Manuscript accepted for publication in
Cognitive Processing, 2017
The final publication is available at Springer
via http://dx.doi.org/10.1007/s10339-017-0829-2

Alberto Greco*
Stefania Moretti*

USE OF EVIDENCE IN A CATEGORIZATION TASK:

analytic and holistic processing modes

* Lab of Psychology and Cognitive Sciences,
Dept. of Social Sciences,
University of Genoa

greco@unige.it
stefania.moretti@edu.unige.it
tel. +39 010 209 9852
fax +39 010 209 9846

Abstract

Category learning performance can be influenced by many contextual factors, but the effects of these factors are not the same for all learners. The present study suggests that these differences can be due to the different ways evidence is used, according to two main basic modalities of processing information, analytically or holistically. In order to test the impact of the information provided, an inductive rule-based task was designed, in which feature salience and comparison informativeness between examples of two categories were manipulated during the learning phases, through the introduction and the progressive reduction of perceptual biases. To gather data on processing modalities we devised the Active Feature Composition task, a production task that does not require classifying new items but reproducing them by combining features. At the end, an explicit rating task was performed, which entailed assessing the accuracy of a set of possible categorization rules. A combined analysis of the data collected with these two different tests enabled profiling participants in regards to the kind of processing modality, the structure of representations and the quality of categorial judgments. Results showed that despite the fact that the information provided was the same for all participants, those who adopted analytic processing better exploited evidence and performed more accurately. Whereas with holistic processing categorization is perfectly possible but inaccurate. Finally the cognitive implications of the proposed procedure, with regard to involved processes and representations, are discussed.

Keywords: categorization; classification; analytic and holistic processes

1. Introduction

In everyday life, one of the most basic mechanisms that drives categorization is inductive learning, which enables telling apart or aggregating cases encountered on the basis of *evidence*, that is the information available in a given moment. Evidence is often manipulated in experimental research, by systematically varying perceptual dimensions of stimuli, or the conditions of stimuli presentation.

Many experimental paradigms offer participants the possibility of comparing examples, assuming that individuals, by effect of these manipulations, are able to gather information about relevant and irrelevant dimensions of categories. Experimental manipulations, however, do not guarantee that available information is always exploited the same way by all learners. One factor of distinction between individuals is whether stimuli are taken in their entirety or single features are analyzed and used for categorization. This distinction is commonly taken as referring to two cognitive styles, known as “analytic” and “holistic”, which are often considered as personal qualities. According to this approach, the main interest is placed on complex factors that can explain the adoption of either style. In the present study we consider them, rather than more or less permanent individual styles, as “modes of processing” that can be exhibited in a categorization task, irrespective of their roots. We are here interested in studying the effects that the adoption of one or the other modality can have on the way evidence is exploited and used, particularly when the available information is fraught with irrelevant aspects to be ignored.

We devised a task in which the criterion for distinguishing between two categories had to be found, and manipulated the evidence by introducing perceptual biases, i.e. by making irrelevant features salient during the first training phases and by progressivley eliminating them. We asked participants to rate different possible rules for distinguishing between categories and, according to the degree of correctness and precision of ratings, we were able to identify two groups, for which we analyzed differences in the processing modality, which turned out to be analytic or holistic, respectively.

In order to analyze how participants managed the elimination of perceptual biases and how accurate their representation of the categories was, we designed a novel method for testing performance, called Active Feature Composition (AFC) task, which asks participants to build examples of a category instead of classifying examples provided by the experimenter. This method is aimed at avoiding the shortcomings of the classification task, a standard test used in categorization researches.

The paper is organized as follows. In Section 1.1 we will first consider how evidence gathered by example comparison may be more or less informative in different cases. We will show that this difference can be due both to task conditions and to different processing modalities, involving explicit feature attending (analytic) or global perception of the entire stimulus (holistic). In Section 1.2. we will then examine the reasons that motivate going beyond classification tasks and introduce the AFC task for testing learning performance. Section 2 describes the experimental framework, Section 3 presents the results, and the following sections are devoted to the general discussion and conclusions.

1.1 Different processes in exploiting evidence

There can be no doubt that categorization is an essential cognitive ability. Distinguishing and grouping the data of experience is needed to reduce environmental complexity and organize reality (Smith and Medin 1981; Murphy 2002). Consider the pictures shown in Fig.1: mushrooms (a) belong to

the *Agaricaceae* family (an edible mushroom), while mushrooms (b) and (c) belong to the family of *Amanitae* (deadly poisonous mushrooms). Now, imagine a non-expert in mycology who is asked to identify which of the three different types of mushrooms (a,b,c) belong to the same family. It is plausible to expect that the mushrooms in Fig.1a and Fig.1b would be grouped together because they look very similar, but this would not be a proper categorization.

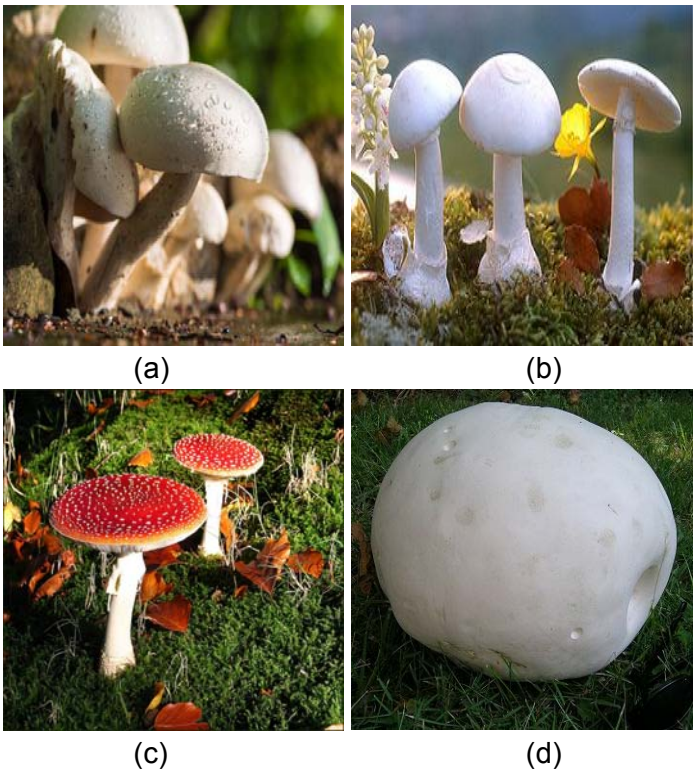

Fig.1 Different types of mushrooms: (a) White edible mushrooms from the *Agaricaceae* family; (b) White poisonous mushrooms from the *Amanitae* family; (c) Red poisonous mushrooms from the *Amanitae* family; (d) White edible mushroom (commonly named *Giant Puffball*) from the *Agaricaceae* family.

Salient features of a stimulus can guide our behavior in the absence of prior knowledge, but this mechanism is not always effective, especially with natural categories, when relevant features are few and relatively non-salient, and many differences between members must be ignored. In these cases some form of supervision is needed in order to achieve a correct categorization. One possibility is to provide implicit information by presenting examples of the same class or of different classes for comparison (Hammer 2015). Category learning by comparison is generally used in everyday life, for example by parents with their children, and has also been investigated in several experimental studies on categorization (Gentner and Markman 1994; Goldstone and Medin 1994; Spalding and Ross 1994; Gentner and Namy 1999; Kurtz and Boukrina 2004; Oakes and Ribar 2005; Boroditsky 2007).

One can ask whether differences in examples proposed for the comparison can predict alone how information will be used. Let us consider Fig.1 again and imagine that non-experts have been informed that mushrooms in Fig.1b and Fig.1c belong to the same class. Now evidence should be more informative, since the comparison would indicate that color is an irrelevant feature. By contrast,

informing them that Fig.1a and Fig.1c mushrooms belong to different categories would not be equally effective, since color still appears to be an important distinctive feature. This example shows that different comparisons can give different information. Considering the reason for this is a relevant question.

Several recent studies have just focused on the factors that can affect the informativeness of the available evidence (Hammer et al. 2008; Hammer et al. 2009; Mathy and Feldman 2009; Andrews et al. 2011; Carvalho and Goldstone 2015; Hammer 2015; Hammer et al. 2015; Palmeri and Mack 2015; Meagher et al. 2017). Some of these studies have shown that the way in which examples are presented can contribute to differences in category learning. For example, Meagher and colleagues (2017) found that the simultaneous presentation of items facilitates the ability to differentiate between perceptually confusable categories; Hammer and colleagues (2008; 2009) showed that learning from same-class comparison can be more informative than learning from different-class comparison.

These studies are not mainly focused on subjective factors. However, some of them (Hammer et al. 2008; Hammer et al. 2009; Carvalho and Goldstone 2015) have found that, even when maximal care is taken in experiments in the way examples are provided, and when training informativeness is systematically manipulated, categorial performance can differ dramatically across learners. These differences clearly show that people do not always exploit the available information. This can be explained assuming that people make use of different processing modalities, which can have effects on the comparison outcome.

One of these processing modalities can be considering stimuli in their entirety. Everyday life experience shows that categorizing on the basis of overall similarities, without the need for a representation of single features can be successful and adaptive. For instance, mushrooms like the one depicted in Fig.1d (named Giant Puffball) could be easily considered edible just on the basis of their global appearance, because there are no dangerous lookalikes. Some studies also examined natural conditions where a holistic processing of evidence, relying on the overall aspect of encountered exemplars, is adopted. The most notable case, for example, is human face processing, which an extensive literature generally considers based on holistic processes (see e.g. Tanaka and Farah 2003; for a review, see Piepers and Robbins 2012).

One different processing modality can be the explicit analysis of single features. Consider once again the example in Fig.1: as we have seen, knowing that the mushrooms in Fig.1b and Fig.1c belong to the same class can be highly informative vis-à-vis the features that are irrelevant. But if the comparison is not made explicitly and the features are not really heeded by the learner, it is unlikely that they will be considered irrelevant and consequently ignored for categorizing properly. Considering again studies on face processing, it has been also shown that in different cultures analytic processing is needed for nonfamiliar or other-race faces (Tanaka et al. 2004; Michel et al. 2006; Michel et al. 2007; Ramon and Van Belle 2016).

