Does practice in category learning increase rule use or exemplar use—or both?

Jean-Pierre Thibaut, Sabine Gelaes & Gregory L. Murphy

Memory & Cognition

ISSN 0090-502X

Mem Cogn
DOI 10.3758/s13421-017-0782-4

ONLINE FIRST

Memory & Cognition

VOLUME 46, NUMBER 1 ■ JANUARY 2018

M&C

A PSYCHONOMIC SOCIETY PUBLICATION
www.psychonomic.org
ISSN 0090-502X

PSYCHONOMIC SOCIETY®

Springer

Springer

Memory & Cognition
https://doi.org/10.3758/s13421-017-0782-4

Does practice in category learning increase rule use or exemplar use—or both?

Jean-Pierre Thibaut[1] · Sabine Gelaes[2] · Gregory L. Murphy[3]

Abstract

Categorization research has demonstrated the use of both rules and remembered exemplars in classification, although there is disagreement over whether learners shift from one to the other or use both strategies simultaneously. Theoretical arguments can motivate predictions for both rule use and exemplar use increasing with more practice. We describe a single large experiment (n = 190) that manipulated the number of training items (category size), the number of presentations of each training item, and the similarity between the training and the transfer stimuli in order to discover when rules and exemplars are most likely to be used. Results showed that rules and exemplars both influenced classification and that exemplars were used more often with smaller categories, with more training on items, and when test items were similar to training items. There was no consistent evidence of a shift from rule-based to exemplar-based categorization with more learning. Importantly, we found a number of conditions in which rules and exemplars were both used, even within individual participants. We discuss our results in terms of hybrid models of classification.

Keywords Category learning · Concepts · Rules · Exemplars · Abstraction

Studies of category learning have contrasted rule-based and exemplar-based processes. According to rule-based accounts, people learn defining rules for categorization and apply them in subsequent categorizations (see the seminal studies by Bruner, Goodnow, & Austin, 1956; Bourne, 1970). Rule application is generally defined as attending to a subset of the stimulus features, with one or more features combined deterministically to decide membership. On the other hand, in exemplar-based categorization, people are assumed to evaluate the similarity of a novel item as a whole to remembered exemplars of known categories (Estes, 1986, 1994; Medin and Shaffer, 1978; Nosofsky, 1984; Nosofsky & Palmeri, 1997; see also Wills, Inkster, & Milton, 2015). This debate has paralleled similar controversies in other cognitive domains such as language (e.g., Pinker, 1999), problem solving (e.g., Medin & Ross, 1989), skill acquisition (e.g., Anderson, Fincham, & Douglass, 1997), and reasoning (e.g., Norenzayan, Smith, Kim, & Nisbett, 2002; Sloman, 1996).

Recently, several hybrid theories based on abstract representations (e.g., rules or prototypes) and similarity to remembered exemplars have been proposed. One key question has been over the respective role of rules and similarity to prior exemplars when they are *both* available (see Allen & Brooks, 1991; Erickson & Kruschke, 1998, 2002; Johansen & Palmeri, 2002; Love, Medin, & Gureckis, 2004; Nosofsky, Palmeri, & McKinley, 1994; Rips, 1989; Smith & Minda, 1998; Smith, Patalano, & Jonides, 1998; Thibaut, Dupont, & Anselme, 2002; Thibaut & Gelaes, 2006). Hahn and Chater (1998, p. 224) suggested that "rules and similarity both have their respective roles, not just side by side, with similarity covering some domains and rules others, or doubling up in parallel, but in an active interplay within a single task."

Is an explicit rule provided or not?

One can distinguish at least two kinds of studies investigating rules and categorization. In the first and most popular kind,

Electronic supplementary material The online version of this article (https://doi.org/10.3758/s13421-017-0782-4) contains supplementary material, which is available to authorized users.

✉ Jean-Pierre Thibaut
jean-pierre.thibaut@u-bourgogne.fr

[1] University of Bourgogne Franche-Comté, LEAD CNRS UMR 5022, 11, Esplanade Erasme, 21065 Dijon, France

[2] Liège, Belgium

[3] New York University, New York, NY, USA

Published online: 08 January 2018

participants do not receive any classification rule at the onset of the experiment. The stimuli are introduced one by one, and the rule is learned through corrective feedback. Most authors assume that the abstracted rules, if any, are single-dimension categorization rules that do not work perfectly (e.g., Johansen & Palmeri, 2002; Nosofsky et al., 1994; Ward & Scott, 1987). For example, Johansen and Palmeri found that when the experience with the category was limited, participants tried simple rules and generalized on the basis of these single diagnostic dimensions, even though the rules were not perfect. Later in learning, generalization was mostly driven by similarity to exemplars. Johansen and Palmeri described this change as a shift from rule-based to exemplar-based classification (see also Raijmakers, Schmittmann, & Visser, 2014). As noticed by Smith, Murray, and Minda, (1997, p. 667), this might result from the category structure of the categories, which is often very difficult to grasp, and push participants towards exemplar encoding, especially if one considers the (often) very limited number of training stimuli. Note that in this kind of paradigm, it can be difficult to disentangle exemplar and rule influence because one does not know whether and which rule or imperfect rule plus exemplar encoding learners follow.

In a second kind of category learning design, participants receive an explicit, perfect rule for classification. Using this design, Lee Brooks and colleagues (Allen & Brooks, 1991; Brooks, Norman, & Allen, 1991; Regehr & Brooks, 1993) showed that classification of transfer stimuli was still influenced by their similarity to the training instances. Participants were trained in applying a rule to a limited set of items (eight stimuli presented five times). The key result was that in a following transfer phase, "good transfer items" (*GoodT*), items that were similar to training items and belonged to the same category, were categorized faster and far more accurately than "bad transfer items" (*BadT*), items similar to training items but belonging to the opposite category according to the rule. Because both types of items were equally classifiable by the rule, a difference between GoodT and BadT items is interpreted as being due to exemplar-based processing.

