": it can't account for relations between attributes. Even in studies inspired on analogical reasoning, such as the SEQL model (Kuehne et al., 2000), the definition of categories is based solely on the presence or absence of some attributes.

In more recent years, it has been proposed that categories can be defined by relations (e.g. Barsalou, 1983; Gentner & Kurtz, 2005; Kittur et al., 2004; Murphy & Medin, 1985; Rips, 1989; Ross & Spalding, 1994) and that models of categorization should use schemas and rules to discover these kind of categories. Some work has already been done (Gasser & Colunga, 2001; C. Preisach, S. Rendle, & L. Schmidt-Thieme, 2008; S Rendle, C Preisach, & L Schmidt-Thieme, 2009), which focuses on the extraction of the relations. Although these works use different methods, they share the use of intersection discovery, in which a schema is learned from examples by keeping what the examples have in common and discarding details on which they differ (as proposed by Hummel & Holyoak, 2003; see also Doumas, Hummel, & Sandhofer, 2008).

In the existing models, the learning of each category is a separate process. These models could be extended with the proposed heuristic in order to account for the help provided by similarities between different categories which are simultaneously learned.

Also machine learning systems, and in particular inductive logic programming (Lavrac & Dzeroski, 1994; Muggleton, 1991) use relational rules (in the form of logic predicates) to discover the common structure of presented exemplars. This approach is clearly very generic. In fact predicates can be of any kind, they can represent all the possible expressible rules, and they can vary in their complexity, thus accounting for simple or complex structures. One limit of these systems is that they are created to be powerful on computers and not to reproduce human learning. Another limit, as for other models, is that the learning of each category is a separate process.

6.2.1. Brief description of the existing models

The model proposed in this thesis, based on the heuristic of "form and test partial hypotheses and refine and re-use them", is not incompatible with the existing theories and models of category learning and analogical reasoning. Instead, they can well be complemented by this heuristic to create more powerful and complete models of human analogical learning. This model could complement the existing models in predicting the results of the presented experiments, because it accounts for the facilitation resulting from the simultaneous learning of similar categories. The next paragraphs will summarize the most important existing models which could solve the tasks of the experiments, and will show that they could benefit from being extended with the proposed heuristic.

6.2.1.1. SEQL

The SEQL model (Kuehne et al., 2000) is perhaps the model that more resembles the one proposed in this thesis. When presented with a new exemplar, the SEQL model looks in memory for a similar generalization or a similar exemplar. For example, for experiment 1 (an analogous discourse can be done for the other two experiments), a first exemplar of category A could be composed by 2 blue circles, 2 red circles and 3 green squares. Having no previous exemplar or generalization in memory, SEQL stores it. A second exemplar of category A could be composed by 2 blue circles, 2 red circles and 4 yellow crosses. SEQL notices the similarity with the first exemplar and creates the generalization "2 blue circles and 2 red circles".

The base version of SEQL model then finds many distinct categories (e.g. one for "2 blue circles and 2 red circles", another for "3 blue circles and 3 red circles", another for "2 blue circles and 4 red circles", and so on) and eventually settles on a single very inclusive category "presence of blue and red circles". A SEQL model extended with the right relational predicates (e.g. "same/different number of elements with the same shape and different colour") would succeed in correctly classifying the exemplars. But a SEQL model extended also with the proposed heuristic of "form and test partial hypotheses and refine and re-use them [for other categories]" would become more powerful and more able to reproduce the results from the experiments. In fact the basilar SEQL model finds each category separately, thus it does not reproduce the temporally relation of the learning of similar categories, nor the participants' error patterns.

6.2.1.2. Dora

The functioning of the Dora model (Doumas & Hummel, 2005; Doumas et al., 2008) is very complex, and will be summarized here only in its essentials. It is a connectionist model at the base of which are (sub)semantic units. Each object is represented by the simultaneous firing of a collection of these units. When two objects are simultaneously presented to the model, it creates a new predicate unit based on the semantic units which pertain to both objects. It has also a "comparator" to notice the simultaneous activation of predicates describing values along the same metric dimension (e.g. size, colour, etc.). In this case, it creates a new predicate with a relation (e.g. "bigger than") between the two objects. In summary, it can learn new relations and can also learn how to classify exemplars through the intersection of their features.

Given these features, Dora succeeds in correctly classifying the exemplars of the experiments. But the learning of the similar categories is unrelated, thus Dora doesn't reproduce any of the results found for the real participants. In fact it learns all the categories separately, therefore the learning times are randomly disposed, as well as the patterns of error before learning (i.e. no partial learning). The similarities between the categories are therefore of no help to Dora, as they are in contrast for people. Dora also could be extended with the proposed heuristic, to become more powerful and more able to reproduce the results from the experiments.

6.2.1.3. Other models

Regarding the models based on the extraction of relations (Gasser & Colunga, 2001; C. Preisach et al., 2008; S Rendle et al., 2009), they also have the limitation of finding each rule independently. Even if one extends them with the standard theories of analogical reasoning (based on structural alignment and transfer of knowledge) not all of the results would be reproduced. A model constructed in this way would first find the rule for one final category, and then would use structure mapping to find the other category. The presence of the initial partial learning would remain unexplained. This is also the case for inductive logic programming (Lavrac & Dzeroski, 1994; Muggleton, 1991), which would find each category independently. If extended with the proposed heuristic, also these models could become able to reproduce the results from the experiments.

6.2.1.4. Summary

To summarize, this proposed new heuristic is not incompatible with the existing models of category learning and analogical reasoning, although it is not based on the existing theories. Instead, it is based on the assumptions that people tend to learn various categories simultaneously (without necessarily using structural mapping), and that this process is usually split in two phases (partial learning, then refinement). These assumptions are in agreement with the results of the presented experiments. This new heuristic could complement the existing models and theories in explaining a large class of learning phenomena.

From this discussion it is clear that many of the existing models are already able to correctly learn to classify the exemplars from the three experiments (i.e. to solve the tasks). If extended with the proposed heuristic of "form and test partial hypotheses and refine and re-use them" those models would also reproduce the partial learning phenomena, and therefore all the results of the real participants.

Thus, while the separate learning of each category (independently from how good and powerful is the model) cannot reproduce some of the found results, the addition of a heuristic to make the learning a single process seems to be the key to give a complete explanation.

At this point, it is reasonable to question if this single heuristic can, on its own, reproduce all the results without the addition of any other model. This is exactly what this chapter wants to test.

6.3. An additional heuristic

In order to have a relation between the learning of similar categories, and to reproduce the experimental results, a model based on the proposed heuristic of "form and test partial hypotheses and refine and re-use them" should be proposed. That is, a model which can change ideas, and *which has a unified process of reasoning*.

It is a common experience to reason in steps: first to have a hypothesis, then to restrict it, and then to widen it again, to include something, exclude something else, etc. *And if learning various things at the same time, to use all the available information from all of those things.*

What I propose is exactly such a model which can create hypotheses and then modify them, *to create hypotheses also for other categories*. It isn't necessary that the first hypotheses are completely correct: they are retained even if they are only partially correct, so that they can undergo a process of modification. As said, these changes can be in any direction (both "general-to-specific" and "specific-to-general"), differently from some other models. In this way partial learning can be obtained, and then subsequently refined to form the final categories.

It is worth to mention that no assumption has been made on how the said modifications are obtained or what is the form of the hypotheses made and refined. Any algorithm could be used to modify the hypotheses, therefore this model will use the simplest and less intelligent form of modification algorithm, that is random modifications. If the model is able to work with this trivial algorithm, it is reasonable that it would only work better with more intelligent modification algorithms, such those implemented in other analogical reasoning models or other artificial intelligence models.

For what concerns the form of the hypotheses to make and modify, given the results from the experiments and the suggestions from the participants, in the proposed model the hypotheses have the form of predicates. As demonstrated by inductive logic programming (Lavrac & Dzeroski, 1994; Muggleton, 1991), predicates can represent any form of classification criteria, from the trivial presence of a distinctive feature to very complex relations. Predicates can be of any kind, and the model will be built to support any kind of predicates. Anyway, the kinds of predicates which will be actually implemented for this work are only those which could be imagined watching the exemplars of the three experiments. For other kinds of categories and tasks, other predicates could be implemented. The model could be even extended with other models able to find and invent new predicates, like Dora (Doumas & Hummel, 2005; Doumas et al., 2008), inductive logic programming (Lavrac & Dzeroski, 1994; Muggleton, 1991), or the models for the extraction of relations (Gasser & Colunga, 2001; C. Preisach et al., 2008; S Rendle et al., 2009).

6.3.1. The Emergence of Analogical Reasoning

This simple mechanism can explain and reproduce the experimental results, and can even give an emergent account of some forms of analogical reasoning. In addition to (and integrated with) the existing models, it can give a more powerful and complete account of analogical reasoning. Before going on, some examples can be useful to illustrate the various possibilities.

The simplest case is just a variation of the standard theories of analogical reasoning. In this case, the model first finds the correct rule for one of the similar categories (for example, for experiment 1, same number of blue and green circles). Then, using some random changes to this rule, it creates some other similar rules, until after a few steps it finds the correct rule for the other category (different number of blue and green circles). Thus, using the modification process, things that have already worked are implicitly reused, "lazily" trying to adapt them to other situations. This is analogical reasoning, but no explicit mapping of knowledge has taken place.

A more interesting case is when some partial rule is found first: that is, a rule which can be applied to two (or more) categories. In contrast to the former case, this one can also explain the phenomenon of the initial partial learning. For example, for the experiment 1, a partial rule for the similar categories could be "some relation between blue and green circles". Although the correct criteria wouldn't yet be found, such a rule would limit the range of possible mistakes. This is exactly what happens during the experiments. After this initial stage of partial learning, the model tries to modify this rule until after a few steps it finds the two final rules, which are similar, for they descend from a common ancestor. As for the former case, no explicit mapping has taken place, but the process can be described as analogical reasoning.

6.3.2. Satisficing Problem Solver

In many aspects the proposed model is similar to a problem solver, specifically in the field of inductive logic programming. It uses relational predicates to represent hypotheses. It follows a path of deductions, using some heuristics to branch and prune, until it finds a solution.

It is worth remarking that, while it is common for some problem solvers to reuse solutions to problems *previously* solved, the present model is able to find partial solutions and reuse them, for problems *being* solved. Obviously this approach gives an advantage only in the case of the simultaneous solving of problems with similarities between them, which is exactly the case investigated in this work. In this respect, for the class of problems being studied, the present model can satisficingly integrate the existing machine learning with the help of this additional heuristic.

6.4. Model architecture

6.4.1. Overview

The general structure and working of the model is very simple. It has a memory for hypothesized rules, and another memory for discarded rules (so they won't be recreated). Each rule consists of a predicate, a weight and a list of valid categories. The list of valid categories is the first main difference from the existing models: rules can be partial or final. Obviously, a rule that is valid for all of the categories or for none of them, is discarded. But it is possible to form temporary "drafts" of rules, valid for more than one category. These temporary rules will be subsequently refined through modification, which is another salient characteristic of this model. In fact, one of the ways in which a rule can be created is by randomly modifying an existing rule.

The model accounts for all the possible ways a rule can be created:

- it can be based on a shown exemplar, taking some of its properties or hypothesizing a relational structure between them

- it can be created randomly

- it can be created by modifying an existing rule

- it can be created by unifying (intersecting) two existing rules.

Therefore it can account for almost all kinds of strategies, such as experience, fantasy, trial and error, and logic.

The use of each method, as well as other aspects of the model, is parameterized (as shown below), so with different parameter values the model can reproduce the behaviour of different kinds of participants. Anyway, the overall functioning of the model remains always the same. The parameters can mimic the inclination of a participant to create more often hypotheses based on seen exemplars instead of randomly, but the basic idea of creating partial hypotheses which can be valid for more than one category is a constant. The most interesting parameter, in this respect, is the use of the creation of new rules modifying an existing rule's predicate. If for some participants it is found that the model can reproduce their behaviour without using the modifying method, that would mean that the proposed heuristic is not used by those participants.

A typical test would act very similarly to a human participant. For experiment 1, the model is shown one exemplar, and is asked to give an answer, then it is given feedback by being told the correct answer. For experiments 2 and 3, it is asked to give answers until the correct one is found. In either cases, from the feedback it can learn about the correct classification.

6.4.1.1. Answering phase

The answering phase is quite simple. Starting from the rule with greatest weight (i.e. the oldest one), the model seeks in memory a rule which has a predicate which is true for the given exemplar. If no rule is found, a random answer is given. If a rule is found, it can be final or partial (that is, valid for more than one category). In this second case the answer is randomly chosen between the categories for which the rule is valid.

For experiments 2 and 3, since the model is asked again to give answers until it finds the correct one, for the subsequent requests it also checks the answers already given for the current exemplar. If the chosen answer has already been given, it goes on to the next one, and if the current valid rule has no other answers, it continues with the next valid rule. If no other valid rule is found, a random answer is given, chosen from the ones not already given.