These points support a general distinction between two different ways of perceiving stimuli and using them to build a representation. One is characterized by a local analysis of the stimulus, involving attentional focus on features that are represented singularly; the other one involves a global processing of the stimulus and an integral representation of it as a whole, without a specific representation of features.

The distinction between these two modalities has historically been well recognized by many studies in perception and categorization (Lockhead 1972; Garner 1974; Navon 1977; Brooks 1978; Foard and Kemler Nelson 1984; Smith and Kemler Nelson 1984; Ward and Scott 1987; Ward 1988; Smith and Shapiro 1989; Regehr and Brooks 1993; Williams et al. 1994) and it is still investigated in some respects (Maddox and Ashby 2004; Pothos 2005; Davis et al. 2009; Minda and Miles 2010; Byrom and

Murphy 2014; Wills et al. 2015; Mooner et al. 2016; Murphy et al. 2016). In the literature, this contrast has assumed several meanings and has been named differently according to its focus on different but related facets (some regarding perceptual aspects, that is how a stimulus is attended, others concerning higher cognitive aspects, by identifying the two modalities with different categorization strategies). Here we shall refer to this distinction as *analytic* and *holistic*, meaning these terms refer to two main modes of processing a stimulus.

The general aim of our study is to examine the relation between these two basic modalities in the higher cognitive processes of category formation and rule extraction. More specifically, we aim to investigate how information provided can influence and drive category learning when one of the two modalities of processing information, analytically or holistically, is adopted.

The fact that learning can be influenced by contextual factors has been well established; however, it would be of great interest for the literature on categorization to know the aspects that are independent of that context. We suggest that the way a stimulus is heeded, by paying attention to all its features or perceiving it as an undifferentiated whole, alone can determine the extent to which contextual factors impact category learning. That is, the influence of the information provided in a given context depends on how that information is processed: if it is processed accurately it will have more impact.

Hence, in order to test how evidence is actually exploited according to analytic and holistic modalities, we designed a rule-based task in which we specifically changed the information obtainable from example comparison across training blocks, by manipulating feature salience through the introduction and the progressive elimination of perceptual biases. In this way, we could test whether only participants who accurately exploited the information provided could ignore irrelevancies and discover the proper categorization rule.

1.2 Active Feature Composition task as a method for testing learning performance

Before discussing the method implemented in this study, some general questions about the way learning performance is generally tested need to be addressed. Scholars have recently stressed the need to be cautious in generalizing the results of empirical studies on categorization which do not take into consideration some methodological issues about the experimental paradigm used (Ross and Murphy 1999; Ashby and Maddox 2005; Zaki and Kleinschmidt 2014) or the test and analysis procedures implemented (Rips and Collins 1993; Johansen and Palmeri 2002; Tunney and Fernie 2012; Donkin et al. 2014; Wills et al. 2015).

The most important criticisms relevant for our study are ones that affect the classification task, a paradigm that has become almost a standard in the assessment of categorization. In this kind of task, classification choices on test material, i.e. transfer items, are analyzed. These items are selected in advance and built accurately by experimenters, so the choice of this material can depend, more or less deliberately, on the hypothesis that the experimenter wants to test. Therefore, being predetermined, this material could bias the interpretation of participants' categorization processes (Rips and Collins 1993; Johansen and Palmeri 2002; Donkin et al. 2014; for similar interpretative problems of other methods, like the triad or the criterial-attribute procedures, see Wills et al., 2015).

Alternative paradigms have been explored but they don't seem able to address satisfactorily these criticisms. In some studies the classification task has been replaced by a feature inference task, which requires the detection of missing features of a stimulus (Yamauchi and Markman 1998; Markman and Ross 2003; Johansen and Kruschke 2005; Nilsson and Olsson 2005; Hoffman and Rehder 2010). Other recent studies have tried to improve such tests, for instance, by recording eye movements during

the classification task (Richardson and Spivey 2000; Richardson and Kirkham 2004; Scholz et al. 2015). However, these alternative procedures are still affected by the problem that transfer stimuli must be selected by experimenters. Moreover, even with improvements, classification tests only investigate the processes implemented in the transfer of knowledge when evaluating test material. In this respect, Ross and Murphy (1999) have argued that the classification task is not the best way to address the question of how categories are formed. This task alone can account only for classification processes, leaving out other important processes such as data induction or category formation. According to these authors, classification is only one of the functions of categorization and "a full picture of concepts and their uses requires considering other functions as well" (Ross and Murphy 1999, p.496).

We strongly agree with these criticisms and since the aim of the current work is to investigate how evidence is analyzed and represented, and not only how it is used for classifying, we believe that a classification task would be inappropriate for our investigation. For this reason, we devised a novel task that we called the Active Feature Composition (AFC) task, in which learners are asked to choose from a set of features and combine them, in order to create items which are considered to be members of the categories shown during the training phase. In this kind of production task, participants are actively involved in using the information acquired during learning, by making choices in the test phase that are dependent on the way they have processed and represented evidence. This can be considered an unsupervised method for gathering data, which can be clearly analyzed, without transfer items coming into play.

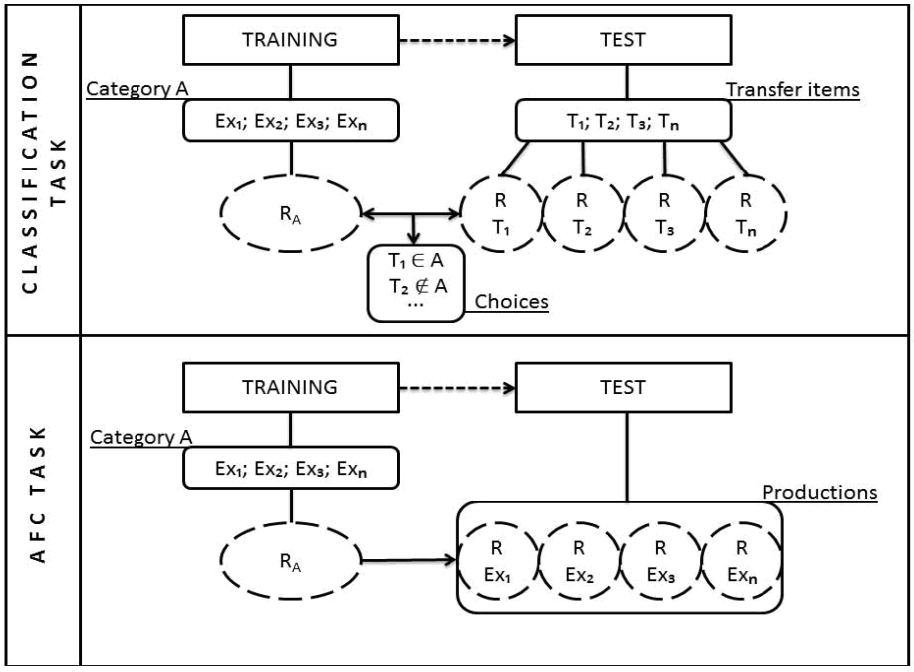

Fig.2 Different cognitive processes involved in classification and AFC tasks

The cognitive mechanisms involved by the two methods are indeed different. Standard classification tasks require a process of comparison between two representations: one (built on the spot) of items to be classified and one of the category acquired during previous learning. Learners' choices, thus, depend on the cognitive process of comparing the representation of a category A (R_A) created from examples (Ex_1; Ex_2; Ex_3; Ex_n) with representations of each transfer item (R_{T1}; R_{T2}; R_{T3} R_{Tn}) in order to

establish their membership [Fig.2, upper]. In the AFC task, instead, learners' choices depend exclusively on the representation of the category acquired during training (R_A) which is used to produce examples (R_{Ex1}; R_{Ex2}; R_{Ex3} R_{Exn}) that are considered members of the category [Fig.2, lower].

In light of these remarks, we assume that choices made in the AFC task make it possible to trace back how evidence is exploited during learning, that is the types of processes implemented in the acquisition and representation of a category. Details of the task design and the test method are discussed in the Method section.

2. Method

2.1. Design

An inductive rule-based categorization task was devised, in which participants were provided with pairs of exemplars from two different categories, during three differently informative training blocks, each followed by an AFC task. To assess the quality of the categorization achieved, at the end of the experiment a final rating test was administered, in which participants were asked to rate the accuracy of a set of rules which define correctly, inaccurately or wrongly the categories acquired.

The stimuli to be categorized were stylized images of mushrooms, with a stem, a cap, and a number of dots on their caps. The number of dots could vary from 1 to 10 and the cap size could vary in proportion with the number of maximum possible dots on it. The stem size could vary on only two dimensions, either short or long. Each mushroom was labeled with the fictitious name of the respective membership class: DAX or MED. The relevant dimension for distinguishing between the two classes was the number of dots, according to a simple, verbalizable and one-dimensional rule: "If the number of dots on the cap is between 1 and 5, then the mushroom is DAX; if it varies from 6 to 10, it is MED". The choice of dots as the criterial feature allows us to gather precise data on how they are processed. Indeed dots can be analyzed in two ways: they can be counted one by one, analytically, or perceived holistically as a whole. Thus, the same rule can be evaluated as: "DAXs have fewer dots than MEDs", or "MEDs have more dots than DAXs", which are also valid criteria, but more general and inaccurate for the task.

During each training block, pairs of exemplars were presented sequentially, one belonging to the DAX category and the other to MED category. Thus, the task allowed both types of comparison: between the two categories (at the same time in each trial) and within the same category members (sequentially across trials).

Three training blocks were devised in order to manipulate feature salience and comparison informativeness sequentially. In the first block, within-category similarities and between-categories differences were maximized by introducing a high perceptual contrast between the examples in the pair. In the second block, within-category similarities and between-category differences were minimized by reducing the contrast. Finally, in the last block the criterial feature was made salient and all previous saliences were made irrelevant. Thus, a careful exploitation of evidence would lead participants to the progressive elimination of the perceptual biases introduced in the first block.