These early studies revealed that similarity to exemplars had a large effect on performance that was most likely to exert its influence when exemplars were both featurally and holistically individuated (i.e., the instances of a given part are different from one stimulus to the other), and the overall shapes of the animals are distinctive. For example, in our Fig. 1, all the instances' legs are different across stimuli (see Regehr & Brooks, 1993, Experiments 3A and 3B) with rules involving several features (e.g., two out of three among F1, F2, and F3). In contrast to Regehr and Brooks, Lacroix, Giguère, and Larochelle (2005) extended these results to stimuli made up with features which had the same perceptual implementation across stimuli (e.g., the same round or square body shape appeared in the stimuli). However, even though significant, the targeted difference was very tiny (e.g., 7 % errors for BadT items vs. 3 % errors for the training twin in Experiment 2) and this percentage decreased with more trials.

Thibaut and Gelaes (2006) showed that similarity effects were also obtained with features that were spatially separated rather than integrated into a whole. Interestingly, they also found that the exemplar effect depended on the similarity between transfer stimuli and their training twins. We will come back to this issue in our experiment. Finally, Hahn, Prat-Sala, Pothos, and Brumby (2010, Experiment 2) extended exemplar effects to a design in which simple, one-feature rules (e.g., "all As have a triangle") were used and with stimuli which were collections of features, rather than integrated wholes as in Brooks and colleagues' studies. However, again, the exemplar effects were small (around 5 %, vs. 20–40 % or more in Regehr & Brooks or Thibaut & Gelaes) and were obtained with a different design.

Overall, there is strong evidence for exemplar effects with strongly individuated stimuli or features in the case of additive rules (Regehr & Brooks, 1993; Lacroix et al., 2005; Thibaut & Gelaes, 2006), and there is consistent, but much smaller, evidence from single-feature rules with stimuli that appeared as a collection of features rather than as an integrated whole (Hahn et al., 2010) or that were composed of interchangeable features (Lacroix et al., 2005). A generalization of these results is that when memory for learning items is made stronger and more distinctive, the exemplar effects are larger.

Does more training increase rule-based or exemplar-based processes?

In the present paper, we used Regehr and Brooks's paradigm, which gave the largest exemplar effects. Starting with conditions that are already known to give an exemplar effect, the key issue was whether rule or exemplar similarity would take over when more trials or more training stimuli were provided during the training phase or whether there was evidence for simultaneous use of both sources of categorization.

Previous theories lead to opposite predictions. Indeed, the effect of learning has been conceptualized either as a shift from abstraction-based classifications (rule-based in Johansen & Palmeri, 2002, or prototype-based in Smith & Minda, 1998) to exemplar-based classifications or as a shift from exemplar-based classifications to abstraction-based classification (especially when no rule is available during learning) (Homa, Sterling, & Trepel, 1981).

One main hypothesis is that when exemplars are encoded more distinctively and accurately, exemplar-based classification is more likely to occur. As a result, increasing the *number of presentations* of the stimuli during the training phase could increase the difference between good transfer (GoodT) and bad transfer items (BadT). As Smith et al. (1998) put it, at the beginning of the training phase, when subjects are still

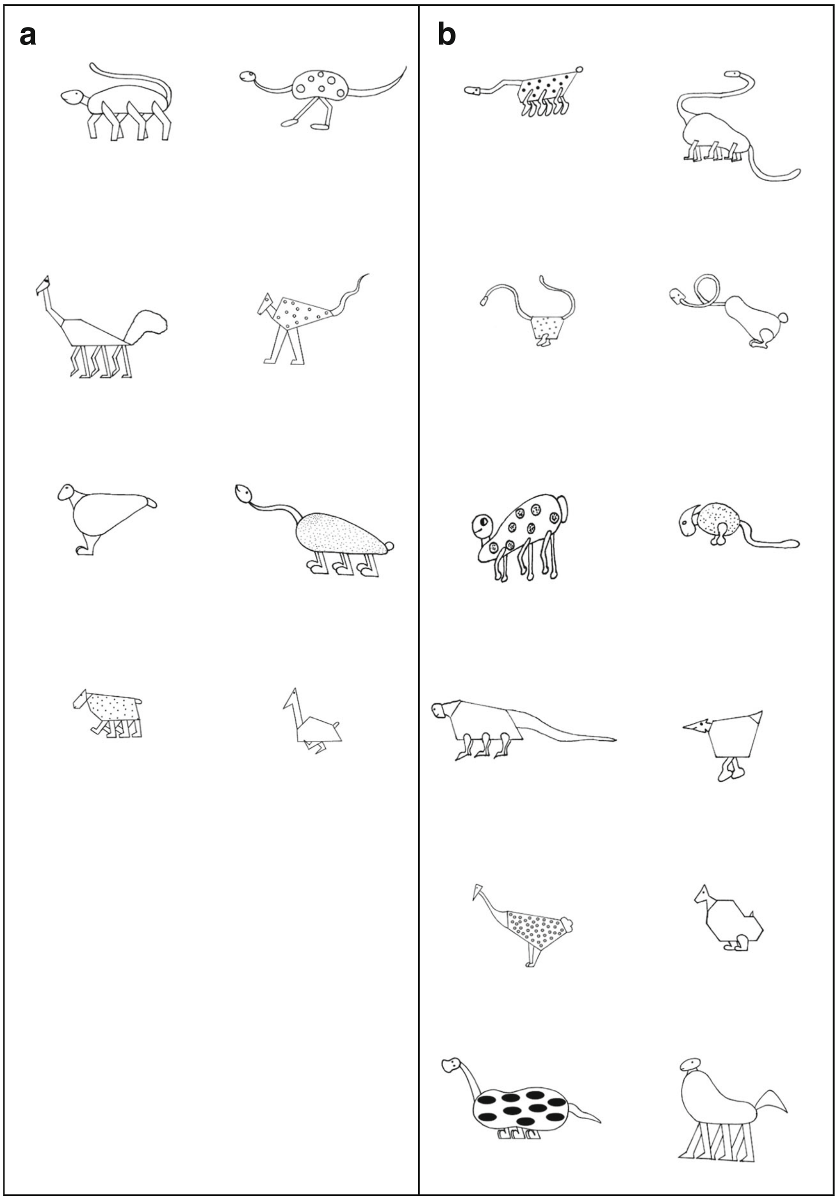

Fig. 1 **a** Eight original training stimuli used in presentation types 8-5 and 8-30. **b** The 12 additional training stimuli used in presentation types 20-2 and 20-12. Their status (positive or negative training item) depends on which of the four rules is used. Note that depending on the rule, the same stimulus will be classified as a Digger or as a Builder. This is true for both training and transfer stimuli. For example, according to Rule 1, "Six legs, spots present, and angular body" Stimulus 1 (upper left) on Figure 1 would be a Digger, but according to Rule 3, "Six legs, spots absent, and round body" it would be a Builder