In a first phase, in which the rules are generic or wrong, the model is expected to give wrong or random answers. Then, if it finds partial rules, it will make mistakes similar to the ones of the participants, in their partial learning phase (i.e. giving answers inside the partial-category). When eventually the correct rules are found, and the partial rules deleted, it will always give the correct answer at the first attempt.

6.4.1.2. Learning phase

The learning phase is more complex, and consists of various steps.

The model starts recording, for each rule in memory, if its predicate is true for the current exemplar. Each rule can be valid, not valid or unknown, for each category. It starts being unknown for all the categories, then, when an exemplar is true for its predicate, the rule is marked as valid for the exemplar's category (unless it was previously marked as not valid for that category). If the exemplar is false for the predicate, the rule is marked as not valid, with a probability proportional to a parameter. The use of this parameter was decided because this kind of counter-factual reasoning is not common, and many participants probably wouldn't use it. Finally, if the predicate is true for the current exemplar, but the rule is final for another category, the rule is removed. This is the simplest form of counter-factual reasoning. If a rule says that an exemplar is surely of one category, but it is of another category, the rule is clearly wrong.

A second step of the learning phase is to check if there are final rules for some categories. If a final rule is found for a category, all the other partial rules are marked as not valid for that category, so they won't produce confusion. Moreover, if more than one final rule exists for a category, only the oldest is retained.

A third step consists of removing the useless rules, that is the rules which are valid for all or none of the categories, or the rules which are partial but too old (according to a parameter). The removed rules are placed in the memory for the bad rules, whose size is controlled by a parameter.

For experiments 2 and 3, all of these recording steps are repeated also for some of the previously seen exemplars, randomly chosen. For experiment 1 there is instead a "virtual notepad" in which some of the previously seen exemplars are annotated, and then used for these recording steps. In either case, the frequency of this "going back" is controlled by a parameter.

Finally, some new rules are created, by the methods previously explained (i.e. based on the shown exemplar, randomly, by modifying an existing rule, or by unifying two existing rules). If a rule is already in the memory for the bad rules, it isn't created again.

As a last step, if the memories exceed their maximum sizes (controlled by two parameters), rules are randomly deleted from both memories (except for final rules in the memory for good rules) until their sizes are as required.

6.4.2. Structure of the model

6.4.2.1. Kinds of predicates

What has been stated until now is very general, and could work with any kind of predicate. And in fact the model is open in respect to which predicates are actually used. A predicate is abstractly defined as something having two functions: "IsTrue" (referred to an exemplar) and "ChangeRandom" (which returns a new predicate). Therefore, virtually all kinds of predicates can be implemented; they must simply provide these functions.

In practice, only the kinds of predicates useful for the presented experiments are currently implemented, but the model can be extended with other predicates. As said above, it can be even merged with some other model of discovery of relational structures, to be able to produce new predicates from scratch. But this is not the aim of my research; as said, my interest is just to account for the phenomena found in the experiments.

Therefore the predicates currently implemented are:

- the presence (or absence) of a feature or a set of features
- abstract quantity (a lot/a few) of elements with a given feature
- similarity between groups of elements (same/difference colour, shape, number)
- causal interaction (clicking on some element[s] causes reaction[s] of some kind)
- intersection of two other predicates

The model contains a predicate factory which creates new predicates. If the factory is asked to create a new predicate based on a given exemplar, the kind of predicate to create is randomly chosen, and the created predicate will be valid for the given exemplar. If the factory is asked to create a random predicate, not based on any exemplar, the kind of predicate is randomly chosen, but the predicate is also randomly created.

The factory can also create an intersection predicate (i.e. the logical operation "AND" between two predicates) based on two existing predicates.

A central feature of the model is the creation of a predicate modifying an existing predicate. In that case, the existing predicate is directly asked to create a randomly changed clone, using its function "ChangeRandom". As said above, better modification algorithms could be implemented, but since the core of the theory is the usefulness of modification in general, if the trivial random modification is found to be useful, any better algorithm could only improve the performance.

As said, the important feature of the model is not the specific implementation of the predicates (any implementation is good as long as exist predicates which can represent the solutions to the tasks) but the ability to randomly modify a predicate to create a similar predicate, and all the implemented predicates have this ability.

6.4.2.2. Rules

Predicates are the central part of the classification criteria hypotheses, which for simplicity in this model are called "rules". Thus each rule will contain its predicate, against which each exemplar will be evaluated to know if the predicate is true or false for that shown exemplar.

The proposed heuristic states that hypotheses can also be partial, i.e. be valid for more than one category. Therefore each rule will have an array of validity statuses, one for each category of the task. When the rule is created, the validity for all the categories for that rule is set to "unknown". Then the validity for the shown exemplar's category can be updated during the learning phase. If the predicate is true for the shown exemplar, the validity is set to "valid", if it isn't already set to "not valid". If the predicate is false for the shown exemplar, the validity can be set to "invalid".

A rule is said to be "final" if it is valid for only one category and its weight (see below) is greater than a given parameter (see the parameters paragraph). In that case, if an exemplar of another category is found for which the predicate is true (i.e. the rule would say that the exemplar is for sure of category A, but it is of category B), the rule is discarded.

A final attribute of the rules is their weight. It is increased by one, for each rule in memory, every time a new exemplar is shown to the model (i.e. it says how long a rule have existed). It is called "weight" instead of "age" because it has a role in the answering phase. In fact the rules are ordered from the oldest to the newest, and then are tested in this order. Therefore the oldest have greater weight because it is more probable that they are chosen to give the answer.

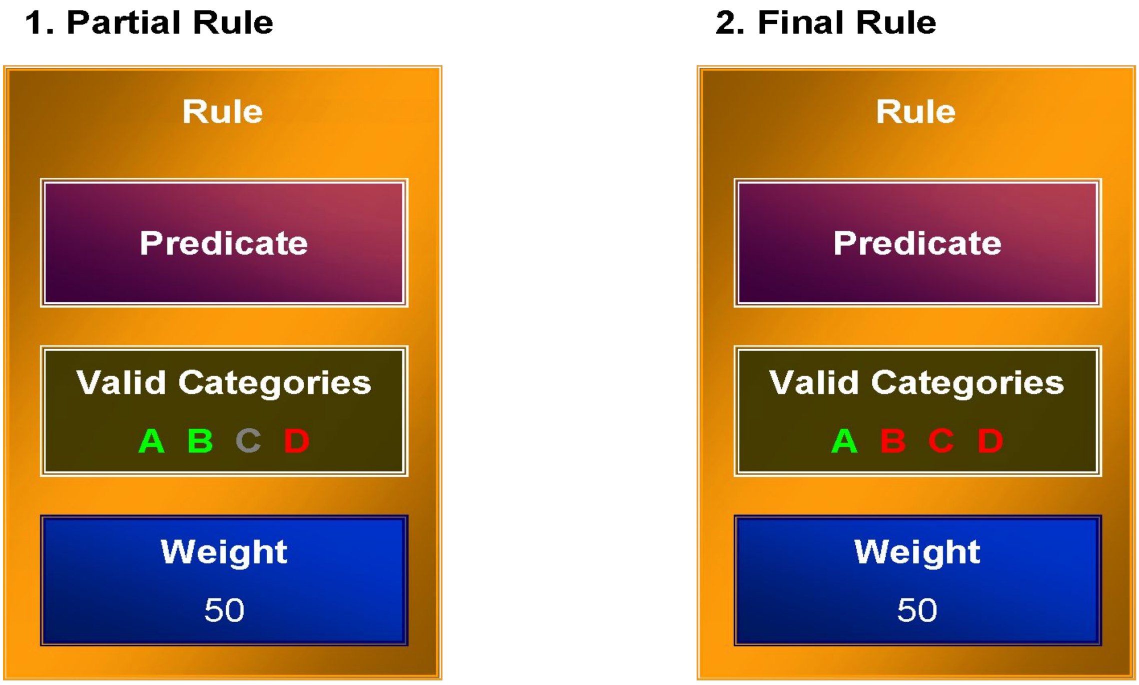


Figure 6.1 Partial and final rules. A partial rule (1) is valid for more than one category (in the example, A and B categories are green, i.e. valid, C is grey, i.e. unknown, D is red, i.e. not valid). A final rule (2) is valid only for one category (in the example, A). All rules are composed by a predicate, the weight and an array of valid/invalid/unknown categories.

6.4.2.3. Memories

The model has two distinct memories, one for good rules and another for bad rules. It is designed this way because from the participants' reports it was clear that they tended to remember separately the rules which they felt "promising" and the rules they had discarded. Moreover, while the participants consciously and actively remembered the good rules, they recalled a bad rule only when they risked to create it again as a good hypothesis. It was like a sentinel which was activated only in a specific case, to tell them not to waste time again on a bad idea. This is also the function the memory for bad rules has in the model. In this memory go the discarded rules, and when a predicate is hypothesized, before becoming a rule it is searched in all the remembered bad rules. If the same predicate was already created and discarded, the new one is automatically discarded.

These two memories are therefore only containers for rules, with their maximum sizes decided by two parameters. If the good rules memory exceeds its maximum size, it is resized by randomly removing rules, except those marked as final. It the bad rules memory exceeds its maximum size, it resized by randomly removing rules. As said above, the good rules memory is ordered from the oldest to the newest, so when searching for a predicate true for a shown exemplar, the oldest rules are tested first.

6.4.3. Representation of the input

Also for exemplars the level of detail in the chosen implementation is quite abstract.

For experiment 1, they are 3 groups of elements, each group represented by its colour, shape and quantity. The spatial distribution has been ignored, since in the experiment the elements were randomly distributed, and the participants reported that the lack of spatial structure was clear from the very first exemplars.

For example, an exemplar composed by 3 blue squares, 3 red squares and 2 green crosses would be represented as an object composed by three groups:

```
[{shape: square, colour: blue, number: 3},
{shape: square, colour: red, number: 3},
{shape: cross, colour: green, number: 2}]
```

For experiments 2 and 3, each element of the exemplar is represented individually. In fact in some cases each of them elicits different reactions, so in this respect it would be impossible to group them. But an algorithm can also group them according to their shape and colour, so the predicates can also work with groups of elements, in addition to the single elements. Thus in this case too, only the relevant features (as reported by the participants) are retained.

For example, for experiment 2, an exemplar of a Dual category, with one group of 5 green squares which play music #3, a second group of 4 blue crosses which react all with the same randomly chosen action at the same time (U_1 criterion), a third group of 5 red triangles (some of which are distractors), 1 yellow circle and 1 orange star, could be described as follows:

```
{element#1: {
        shape: blue, colour: cross, melody: null, loop: false,
        reactions: [{element: 1, action: 2},
           {element: 6, action: 2},
           {element: 9, action: 2},
           {element: 5, action: 2}]},
element#2: {
        shape: square, colour: green, melody: 3, loop: false, reactions:
        []},
element#3: {
        shape: square, colour: green, melody: 3, loop: false, reactions:
        []},
element#4: {
        shape: square, colour: green, melody: 3, loop: false, reactions:
        []},
element#5: {
        shape: blue, colour: cross, melody: null, loop: false,
        reactions: [{element: 1, action: 1},
           {element: 6, action: 1},
           {element: 9, action: 1},
          {element: 5, action: 1}]},
element#6: {
        shape: blue, colour: cross, melody: null, loop: false,
        reactions: [{element: 1, action: 4},
          {element: 6, action: 4},
          {element: 9, action: 4},
          {element: 5, action: 4}]},
element#7: {
        shape: square, colour: green, melody: 3, loop: false, reactions:
        []},
element#8: {
        shape: square, colour: green, melody: 3, loop: false, reactions:
        []},
element#9: {
        shape: blue, colour: cross, melody: null, loop: false,
        reactions: [{element: 1, action: 6},
          {element: 6, action: 6},
          {element: 9, action: 6},
          {element: 5, action: 6}]},
element#10: {
        shape: triangle, colour: red, melody: 6, loop: false,
        reactions: [{element: 10, action: 3},
        {element: 12, action: 9}]},
element#11: {
        shape: triangle, colour: red, melody: null, loop: false,
        reactions: []},
element#12: {
        shape: circle, colour: yellow, melody: null, loop: false,
        reactions: [{element: 14, action: 11}]},
element#13: {
        shape: star, colour: orange, melody: 10E, loop: false, reactions:
        []},
element#14: {
        shape: triangle, colour: red, melody: 5, loop: false, reactions:
        []},
element#15: {
        shape: triangle, colour: red, melody: null, loop: false,
        reactions: []},
element#16: {
        shape: triangle, colour: red, melody: 10A, loop: false,
        reactions: [{element: 16, action: 10}]},
groups: [[1, 5, 6, 9], [2, 3, 4, 7, 8],
        [10, 11, 14, 15, 16]]
}
```

It can be disputed that these representations of the exemplars are no more than a set of features, exactly as many other models. But I don't argue that single exemplars can be represented by their features, only that some kind of categories are better described by relations (or even better by predicates). The representation chosen has the level of detail needed to solve the test without making things too complex: my interest is on the kind of reasoning used, not on perception and the discovery of features.