Fig.3 Examples of pairs presented during the training: in the first block (Training 1), the difference between DAXs and MEDs dots was higher than five, the cap was correlated with the number of dots, and the stem was long for MEDs and short for DAXs; in the second block (Training 2), the difference between dots was lower than five, the cap correlated with dots, while the stem was long and short for both categories; in the last block (Training 3), dots, arranged in a symmetrical fashion, included the entire range of the class, the size of the cap was fixed and the stem was long or short for both.

Biased pairs were obtained by calculating the absolute difference between the number of dots of the two categories and by manipulating the correlation with the cap and the stem [Fig.3].

In the first block, DAXs with 1 to 4 dots and a short stem, and MEDs with 7 to 10 dots with a long stem were shown, both with the size of the cap correlated with the number of dots. For each pair, the difference between the number of dots in DAXs and MEDs was higher than five. Thus, the global size of mushrooms was made salient in such a way that DAXs looked smaller than MEDs and so that participants were led to take it as a relevant difference.

In the second block, the covariance cap-dots was maintained, while the stem size was made non-relevant (long and short for both classes). DAXs with 2 to 5 dots and MEDs with 6 to 9 dots were shown. For each pair, the difference between dot numbers in DAXs and MEDs was lower than five. In this block, mushrooms in the pair appear less different and participants had to revise any hypothesis about the size contrast.

In the last block, the cap of both classes had the same fixed size, the stem was still irrelevant and the number of dots included the entire range of the class (1 to 5 for DAX and 6 to 10 for MED), so that the perceptual biases introduced at the beginning were totally eliminated. In addition, dots were arranged in a symmetrical fashion in order to make them salient and easier to count.

In order to test how evidence affected the representations of the two categories, and to assess if and how representations changed as the available evidence changed, the Active Feature Composition task (AFC) was provided. In the production screen [Fig. 4], shown after each training, each feature (caps with 10 different sizes, dot number from 1 to 10, and two stem types) was represented inside a clickable box. When all features had been selected, the combined mushroom appeared in a box and then participants had to select the category membership by clicking on the DAX or MED labels.

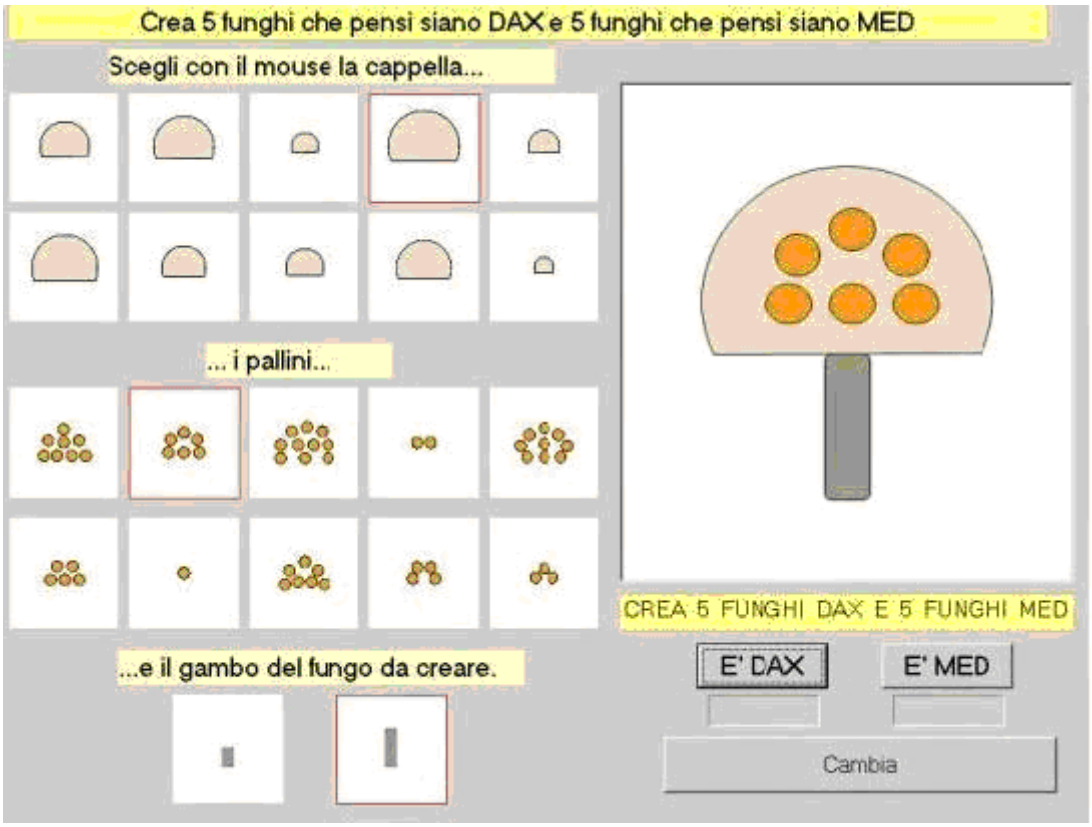

Fig. 4 Production screen for the AFC task. Captions (in Italian) say: "Create 5 mushrooms that you think are DAX and 5 mushrooms that you think are MED", "Select with the mouse the cap…", "...dots…", "...**and** the stem of the mushroom to be created"

The final rating task was designed to gather information about possible rules used for distinguishing between categories. Rules to be rated differed according to their level of strictness, completeness, and the feature they referred to:

1) MEDs have more dots than DAXs (inaccurate, complete, on dots)
2) MEDs have a longer stem than DAXs (wrong, complete, on stem)
3) DAXs have from 1 to 5 dots (correct, incomplete, on dots)
4) DAXs have from 1 to 5 dots and MEDs from 6 to 10; (correct, complete, on dots)
5) MEDs' caps are larger than DAXs' (wrong, complete, on caps)
6) DAXs are smaller than MEDs (wrong, complete, on global dimension)
7) DAXs have few dots (inaccurate, incomplete, on dots)
8) DAXs have a more orderly arrangement of dots (wrong, incomplete, on spatial arrangement of dots)
9) DAXs have short stems (wrong, incomplete, on stem)

The first and the seventh rule concern the estimation of the number of dots, the third and fourth involve dot counting, the eighth pertains to the perceptual bias on the criterial feature, and the remaining rules concern perceptual biases on irrelevant features.

Having devised a new task for testing category performance, a new method of analysis was required. The general idea was to relate the type of appraisal of rules shown at the final test with the kind of use of evidence that can be traced back according to the choices made in the AFC task.

Therefore, explicit final ratings given by participants on the nine final rules were analyzed first. For this purpose, we initially used a Principal Components Analysis to reduce the data. Afterwards, scores given by participants to the nine rules were submitted to a PCA with a Promax rotation. Then, a K-means cluster analysis was computed (SPSS Quick Cluster procedure) in order to separate participants in different groups on the basis of their scores in each component. We expected that differences relating to rule appraisal would depend on the kind of analysis and representation of evidence. So, in order to understand how participants acquired and represented the two categories across stages, we analyzed productions in the AFC considering the following 5 dependent variables:

(1) *Global size bias*: by comparing produced items (the combination of all features: caps, dots, stems) with the training items, we computed how many items had been created exactly the same as the items seen in corresponding and previous training blocks. This enabled us to detect the persistence of the global size bias introduced in the first training.

(2) *Stem bias*: by considering the type of stem chosen for each item of the two categories, we aimed to identify the presence or absence of this bias across training phases.

(3) *Class completeness*: by analyzing the compliance with the criterial feature (number of dots) in each production we measured the completeness of the produced class compared to the rule-defined class. Series were considered complete when they included all items with the number of dots required by the correct rule (e.g. 1,2,3,4,5 for DAXs).

(4) *Errors*: by counting the number of items built with a number of dots different from the criterial one, we computed the number of mistakes across phases. Errors show that the representation of the class was not well defined or was based on irrelevant features.

(5) *Viewing times*: by recording observation times of pairs of items during training blocks we computed the depth of the visual analysis. Even if longer observation times could indicate a greater analysis, this may not always be the case since other factors could affect observation time lengths. But we can safely assume that shorter viewing times necessarily indicate that an accurate analysis was not made.

2.2. Experiment

Participants

A total of 31 undergraduate students (24 female, mean age 22.1, sd 3.5) participated in the experiment for course credit. They had normal color vision and normal or corrected-to-normal visual acuity. Informed consent was obtained at the beginning of the experiment.

Procedure

All participants were tested individually. They were seated in a quiet room at a comfortable viewing distance from the monitor. The instructions were displayed on the screen. Full sessions were conducted using a dedicated computer program; all instructions and stimuli were presented on a VGA flat screen color computer monitor. Responses were recorded by the same program; only a mouse (no keyboard) was provided to participants. The following short cover story was shown in order to allow participants to fully understand instructions, materials, and the task: “A species of mushroom has been classified by a

group of scientists as poisonous and belonging to an island named DAX, while another species has been classified as non-poisonous and belonging to an island named MED. Now you will see some mushroom pairs, one of which has been recognized as poisonous from the DAX island, while the other is from the MED island. In this phase, you should only observe the mushrooms and try to understand what differentiates them."

Experimental sessions were divided into three stages, each composed of a training block and a test phase. During the first two training blocks, 40 pairs of stimuli at a time were presented to each participant: 20 high-contrast pairs in the first block and 20 low-contrast pairs in the second. In the last block, 100 pairs of stimuli were prepared, and participants had the opportunity to view pairs until they thought they had discovered the rule and then they could end the training session. Stimuli viewing was self-paced without any time limit at each training stage. Viewing times were recorded, from the appearance of the pairs on the screen to the click on the advancement button. The order of pairs and their arrangement on the screen (right, left) was randomized differently for each participant. At the beginning of the last two training blocks it was specified that the criterion for distinguishing the two types of mushrooms did not change.