learning to categorize the objects, the representation of the entire object will be imperfectly retrieved. Consequently, exemplar retrieval should have little influence on categorization until automatic memory retrieval starts to take over. Logan (1988) posits the development of automaticity as a shift from algorithm-based to instance-based processing. Encoding and retrieval from memory of all information associated with a stimulus are unavoidable consequences of attention. As people gain experience with the training exemplars, their later performance should become more influenced by these exemplars (see also Lamberts, Brockdorff, & Heit, 2003; Nosofsky & Palmeri, 1997). Nosofsky (1988) found that instance frequency influences category knowledge directly, positing that repetition of the training exemplars produces multiple traces of the item that provide more evidence for category membership and result in stronger exemplar similarity effects (see also Logan & Etherton, 1994). This was confirmed by Smith and Minda (1998) who demonstrated an early advantage for the prototype model and a late advantage for the exemplar model under the assumption that exemplars tend to retrieve themselves with more practice (see Johansen & Palmeri, 2002, and Nosofsky & Zaki, 2002, for discussion).

Following this logic, within our design, increasing the number of presentations should maximize the priority of exemplar influence, because the training exemplars should become more distinctively encoded, which should promote exemplar-based categorization. Increasing similarity between the training and transfer stimuli should also increase exemplar influence, because the transfer stimuli will evoke their training twin memory trace more strongly. For the same reasons, a smaller number of training exemplars should lead to a larger exemplar effect: Each exemplar should be more distinctively stored in memory and be more directly evoked by its transfer twin. In this vein, Smith and Minda (1998) argue that small exemplar sets favor exemplar-based processes, presumably by reducing interference among similar category members. Homa and colleagues (Homa & Vosburgh, 1976; Homa, Burruel, & Field, 1987; Homa, Sterling, & Trepel, 1981) found evidence that the advantage of old over new exemplars decreases when the number of training instances increases. However, in their experiments, the categories could not be identified by a rule (see also Smith & Minda, 2000).

When a perfect rule is available, increasing the number of stimulus presentations increases the expertise of using that rule. With more practice, the explicit rule becomes more automatized, consuming fewer resources than at the beginning of the experiment (Shiffrin & Schneider, 1977). People focus on the defining features, becoming more efficient at extracting the abstract features for categorization and at ignoring the idiosyncratic aspects of each exemplar, as confirmed in eyetracking studies (Rehder & Hoffman, 2005). As a result, exemplars cease to influence performance (e.g., Shiffrin & Schneider, 1977; see also the ACT-R theory of skill acquisition, Anderson, 1993; Anderson, Fincham, & Douglass, 1997; Smith et al., 1998).

A third possibility is that both strategies are used. However, there is a further theoretical issue at stake, namely whether these strategies compete. It is possible that any given individual will carefully follow the provided rule or else will use exemplar similarity, but not both. Thus, both strategies could be used across participants but not within them. The COVIS model suggests this possibility by its name, *COmpetition between Verbal and Implicit Systems*, but its workings are actually more complex. Ashby, Alfonso-Reese, Turken, and Waldron (1998) propose that an implicit associative system and rule-learning occur in parallel. Learners discover which system is more accurate, and the two systems are weighted accordingly. On any one trial, both systems compute their answer and the strongest one determines the response. As this is a function of the weighting of the two systems and the identity of the particular stimulus, both systems can contribute to responses over trials, even if one is dominant. Ashby et al. propose that rule use is dominant at the beginning, given subjects' expectations about such experiments and also the absence of any implicit learning. Their theory suggests that rule learning and other forms of learning can co-exist and potentially both have an effect on a given person's behavior, even in individual trials (see Hahn et al., 2010). We attempt to identify whether a given individual in our study uses both.

Goals of the experiment

Previous studies have sought to establish under which conditions exemplar effects can be obtained. The major goal of the present experiments is to explore, within the Regehr and Brooks paradigm, how training conditions influence the use of both rules and exemplars and whether we can find simultaneous use of both strategies. This is a crucial issue that has not been systematically studied. Most past studies have sought evidence for exemplar effects but have not explicitly addressed the possibility that rules might also be used at the same time. Here we focus on distinctive stimuli (i.e., holistically individuated) for which the evidence for exemplar effects was the strongest (as reviewed above). We manipulated variables likely to increase or decrease the use of exemplars: category size, the number of stimulus presentations, and the similarity between training and transfer stimuli.

Knowing the rule is a critical component of these predictions. People do not always spontaneously notice and use rules even under advantageous conditions (Murphy, Bosch, & Kim, 2017). In our experiment, participants are explicitly informed of the rule at the start. Our study then asks whether exemplar processing has an effect in spite of knowledge of the rule, and under what training conditions such effects are most likely to be found.

Design

Exemplar effects are evidenced by differences between old and new (transfer) stimuli or between transfer stimuli similar to an exemplar from the same category (GoodT) and stimuli similar to an exemplar from the opposite category (BadT). Rule use would be confirmed by successful classification and the absence of these exemplar effects. However, both patterns could be found, as when there is above-chance classification even for BadT items, suggesting rule use even if there is a similarity effect.

Our starting point was the presentation type 8-5 (eight training stimuli presented five times) used in Experiments 1 and 2 from Thibaut and Gelaes (2006). We compared four category structures varying in the number of exemplars and their repetition during learning: 8-5, 8-30 (i.e., eight training stimuli presented 30 times), 20-2, and 20-12.