The level of representation used in this model is based on the notes and debriefings of the participants. From their insights it was clear that many perceptual attributes of the exemplars, like for example the spatial disposition of the elements in the first experiment, were simply ignored, and that only the "logical" attributes were noted.

6.4.4. Summary of the parameters

The parameters used by the model can be grouped in three categories: use of creation methods, available working memory and rationality. They are summarized in Table 6.1.

Table 6.1: Parameters used by the model, with ranges used by the simulated annealing algorithm.

		Min	**Max**
Creation methods			
CreateFromExemplar	probability of using the creation of new predicates based on a shown exemplar	0	10
CreateModifying	probability of using the creation of new predicates by modifying an existing one	0	10
CreateRandom	probability of using the random creation of new predicates	0	10
CreateUnifying	probability of using the creation of new predicates by unifying two existing ones	0	10
Memory			
GoodMemSlots	size of the memory for good rules	0	15
BadMemSlots	size of the memory for bad rules	0	15
NotepadSlots	only for experiment 1: size of the virtual notepad	0	50
GoingBackProb	only for experiments 2 and 3: the probability of going back to learn from previous exemplars	0	100
Rationality			
RecordCorrectProb	probability of using counter-factual reasoning in recording the goodness of a rule	0	100
PartialRuleLimit	maximum age of a rule before it can be removed if partial or it is considered final if valid for only one category	1	40
PartialRuleRemoveProb	probability of removing a partial rule when too old	0	100

As explained in the following, in order to find the combination(s) of parameters which best simulate the preferences and best mimic the inclinations of each participant, a simulated annealing algorithm is used to vary the parameters. Each parameter can vary between a range, which is shown in the table.

It is worth to notice that it is possible to completely disable the proposed heuristic by setting to 0 the probability of using the creation of new predicates by modifying an existing one (CreateModifying). While all the rest of the model is only a simple reproduction of the reasoning adopted by participants (and the parameters can vary this reasoning to best adapt the model to the peculiar preferences and inclinations of each participants), this single feature (the use of the modification of hypotheses) is the core of our theory. Clearly, it was impossible to test this modification mechanism without a structure in which to put it, and which could reproduce the functioning of the tests and the reasoning of the participants.

In addition to those parameters, some values (6.2) are computed about the effective use of some features in each repetition of the test. The parameters give the model the probability of using some creation method, or constraints on memory. But the actual use of these features probably changes with different repetitions of the test, so it can be useful to know how much each feature has been actually used.

When each repetition ends, for the final rules in memory it is counted how many have been originally created from the shown exemplars, how many randomly and how many by unifying two other rules. It is also counted how many times a rule has been changed to become the final rule, as a measure of the use of modification.

Table 6.2: Computed effective use of model's features.

Creation methods	
FromExemplarAvgUse	Average use of creation from the shown exemplar
ModifyingAvgUse	Average number of changes that final rules have undergone
FromRndAvgUse	Average use of random creation
UnifyingAvgUse	Average use of creation by intersection
Memory	
GoodSlotsAvgUse	Average use of memory for good rules
BadSlotsAvgUse	Average use of memory for bad rules

6.4.5. Functioning of the model

The overall functioning of the model has been described in the overview. Here it will be illustrated in a more algorithmic form. Since the three experiments have some differences, in some aspects the functioning must be different for each experiment. When the functioning differs between the experiments, the algorithm will be illustrated for each experiment separately.

The model's core lies in the predicate's ability to be modified (block F - "Create new rules") and in the rule's possibility to be final or partial (i.e. valid for more than one category). All the remaining (which is the largest portion of the algorithm) is only a structure to organize the predicates and make it possible to reproduce the tests as shown to the participants. In other terms, the model must obviously be able to interact with the test in order to tell it the answers and get the feedback, and must be able to record the good and bad rules and manage them in a memory. Although these clearly are computationally complex tasks, they are not the important features.

Because the simulated annealing algorithm (see below) was used to find the combination(s) of parameters which best mimic the inclinations of each participant, the model received as input the same exemplars which had been shown to the participant, in the same order. During the development, it was also tested with sets of newly created exemplars to further validate its functioning.

A1. Answering phase, experiment 1

```
1. get a new Exemplar
2. sort GoodMemory from greatest to lower weight
3. for each Rule in GoodMemory
   3.1. if the Rule's Predicate is true for the Exemplar
         3.1.1. give answer(random([Rule's Valid categories]))[3]
4. if no answer given
   4.1. give answer(random([all categories]))
```

A2. Answering phase, experiments 2 and 3

```
1. get a new Exemplar
2. sort GoodMemory from greatest to lower weight
3. set PossibleAnswers = [all categories]
4. for each Rule in GoodMemory
   4.1. if the Rule's Predicate is true for the Exemplar[4]
      4.1.1. set RemAnswers = intersection(PossibleAnswers, [Rule's Valid
             categories])
         4.1.1.1. if RemAnswers is not empty
            4.1.1.1.1. give answer(random(RemAnswers))
5. if no answer given
   5.1. give answer(random(PossibleAnswers))
6. get Feedback
7. if Feedback is not correct
   7.1. remove given answer from PossibleAnswers
   7.2. repeat from 4
```

B. Record answer for shown Exemplar

```
1. get CorrectAnswer for the Exemplar
2. for each Rule in GoodMemory
   2.1. if the Rule's Predicate is true for the Exemplar
      2.1.1. if the Rule is final but Rule's Valid category is not the
             CorrectAnswer
         2.1.1.1. move Rule from GoodMemory to BadMemory
      2.1.2. if the Rule is Unknown for CorrectAnswer
         2.1.2.1. set Rule Valid for CorrectAnswer
   2.2. else
      2.2.1. if random(0-100) < RecordCorrectProb
         2.2.1.1. set Rule NotValid for CorrectAnswer
```

C1. Record answers from Notepad (experiment 1)

```
1. for each NpExemplar in Notepad
   1.1. repeat block B for NpExemplar
2. add Exemplar to Notepad
3. while size(Notepad) > NotepadSlots
   3.1. remove a random exemplar from Notepad
```

C2. Go back to record answers from previous exemplars (experiments 2 and 3)

```
1. for each PrevExemplar in PreviousExemplars
   1.1. if random(0-100) < GoingBackProb
      1.1.1. repeat block B for PrevExemplar
2. add Exemplar to PreviousExemplars
```

D. Check definitive rules

```
1. for each RuleA in GoodMemory
   1.1. if RuleA is final
      1.1.1. for each RuleB in GoodMemory (except RuleA)
         1.1.1.1. set RuleB NotValid for RuleA's Valid category
```

[3] This includes the case of final rules. If the rule is final, it will obviously have only one valid category. For the definition of "final rule" see the paragraph on rules.

[4] If the predicate is a causal interaction, it can check the Exemplar also by virtually clicking on its elements to see their reactions.

```
E. Remove useless rules
1. for each Rule in GoodMemory
   1.1. increment Rule's Weight by 1
   1.2. if Rule is Valid for all categories or Rule is Valid for no
         category
      1.2.1. move Rule from GoodMemory to BadMemory
   1.3. if Rule is partial and Rule's Force > PartialRuleLimit and
         random(0-100) < PartialRuleRemoveProb
      1.3.1. move Rule from GoodMemory to BadMemory

F. Create new rules
1. NewRules = []
2. add to NewRules #CreateFromExemplar new Rules with predicates created
   from shown Exemplar
3. add to NewRules #CreateModifying new Rules created randomly modifying
   predicates of rules randomly taken from GoodMemory
4. add to NewRules #CreateRandom new Rules with randomly created predicates
5. add to NewRules #CreateUnifying new Rules created unifying predicates of
   two rules randomly taken from GoodMemory
6. shuffle NewRules
7. for each Rule in NewRules
   7.1. if GoodMemory does not contain Rule and BadMemory does not contain
         Rule and Rule is true for shown Exemplar
      7.1.1. set Rule Valid for Exemplar's category
      7.1.2. add Rule to GoodMemory

G. Resize memories
1. while size(GoodMemory) > GoodMemSlots
   1.1. remove a random rule from GoodMemory except final rules
2. while size(BadMemory) > BadMemSlots
   2.1. remove a random rule from BadMemory
```

6.4.6. A typical session

A typical session would run as follows. When the first exemplars are shown, the model has no rule in memory, so it answers randomly. From the given feedbacks, it creates some rules with both the "From exemplar" and "Random" methods. With subsequent exemplars it can test the created rules, to discover if they are valid for more than one category (i.e. partial), or for none. It can also create new rules by modifying the ones already in memory, and can test all the rules in memory against some previous exemplars. Because the model considers one rule at a time, it does not use direct comparison of exemplars, as discovered for the participants.

If some partial rule is too generic, the model can discover it is valid for all the categories (i.e. useless), and delete it. At any rate, if a partial rule survives too long, its practical usefulness decreases, because the confusion it produces will exceed the little piece of information it provided, so it will be deleted.

When the model has a partial rule in memory, if presented exemplars apply to that rule, the model chooses randomly one of the categories of that rule. Thus the answers will be no longer completely random, but they will follow a pattern similar to the partial learning discovered in human participants.

At some point, a final rule will be discovered, or a partial rule for similar categories. From one of those rules, with the modification method the model can "short-cut" the process of discovering the other final rule (or both of the final rules). This will reproduce the temporal relation of the learning of similar categories, and the tendency to discover similar criteria when given an alternative.

When a final rule is discovered, all the partial rules are set as not valid for that category. Thus, after a while, the partial rules will be marked as not valid for all the categories, and deleted. At this point, the model will give always the correct answer and the test will end.

6.4.7. The simulated annealing

As mentioned above, there are various parameters used in this model, accounting for various strategies and aspects of reasoning. Using a simple form of simulated annealing, it is possible to find the set of parameters that best reproduces the answers of each participant.

This paragraph is an overview of the simulated annealing algorithm used to find the best sets of parameters. It is worth noting that the simulated annealing is not a central part of the model, and that any other algorithm could be used to find the sets of parameters which accurately reproduce the participants' answers. Therefore the algorithm used will be presented only in its outline.

For all the test sessions administered to the human participants, the details about the composition of the exemplars had been recorded. Therefore it was possible to submit to the model the same exemplars shown to the participants, in the same order. If the answers given by the model reproduce those given by the participants, it can be said that the model reliably mimics their behaviour, and thus probably their reasoning also.

Moreover, if the model gives the same set of answers to the same exemplars shown to a participant, it also makes the same mistakes and learns the categories in the same order. Thus, this implicitly confirms all the results (as well as the "form and test partial hypotheses and refine and re-use them" heuristic) found for human participants in the previous chapters.

In order to test the participants' answers reproduction accuracy, an agreement measure is needed. For Experiment 1, the Cohen's Kappa (1960) was chosen. Since its distribution is not stable when the quantity of exemplars changes (see Appendix B), a MonteCarlo simulation was implemented to compute the significance of each case.

For experiments 2 and 3 the participants must keep answering until they give the correct response. But with multiple responses the standard version of Cohen's Kappa cannot be used. Two alternative measures have been tested (see Appendix C): a fuzzy version of Kappa (Dou et al., 2007), and the average rank difference. Since they are equally sensitive to the actual agreement (R = 0.955 and R = -0.953), and since they are highly correlated (R = -0.986), the Fuzzy Kappa has been chosen, to be consistent with the measure previously used.

In order to test only the learning phase, the interval of answers being compared ended at the last learning point, thus not including all the following correct answers needed to finish the test (which would have positively biased the agreement).

The simulated annealing algorithm used was very simple, and was based on the narrowing of the space of solutions (i.e. the possible sets of model's parameters). For each participant, 200 sets of parameters were initially created, randomly but homogeneously covering all the space of solutions (see the parameters ranges in the parameters paragraph). Each session was repeated 10 times for each set of parameters. Then a two-phases annealing was used.

In both phases, for each participant the space of solutions was repeatedly narrowed by selecting the best repetitions, and then creating 200 new sets of parameters inside the new space of solutions, until no improvement could be achieved. At the end of the first phase the cycle was repeated for the second phase. The difference between the two phases was the criterion of selection of the best repetitions (with their associated sets of parameters).

In the first phase the criterion of selection of the best repetitions was to minimize the difference, between the model and the participant, of the quantity of exemplars needed to solve the task. In the second phase the criterion was to maximize the agreement of the answers given by the model with those given by the participant.

This two-phase simulated annealing was used to quicken the process. In fact it is pointless to estimate the agreement of the answers if the quantity of exemplars needed to solve the task differs too much between the model and the participant. Only when the space of solutions is narrowed enough to reproduce the participant's performance, it is useful to check the agreement to further narrow the space of solutions in order to find the solution(s) which have the best stable agreement with the participant's answers.

It is worth to restate that for each set of parameters, the participant's session is repeated 10 times, because the model has stochastic processes and therefore two repetitions are never exactly the same. Both the comparison of quantity of exemplars and the agreement of answers are computed as the average over all the 10 repetitions of the session. Therefore, only the most stable (and not only best mimicking) sets of parameters are selected to create new sets. For this reason, at the end of the simulated annealing process, the selected set(s) of parameters, for each participant, are guaranteed to have the best achievable stable agreement with the participant's answers.