After each training block, the production screen [Fig.4, see 2.1 above for details] was shown for the AFC task and participants were asked to produce five DAXs and five MEDs by selecting a component feature at a time. The order for creating mushrooms was not imposed, no feedback on choices and no time limit was given; a message saying "can't create this mushroom" was only shown when the participant attempted to create a mushroom with a number of dots greater than the size of the cap selected.

At the end of the last test, a screen was shown presenting nine rules as possible criteria used by scientists for distinguishing the two classes of mushrooms. Participants were asked to rate each one from 0 (wrong) to 3 (correct).

At the end of the experiment participants were debriefed about their possible difficulties in the task and were asked to restate the criterion they had used to distinguish between the two types of mushrooms.

3. Results

The correct and complete rule was identified by 42% of participants (by giving the highest value of accuracy to rule 4), revealing that they correctly counted both DAX and MED dots; 19% identified the correct rule for only one category (by giving the highest value to rule 3); 23% gave the highest value to the complete but inaccurate rule 1; 10% to the incomplete and inaccurate rule 7; the remaining 6% gave a low value to the rules involving dot numbers. However, only 19% of all participants completely eliminated all the irrelevancies (by giving the value of "0" to rules 2,5,6,8,9).

On the basis of different ratings given to the nine final rules, four components were extracted using the PCA, together accounting for 82.98 % of variance [Fig.5, see also Table 1 in Appendix]: the first component had high loadings for wrong rules concerning irrelevancies, thus expressing a *perceptual bias*; the second component had high loadings for rules about the exact number of dots, thus expressing *counting*; the third component loaded for rules on the estimated number of dots, thus expressing *estimate*; the fourth component for rules about a bias on the criterial feature, namely the arrangement of dots, thus expressing *arrangement bias*.

A K-means cluster analysis was then computed (SPSS Quick Cluster procedure) from participants' scores in each component. Two groups resulted [Fig.6, Table 2 in Appendix]: Group A consisting of 16 participants and Group B of 15 participants.

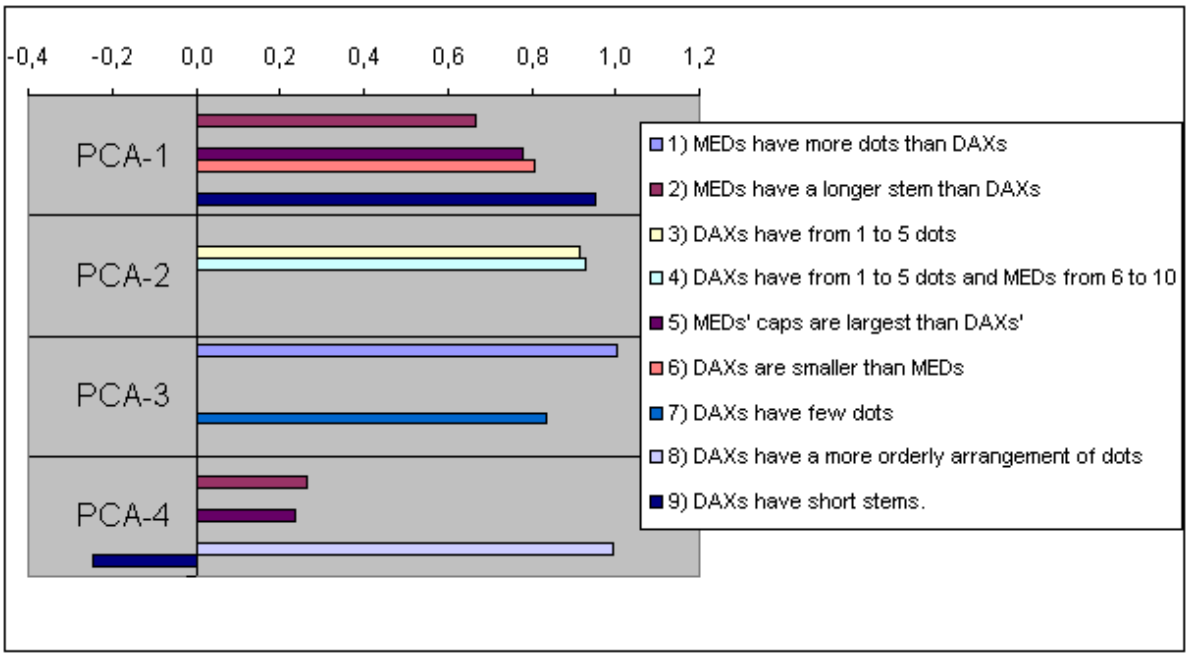

Fig.5 Loadings of components extracted by a Principal Component Analysis from final rules rates. PCA1 = perceptual bias; PCA2 = counting; PCA3 = estimate; PCA4 = arrangement bias

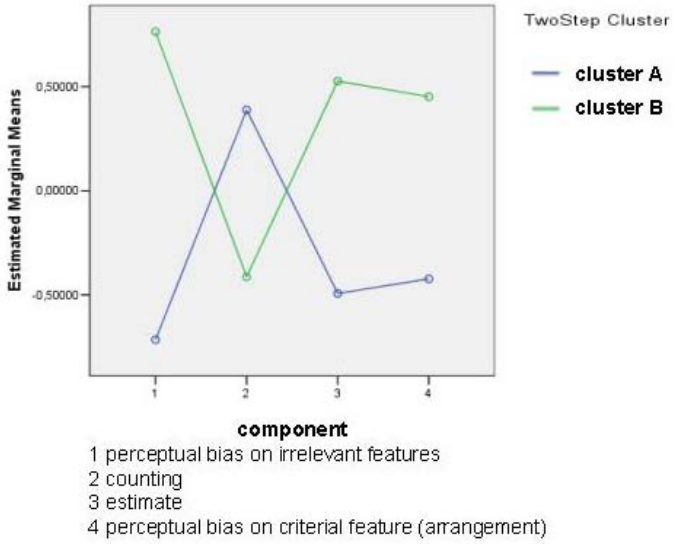

Fig.6 Factor scores of two groups for each component

The major factor score for Group A was related to the *counting* component, having high loadings for rules 3 (partially correct) and 4 (correct), while for Group B the highest factor score stemmed from rules expressing a *perceptual bias on irrelevant features*. Our main focus at this point was to test whether performance in the rule rating task was related to the choices made in the AFC production task.

Thus, we proceeded by looking for differences in production phases among the five dependent variables (Global size bias**,** Stem bias, Class completeness, Errors, Viewing times). We used a

repeated-measures ANOVA for Global size bias, Stem bias, and Viewing times, with 3 independent variables: *Types of mushrooms* (DAX-MED) and *Test phase* (1-2-3) as within-subject variables, and *Group* (A-B, resulting from the cluster analysis) as between-subjects variable. For each variable, we analyzed the main effects and the differences between groups. The Huynh-Feldt correction was applied in cases where the Mauchly's sphericity test was significant; effect sizes are reported using partial eta squared values (η_p^2). Chi-square tests were used for the other two variables (*Class completeness* and *Errors*).

Global size bias

A decrease in number of equal items across stages was found for all participants: the effect of *Test* phase was significant, $F(2,58) = 19.67$, MSE = .9, $p<.0001$, $\eta_p^2 = .40$) . The mean number of items, built exactly the same as the *corresponding training* items, decreased in Test2 (1.24) compared to Test1 (2.33) (post-hoc pairwise comparison Bonferroni corrected, $p<.0001$) and Test3 (.98) compared to Test1 ($p<.0001$). Thus, all participants reduced the production of items identical to ones seen in the corresponding training block, in the last two stages compared to the first stage. These data show that evidence can be reproduced easily when there is little information, whereas with increasing information features evidence is difficult to accurately remember and combine.

We also analyzed the number of items built exactly the same as ones seen in *previous training blocks* (Test2 compared with Training1, Test3 with Training2 and Training1). In line with expectations, a significant difference between Groups was found [Fig.7]: Group B in Test2 made more items equal to ones met in Training1 (1.25) than Group A (.67) [independent samples $t(29) = 1.86$, $p=.038$, one-tailed]; in Test3 also made more items equal to Training1 (1.06) than Group A (.46), [$t(29) = 2.19$, $p=.019$, one-tailed]. Thus, participants in Group B showed a difficulty in eliminating the global size bias because in the subsequent stages, in which biases were removed from shown examples, they continued to build significantly more items equal to ones met at the first stage, when the global size bias had first been introduced.

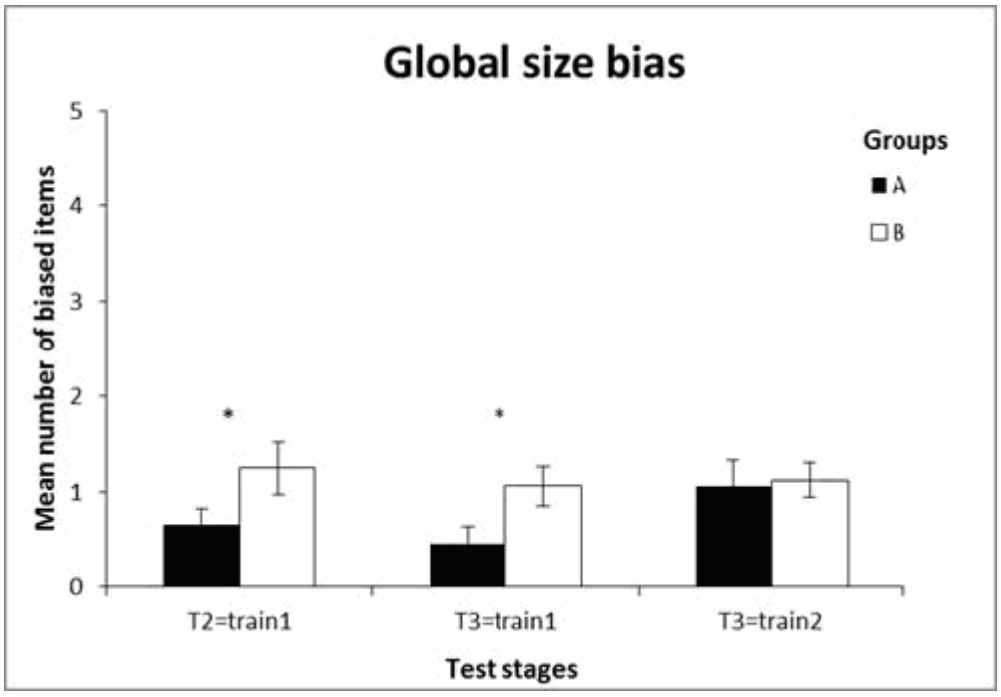

Fig.7 Average number of items created at Tests 2 and 3 exactly like ones seen in previous training blocks (error bars indicate standard error of the mean; * = p < .05)