The contribution of rule-based and exemplar influence was assessed through two effects. First, a difference between BadT and GoodT items, which we refer to as the *BadT-GoodT effect* and, second, a reliable difference between training and transfer phases (i.e., old vs. new items) would each suggest an exemplar influence on the classification of transfer items. Indeed, if performance is entirely rule-based, classification accuracy and response times (RTs) of old and new items should be equivalent.

We expected three manipulations to reveal differences in exemplar use. First, more presentations of the same set of stimuli should strengthen exemplar memory. Thus, finding a similar pattern of results across stimulus exposures will be a sign of rule use. The second manipulation was category size, when the number of training exemplars increases but the total number of trials remains the same (e.g., 20 training stimuli presented 12 times vs. eight training stimuli presented 30 times). Again, no category size influence on the exemplar effect, would suggest the dominance of rule use across conditions. Third, we varied the similarity between training and transfer items. If participants' performance is influenced by rule use only, then similarity between training and transfer items should have no effect. Thibaut and Gelaes (2006) found no exemplar effect in Experiment 1 but a significant one in Experiment 2, when the similarity between test items and the learning items was greater. Thus, we will focus on how increasing the number of trials will differentially affect the exemplar effect in both similarity conditions.

Of particular interest is a pattern showing evidence of both rule and exemplar use within individual participants. As noted above, when classification on BadT items is worse than that of GoodT items, it indicates that people are relying on similarity to learned exemplars. If performance on BadT items is simultaneously above chance, however, that also indicates rule use, as reliance solely on similar exemplars would result in 0 % accuracy (BadT items are similar to an exemplar in the "wrong" category). Thus, we looked in particular for this pattern, which would give evidence for simultaneous rule and exemplar use, not only within specific conditions but also within individual participants. We also looked for individuals who displayed (virtually) perfect accuracy—suggesting rule use—together with longer RTs for BadT items than for GoodT items—suggesting that despite using the rule, they were slowed down for BadT items similar to examples from the opposite category. To the best of our knowledge, past research has not done this. Overall, we investigated under which conditions the exemplar effect would be the largest despite evidence for rule use, and we sought individual data that would witness unambiguously the simultaneous use of rule and exemplars. This is a considerable extension of previous studies devoted to these issues.

Method

Participants

One hundred and ninety university undergraduates participated as unpaid volunteers, 24 per category type, except the category type 20-2, in the lower similarity condition, which had 22 participants. The lower and higher similarity conditions were not run at the same time, so participants were not randomly assigned to the high and low similarity category types. Rather, they were randomly assigned to each category type within the two similarity conditions. However, all subjects came from the same population.

Materials

Training stimuli Two sets of training stimuli were introduced in two similarity conditions. The first set, for category types 8-5 and 8-30, was composed of the eight original stimuli created by Regehr and Brooks (1993) in their experiment 3A. They were line drawings of imaginary animals (see Fig. 1, Cell A). The animals were made up from five binary dimensions: number of legs (two or six), body shape (round or angular), spots (present or absent), tail length (short or long), and neck length (short or long). The second set of stimuli, used in the category types 20-2 and 20-12, contained twenty training items, the eight original stimuli from Regehr and Brooks, plus twelve new training items (Fig. 1, Cell B) constructed according to the same specifications as the original stimuli (see Table 1 for a logical description of the stimuli).

Transfer stimuli There were eight transfer stimuli. Each transfer stimulus was based on one of the original set of eight training stimuli (the *twin training items*). The difference between a transfer stimulus and its training twin was either on the dimension of spots (lower similarity condition) or of body shape (higher similarity condition, see Fig. 2 A). For example, if body shape was round on the training item, the body shape of its transfer twin was angular, and vice-versa. These manipulations gave rise to four types of items:

1. Positive training items: twins of GoodT.
2. Negative training items: twins of BadT.
3. Positive transfer (GoodT) items: A stimulus seen in the transfer phase that, according to the rule, was in the same category as its twin training stimulus.
4. Negative transfer (BadT) items: A stimulus seen in the transfer phase that, according to the rule, was in the category opposite to its twin training stimulus.

It is important to note that the training stimuli get the names *positive* and *negative* only by reference to the status of their transfer twin, a positive training item being the twin of a transfer item belonging to the same category, whereas a negative training item is the twin of a transfer belonging to the other category in terms of the rule. Thus, there is no reason to expect that they will lead to different levels of performance.

Table 1 Logical description of the stimuli

Item n°	No. of legs	Body shape	Spots	Neck length	Tail length	Category according to RULE 1 (Builder = Six legs, Spots, Angular body)
Training stimuli						
1	1	0	0	0	1	Digger
2	0	0	1	1	1	Digger
3	1	1	0	1	1	Builder
4	0	1	1	0	1	Builder
5	0	0	0	0	0	Digger
6	1	0	1	1	0	Builder
7	1	1	1	0	0	Builder
8	0	1	0	1	0	Digger
9	1	1	1	1	0	Builder
10	1	0	0	1	1	Digger
11	0	1	1	1	1	Builder
12	0	0	0	1	0	Digger
13	1	0	1	0	0	Builder
14	0	0	1	0	1	Digger
15	1	1	0	0	1	Builder
16	0	1	0	0	0	Digger
17	0	1	1	1	0	Builder
18	0	1	0	0	0	Digger
19	1	0	1	1	1	Builder
20	1	0	0	0	1	Digger
Transfer stimuli in the low similarity condition						
1	1	0	***1***	0	1	Builder
2	0	0	***0***	1	1	Digger
3	1	1	***1***	1	1	Builder
4	0	1	***0***	0	1	Digger
5	0	0	***1***	0	0	Digger
6	1	0	***0***	1	0	Digger
7	1	1	***0***	0	0	Builder
8	0	1	***1***	1	0	Builder
Transfer stimuli in the high similarity condition						
1	1	***1***	0	0	1	Builder
2	0	***1***	1	1	1	Builder
3	1	***0***	0	1	1	Digger
4	0	***0***	1	0	1	Digger
5	0	***1***	0	0	0	Digger
6	1	***1***	1	1	0	Builder
7	1	***0***	1	0	0	Builder
8	0	***0***	0	1	0	Digger

Note. For the training stimuli, Stimuli 1–8 are used in presentation types 8-5 and 8-30. Stimuli 9 to 20 were added in presentation types 20-2 and 20-12. 0 and 1 represent the following values: number of legs, 1 = six legs, 0 = two legs; body shape, 1 = angular, 0 = round; spots, 1= present, 0 = absent; neck length, 1 = long, 0 = short; tail length, 1 = long, 0 = short.