6.5. Reproduction of Experiment 1 results

6.5.1. Model's fit

For 86% of the participants, the model was able to accurately reproduce (significance of kappa < 0.05) the participants' answers. It also accurately reproduced the results of the experiment. The sign tests of the learning intervals (see Chapter 3) are significant (p <

0.01) and in the same direction as in the experiment. In other words, also for the model, the learning of one category helps the learning of the other category with similar relational structure.

The results are the same as in the experiment also for the sign tests of the given vs. expected number of partially incorrect answers during learning (see Chapter 3). The partially incorrect answers are more than expected ($p < 0.01$), meaning that also for the model the final learning of similar categories is preceded by a phase of partial learning.

No difference was found between the Paired, Unpaired and Single groups, confirming that direct comparison of exemplars from both similar and/or dissimilar categories does not help or hinder learning.

6.5.2. Parameters values

Examining the averages of the parameter values found by the simulated annealing algorithm, it is possible to assess if and how much each rule creation method is used by the model (Table 6.3).

Table 6.3: Average values of creation methods as found by the simulated annealing algorithm, experiment 1.

Parameter values represent the number of times each method is tried for each presented exemplar. Use values represent the times each method is used to arrive to the final versions of the correct rules.

	Mean	95% Confidence interval		Median
		Lower Bound	Upper Bound	
Parameters				
CreateFromExemplar	5.67	5.50	5.84	5.00
CreateModifying	5.96	5.80	6.11	6.00
CreateRandom	4.09	3.90	4.28	3.00
CreateUnifying	5.13	4.95	5.31	5.00
Effective use				
FromExemplarAvgUse	0.96	0.95	0.97	1.00
ModifyingAvgUse	1.82	1.75	1.90	1.00
FromRndAvgUse	0.0165	0.0121	0.0209	0.0000
UnifyingAvgUse	0.0045	0.0040	0.0050	0.0000

It is clear that the modification of rules is effectively used: rules are changed an average of 1.82 times before becoming the final ones. This confirms that the proposed heuristic of "form and test partial hypotheses and refine and re-use them" is effectively and largely used.

It is not surprising that the creation based on the shown exemplars is used as the initial source of rules almost all the times: the probability to find a good rule randomly is very low, as well as the probability of finding a good rule by unifying two existing rules (at least for the experimental tasks in this work).

It is also interesting to study the effect of the parameters on the difficulty (measured in number of exemplars needed to solve the test). The only significant correlations with the difficulty are summarized in 6.4.

Table 6.4: Significant parameters correlations with difficulty for Experiment 1.

CreateRandom	.167(**)
ModifyingAvgUse	-.318(**)
NotepadSlots	-.169(**)
RecordCorrectProb	-.346(**)
**. Correlation is significant at the 0.01 level (2-tailed).	

The fact that the random creation of new rules is positively correlated with the difficulty isn't surprising, since doing so would produce only confusion.

It is reassuring the negative correlation between the difficulty and the effective use of creation by modification. It means that the more the creation by modification is used the easier is the task. This confirms that the proposed heuristic effectively helps learning.

The negative correlation of the available notepad slots also confirms what found in the experiments.

Finally, the fact that the use of counter-factual reasoning (RecordCorrectProb) helps learning was expected. This confirms that for this kind of tasks, people need to reason by trial and error, forming hypotheses and rejecting the wrong ones.

In summary, the model meets all the requirements: it mimics people's answers, reproduces the experiment's results and uses modification of rules as a heuristic to find the solution more quickly. The results confirm all the hypotheses and predictions.

6.6. Reproduction of Experiment 2 and 3 Results

6.6.1. Model's fit

Given their similarities, Experiments 2 and 3 were reproduced and analysed together.

For these experiments the model was able to accurately reproduce (significance of fuzzy kappa < 0.05) the behaviour of 79% of the participants. It also accurately reproduced the results of the experiment. The sign test of the learning intervals is significant ($p < 0.01$) and in the same direction as in the experiments. In other terms, also for the model, the learning of one category helps the learning of another other category with similar relational structure.

The results are the same as in the experiments also for the sign tests of the given vs. expected number of partially incorrect answers during learning. They are more than expected ($p < 0.01$), meaning that also for the model the final learning of similar categories is preceded by a phase of partial learning.

As for preferring to learn similar criteria than learn dissimilar ones, the sign test of the clicks on analogical and non-analogical elements gives the same result as in the experiments. The clicks[5] on the analogical elements are significantly more than on non-analogical elements ($p < 0.01$).

6.6.2. Parameters values

Examining the means of the parameters values found by the simulated annealing algorithm, it is possible to assess if and how much each rule creation method is used by the model (6.5).

Table 6.5: Average values of creation methods as found by the simulated annealing algorithm, experiments 2 and 3.

Parameters values represent the number of times each method is tried for each presented exemplar. Use values represent the times each method is used to arrive to the final versions of the correct rules.

	Mean	95% Confidence interval		Median
		Lower Bound	Upper Bound	
Parameters				
CreateFromExemplar	6.12	6.01	6.24	6.00
CreateModifying	6.36	6.26	6.47	7.00
CreateRandom	4.80	4.67	4.93	5.00
CreateUnifying	5.05	4.93	5.18	5.00
Effective use				
FromExemplarAvgUse	0.91	0.85	0.97	1.00
ModifyingAvgUse	3.68	3.35	4.01	2.00
FromRndAvgUse	0.0147	0.0117	0.0177	0.0000
UnifyingAvgUse	0.0041	0.0035	0.0047	0.0000

[5] The model exactly reproduces the behaviour of a human participant, so for the analysis it is undistinguishable. To discover the (possible) action associated to an element, the model has to "click" on that element, and only then it is told what "happens", exactly as for human participants.

It is clear also that for these experiments the modification of rules is effectively used; rules are changed an average of 3.68 times before becoming the final ones. Like before, it is not surprising that the creation based on the shown exemplars is used as the initial source of rules almost all the time. Both the probability of finding a good rule randomly and the probability of finding a good rule unifying other two, are very low.

It is also interesting to study the effect of the parameters on the difficulty (measured in number of exemplars needed to solve the test). For these experiments the significant correlations with the difficulty are more (6.6), also because of the greater number of participants.

Table 6.6: Significant parameters correlations with difficulty for Experiments 2 and 3.

CreateFromExemplar	-.188(**)
CreateRandom	.160(**)
ModifyingAvgUse	-.087(**)
PosMemSlots	-.152(**)
GoodSlotsAvgUse	-.087(**)
GoingBackProb	-.074(**)
RecordCorrectProb	-.197(**)
PartialRuleLimit	.402(**)
**. Correlation is significant at the 0.01 level (2-tailed).	

The new results (with respect to the repetition of the first experiment) are not surprising; both the creation of new predicates based on a shown exemplar and the available memory for good rules help learning.

The increasing difficulty level when the age limit for partial rules increases can be easily explained by the fact that the more a person waits to remove a partial hypothesis, the fewer are the opportunities to discover new good rules. The remaining results are the same as for the previous experiment, a fact that confirms the stability of the model. The lower correlations (in comparison with experiment 1) could be simply explained by the fact that different kinds of predicates are used in these two other experiments. Although all the predicates have the same functions, the actual implementations are different: thus it is not surprising that for different classes of problems the sensitivity of the model is different.

The model meets all the requirements for these tests too: it mimics people's answers, reproduces the experiment's results and uses the modification of rules as a heuristic to find the solution more quickly.

6.7. Testing constraints on the use of the heuristic

In order to test if the proposed heuristic of "form and test partial hypotheses and refine and re-use them" was effectively needed by the model to reproduce the participants' results, the model was also tested with some features disabled.

A second run of the simulated annealing was performed with the CreateModifying parameter fixed to the value of zero. This completely disabled the possibility of the model to re-use hypotheses. In this condition, the model was able to accurately reproduce the answers for only the 40% of the participants for Experiment 1, and the 38% for Experiments 2 and 3.

A third run was performed without the possibility to have partial rules. A rule could be valid for only one category or for none (which is the standard account of categorization). In this condition, the model was able to accurately reproduce the answers for only the 36% of the participants for Experiment 1, and the 37% for Experiments 2 and 3.

From these results it is clear that the model of reasoning, in itself, is able to solve the learning task mimicking some of the participants' behaviours, but the proposed heuristic is the keystone.

6.8. Summary

The proposed new heuristic of "form and test partial hypotheses and refine and re-use them" has been implemented and tested in this model. Obviously, the model had to include ways to interact with the tasks, to manage the memory, to create hypotheses and test them, etc. Therefore it needed also parameters to fine-tune its behaviour to best match the kind of reasoning and preferences of each participant. And in order to find the set of parameter which allowed the model to best mimic the participant's behaviour, a simulated annealing algorithm was used.

The model, with at the core the ability to create partial hypotheses and modify them, was able to accurately reproduce the answers of 81% of participants. Without this ability, it was able to accurately reproduce the answers of less than 40% of participants. Which means that the fine-tuning, alone, wasn't sufficient, and that the proposed heuristic was crucial to account for all the phenomena discovered in the three experiments.

The analysis of the sets of parameters which best reproduced the participants' answers showed incontrovertibly that the modification of the hypotheses was largely used.

These results clearly indicate that the experimental findings of the three experiments can be explained by the theory implemented in this model. The computational model, in order to accurately reproduce the participants' answers, needed the proposed heuristic. It can be thus inferred that the participants effectively used that heuristic.

More in general, it can be inferred that, when categories with structural similarities between them are simultaneously learned, partial hypotheses are formed and then refined and re-used. In addition, since this model didn't use structure mapping (to create, to modify or to test hypotheses, or in any other form), this demonstrates that an analogical reasoning process emerges anyway from the use of the proposed heuristic without the need of structure mapping. This is also consistent with the findings of Experiment 1, which hinted that structure mapping didn't have a role in the learning process.

Because this model used random modification of rules, it is surely the most general case, and any better algorithm of modification could be included to further improve the model's performance.

This is also true for the kinds of predicates. The proposed model included only the predicates which could be effectively used to represent classification hypotheses on the exemplars of the three experiments. But the model is open to any other kind of predicate, and can be extended with algorithms for the generation *ex-novo* of new kinds of predicates. The only constraint on predicates is that they must be randomly modifiable to create new predicates with the modification method.

In summary, the proposed implementation of the heuristic of "form and test partial hypotheses and refine and re-use them" can accurately mimic the participants' behaviour and reproduce all the results of the three experiments: learning of similar categories is related, similar categories are learned in a single process, final learning is preceded by a phase of partial learning, direct comparison of exemplars is not used, people prefer similar criteria over easily discernible ones, and people form hypotheses and test those hypotheses in order to reject or refine them.

As said in the introduction, these findings are not in contrast with other models of analogical reasoning or category learning, which could be complemented by the proposed heuristic. The ideas which form this model can be generalized to other domains of learning and reasoning. In order to do so, more experiments could be needed, and this would be outside the current scope. This model could help expand our knowledge of human reasoning by complementing the explanation of a large class of phenomena in learning.

Chapter 7

Conclusions

This work describes a model of analogical reasoning that diverges from all those which have been produced so far.

The standard models of analogical reasoning (e.g. Gentner, 1981; Gentner, 1983; Holyoak & Koh, 1985) allow the common structure between two domains to be exploited. One structure is methodically mapped onto the other, in order to find the exemplars in the target which correspond to the roles inferred from the base domain. These models therefore can't predict what happens when analogy occurs between two partially understood domains.

The simultaneous learning of categories with similar structures is an aspect of analogical reasoning which has never been explicitly studied in more than twenty years of the study of analogical reasoning. For decades the attention has been focused on the similarities between exemplars of the same category ("within category" similarities), and on analogy between structures, at least one of which is assumed to be complete. Consequently all the models that have been based on this assumption dealt with high level problems like mapping, concept alignment, and their computational complexity. These kinds of problems are central for that kind of analogical reasoning, but they become immediately superfluous when one considers that in human learning analogical reasoning isn't necessarily of that type: analogy between complete structures is instead a rare case.

On the contrary, in order to fill this gap, this thesis intended to investigate the case of similarities between different categories, all learned simultaneously. It seemed much important to lead the studies on analogical reasoning toward the investigation of how analogy happens when simultaneously-learning various similar concepts. This is in fact a more general phenomenon, which happens more frequently than the analogy between completed structures. It is even more important to stress the importance of this aspect, considering that for decades it has been overlooked, while focusing on studies that can be applied only to a few cases of human reasoning.

Another essential problem that all research in the area of cognitive science must take into account is time and memory constraints. These constraints have a big impact on analogical reasoning and generally on human learning. Even so, the approaches that dominated the study of analogical reasoning haven't yet devised a satisfying solution to the problem.

The proposed heuristic of "form and test partial hypotheses and refine and re-use them", at the core of the proposed model, can maximize the amount of extracted information, thus minimizing the memory and time efforts.