Stem bias

A main effect of *Type*, F (1, 29) = 160.89, MSE = 1.53, p<.0001, η_p^2 = .85. and an interaction of *Test xType* resulted: F (1.51,44) = 66.61 (Huynh-Feldt corrected), MSE = 1.31, p<.0001, η_p^2 = .70. Short stems were chosen less for DAXs in Test2 (2.79) compared to Test1 (4.75) (post-hoc pairwise comparison Bonferroni corrected, p<.0001), and in Test3 (3.05) compared to Test1 (p<.0001). Short stems were, on the contrary, chosen more for MEDs in Test2 (1.69) compared to Test1 (.06) (p<.0001) and in Test3 (1.91) compared to Test1 (p<.0001). For long stems the situation was complementarily reversed. These results show that participants changed their production on the basis of the available evidence and, since the stem bias was eliminated in the second training block, all participants reduced the stem bias for both categories in the last two stages compared to the first stage.

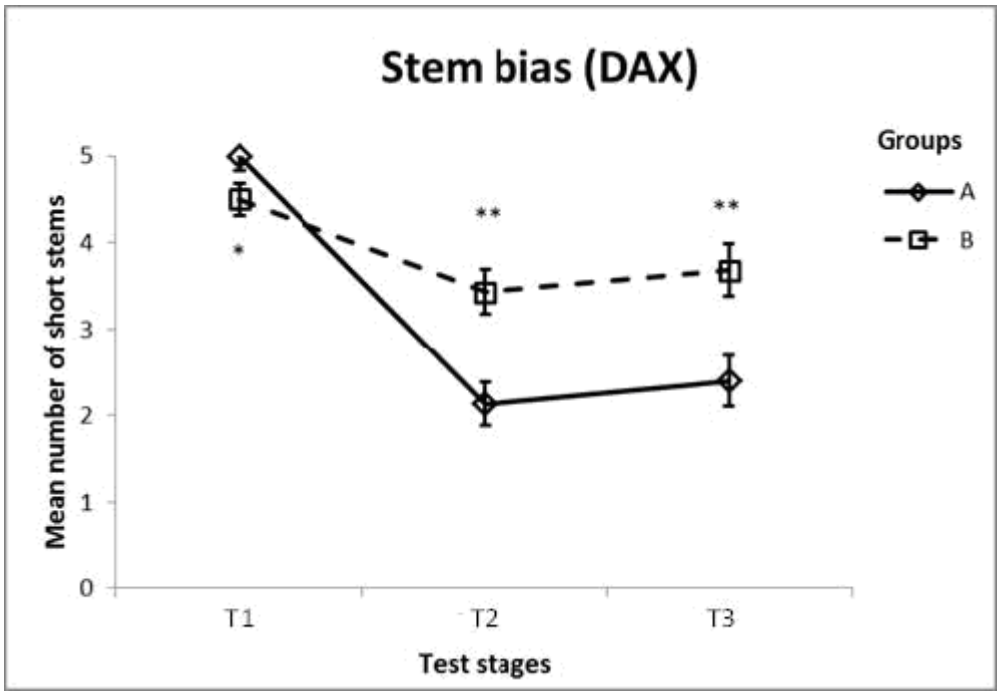

Fig. 8 Average number of DAX items created with short stems at different test stages (error bars indicate SEM; * = p ≤ .05, ** = p < .01)

However, this improvement in performance did not occur in both groups equally. In fact, a significant interaction *Test X Type X Group* resulted [F(1.51,44) = 8.75 (Huynh-Feldt corrected), MSE = .99, p=.002, η_p^2=.23]. A post-hoc pairwise comparison (Bonferroni corrected) revealed significant differences in choice of stems between groups for DAXs [Fig. 8]: in Test1, Group B made less DAXs with short stems (4.50) than Group A (5.00) (p=.05). In Test2, Group B made more DAXs with short stems (3.44) than Group A (2.13) (p=.001) and likewise in Test3 more DAXs with short stems (3.39) than Group A (2.40) (p=.005).

These results revealed the presence of the stem bias for Group B in the representation of DAX category. Furthermore, at the first stage, even if DAXs were shown exclusively with short stems, some participants of Group B revealed less precision in focusing attention on both exemplars, by making some DAXs with long stems.

In fact, even if the two groups reduced the effect of biases introduced in the first training block, participants in Group B were more attached to their first perception of size differences as a salient difference between the two classes of exemplars; for this reason they continued to choose more short stems for DAXs, even after the first stage.

Class completeness

For this variable we had dichotomous data, considering, for each participant, in each test stage, and for each type, whether a complete series (i.e. including the full range of number of dots) was produced or not. The number of complete series (for both types) progressively decreased in Test2 (26%) and Test3 (19%) compared to Test1 (55%), The proportion of incomplete over complete series in Test2 and Test3 was significantly above chance by a binomial test (Test2, p = .011; Test3, p = .001). This result may be explained by considering that the majority of participants could not avoid focusing more on changes in the irrelevant features than on changes in the relevant ones.

For each combination of *Test and Type*, the proportion of complete and incomplete series was analyzed by *Group*, and submitted to a Chi-Square test. A significant difference [$\chi^2(1) = 4.05$, $p = .044$] in the proportion of incomplete over complete series was observed in Test1 [Fig.9], in which Group B produced less MED complete series (18%) than Group A (53%), revealing less precision in the early representation of this class. This was presumably due to the difficulty of counting a greater number of dots without an explicit analysis.

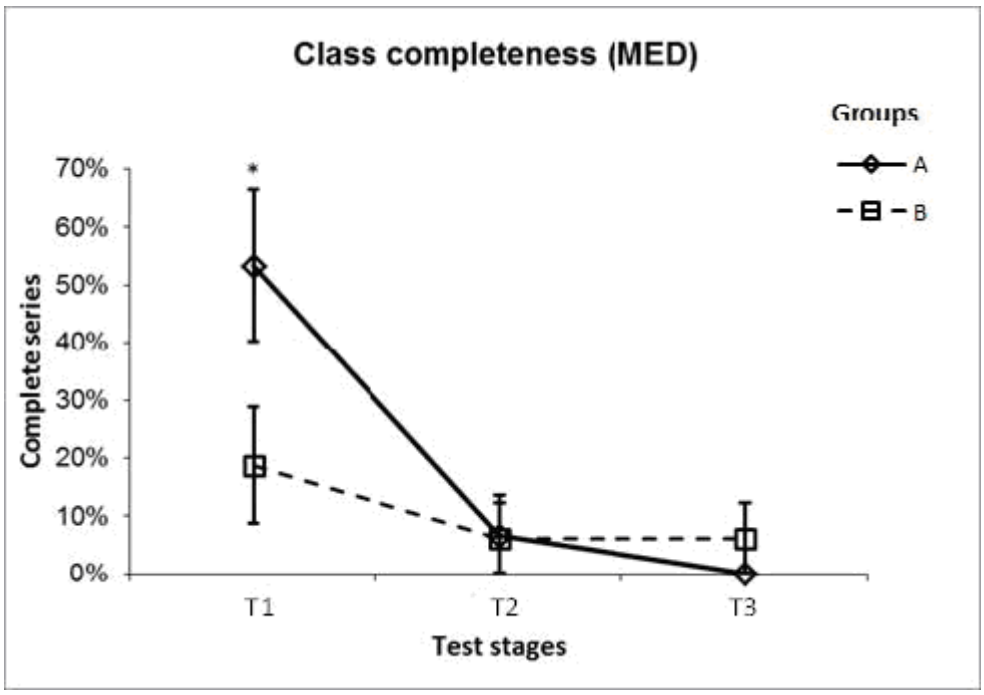

Fig.9 Percent of MED complete series produced at different test stages
(error bars indicate SEM; * = $p < .05$)

Errors

The number of errors did not decrease across stages, but it remained steadily low for all participants. For each combination of *Test and Type*, the number of errors was analyzed by *Group*, and submitted to a Chi-Square test. A significant difference [$\chi^2(2) = 6.06$, $p = .048$] was observed in T3, in which for the MED category Group B made 11 errors while Group A made only one error [Fig. 10].

We found that this result could be a consequence of the inaccurate representation of the MED class for Group B due to a holistic processing of the stimuli.