The eight transfer stimuli were transformations of the training stimuli 1–8 on the dimension of "spots" in the low similarity condition and on the dimension of "body shape" in the high similarity condition.

The last column gives the category membership (Builder or Digger) of the training stimuli and of the eight transfer items when Rule 1 is used

The difference between the body shape and the spots transformations gave rise to the Similarity factor (see Thibaut and Gelaes, 2006, p. 1408-9). In order to establish whether body shape transformations would give more similar twin stimuli than transformations on spots, Thibaut and Gelaes asked participants to choose which of the two types of transfer stimuli

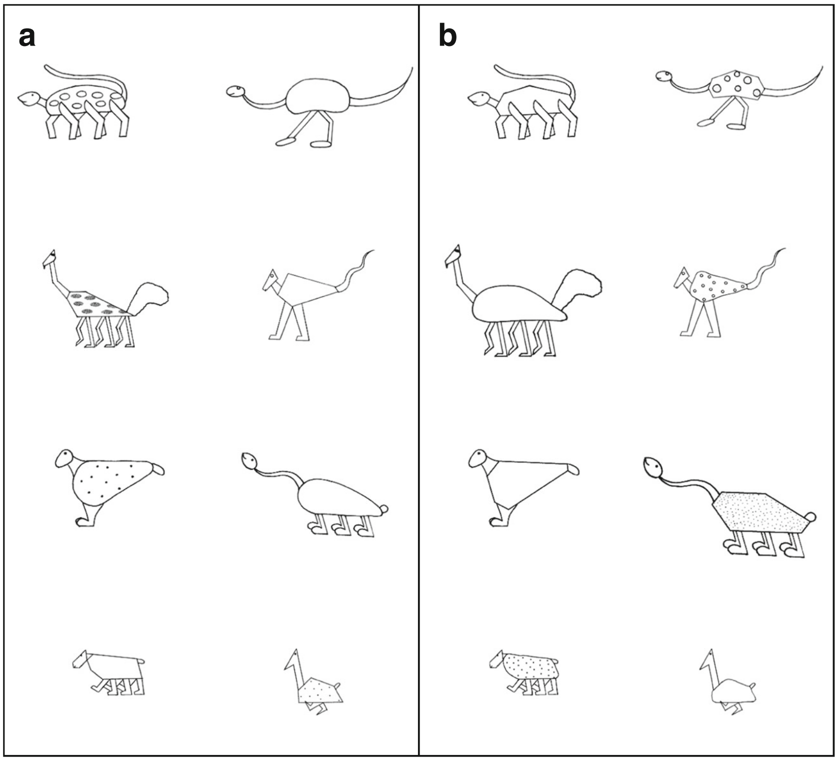

Fig. 2 **a** Transfer stimuli used in the low similarity condition. **b** Transfer stimuli used in the high similarity condition. Compare to corresponding cells of Fig. 1. The eight transfer items were transformations of the training stimuli (Fig. 1) on the dimension of spots in the low similarity case and on the dimension of body shape in the high similarity case

was the most similar to the corresponding training stimuli (the standard). The mean percentage of "body shape" choices was 76 %, which differed significantly from 50 %, $t(9) = 3.8$, $p < .005$. Then, they found large classification differences in their main experiments due to this similarity manipulation.

Participants categorized the stimuli into two categories, Builders and Diggers, using a three-feature additive rule that was a combination of three dimensions: number of legs, body shape, and spots. An animal was classified as a Builder if it possessed a majority of the Builders' features (at least two of the three Builder features, see below); the four other animals were deemed as Diggers. The values of the two irrelevant dimensions (tail length and neck length) appeared equally often in the two categories and were not diagnostic. Four rules were used across participants to counterbalance the stimuli across category types. This ensured that each transfer stimulus served as both a GoodT and a BadT item, depending on the rule. Thus, any difference between GoodT and BadT conditions could not result from differences associated to irrelevant characteristics of the stimuli (see Thibaut & Gelaes, 2006, for a methodological discussion).

Builders were defined as follows:

Rule 1 - Six legs, spots present, and angular body. (For example, according to this rule the first item (upper left) in Fig. 1 would be a Digger [and Negative training] and stimulus Transfer 1 in Fig. 2 would be a Builder, thus a BadT item.) See Table 1 for the classification of training and transfer stimuli according to Rule 1.

Rule 2 - Two legs, spots present, and angular body.

Rule 3 - Six legs, spots absent, and round body. (For example, the first item (upper left) in Fig. 1 would be a Builder and stimulus Transfer 1 (upper left) in Fig. 2 would also be a Builder, thus a GoodT item.)

Rule 4 - Two legs, spots absent, and round body.

Procedure

Participants were tested individually. They were seated at about 70 cm from the screen of an Apple Macintosh computer. Superlab was used to control the experiment, present the instructions and the stimuli, and record the answers. Participants had to press one of two keys (Builder = key 4 and Digger = key 5) on the numerical keyboard to make their classifications. The reaction time was the interval between the onset of stimulus presentation and the response. The stimuli were displayed until the answer was given. The experiment was composed of two phases, a training and a transfer phase.

Training phase Participants were told that they were to learn to classify line drawings of imaginary animals into two categories according to the explicit categorization rule provided to them. The rule was written on a sheet of paper placed between the keyboard and the screen. It remained in view during the entire experiment. Participants had to categorize the training stimuli as quickly and as accurately as possible. In the 8-5 and 8-30 category types, the eight training stimuli appeared five and thirty times respectively (i.e., 40 and 240 trials). In the 20-2 and 20-12 category types, the 20 training stimuli appeared two and 12 times (i.e., 40 and 240 trials). Within each block of eight or 20 items, stimuli were presented randomly. Feedback followed each response.