It is because of the limits imposed by memory and time, that the human mind chooses to resort to analogy to learn unknown concepts. Our mind does so also when simultaneously learning unknown concepts, as already suggested by some studies (e.g. Keane, 1995) and further shown by the results of my model.

Another aspect overlooked by current research on analogy is the contiguity between analogical reasoning and scientific thinking. Although research on analogy started from this aspect, in recent years it has been abandoned. My study rediscovers this link, for my model highlights that human learning proceeds through the formulation of a series of hypotheses that are tested and accepted, rejected or partially modified, with a process very similar to scientific reasoning. As said, this aspect was central in the first studies on analogical reasoning (Clement, 1981; Del Re, 2000; Hoffman, 1980; Oppenheimer, 1956), but the subsequent developments progressively diverted the attention from this interesting problem, whose investigation potentiality seemed already exhausted a few years ago. In the light of the new implications which have emerged from this study, it is of great importance for the future development of cognitive science to return to investigate the link between analogy and scientific thinking, with new models based on the process of partial learning and subsequent refinement as proposed in the present work.

7.1. Future work

Given the salience of the discoveries presented in this work for the future development of cognitive science, some suggestions arise about what needs to be studied. They are presented here only briefly, since they directly descend from the discussion above.

The idea that simultaneous learning of similar concepts is in reality a single process could be further expanded, as well as the idea that analogy can emerge from modification of a concept. This last idea is related to Keane's proposal of incremental analogy (1995; 1994) which could be expanded to form a new framework of analogy-making which doesn't need explicit mapping.

Other interesting work could be done on the early use of analogy, in the cases in which there isn't enough time or there aren't enough resources to learn concepts separately and then use structure mapping.

It could be interesting, for example, to investigate, in the light of the present study, the use of analogy in irony, in idiomatic expressions, in insults, and even in dreams.

In summary, this present work opens a new field of research on analogy and category learning, and proposes that analogy can also be an emergent ability. Analogy is not necessarily used explicitly, and it can emerge from a heuristic of "form and test partial hypotheses and refine and re-use them".

Acknowledgements

It will be difficult to find the right words to thank all the people who helped me in various ways through my PhD and the writing of this thesis.

Thanks to my wife Isabella, without whose help, support and encouragement this thesis would have never seen the light. There are no words to express how grateful I am to her.

Thanks to Dervla, for helping me with my English and for sharing many aspects of our PhDs, but mainly for her friendship, her sagacious remarks, and her introduction to the Irish culture.

Thanks to my supervisor Dr. Fintan Costello, for his insights and suggestions, but also for his patience and modesty. I hope I have learned from him as a person as much as I have learned from him as a scholar.

Thanks to all the people of the Analogy Project: Boicho, Elisabetta, Valentina, Gloria, Irena, Denis, Robert, Sabina, Usha, Vicky, Milena, Susanna, Svetlana. Not only they permitted my PhD to exist, but it has been a big opportunity to be able to avail of their culture and experience.

Thanks to the many volunteers who participated to my experiments, often only for the glory. They have sacrificed time and grey matter to let me do my research. Without their good will this thesis would have been simply impossible.

Thanks to Rita for her support in a decisive moment of my PhD, and for her advice in all these years.

Thanks to all the other people who gave me advice, support, help of various kind in these years. I feel I am in debt with many people who contributed to my research, and without whom it would have been much more difficult, if not impossible.

I'd also like to thank the few people I met in Dublin and who became my friends. They made my survival in Ireland more tolerable. Thanks for the chats and for your support.

Finally, thanks to the funding of the European Commission: the FP6 NEST Programme of the European Commission (ANALOGY: Humans – the Analogy-Making Species: STREP Contr. No 029088).

Appendices

Appendix A

Statistical distributions of interval averages and differences

The analysis of learning intervals, as proposed for the three experiments, creates the problem of analysing their distribution, and the distribution of their differences. The analysis can be summarized as follows. Having 3 or 4 points in time (A, B, C and D), in random order, how is the probability of the following differences distributed?

- Experiment 1:

$$d=\overline{AB}-\frac{\left(\overline{AB}+\overline{AC}+\overline{AD}+\overline{BC}+\overline{BD}+\overline{CD}\right)}{6} \tag{A.1}$$

  or equivalently:

$$d=\overline{CD}-\frac{\left(\overline{AB}+\overline{AC}+\overline{AD}+\overline{BC}+\overline{BD}+\overline{CD}\right)}{6} \tag{A.2}$$

- Experiment 2:

$$d=\frac{\left(\overline{AB}+\overline{AC}+\overline{BC}\right)}{3}-\frac{\left(\overline{AB}+\overline{AC}+\overline{AD}+\overline{BC}+\overline{BD}+\overline{CD}\right)}{6} \tag{A.3}$$

- Experiment 3:

$$d=\overline{AB}-\frac{\left(\overline{AB}+\overline{AC}+\overline{BC}\right)}{3} \tag{A.4}$$

A MonteCarlo simulation was performed, assigning random values to the points, and then normalizing them so that the smallest number was 0 and the highest was 1 (this step was needed so the various repetitions could be compared). The computation of the intervals and their differences was repeated 1e+6 times, and the statistical distribution recorded.

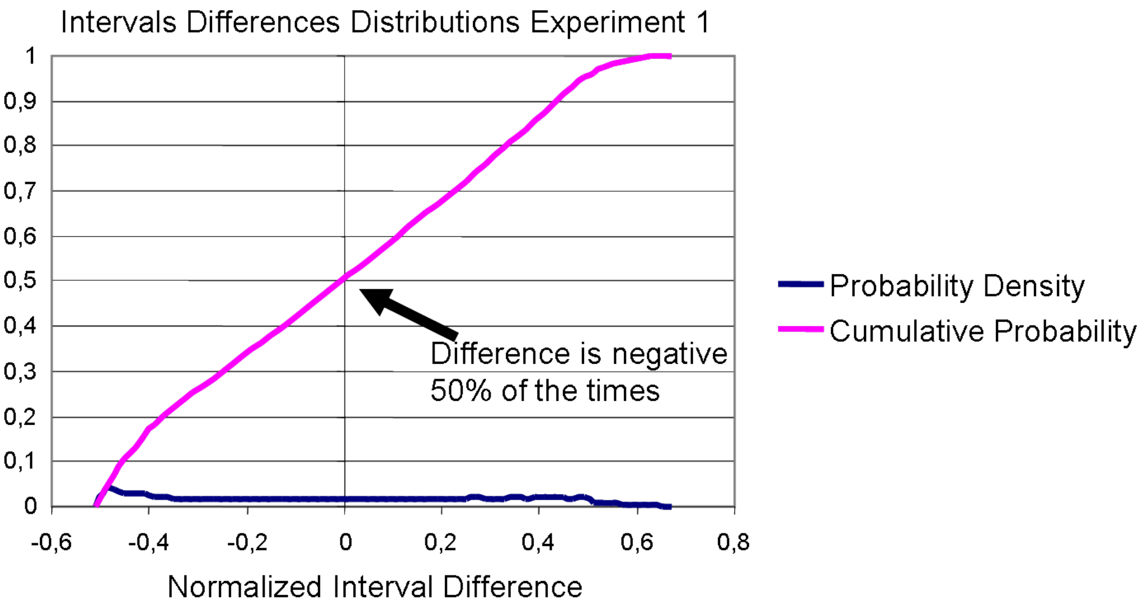


Figure A.1: Distributions of Intervals and Differences for Experiment 1.

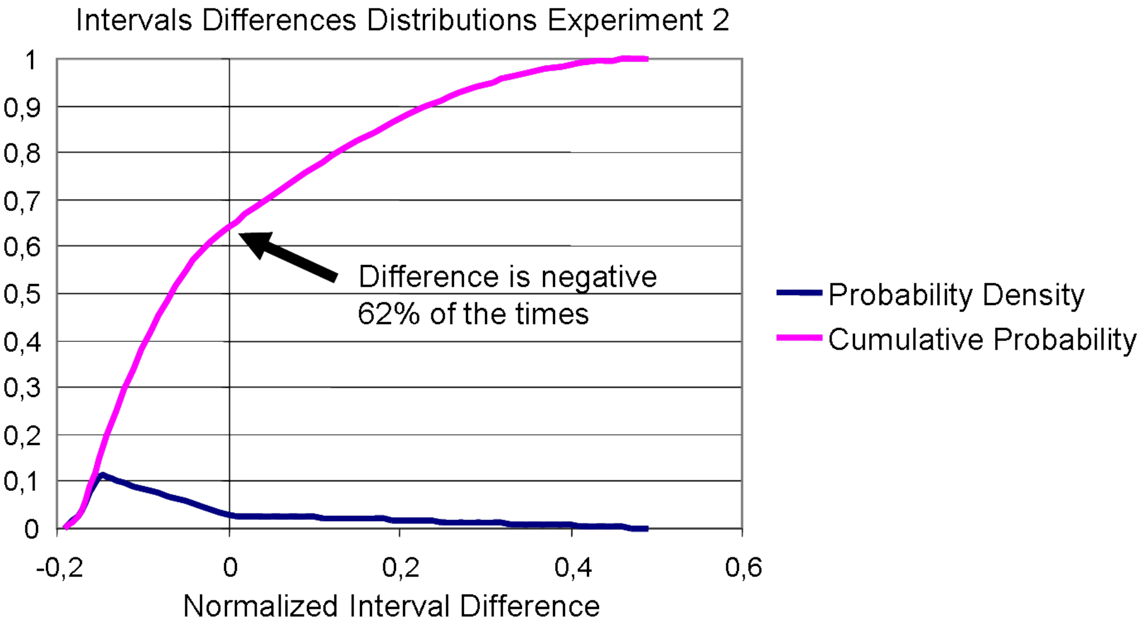


Figure A.2: Distributions of Intervals and Differences for Experiment 2.

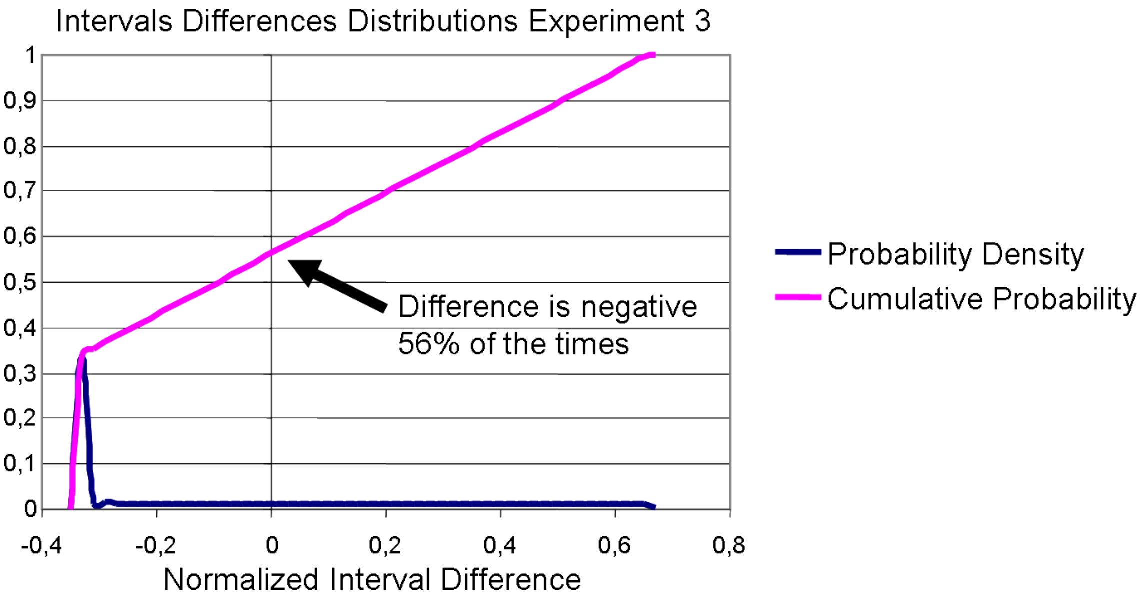


Figure A.3: Distributions of Intervals and Differences for Experiment 3.

It is evident from Figures A.1, A.2 and A.3 that the distributions don't approximate to a Gaussian curve. Therefore they cannot be analysed using parametric statistics like the Student's *t*-test.

Anyway it is possible to use a non parametric test like the binomial test, to compare the times the differences are positive to the times they are negative. In fact, knowing the expected distributions of negative and positive differences of random points, the null hypothesis is that the real distributions are not different from random.

It can be argued that the distributions would change without the normalization of the points to fit the interval [0-1]. Although this is true, it can be easily demonstrated that the proportion of negative and positive differences is independent of any linear transformation of the points' coordinates. Therefore, even without normalization, the binomial test can be safely used.