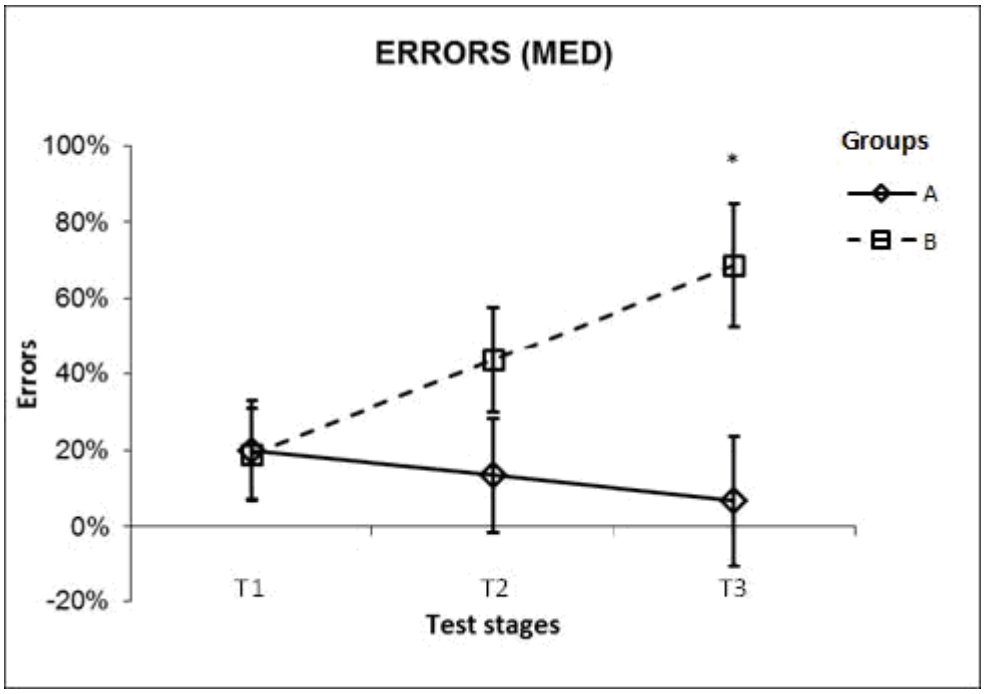

Fig. 10 Percent of errors for MED category at different test stages
(error bars indicate SEM; * = p < .05)

Viewing times

In addition to the analyses mentioned above, we computed the observation times of items during training stages. We first eliminated outlier higher times (> 19,422 msec) identified using the adjusted boxplot procedure (Hubert and Vandervieren 2006) and software R. We then computed, for each subject, the mean viewing times only for pairs whose values were lower than the average display time for each training. Participants in Group B had shorter viewing times than Group A in the first training block (3297 vs 4119 msec) and this difference was progressively reduced at subsequent stages, becoming null at the last stage. Considering only the first two stages, a 2 x 2 repeated measures ANOVA yielded a significant interaction of *Training* stage X *Group*, F(1,29) = 6.60, MSE=127,671, p=.016, η_p^2 = .19. This means that participants in Group B approached the task allocating less time for analysis.

Overall result discussion

From the analysis of the final ratings, it turned out that 16 of all participants correctly counted the dots and eliminated most irrelevancies, while the other 15 approximately estimated the number of dots and considered relevant most saliencies on perceptual features. We refer to these two kinds of participants as Group A and Group B, respectively. Then, from the analysis of productions in the AFC task, we sought to identify whether these differences in the explicit evaluations were due to different modalities of exploiting evidence. The result is that evidence influenced all participants' performance but with significant differences.

That is, all participants similarly changed their productions on the basis of what changed during the training. As biases were reduced and eliminated from training items, by presenting different feature combinations stage by stage, participants did likewise with their produced items, choosing and

combining features differently among stages, but in a more or less accurate way.

Participants in Group A, by relying on the explicit comparison between exemplars, focusing on features and analyzing differences, were more able to identify relevant and irrelevant features, and consequently to better abstract the target rule. In fact, they successfully avoided the biases, produced more complete series, made fewer errors and relied on explicit counting of dots which helped them to recognize the correct rule.

Participants in Group B, on the contrary, were able to produce what they had seen at each training block but in a more approximate way, and they could not avoid reproducing particular perceptual salient feature combinations, especially those seen during the first training block when the difference between the two categories concerned the global size. Thus they were highly sensitive to this global bias and, without analyzing differences among stages and focusing on single features, they could only identify an approximate rule. In this way, they were able to distinguish between categories and make relatively few errors (albeit a greater number than Group A) on the basis of a correct but approximate criterion. This was the reason why, in the final test, Group B assigned high scores to the rules including irrelevant aspects, while Group A gave them low scores. Hence, differences found between the two groups suggest that salient changes on the irrelevant components were perceived and processed in a more or less accurate way according to the processing modality adopted.

But this does not mean that irrelevant features were always completely ignored or eliminated when distinguishing between classes. We designed the last training block with the aim of eliminating any participants‘ assumptions arising from a shallow distinction based on the salient features. However, this elimination only partially took place, because most of the participants who correctly understood that the criterial feature was “number of dots”, at the same time were not able to ignore salient perceptual distinctions made and learned earlier. From the analysis of the explicit rating test, we found, in fact, that very few participants completely eliminated irrelevancies and understood that salient irrelevant features were absolutely non discriminative between the two classes. Furthermore, even among those who when debriefed verbalized the correct rule, some attributed a score above zero to the rules that expressed perceptual biases concerning irrelevant features.

This result shows that although the accuracy of categorization inductively achieved from exemplars depends on how evidence is used, salient irrelevant variability cannot be completely ignored, neither holistically nor analytically. This confirms what Hahn et al. (2010) found within a different paradigm, that the processing of irrelevant features is to some extent unavoidable, especially when they are correlated with the criterial feature. In fact, in our experiment the global size bias was created by correlating cap and stem size with the number of dots. A confirmation of this influence comes from the literature on visual perception and attention, which defines the inability to ignore or filter changes on irrelevant dimensions, when they are perceived holistically and are not separable, as the "Garner interference" (Garner 1974, 1976; Wagemans et al. 2012; Wang et al. 2013).

However, in our experiment participants were trained with unbiased pairs too, and moreover in the production phases they were requested to choose a single feature at a time. Thus, it is possible that the importance given to the irrelevant features was due to an effect discovered by Lewicki (Lewicki et al. 1989), known as the phenomenon of “*Self-perpetuating* of encoding biases". It occurs when participants not only learn without being aware of the rules, such as those concerning some covariation, but these rules continue to influence them even when they are not present anymore. Thus, people continue to encode new information on the basis of a distortion from previous inferences, which is what we observed significantly for Group B.

Evidence manipulation still had its role in changing representations: in both groups the number of complete series decreased and the number of errors increased in the last two phases compared to the

first. The fact that fewer errors and more complete series were produced when size contrast was high can be due to the high perceived dissimilarity between the two categories compared to the last phases in which differences were minimized and a more complex criterion compared to the "DAX are smaller than MED" had to be found. This confirms that, like in this specific case, categorization based on the overall aspect of a stimulus can be effective, and it supports the idea that people usually implement heuristics that look for the simplest representation possible and with the minimum effort (Nosofsky et al. 1994; Hammer et al. 2009). However, we observed that participants who did not change their initial hypothesis in accordance with the information provided during learning revealed simplified representations of the two categories.

Another interesting finding concerns a worse processing of the MED category in our experiment. It may be due to the fact that the number of dots for DAX was in the range of subitizing, so they were easy to instantly and precisely count, while MED dots needed to be explicitly and serially enumerated. Furthermore, DAX mushrooms were presented as poisonous while MEDs were not. As hypothesized by Hahn et al. (2010), when presenting two complementary categories, features can inadvertently acquire a positive association with one of the two categories, which would increase their relevance compared to other features. However, these effects should have concerned the entirety of participants, whereas differences clearly emerged in the two groups.

4. General discussion

The present study developed from findings showing that inductive learning is profoundly influenced by the characteristics of available information, both in real-life contexts and in experimental settings. In the latter, the way stimuli and learning conditions are manipulated can have a deep impact on what and how participants learn (Hammer et al. 2008; Mathy and Feldman 2009; Hammer et al. 2009; Andrews et al. 2011; Carvalho and Goldstone 2015; Hammer 2015; Hammer et al. 2015; Palmeri and Mack 2015; Meagher et al. 2017). Our study supported the idea that the extent of this impact depends on different modalities of processing information, analytic or holistic.

We manipulated feature salience by introducing and progressively reducing perceptual biases, i.e. making irrelevant features more salient at the start and gradually eliminating such saliences. In order to overcome some difficulties in the classification tasks based on categorial decisions about examples, in our study we created a novel test (the Active Feature Composition task), that does not require classifying items but producing them by combining features. Also, in a final explicit rating task, we asked participants to assess the accuracy of a set of possible categorization rules. Data coming from these two different tests were used to profile participants in relation to the kind of processing mode, the structure of representations, and the quality of categorial judgments.

What we found is that despite the fact that the information provided was the same for all participants, not all participants exploited it in the same way. About half of them correctly processed the changes in available evidence across phases, by revising their biased representations, and accurately updating them with the new information. These participants revealed a final analytic representation of the categories and a correct rule induction. The other half, instead, processed the information from new evidence poorly, and remained attached to the perceptual biases of the first phases. They had a holistic representation of the categories and induced the approximate rule. Hence, all participants based their representations on the evidence resulting from experimental manipulations made across training

blocks, but some of them exploited information more efficiently than others.

Individual differences in category learning have been already accounted for by research focused on the influence of objective factors on categorial performance. In particular, recent studies on visual category learning (Hammer 2015; Hammer et al. 2015) have explained performance differences in light of the distinction between attentional learning and perceptual learning. The authors discuss that, although the interaction between these two processes is needed in order to achieve a good categorial performance, people do not always rely on both, and individual differences are explained as a consequence of the fact that this interaction depends on context, brain maturation, and subjective factors (Gazzaley and Nobre 2012; Weissman and Prado 2012; Hammer et al. 2015). We found this explanation very interesting and it could be somewhat relatable to our results, since feature salience and relevance were main variables in our design. However, in our work we decided to investigate two more basic processes that concern the way a stimulus is processed. These processing modalities are prior to the identification of relevant features involved in attentional learning or to the recognition of less salient differences involved in perceptual learning. That is, analyzing all the features of a stimulus, or perceiving it as a whole, determines the way information on relevance is gained and exploited. Indeed, categorization does not only rely on the type of information obtainable from different comparisons, but also on what one chooses to compare (features or the whole).