Transfer phase The eight transfer stimuli were presented randomly. There was no feedback during this phase. However, the transfer phase started with four training stimuli in order to familiarize participants with the absence of feedback. As in Regehr and Brooks (1993), we asked participants to classify the animals according to the rule as quickly and as accurately as possible.

Results

The raw data are archived at https://osf.io/tnqjg/ . The main aim of our analyses is to provide two complementary types of evidence of exemplar effects. Recall that participants were given the explicit categorization rule and that a high level of accuracy depends on the use of this rule. Pure rule use is evidenced by cases in which there is little or no evidence of a difference between GoodT and BadT items or between training and transfer items. Additionally, if performance is governed by rule use only, there should be no interaction involving the factor of similarity between training and transfer items.

In the first analyses, we looked at the exemplar effect as a function of training conditions and level of similarity between training and transfer items. In order to compare these conditions, we carried out two analyses, the first with the difference between BadT and GoodT test items as the dependent variable, and the second one with the difference between Transfer and Training items as the dependent variable. These analyses will tell us in which conditions the differences we are looking for, i.e., revealing exemplar influence, are larger or smaller. If performance is totally under rule control, no differences should be observed.

In the second analyses, we compared higher performing participants (thus who followed the rule) with lower performing participants (presumably relying on exemplar similarity). Given that rules presumably control the most accurate performers, would they still show evidence of exemplar influence?

Comparing training conditions

Our first analyses will focus on proportions of errors and RTs as a function of training condition and similarity level. As mentioned above, we focused on two difference scores, each directly reflecting exemplar use: the difference between BadT and GoodT test items and the difference between old (training) and new (test) items. (We also carried out a global analysis of variance on the raw classification and RT data, in which the variables similarity (higher, lower), stimulus type (positive, negative), phase (training, transfer), and category type (8-5, 8-30, 20-2, 20-12) were crossed (see Table 2 for the resulting means). However, as a number of critical results involved higher-level interactions followed up by contrasts, the result is very difficult to follow. Therefore, we have placed this analysis in the Supplementary Materials for those who are interested. Here we report the theoretically significant analyses using the difference scores that directly reflect exemplar usage.)

Proportion of errors

We first performed a two-way ANOVA on the proportion of errors for BadT minus proportion of errors for GoodT, with category type (8-5, 8-30, 20-2, 20-12) and similarity (higher, lower) as between factors. Exemplar influence is measured by the size of the difference between BadT and GoodT items. The ANOVA revealed a significant effect of category type, $F(3, 182) = 6.17, p < 0.001, \eta_P^2 = .09$. The Tukey HSD test revealed that the difference was significantly larger in 8-5 and 8-30 than in 20-2. The difference between 8-30 and 20-12 was only marginally significant (p < .07). It also revealed that the BadT-GoodT effect was significantly larger in the higher similarity condition (M = .38) than in the lower condition (M = .12) $F\,(1, 182) = 43.52, p < 0.00001, \eta_P^2 = .19$. The interaction did not reach significance (p < .5). Thus, the exemplar effect is more pronounced with higher similarity and with smaller numbers of training items.

The second analysis examined the difference between each transfer item and its training twin, that is GoodT minus Positive training (*Pos*) and BadT minus Negative training (*Neg*). Since similarity of transfer items is helpful in the first comparison (GoodT-Pos) but misleading in the second (BadT-Neg), the difference between training and transfer should be greater in the latter comparison. Thus comparing these two differences (hereafter, the *Positive-Negative* factor) provides another direct measure of exemplar influence.

The three-way ANOVA on the transfer-training difference scores, with category type and similarity as between factors and negative-positive (GoodT-Pos, BadT-Neg) as a within factor, revealed a main effect of category type, $F(3, 182) = 19.72, p < 0.0001, \eta_P^2 = .24$, of similarity, $F(1, 182) = 44.35, p$

Table 2 Mean response times and proportions of errors for training and transfer stimulus across category types and similarity conditions (standard deviations in brackets)

	Low similarity					High similarity				
	Training			Transfer		Training			Transfer	
Category Type		Pos	Neg	GoodT	BadT		Pos	Neg	GoodT	BadT
8-5	**RTs**	1,205	1,294	1,373	1,599	**RTs**	1,233	1,294	1,501	2,109
	(*N*=24)	(414)	(445)	(421)	(511)	(N=24)	(623)	(502)	(510)	(687)
	Errors	0.10	0.14	0.02	0.177	**Errors**	0.09	0.05	0.03	0.47
	(*N*=24)	(.18)	(.22)	(.07)	(.24)	(N=24)	(.14)	(.13)	(.08)	(.38)
8-30	**RTs**	598	642	1,322	1,533	**RTs**	610	628	1,071 (556)	1,818
	(*N*=24)	(123)	(122)	(561)	(873)	(N=20)	(506)	(736)		(883)
	Errors	0.03	0.02 (.07)	0.01	0.3	**Errors**	0.02	0.04	0.08	.604
	(*N*=24)	(.08)		(.15)	(.26)	(N=24)	(.07)	(.12)	(.14)	(.24)
20-2	**RTs**	1,241	1,398	1,337	1,430	**RTs**	1,839	2,047	1,619	1,964
	(*N*=24)	(437)	(646)	(406)	(432)	(N=24)	(506)	(736)	(556)	(883)
	Errors	.073	.21	.052	.094	**Errors**	.1	.052	.03	.25
	(*N*=24)	(.14)	(.25)	(.10)	(.09)	(N=24)	(.18)	(.13)	(.08)	(.21)
20-12	**RTs**	828	857	1,168	1,343	**RTs**	898	742	865	1,463
	(*N*=22)	(218)	(214)	(324)	(438)	(N=24)	(429)	(276)	(224)	(869)
	Errors	.06	.07	.06	.15	**Errors**	.05	0	.06	.42
	(*N*=22)	(.11)	(.18)	(.13)	(.18)	(N=24)	(.10)	(.00)	(.11)	(.29)

* Note. Pos stands for positive training items Neg for negative training items, GoodT for good transfer items, and BadT for bad transfer items. In the Category Type column the first number refers to the number of training items and the second to the number of presentations of each training set

< 0.0001, η_{P}^{2} = .20, and of negative-positive, $F(1, 182)$ = 108.29, $p < 0.0001$, η_{P}^{2} = .37. These effects were subsumed by two significant interactions. The first interaction was the negative-positive × category size interaction, $F(3, 182) = 6.10$, $p < 0.001$, η_{P}^{2} = .09. Exemplar use would be shown by larger differences for negative than positive items, due to the effect of BadT items. As can be seen in Fig. 3, the differences are small in Pos items but large in Neg items, especially with smaller categories.