Appendix B

Significance of Cohen's Kappa

The most used measure of inter-rater agreement for qualitative items is Kappa, proposed by Cohen in 1960. Differently from simple percent agreement, Kappa takes into account the agreement occurring by chance. It is computed with the formula:

$$k = \frac{P(a) - P(e)}{1 - P(e)} \tag{B.1}$$

where *P(a)* is the relative observed agreement among raters, and *P(e)* is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly choosing each category. Theoretically, Kappa ranges from -1 (complete disagreement) to 1 (complete agreement). In reality, the actual possible range of values is a function of the number of items being rated and the number of categories. Notwithstanding this, Landis and Koch (1977) proposed the following (arbitrary) table to interpret Kappa values:

Table B.1: Landis and Koch (1977) proposed interpretations of Kappa values

Kappa statistic	Strength of agreement
< 0.0	Poor
0.0 - 0.2	Slight
0.2 - 0.4	Fair
0.4 - 0.6	Moderate
0.6 - 0.8	Substantial
0.8 - 1.0	Almost perfect

Instead, as observed by Gwet (2001), the Cohen's Kappa (1960) distribution varies with the number of categories and items. That is, the generally accepted values proposed by Landis and Koch (1977) to assess its significance are arbitrary and can't be used in all situations.

Because the mathematical derivation of the distribution of Kappa is prohibitive, two other approaches are generally used to evaluate its significance.

The first approach is to assume, based on the central limit theorem, that Kappa approximates to a normal distribution. Various methods have been then proposed to estimate the variance of a computed kappa. With the variance and the assumption of normal distribution, the significance is then calculated. Unfortunately this approach is generally wrong. The approximations to a normal distribution holds (Gwet, 2001) only for large ($n > 30$) quantities of items and when the number of items is greater than the number of categories ($n > c$), but usually this is not the case.

The second approach for the estimation of the significance of a given kappa is using MonteCarlo simulations. A large ($n = 1e+5$) number of kappas is computed using random answers to a given number of items between a given number of categories. The distribution of these random computed kappas can then be used to assess the probability that an observed kappa has occurred only by chance.

Using this method, it is also possible to estimate the *real* variance of kappa for a given number of items and categories. Moreover, with a chi-square test it is possible to assess the goodness of the approximation to a normal distribution.

Figure B.1 shows the comparison of distributions of kappa, computed for 4 categories (i.e. the case of Experiment 1) and for various numbers of items. For small numbers of items the distributions poorly approximate a normal function. Moreover, the variance decreases with increasing numbers of items.

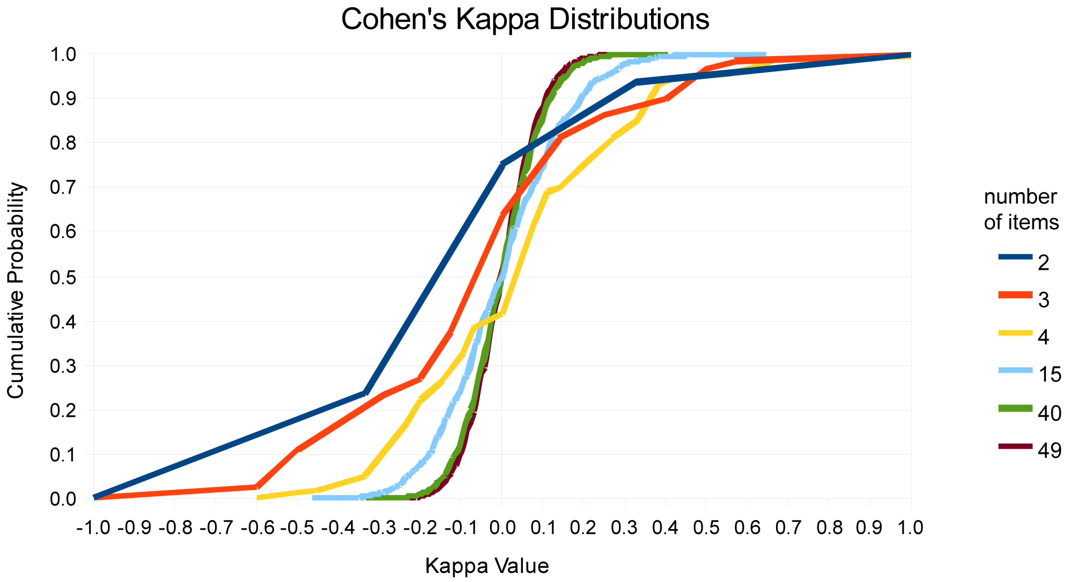


Figure B.1: Cohen's Kappa cumulative distributions for 4 categories and various numbers of items (MonteCarlo simulation).

In summary, the values proposed by Landis and Koch (1977) are totally arbitrary, and can't be used to assess the real significance of a computed kappa. Also the method based on the approximation to a normal distribution can't be used for small sets of items. Therefore, for this work the MonteCarlo method was chosen to estimate the significance of the model's ability to reproduce the human behaviour.

Appendix C

Measures of inter-rater agreement for multiple responses

The most used measure of inter-rater agreement is Cohen's Kappa, but in a case with multiple responses for each item, clearly it cannot be used. Therefore two other measures have been tested: a fuzzy version of Kappa, as proposed by Dou et al. (2007), and the average rank difference.

C.1. Fuzzy Kappa

The Fuzzy Kappa is a generalization of Cohen's Kappa for fuzzy sets. Therefore, in a case with four categories, a typical rating situation could be as follows:

Category	**A**	**B**	**C**	**D**	**Tot**
Rater 1	0	0,2	0,2	0,6	1
Rater 2	0,4	0	0,2	0,5	1
Agreement	0	0	0,2	0,5	0,7

The agreement is computed for each category with an intersection function, which usually is the minimum function. The total agreement on a given item is the sum of the agreements for each category:

$$f^F(x)=\sum_{i=1}^{N} \mu_i^A(x) \wedge \mu_i^B(x) \tag{C.1}$$

where *x* is the item, *A* and *B* are the raters, and $\mu_i(x)$ is the fuzzy value for category *i.*

The total agreement between the two raters is the sum of the agreements for each item:

$$P_0^F=\frac{1}{M}\sum_{m=1}^{M} f^F(x_m)=\sum_{m=1}^{M}\sum_{i=1}^{N} \mu_i^A(x_m) \wedge \mu_i^B(x_m) \tag{C.2}$$

Dou et al. (2007) derive the expected random agreement as:

$$P_e^F=\sum_{i=1}^{N} \int_{\mu_i^A=0}^{1} \int_{\mu_i^B=0}^{1} p(\mu_i^A)\, p(\mu_i^B)(\mu_i^A \wedge \mu_i^B)\, d\mu_i^A\, d\mu_i^B \tag{C.3}$$

where $p(\mu_i)$ is the probability of fuzzy value μ for category *i.*

The Fuzzy Kappa is then computed as the *standard* Kappa:

$$K^F=\frac{P_0^F-P_e^F}{1-P_e^F} \tag{C.4}$$

The Fuzzy Kappa has the same properties as the *standard* Kappa. Moreover, for the special case in which one fuzzy value = 1 and all the others = 0 (that is, absolute certainty on one classification), the Fuzzy Kappa will retrogress to a *standard* Kappa.

For this thesis, an algorithm for the computation of the Fuzzy Kappa has been implemented in C# (Figure C.1).

```
public class FuzzyKappa
{
   private Dictionary<double, int>[] margAnsw1, margAnsw2;
   private double answAgreement;
   private int totItems, totCategories;

   public FuzzyKappa(int totCategories)
   {
      this.margAnsw1 = new Dictionary<double, int>[totCategories];
      this.margAnsw2 = new Dictionary<double, int>[totCategories];
      this.answAgreement = totItems = 0;
      this.totCategories = totCategories;

      for (int i = 0; i < totCategories; i++)
      {
         margAnsw1[i] = new Dictionary<double, int>();
         margAnsw2[i] = new Dictionary<double, int>();
      }
   }

   public void AddClassification(double[] answ1, double[] answ2)
   {
      totItems++;

      //for each category
      for (int i = 0; i < totCategories; i++)
      {
         //increase the general agreement
         answAgreement += Math.Min(answ1[i], answ2[i]);

         //by default a Dictionary doesn't contain a value
         if (!margAnsw1[i].ContainsKey(answ1[i]))
            margAnsw1[i][answ1[i]] = 0;
         if (!margAnsw2[i].ContainsKey(answ2[i]))
            margAnsw2[i][answ2[i]] = 0;

         //increase the probability of getting that fuzzy value
         //for that category
         margAnsw1[i][answ1[i]]++;
         margAnsw2[i][answ2[i]]++;
      }
   }

   public double CalcFuzzyKappa()
   {
      if (totItems == 0) return 0;

      //calculate the random agreement
      double pe = 0;
      for (int i = 0; i < totCategories; i++)
         foreach (KeyValuePair<double, int> m1 in margAnsw1[i])
            foreach (KeyValuePair<double, int> m2 in margAnsw2[i])
               pe += m1.Value * m2.Value * Math.Min(m1.Key, m2.Key);

      pe /= (double)totItems;

      //in case of perfect agreement, avoid division by zero
      if (totItems == pe) return 1;

      return ((double)answAgreement - pe) /
         ((double)totItems - pe);
   }
}
```

Figure C.1: Code of the implementation in C# of Fuzzy Kappa

C.1.1. Quantification of uncertainty

There could be many different solutions for the derivation of the uncertainty (and therefore the fuzzy values) from the participants' answers. Take for example an item of category A, to which a participant gave the answers B, C and finally A. Obviously, the final answer must be the correct one (and no other answer can be given afterwards). But the estimation of the participant's uncertainty changes, if the participant was answering randomly, or was in doubt between A, B and C, or for example in the beginning he was certain about B. As it is impossible to ascertain the participant's knowledge, a general method must be decided.

The chosen method was the simplest one: give the same weight to each answer. So, in the case above, both A, B and C would be given 0.333 as fuzzy values. This method creates a bias toward agreement, but the method (see below) chosen for the estimation of the significance takes this into account.

C.1.2. Sensitivity to disagreement

In order to assess the sensitivity of fuzzy kappa to different levels of disagreement, a simple method was used. Taking the real sets of answers in a participant's test session, some of them were randomly replaced, and the new session with random replacements was compared to the original one. Increasing the number of random replacements, the kappa value is expected to decrease.

The replacement method took into account the peculiarities of the test. That is, the final answer had always to be the correct one. In 1 there is an example of the possible replacements.

Table C.1: Examples of random replacement of answers

Item's category	**A**	**B**	**A**	**C**
Original answers	B **A**	**B**	C D **A**	A **C**
Random replaced answers	C D **A**	A **B**	**A**	B **C**

Each participant's session was compared to randomly modified versions several times, for various quantities of random replacements. The correlation between the fuzzy kappa and the number of random replacement is R = -0.955, p< 0.01 (see Figure C.2). This measure is therefore very sensitive to increasing levels of disagreement, and it is a good candidate to test the model's ability to reproduce the participants' behaviour.

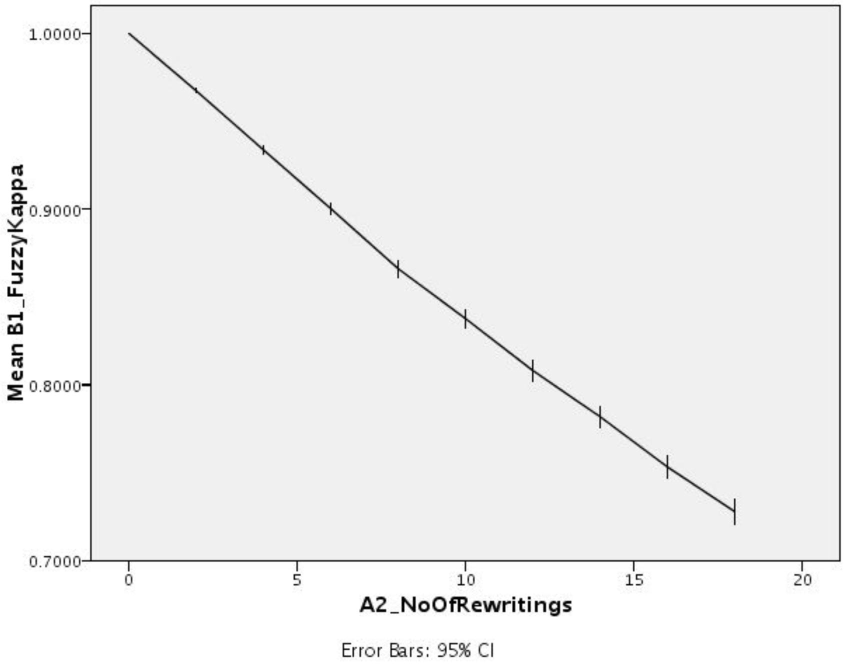


Figure C.2: Correlation between Fuzzy Kappa and number of random replacements.

C.1.3. Significance

Since the distribution of the fuzzy kappa changes with different quantities of items (as the standard kappa), a method to assess the significance of an obtained value is needed. Moreover, given the particular method used to quantify the fuzzy values from the given answers, each session has its own distribution of kappa.