At this point, it should be clarified that we do not consider analytic and holistic ways of processing as personal cognitive styles. Cognitive styles are the ways in which an individual usually organizes and processes information, they are processes that develop over time, and are relatively permanent (Goldstein and Blackman 1978; Ford et al. 1994; Sternberg and Grigorenko 1997; Riding and Rayner 1998; Shi 2011). As we have explained earlier, in our work we are not interested in accounting for factors that could determine the adoption of one or the other modality. Whether it depends on a personal tendency or some momentary individual disposition is beyond the scope of the present study. The analytic-holistic contrast is simply aimed at distinguishing between subjects who *in this task* actually analyzed more single features and those who mostly based their process on the whole stimuli. Our primary interest, indeed, is to understand how evidence provided is actually exploited and to assess the changes in the processing, representation, and use of information, when one modality or the other is adopted.

One novelty of our contribution is the Active Feature Composition task, that enables gathering clear data without the methodological limits affecting classification test procedures. The AFC task consists in building examples belonging to the categories learned during training blocks, by combining single features. Given that in this task participants were simply required to start from the categorial representation acquired, and use it in order to build examples, this enabled us to gather valuable information about the processing and retrieval of stimuli, in terms of: which features are selected and which are ignored, how many times they are chosen, which other features they are combined with, at the same stage and across stages, for each subject. Thus, by comparing exemplars produced with the observed ones, it is possible to detect how evidence influences representations and how it is processed. Also, comparing examples built among different phases makes it possible to keep track of the changes in learning over time.

The AFC task has also led us to find interesting outcomes that could have some methodological implications. That is, in traditional classification tasks, categorial processes and representations are inferred by comparing the items selected by the learner in the test with those shown during training sessions, implicitly assuming that training items and their representation would match. Nevertheless, in our AFC task we were able to observe that participants' productions were often different from the examples shown in the corresponding training stage. Productions changed during learning, taking into

account not only the present evidence but also previous evidence, and previous representations. As a consequence, examples produced did not exactly match exemplars encountered. This is one reason supporting our claim that inferring categorization relying on the classification of test materials, as is usually done, could be misleading. Furthermore, with our method we were able to show that, in many cases, even participants who correctly identified the categorization rule might reveal the persistence of a bias on irrelevant features. This information is usually lost with standard methods.

The rating task administered in our experiment at the end of the last stage had the purpose of investigating the accuracy of the achieved categorization. In this task we asked participants to explicitly evaluate the correctness of different rules. In this way it was possible to find if and to what extent the criteria used by participants to distinguish between the two categories were correct. The set of rules, indeed, was designed to include possible factors that could give important information about the quality of rule induction, like the type of processing of the criterial feature, the degree of influence of irrelevant features, and if the focus was on both categories or just one. Our main goal was to test whether biases could be eliminated from representations, criterial features discovered, and the rule accurately identified, depending on how the evidence available at each of the different phases of learning was exploited. That is, whether the effect of evidence on the ability of eliminating irrelevancies and identifying relevant features was stronger or weaker depending on participants' mode of processing: analytic or holistic.

Hence, in our study we have investigated the relationship between the most basic process of perceiving a stimulus with higher cognitive processes like representation and rule induction, within a setup that allows keeping track of the changes in learning over time and according to the evidence available in a given moment. To the best of our knowledge, this complex relationship has not yet been widely examined, except for some recent research on the link between learning, time and generalization (Perry et al. 2015; Perry and Saffran 2016; Vlach 2016).

Future research should continue to examine several different mechanisms involved in categorization, considering the influence of both objective and subjective factors. For example, we limited the investigation to five variables because we considered them as good detectors of the kind of processing made based on evidence, but we do not exclude that other factors could be identified by further research. Indeed, on a large amount of data, like the one gathered with our method, many different analyses are possible.

In general terms, for example, using the AFC task, it would be possible to analyze productions to infer the presence of prototypes, or follow the feature selection process to detect which one is chosen first and regularly, in order to assess its relevance in the category formation. For the training-test item comparison, several other analyses are possible, too. For example, we determined 'Global size' by counting the number of items produced that exactly matched training items, leaving out a more complex analysis of degrees of similarity. Likewise, for 'Class Completeness' we did not measure how much a series was complete but only if it was or not.

Similarly, other rules could be presented for evaluation in addition to the nine we have identified in order to investigate several other possible distinguishing criteria used by participants. Other significant effects could be found by using the AFC task in combination with different tools, like eye tracking, to detect attentional focus in the exploration of the stimuli pair, or with different learning paradigms used in the literature.

Interesting future studies could integrate the production task with a classification task, and compare the chosen transfer items with the produced items, instead of the training items, in an attempt to provide a more complete and accurate picture of how information is acquired and used in classifying new material. This integration could give interesting results if some classical paradigms are tested. Finally, it

would certainly be of interest, especially in the field of developmental psychology or in educational research, to see how the AFC task correlates with questionnaires of analytic and global cognitive styles.

We have shown that the two modes of processing we have studied may have a relevant role in the way evidence is exploited, but it goes without saying that there are many other subjective factors interacting in a significant way with objective stimuli features, like prior knowledge, goals, expectations, and so on. We cannot certainly predict, based on the observation of analytic or holistic processes alone, whether a person eating mushrooms personally collected in the woods will safely return to pick them again.

References

Andrews JK, Livingston KR, Kurtz KJ (2011) Category learning in the context of co-presented items. Cognitive Process 12(2):161–175. doi:10.1007/s10339-010-0377-5

Ashby FG, Maddox WT (2005) Human category learning. Annu Rev Physiol 56:149-178. doi:10.1146/annurev.psych.56.091103.070217

Boroditsky L (2007) Comparison and the development of knowledge. Cognition, 102(1), 118–128. doi:10.1016/j.cognition.2002.08.001

Brooks LR (1978) Nonanalytic concept formation and memory for instances. In: Rosch E, Lloyd BB (ed), Cognition and Categorization. Lawrence Erlbaum Associates, pp 3-170.

Byrom NC, Murphy RA (2014) Sampling capacity underlies individual differences in human associative learning. J Exp Psychol: Anim Learn Cognition, 40, 133–143. doi:10.1037/xan0000012

Carvalho PF, Goldstone RL (2015) What you learn is more than what you see: what can sequencing effects tell us about inductive category learning?. Front Psychol, 6, 505. doi: 10.3389/fpsyg.2015.005055

Davis T, Love BC, Maddox WT (2009) Two pathways to stimulus encoding in category learning? Mem Cognition, 37, 394–413. doi: 10.3758/MC.37.4.394

Donkin C, Newell BR, Kalish M, Dunn JC, Nosofsky RM (2014) Identifying strategy use in category learning tasks: A case for more diagnostic data and models. J Exp Psychol Learn 41(4):933–948. doi: 10.1037/xlm0000083.

Foard CF, Kemler Nelson DG (1984) Holistic and analytic modes of processing: the multiple determinants of perceptual analysis. J Exp Psychol Gen 113(1):94-111. doi: 10.1037/0096-3445.113.1.94

Ford N, Wood F, Walsh C (1994) 'Cognitive styles and searching', Online and CDROM Review, vol. 18, no. 2, pp. 79 – 86. doi: 10.1108/eb024480

Garner WR (1974) The processing of information and structure. Potomac Md: Erlbaum Associates.

Garner WR (1976) Interaction of stimulus dimensions in concept and choice processes. Cognitive Psychol 8:98-123.

Gazzaley A, Nobre AC (2012) Top-down modulation: bridging selective attention and working memory. Trends Cogn Sci, 16, 129–135. doi:10.1016/j.tics.2011.11.014

Gentner D, Markman AB (1994) Structural alignment in comparison: No difference without similarity. Psychol Sci 5(3), 152–158. doi: 10.1111/j.1467-9280.1994.tb00652.x

Gentner D, Namy LL (1999) Comparison in the development of categories. Cognitive Dev 14, 487–513.

Goldstein KM, Blackman S (1978) Cognitive style: Five approaches and relevant research. John Wiley & Sons, New York.

Goldstone RL, Medin DL (1994) Time course of comparison. J Exper Psychol Learn 20, 29–50. doi: 10.1037//0278-7393.20.1.29

Hahn U, Prat-Sala M, Pothos EM, Brumby DP (2010) Exemplar similarity and rule application. Cognition 114(1):1-18. doi: 10.1016/j.cognition.2009.08.011

Hammer R (2015) Impact of feature saliency on visual category learning. Front Psychol, 6, 451. doi: 10.3389/fpsyg.2015.00451

Hammer R, Bar-Hillel A, Hertz T, Weinshall D, Hochstein S (2008) Comparison processes in category learning: from theory to behavior. Brain Res, 1225, 102-118. doi: 10.1016/j.brainres.2008.04.079

Hammer R, Diesendruck G, Weinshall D, Hochstein S (2009) The development of category learning strategies: What makes the difference?. Cognition, 112(1), 105-119. doi: 10.1016/j.cognition.2009.03.012

Hammer R, Sloutsky V, Grill-Spector K (2015) Feature saliency and feedback information interactively impact visual category learning. Front Psychol, 6, 74. doi: 10.3389/fpsyg.2015.00074

Hoffman AB, Rehder B (2010) The costs of supervised classification: The effect of learning task on conceptual flexibility. J Exp Psychol Gen 139(2):319. doi: 10.1016/j.cognition.2014.11.019

Hubert M, Vandervieren E (2006) An Adjusted Boxplot for Skewed Distributions. Techn. Rep.TR-06-11, KU Leuven, Section of Statistics, Leuven.