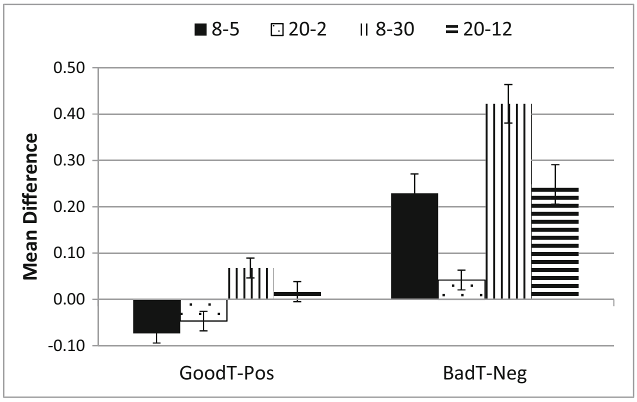

Fig. 3 Interaction between negative-positive and category type. The dependent variable is in terms of proportions. Error bars are standard errors of the mean. GoodT-Pos stands for difference between GoodT items and Positive training items. BadT-Neg stands for difference between BadT items and Negative training items

Tukey HSD tests showed that differences were larger for negative than positive items in the 8-5, 8-30, and 20-12 conditions. The same analysis also showed that the four conditions did not differ significantly for the GoodT-Pos difference. For the BadT-Neg difference, 20-2 was significantly smaller than all the other conditions, and 8-5 and 20-12 were significantly smaller than 8-30 and did not themselves differ significantly. Again, more trials with a small set of training items led to larger exemplar effects.

The second interaction, shown in Fig. 4, was between negative-positive and similarity, $F(1, 182) = 52.52$, $p < 0.0001$, η_{P}^{2} = .22. Tukey HSD tests showed that the positive and negative differences were not significantly different in the low similarity case but were in the high similarity case ($p < .05$) and that the low and higher similarity conditions did not differ significantly for the positive differences, but differed significantly in the negative difference case.

Response times

We conducted the same analyses on RTs as we performed on the errors. Following standard procedure, we analyzed RTs of correct trials only, as the time to record an incorrect response is not a measure of how long it actually took to calculate the answer. As a result, we expect to find similar patterns as with error rates. (If we had included errors, then no clear predictions could be made, as fast

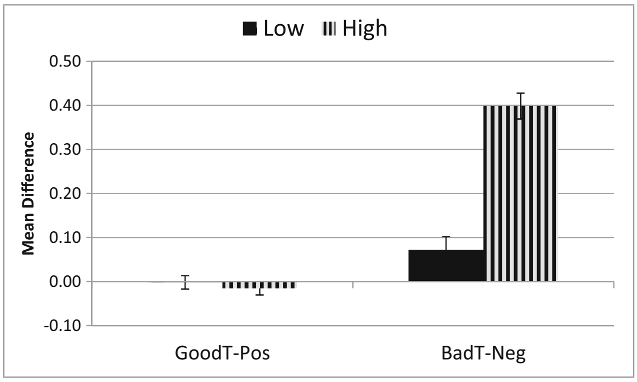

Fig. 4 Interaction between similarity and negative-positive. The dependent variable is in terms of proportions. Error bars show standard errors of the mean. GoodT-Pos stands for difference between GoodT test items and Positive training items. BadT-Neg stands for difference between BadT test items and Negative training items

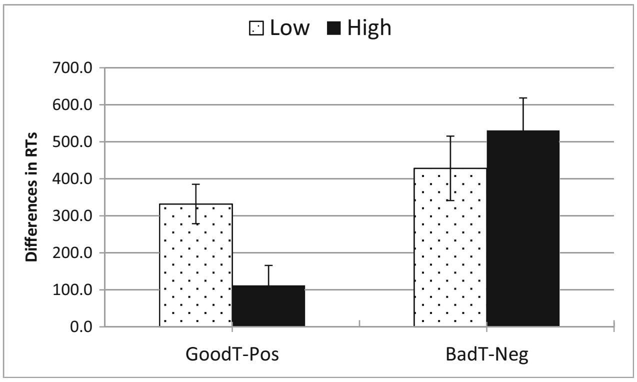

Fig. 5 Interaction between similarity and negative-positive. The dependent variable is transfer minus training response times. Error bars show standard errors of the mean. GoodT-Pos stands for difference between GoodT items and Positive training items. BadT-Neg stands for difference between BadT items and Negative training items

errors would be mixed together with slow correct answers in some conditions.) Some participants could not be included in the RT analyses due to missing data (four participants in 8-5, four in 8-30, and two in 20-12 in the high similarity condition), generally due to no correct responses in a BadT condition.

We first ran an ANOVA on the BadT-GoodT difference, with category type and similarity as between factors. Although there is an above-zero difference in all category types, suggesting an exemplar effect (see the global analysis in the Supplementary materials for more details), it did not differ significantly across category types. The ANOVA revealed the difference was significantly larger in the high similarity condition ($M = 458$ ms) than in the low similarity condition ($M = 175$ ms), $F(1, 179) = 8.10$, $p < 0.005$, $\eta^2_P = .04$. The interaction did not reach significance ($p < .5$). Thus, the exemplar effect is more pronounced when test items are similar to the training items.