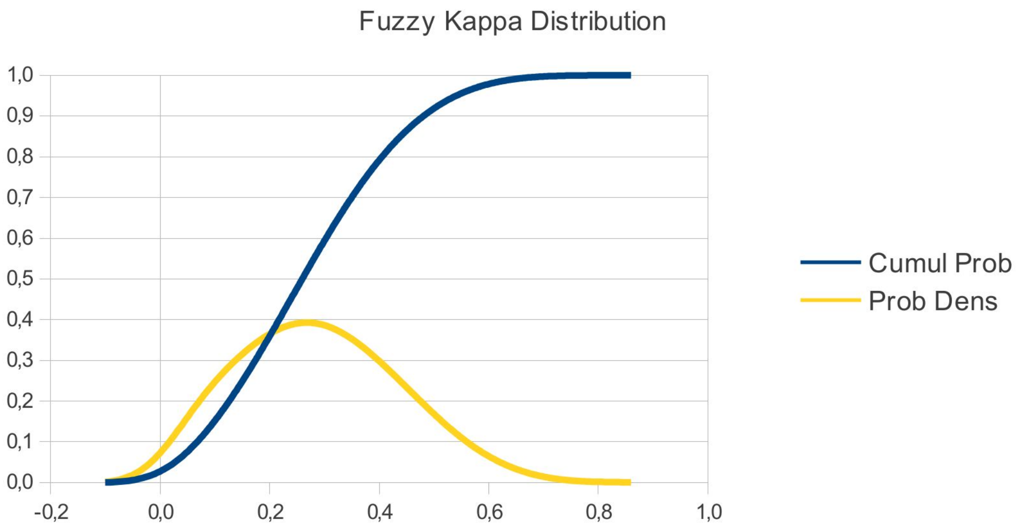


Figure C.3: A typical distribution of Fuzzy Kappa values for random answers

The method chosen to estimate the distribution of the fuzzy kappa for each session takes into account the constraints deriving from the test. That is, each set of answers must end with the right answer. Therefore the same replacement method as shown above was used to randomly replace all the answers. Each session has been compared to 1e+5 new random sessions, and the distribution of the resulting fuzzy kappa values has been computed (Figure C.3 for one of the distributions). In this way, for each session it is possible to estimate the significance of the agreement between the original answers and the model's answers.

C.2. Average rank difference

Another method to assess the inter-rater agreement (or better in this case the disagreement) in a case with multiple responses is the rank difference. The computation is simple and depends on the order of the answers of both the raters.

For example, if rater 1 answers (B, C, A), and rater 2 answers (B, A), the ranks will be assigned as in 2.

Table C.2: Computation of rank difference, based on order of answering

		Ranks				Total
	Answers	A	B	C	D	
Rater 1	B, C, A	3	1	2	4	
Rater 2	B, A	2	1	3,5	3,5	
Difference		1	0	1,5	0,5	3

The average of the rank differences of all the items is then a measure of the disagreement between the two raters.

C.2.1. Sensitivity to disagreement

To assess the sensitivity of the average rank difference the same method of the fuzzy kappa was used. The original sessions were compared to randomly modified copies, and the number of answer replacements was the independent variable measuring the actual level of disagreement.

As for fuzzy kappa, each participant's session was compared to randomly modified versions several times, for various quantities of random replacements. The correlation between the average rank difference and the number of random replacement is R = 0.953, p< 0.01 (see Figure C.4). Thus this measure is also very sensitive to increasing levels of disagreement, and it is another good candidate to test the model's ability to reproduce the participants' behaviour.

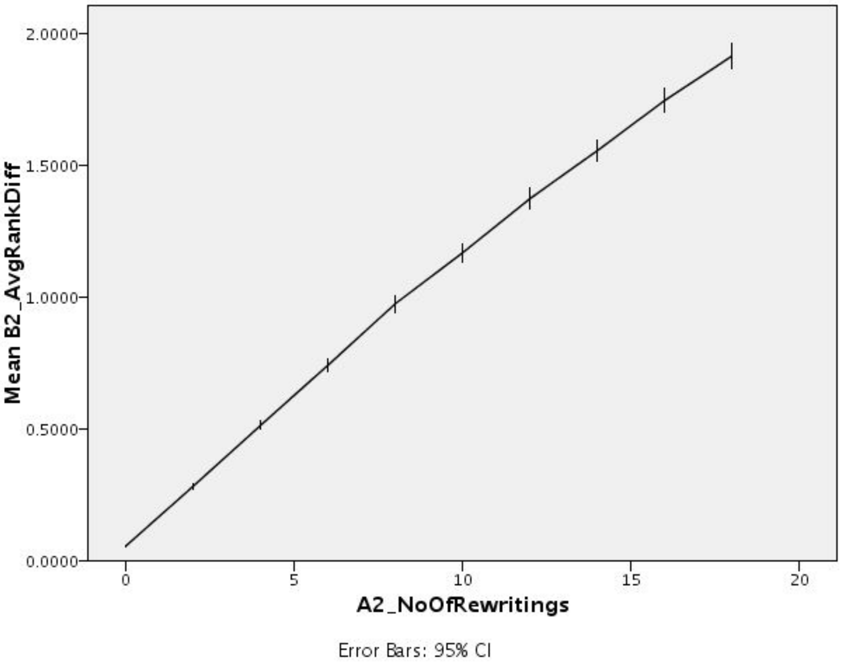


Figure C.4: Correlation between average rank difference and number of random replacements

C.2.2. Significance

To estimate the significance of an obtained average rank difference value, a method similar to the fuzzy kappa was used. The same replacement method was used to randomly replace all the answers. Each session was compared to 1e+5 new random sessions, and the distribution of the resulting fuzzy kappa values was computed (Figure C.5 for one of the distributions). In this way, for each session it is possible to estimate the significance of the agreement between the original and the model's answers.

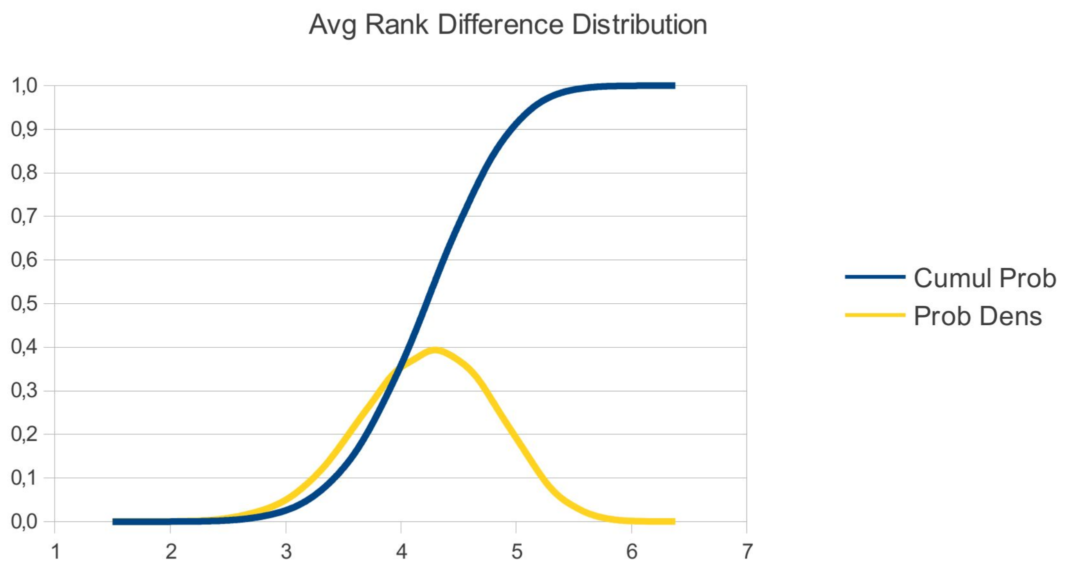


Figure C.5: A typical distribution of average rank difference values for random answers

C.3. Correlation between fuzzy kappa and average rank difference

As a last proof that the two measures effectively quantify the same variable, they have been compared using, for their computation, the same pairs of original and randomly modified sessions. The correlation between the obtained fuzzy kappas and average rank differences is $R = -0.986$, $p < 0.01$ (Figure C.6). Thus they are essentially the same measure, and one of the two can be arbitrary chosen.

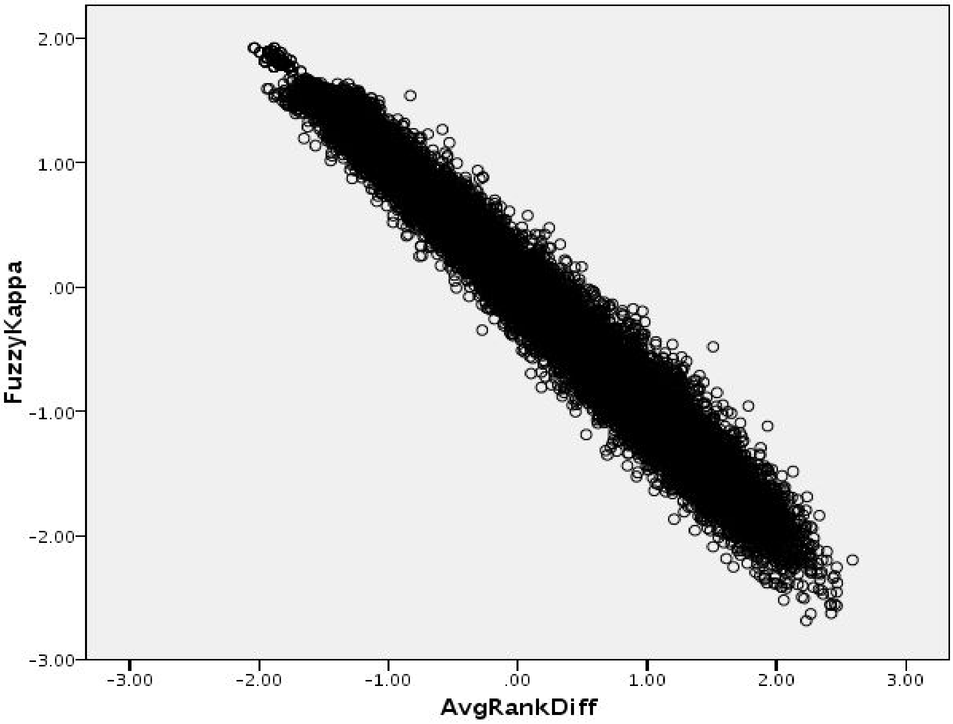


Figure C.6: Correlation between average rank difference and fuzzy kappa

Appendix D

Text of instructions for Experiment 2

A toy factory purchased a machine to produce interactive toys. The manager responsible for the invention of these toys had designed the machine to produce 5 different types of toys. The machine was not working correctly, however, which meant that as well as producing these 5 different types of toys it also produced random toys which didn't work and had to be discarded. Then, overnight, the manager changed his job and took with him all the documentation for this project.

The factory has decided to hire an expert (i.e. you) to reconstruct the missing manager's ideas. The machine in the meantime has produced a lot of toys of all 5 types, and many toys which didn't work. The factory owner has asked you to study these toys and learn to identify which toys are of which type and which toys don't work.

Your task is to learn to classify the toys you will be shown. For each toy you will have to guess which type of toy it is, and the old manager (at great trouble to the factory) will be called to tell you if you guessed right or not. **If you guess wrong, you must try guessing another type until you find the correct answer. In the beginning you will have no idea of the type of any toy and you will have to choose randomly, but slowly you will learn how to classify the toys.** Every mistake you make will be considered a cost by the factory, and your target is to **learn to classify the toys with as few mistakes as possible.**

You will be shown four toys a time, chosen completely at random, and when you have labelled them all correctly, an arrow button will appear allowing you to go on to the next four toys. At any time you can go back to look at the toys already labelled, by clicking on a "back" arrow. You will be able to go back and forth any time you want without losing the work you have already done. This will allow you to look again at previously identified toys and then return to the ones to be labelled. You won't be allowed to take notes with pen and paper.

When you will have learned to correctly classify all the toys (i.e. when you can correctly label them at your first guess), the computer will tell you that your task is completed. You will then have to write a brief report to teach a worker how to identify the different types of toys. This worker will then go on with the boring labelling job.

A brief tutorial will now teach you how the task works. Please follow all the steps to practice with the toys and the labels that are used to identify them. Now close these instructions to start the tutorial. You can reopen these instructions at any time.

References

Ashby, F. G., & Casale, M. B. (2003). The cognitive neuroscience of implicit category learning . in: L. Jiménez (Ed.), *Attention and implicit learning*, 109-141. Amsterdam: John Benjamins Publishing Company.

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

Ashby, F. G., & Maddox, W. T. (2005). Human Category Learning. *Annu. Rev. Psychol*, *56*, 149-78.

Astington, J. W., Harris, P. L. & Olson, D. R. (1990). Developing theories of mind. *Cambridge University Press.*

Barsalou, L. W. (1983). Ad hoc categories. *Memory and cognition*, *11*(3), 211-227.