Johansen MK, Palmeri TJ (2002) Are there representational shifts during category learning? Cognitive Psychol 45:482-553. doi: 10.1016/S0010-0285(02)00505-4

Johansen MK, Kruschke JK (2005) Category representation for classification and feature inference. J Exper Psychol Learn 31(6), 1433. doi: 10.1037/0278-7393.31.6.1433

Kurtz KJ, Boukrina O (2004) Learning relational categories by comparison of paired examples. Proceedings of the 26th Annual Conference of the Cognitive Science Society. http://escholarship.org/uc/item/3943j2wv

Lewicki P, Hill T, Sasaki I (1989) Self-perpetuating development of encoding biases. J Exp Psychol Gen 118:323-337. doi: 10.1037/0096-3445.118.4.323

Lockhead GR (1972) Processing dimensional stimuli: A note. Psychol Rev 79, 410–419. doi: 10.1037/h0033129

Maddox WT, Ashby FG (2004) Dissociating explicit and procedural-learning based systems of perceptual category learning. Behav Process 66, 309–332. doi: 10.1016/j.beproc.2004.03.011

Markman AB, Ross BH (2003) Category use and category learning. Psychol Bull 4:592-613. doi: 10.1037/0033-2909.129.4.592

Mathy F, Feldman J (2009) A rule-based presentation order facilitates category learning. Psychon Bull Rev 16, 1050–1057. doi: 10.3758/PBR.16.6.1050

Meagher BJ, Carvalho PF, Goldstone RL, Nosofsky RM (2017) Organized simultaneous displays facilitate learning of complex natural science categories. Psychon Bull Rev, 1-8. doi: 10.3758/s13423-017-1251-6

Michel C, Corneille O, Rossion B (2007) Race categorization modulates holistic face encoding. Cognitive Sci 31:911924. doi: 10.1080/03640210701530805

Michel C, Rossion B, Han J, Chung CS, Caldara R (2006) Holistic processing is finely tuned for faces of one's own race. Psychol Sci 17:608615. doi: 10.1111/j.1467-9280.2006.01752.x

Minda JP, Miles SJ (2010) The influence of verbal and nonverbal processing on category learning. Psychol Learn Motiv, 52, 117-162. doi: 10.1016/S0079-7421(10)52003-6

Moneer S, Wang T, Little DR (2016) The processing architectures of whole-object features: A logical-rules approach. J Exp Psychol Hum Percept Perform, *42*(9), 1443. doi: 10.1037/xhp0000227

Murphy GL (2002) The big book of concepts. MIT Press, Cambridge.

Murphy GL, Bosch DA, Kim S (2016) Do Americans Have a Preference for Rule□Based Classification? Cognitive Sci. doi: 10.1111/cogs.12463

Navon D (1977) Forest before trees: The precedence of global features in visual perception. Cognitive Psychol 9, 353 – 383. doi: 10.1016/0010-0285(77)90012-3

Nilsson H, Olsson H (2005) Categorization vs. inference: Shift in attention or in representation? In: Bara BG, Barsalou L, Bucciarelli M (ed) Proceedings of the 27th Annual Conference of the Cognitive Science Society Stresa, Italy: Cognitive Science Society, pp 1642-1647.

Nosofsky RM, Palmeri TJ, McKinley SC (1994) Rule-plus-exception model of classification learning. Psychol Rev 101(1): 53. doi:10.1037/0033-295x.101.1.53

Oakes LM, Ribar RJ (2005) A comparison of infants' categorization in paired and successive presentation familiarization tasks. Infancy 7, 85–98. doi: 10.1207/s15327078in0701_7

Palmeri TJ, Mack ML (2015) How experimental trial context affects perceptual categorization. Front Psychol, 6, 180. doi: 10.3389/fpsyg.2015.00180

Perry LK, Saffran JR (2016) Is a pink cow still a cow? Individual differences in toddlers' vocabulary knowledge and lexical representations. Cognitive Sci. 1, 16. doi: 10.1111/cogs.12370

Perry LK, Axelsson EL, Horst JS (2015) Learning what to remember: Vocabulary knowledge and children's memory for object names and features. Infant Child Dev *25*(4), 247-258. doi: 10.1002/icd.1933.

Piepers D, Robbins R (2012) A review and clarification of the terms "holistic,""configural," and "relational" in the face perception literature. Front Psychol, 3, 559. doi: 10.3389/fpsyg.2012.00559

Pothos E (2005) The rules versus similarity distinction. Behav Brain Sci 28:1-14. doi: 10.1017/S0140525X05000014

Ramon M, Van Belle G (2016) Real-life experience with personally familiar faces enhances discrimination based on global information. Peerj, 4, e1465. doi: 10.7717/peerj.1465

Regehr G, Brooks LR (1993) Perceptual manifestations of an analytic structure: the priority of holistic individuation. J Exp Psychol Gen 122 (1):92-114. doi: 10.1037/00963445.122.1.92

Richardson DC, Kirkham NZ (2004) Multimodal events and moving locations: Eye movements of adults and 6-month-olds reveal dynamic spatial indexing. J Exp Psychol Gen 133:46-62. doi: 10.1037/0096-3445.133.1.46

Richardson DC, Spivey MJ (2000) Representation, space and Hollywood Squares: Looking at things that aren't there anymore. Cognition 76:269-295. doi: 10.1016/ S00100277 (00)00084-6

Riding RJ, Rayner SG (1998) Cognitive styles and learning strategies. David Fulton, London.

Rips LJ, Collins A (1993) Categories and resemblance. J Exp Psychol Gen 122(4):468-486. doi: 10.1037/0096-3445.122.4.468

Ross BH, Murphy GL (1999) Food for thought: Cross-classification and category organization in a complex real-world domain. Cognitive Psychol 38(4):495-553. doi: 10.1006/cogp.1998.0712

Scholz A, von Helversen B, Rieskamp J (2015) Eye movements reveal memory processes during similarity-and rule-based decision making. Cognition 136:228-246. doi: 10.1016/j.*cognition*.2014.11.019

Shi C (2011) A Study of the Relationship between Cognitive Styles and Learning Strategies. High Educ Studies, 1(1), 20-26. doi:10.5539/hes.v1n1p20

Smith JD, Kemler Nelson DG (1984) Overall similarity in adults' classification: The child in all of us. J Exp Psychol Gen, 113, 137-159. doi: 10.1037/0096-3445.113.1.137

Smith EE, Medin DL (1981) Categories and concepts. Harvard University Press, Cambridge. doi :10.2307/414206

Smith JD, Shapiro JH (1989) The occurrence of holistic categorization. J Mem Lang 28, 386-399 (1989)

Spalding TL, Ross BH (1994) Comparison-based learning: effects of comparing instances during category learning. J Exper Psychol Learn, 20 (6), 1251–1263. doi: 10.1037//0278-7393.20.6.1251

Sternberg RJ, Grigorenko EL (1997) Are cognitive styles still in style?. Am Psychol, 52(7), 700. doi: 10.1037/0003-066X.52.7.700

Tanaka JW, Farah MJ (2003) The holistic representation of faces. In Peterson MA, Rhodes G (Eds.) Perception of faces, objects, and scenes: Analytic and holistic processes. Oxford University Press, 53-74.

Tanaka JW, Kiefer M, Bukach CM (2004). A holistic account of the own-race effect in face recognition: Evidence from a cross-cultural study. Cognition, 93(1), B1-B9. doi: 10.1016/j.cognition.2003.09.011

Tunney RJ, Fernie G (2012) Episodic and prototype models of category learning. Cognitive Process 13(1):41–54. doi:10.1007/s10339-011-0403-2

Vlach HA (2016). How we categorize objects is related to how we remember them: The shape bias as a memory bias. J Exp Child Psychol 152, 12-30. doi: 10.1016/j.jecp.2016.06.013

Wagemans J, Feldman J, Gepshtein S, Kimchi R, Pomerantz JR, van der Helm PA, van Leeuwen C (2012) A century of Gestalt psychology in visual perception: II. Conceptual and theoretical foundations. Psychol Bull 138(6):1218. doi: 10.1037/a0029333

Wang Y, Fu X, Johnston RA, Yan Z (2013) Discriminability effect on Garner interference: evidence from recognition of facial identity and expression. Front Psychol 4:1-10. doi: 10.3389/fpsyg.2013.00943

Ward TB, Scott J (1987) Analytic and holistic modes of learning family-resemblance concepts. Mem Cognition 15(1):42-54. doi:10.3758/BF03197711

Ward TB (1988) When is category learning holistic? Mem Cognition, 16(1), 85–89. doi: 10.3758/BF03197749

Weissman DH, Prado J (2012) Heightened activity in a key region of the ventral attention network is linked to reduced activity in a key region of the dorsal attention network during unexpected shifts of covert visual spatial attention. Neuroimage 61, 798–804. doi: 10.1016/j.neuroimage.2012.03.032

Williams, D. A., Sagness, K. E., & McPhee, J. E. (1994). Configural and elemental strategies in predictive learning. J Exp Psychol Learn, 20, 69– 709. doi: 10.1037/0278-7393.20.3.694

Wills AJ, Inkster AB, Milton F (2015) Combination or differentiation? Two theories of processing order in classification. Cognitive Psychol, 80, 1-33. doi: 10.1016/j.cogpsych.2015.04.002

Yamauchi T, Markman AB (1998) Category learning by inference and classification. J Mem Lang 39:124-48. doi: 10.1006/jmla.1998.2566

Zaki SR, Kleinschmidt DF (2014) Procedural memory effects in categorization: evidence for multiple systems or task complexity? Mem Cognition 42(3):508–24. doi: 10.3758/s13421-013-0375-9

Appendix

RULES	PCA-1	PCA-2	PCA-3	PCA-4
1) MEDs have more dots than DAXs			1.005	
2) MEDs have a longer stem than DAXs	.668			.262
3) DAXs have from 1 to 5 dots		.916		
4) DAXs have from 1 to 5 dots and MEDs from 6 to 10		.929		
5) MEDs' caps are larger than DAXs'	.777			.238
6) DAXs are smaller than MEDs	.807			
7) DAXs have few dots			.837	
8) DAXs have a more orderly arrangement of dots				.995
9) DAXs have short stems.	.954			-.247
	perceptual bias	counting	estimate	arrangement bias

Table 1 Loadings of components extracted by a Principal Component Analysis from the final rating task

FACTORS	Cluster	
	A	B
perceptual bias on irrelevant features	-.77430	.82591
counting	.21018	-.22419
estimate	-.46519	.49620
perceptual bias on criterial feature (arrangement)	-.46043	.49113

Table 2 Factor scores of two groups for each component