Berns, G.S., Cohen, J.D. & Mintun, M.A. (1997). Brain regions responsive to novelty in the absence of awareness. *Science 276*, 1272–1275

Berry, D. C. & Broadbent, D. E. (1984). On the relationship between task performance and associated verbalizable knowledge. *Q. J. Exp. Psychol. 39*, 585–609

Berry, D. C. & Dienes, Z. (1993). Implicit Learning: Theoretical and Empirical Issues. *Erlbaum*

Boroditsky, L. (2007). Comparison and the development of knowledge. *Cognition, 102*, 118–128

Butterworth, G.E., Harris, P.L., Leslie, A.M. & Wellman, H.M. (1991). Perspectives on the child's theory of mind. *Oxford University Press.*

Cleeremans, A., Destrebecqz, A., & Boyer, M. (1998). Implicit learning: news from the front . *Trends in Cognitive Sciences* , *2*(10), 406–416

Clement, J. (1981). Analogy Generation in Scientific Problem Solving. *Proceedings of the Third Annual Meeting of the Cognitive Science Society*.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37-46.

Del Re, G. (2000). Models and analogies in science. *HYLE–International Journal for Philosophy of Chemistry*, *6*(1), 5-15.

Dou, W., Ren, Y., Wu, Q., Ruan, S., Chen, Y., Bloyet, D., & Constans, J. M. (2007). Fuzzy kappa for the agreement measure of fuzzy classifications. *Neurocomputing*, *70*(4-6), 726-734.

Doumas, L. A. A., & Hummel, J. E. (2005). A symbolic-connectionist model of relation discovery. In *Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society* (pp. 606-611).

Doumas, L. A. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review*, *115*(1), 1-43.

Falkenhainer, B., Forbus, K. D., & Gentner, D. (1986). *The structure-mapping engine*. Department of Computer Science, University of Illinois at Urbana-Champaign.

Feldman, J. (2003). The simplicity principle in human concept learning. *Current Directions in Psychological Science*, 227-232.

Frye, D. & Moore, C. (1991). Children's theories of mind: Mental states and social understanding. *Psychology Press.*

Gasser, M., & Colunga, E. (2001). Learning relational correlations. In *International Conference on Cognitive Modeling* (Vol. 4, pp. 91-96). Citeseer.

Gelman, S. A., Raman, L., & Gentner, D. (2009). Effects of language and similarity on comparison processing. *Language Learning and Development, 5*(3), 147-171.

Gentner, D. (1981). A structure-mapping theory of metaphor and analogy. *Bullettin of the psychonomic society*, *18*(2), 77.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive science*, *7*(2), 155-170.

Gentner, D. (1993). The shift from metaphor to analogy in Western science. In *Metaphor and thought (2nd ed.).* (pp. 447-480). New York: Cambridge University Press.

Gentner, D., & Kurtz, K. J. (2005). Relational categories. *Categorization inside and outside the laboratory: Essays in honor of Douglas L. Medin, ed. WK Ahn, RL Goldstone, BC Love, AB Markman & P. Wolff*, 151–75.

Gentner, D., Loewenstein, J., & Hung, B. (2007). Comparison facilitates children's learning of names for parts. *Journal of Cognition and Development, 8*(3), 285-307.

Gentner, D., Loewenstein, J. & Thompson, L. (2003). Learning and Transfer: A General Role for Analogical Encoding. *Journal of Educational Psychology, 95*(2), 393– 408

Gentner, D., & Markman, A. B. (1994). Structural alignment in comparison: No difference without similarity. *Psychological science*, 152-158.

Gentner, D., & Medina, J. (1998). Similarity and the development of rules. *Cognition*, *65*(2-3), 263-297.

Gentner, D., & Namy, L. L. (1999). Comparison in the development of categories. *Cognitive Development*, *14*(4), 487-513.

Gentner, D., & Namy, L. L. (2006). Analogical processes in language learning. *Current Directions in Psychological Science, 15*(6), 297-301.

Gopnik, A. (1984). Conceptual and Semantic Change in Scientists and Children: Why There Are No Semantic Universals. *Linguistics, 20*, 163-79.

Gopnik, A. (1988). Conceptual and Semantic Development as Theory Change. *Mind and Language, 3*, 197-217.

Gopnik, A. (2000). Explanation as orgasm and the drive for causal understanding: The function, evolution, and phenomenology of the theory-formation system. *In F. C. Keil & R. A. Wilson (Eds.), Explanation and cognition (pp. 299–324). MIT Press.*

Goswami, U., & Brown, A. L. (1990a). Melting chocolate and melting snowmen: Analogical reasoning and causal relations. *Cognition*, *35*(1), 69-95.

Goswami, U., & Brown, A. L. (1990b). Higher-order structure and relational reasoning: Contrasting analogical and thematic relations. *Cognition*, *36*(3), 207-226.

Gwet, K. (2001). Handbook of inter-rater reliability: How to estimate the level of agreement between two or multiple raters. *Gaithersburg, MD, STATAXIS Publishing Company*.

Hahn, U., Chater, N., & Richardson, L. B. (2003). Similarity as transformation. *Cognition*, *87*(1), 1-32.

Hammer, R., Diesendruck, G., Weinshall, D., & Hochstein, S., The development of category learning strategies: What makes the difference? *Cognition, 112*, 105–119

Hazeltine, E., Grafton, S. T. & Ivry, R. (1997). Attention and stimulus characteristics determine the locus of motor sequence encoding: a PET study. *Brain 120*, 123–140

Hoffman, R. R. (1980). Metaphor in science. *Cognition and figurative language*, 393–423.

Holyoak, K. J., & Koh, K. (1985). The pragmatics of analogical transfer. *The psychology of learning and motivation: Advances in research and theory*, *19*, 59-87.

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*, *110*(2), 220-264.

Keane, M. T. (1995). On order effects in analogical mapping: Predicting human error using IAM. In *Proceedings of the seventeenth annual conference of the Cognitive Science Society: July 22-25, 1995, University of Pittsburgh* (p. 449). Lawrence Erlbaum.

Keane, M. T. (1996). On adaptation in analogy: Tests of pragmatic importance and adaptability in analogical problem solving. *Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, *49*, 1062-1085.

Keane, M. T. (1997). What Makes an Analogy Difficult? The Effects of Order and Causal Structure on Analogical Mapping. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(4), 946-967.

Keane, M. T., Ledgeway, T., & Duff, S. (1994). Constraints on analogical mapping: A comparison of three models. *Cognitive Science*, *18*(3), 387-438.

Kittur, A., Hummel, J. E., & Holyoak, K. J. (2004). Featurevs. relation-defined categories: Probab (alistic) ly not the same. In *Proceedings of the Twenty Six Annual Conference of the Cognitive Science Society* (pp. 696-701).

Knowlton, B.J., Ramus, S.J. & Squire, L.R. (1992), Intact artificial grammar learning in amnesia: dissociation of classification learning and explicit memory for specific instances. *Psychol. Sci. 3*, 172–179

Kokinov, B., Bliznashki, S., Kosev, S., & Hristova, P. (2007). Analogical Mapping and Perception: Can Mapping Cause a Re-Representation of the Target Stimulus? In *Proceedings of the 29th Annual Conference of the Cognitive Science Society. Erlbaum, Hillsdale, NJ*. Citeseer.

Kotovsky, L., & Gentner, D. (1996). Comparison and categorization in the development of relational similarity. *Child Development, 67*(6), 2797-2822.

Kuehne, S. E., Forbus, K. D., & Gentner, D. (2000). SEQL: Category learning as progressive abstraction using structure mapping. *Proc. 22nd Ann. Conf. Cognitive Science Soc.*

Kurtz, K. J., & Boukrina, O. (2004). Learning Relational Categories by Comparison of Paired Examples. *Proceedings of the 26th annual conference of the cognitive science society.*

Kurtz, K. J., & Gentner, D. (1998). Category Learning and Comparison in the Evolution of Similarity Structure. *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society,*

Kurtz, K. J., & Loewenstein, J. (2007). Converging on a new role for analogy in problem solving and retrieval: When two problems are better than one. *Memory and Cognition*, *35*(2), 334.

Kurtz, K. J., Miao, C. H., & Gentner, D. (2001). Learning by analogical bootstrapping. *The Journal of the Learning Sciences*, *10*(4), 417-446.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159-174.

Lavrac, N., & Dzeroski, S. (1994). Inductive logic programming. *Journal of Logic Programming*, *19*(20), 629-679.

Lewicki, P., Hill, T. & Bizot, E. (1988). Acquisition of procedural knowledge about a pattern of stimuli that cannot be articulated. *Cognit. Psychol. 20*, 24–37

Lipkens, R., & Hayes, S. C. (2009). Producing and recognizing analogical relations. *Journal of the Experimental Analysis of Behavior*, *91*(1), 105.

Markman, A. B., & Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive Psychology*, *25*, 431-431.

Markman, A. B., & Gentner, D. (1996). Commonalities and differences in similarity comparisons. *Memory & Cognition*, *24*(2), 235-249.

Markman, A. B., & Wisniewski, E. J. (1997). Similar and different: The differentiation of basic-level categories. *Journal of Experimental Psychology*, *23*(1), 54-70.

Michalski, R. S. (1983). A theory and methodology of inductive learning. *Machine Learning: an artificial intelligence approach*, *1*, 83–134.

Muggleton, S. (1991). Inductive logic programming. *New generation computing*, *8*(4), 295-318.

Muggleton, S., & Raedt, L. D. (1994). Inductive logic programming: Theory and methods. *Journal of logic programming*, *19*(20), 629-679.

Murphy, G. (2002). The Big Book of Concepts. *The MIT Press. Cambridge*

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological review*, *92*(3), 289-316.

Namy, L., & Gentner, D. (2002). Making a Silk Purse Out of Two Sow's Ears: Young Children's Use of Comparison in Category Learning. *Journal of Experimental Psychology: General, 131*(1), 5-15

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.

Nosofsky, R. M., Clark, S. E., & Shin, H. J. (1989). Rules and Exemplars in Categorization, Identification, and Recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognilion 15*(2), 282-304

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*, 53-53.

Oakes, L. M., Kovack-Lesh, K. A., & Horst, J. S. (2009). Two are better than one: Comparison influences infants' visual recognition memory. *Journal of experimental child psychology, 104*(1), 124-131.

Oppenheimer, R. (1956). Analogy in science. *American Psychologist*, *11*(3), 127-135.

Preisach, C., Rendle, S., & Schmidt-Thieme, L. (2008). Relational classification using automatically extracted relations by record linkage. In *Proceedings of the High Level Information Extraction Workshop at the European Conference on Machine Learning*.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, *1*(1), 81-106.

Rattermann, M. J., Gentner, D., & DeLoache, J. (1990). The effects of familiar labels on young children's performance in an analogical mapping task. In *Proceedings of the twelfth annual conference of the cognitive science society* (pp. 22-29).

Raush, S.L. et al. (1995). A PET investigation of implicit and explicit sequence learning. *Hum. Brain Mapp. 3*, 271–286

Reber, A.S. (1989). Implicit learning and tacit knowledge. *J. Exp. Psychol. Gen. 118*, 219–235

Reber, A.S. (1993). Implicit Learning and Tacit Knowledge: An Essay on the Cognitive Unconscious. *Oxford University Press*

Reber, P.J. & Squire, L.R. (1994). Parallel brain systems for learning with and without awareness. *Learn. Mem. 1*, 217–229

Reed, J. & Johnson, P. (1994). Assessing implicit learning with indirect tests: determining what is learned about sequence structure. J. Exp. Psychol. Learn. Mem. Cognit. 20, 585–594

Rendle, S., Preisach, C., & Schmidt-Thieme, L. (2009). Learning to Extract Relations for Relational Classification. In *Proceedings of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining* (pp. 1062-1071). Bangkok, Thailand: Springer-Verlag.

Rips, L. J. (1989). Similarity, typicality, and categorization. *Similarity and analogical reasoning*, 21-59.

Rosch, E. (1978). Principles of categorization. *Concepts: core readings*, 189-206.

Ross, B. H., & Spalding, T. L. (1994). Concepts and categories. *Thinking and problem solving*, 119–148.

Skorstad, J., Gentner, D., & Medin, D. (1988). Abstraction processes during concept learning: A structural view. In *Proceedings of the Tenth Annual Conference of the Cognitive Science Society* (pp. 419-425).

Smith, E. E., Patalano, A. L., & Jonides, J. (1998). Alternative strategies of categorization. *Cognition*, *65*(2-3), 167-196.

Thibaut, J. P., French, R., & Vezneva, M. (2008). Analogy-Making in Children: The Importance of Processing Constraints. In *Proceedings of the Thirtieth Annual Cognitive Science Society Conference*.

Thibaut, J. P., French, R., & Vezneva, M. (2010). Cognitive load and semantic analogies: Searching semantic space. Psychonomic Bulletin & Review, 17(4), 569-574

Waxman, S. R., & Klibanoff, R. S. (2000). The role of comparison in the extension of novel adjectives. *Developmental Psychology, 36*(5), 571-580.

Yan, J., Forbus, K., & Gentner, D. (2003). A theory of rerepresentation in analogical matching. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society*.