As in the error analysis, the second RT analysis was a three-way ANOVA on the transfer-training differences for both positive and negative items (i.e., GoodT minus Positive training, and BadT minus Negative training), with category type and similarity as between factors and negative-positive as a within factor revealed a main effect of category type, $F(3, 179) = 14.83$, $p < 0.0001$, $\eta^2_P = .20$, and of negative-positive $F(1, 179) = 18.25$, $p < 0.0001$, $\eta^2_P = .09$. There was a significant interaction between negative-positive and similarity, $F(3, 179) = 7.14$, $p < 0.01$, $\eta^2_P = .04$. Figure 5 shows that the transfer items (GoodT vs. Pos and BadT vs. Neg) were generally classified more slowly, but the effect was greater for negative items. Tukey HSD tests showed that GoodT-Pos and BadT-Neg did not differ significantly in the low similarity case whereas BadT-Neg were significantly higher than GoodT-Pos in the high similarity case ($p < .05$).

Evidence of parallel use of rule and exemplars: Analysis of individual profiles

The analyses above show ample evidence of both rule use and exemplar influence and that the size of the exemplar effect is modulated by the factors we introduced. A question raised in the introduction in the context of mixed theories was whether participants might be using rules while also being influenced by the exemplars. One sign of this would be to find participants who had high level of accuracy together with an exemplar influence. In the following paragraphs, we present two analyses providing evidence of individuals following rules while showing evidence of exemplar influence. The first type of analysis shows that rule-following participants still made more errors on BadT than on GoodT items. The second analysis asked whether the BadT-GoodT effect in RTs is found across accuracy levels, especially in participants who made so few errors that they must have been following rules.

In the first analysis, we counted the number of participants who showed high accuracy (three out of four answers correct) in the BadT items (explainable by rule use) plus no errors in the GoodT items. The better performance in GoodT items suggests exemplar use. There were 38 people who fit this profile, suggesting usage of both exemplars and rules. Of course, with these low numbers of errors, such a difference could result from chance. However, there were only ten people who showed the reverse pattern of 1 or more GoodT errors and 0 BadT errors. This difference is significant ($X^2(1) = 16.3$, $p < .001$). Thus, it seems that there are individuals who generally followed the rules (having high accuracy even on BadT items), but still showed a GoodT-BadT difference.

In the second analysis, we asked whether individuals with high accuracy would still exhibit an exemplar influence as revealed by a significant difference between GoodT and BadT in RTs. We selected the 118 participants who were at least 75 % correct for BadT items, indicating that they had relied on the rule for their classifications. Nonetheless, these subjects were significantly slower in BadT case (M = 1,655 ms) than in GoodT trials (M = 1,331 ms), $t(117) = -6.4$, $p <$ 0.0001. The same result was obtained when we used a more stringent selection rule of 100 % correct in BadT items (suggesting perfect reliance on the rule; these participants were correct on 96 % of GoodT trials). Sixty-eight participants fit this pattern, and they also answered faster in the GoodT trials (M = 1422 ms) than in the BadT trials (M = 1685 ms), $t(67) = -$ 5.01, $p < 0.0005$. Because the RTs are only for correct trials that were presumably answered by using the rule (since using similarity would result in an error), this suggests that both rule and exemplar information were used *in the same trials*.

Following the same logic, we compared the size of the difference between BadT and GoodT items in RTs for participants who made few errors and those who had two or more errors. We ran a mixed two-way ANOVA, with accurate (at least three out of four answers correct in BadT items) versus less accurate (two, or less, correct answers in BadT items) participants as a between factor, and stimulus type (GoodT, BadT) as a within factor. Results revealed no main effect of accuracy group, $p > .20$, a significant effect of stimulus type, $F(1, 178) = 55.17$, $p < 0.00001$, $\eta^2_P = .24$ (GoodT, M = 1268 ms; BadT, M = 1647 ms) and, interestingly, no significant interaction between these two factors, $p > .20$. Indeed, the low and high accuracy groups separately had significant effects of stimulus type, $t(61) = 4.23$, $p < .001$, and $t(117) =$ 6.39, $p < .001$, respectively. These results show that no matter the participant type (accurate or making errors), the exemplar effect remained, suggesting that exemplar effects cannot be avoided even with careful rule use.

General discussion

We explored rule and exemplar influence in the paradigm designed by Brooks and colleagues in which people were informed of a perfect rule for classification. We used the BadT-GoodT effect and the difference between the training and the transfer phases as measures of exemplar influence. We also investigated whether these differences would be equivalent across various conditions manipulating the number of trials, the number of different exemplars, and the similarity between training and transfer items. Finally, we investigated whether there was evidence of simultaneous rule and exemplar influence across our conditions but also within participants.

Our results show that participants did use both types of information and did so simultaneously. Exemplar effects were most in evidence when categories were small, when learning items were repeated more and when the test items were more similar to the learning items. There were also clear indications of rule use. First, the percentage of errors was low at the end of the training phase (see Table 2). Second, there was no significant accuracy difference between old and new items at test when the items were positive (Positive training vs. GoodT) in any category type, and no significant RT differences in category types 20-2, 8-5, and 20-12. Furthermore, in none of the conditions did BadT errors go over 50 %, as would occur if similar exemplars controlled performance (as demonstrated by Allen & Brooks, 1991). The level of accuracy in all category types suggests that rules were being used a significant proportion of time.[1]

Exemplar and rule influence combined

In order to explain our results, we posit that expertise with exemplars and the rule both increase with more practice. On the exemplar side, following Logan (1988), attention towards the individual exemplars may automatically cause their encoding into memory, and each encounter with a stimulus might strengthen the corresponding exemplar's distinctiveness (see also Johansen & Palmeri, 2002; Nosofsky, 1988). On the rule side, as training proceeds, the selection of the relevant dimensions improves, and the memory for and the use of the subset of the rule-defining dimensions become more accurate (e.g., Anderson et al. 1997; Smith et al., 1998).

Our results show that there is no contradiction between improvements in rule use and exemplar encoding. When initially subjecting a viewed exemplar to a rule (as they must do at the beginning of learning), subjects presumably encode the configuration of features constituting the stimulus, including the features that are not involved in the rule (see Hahn et al., 2010). Thus, both strategies may be strengthened with this practice. By comparison with previous results (see introduction), one main contribution of our study is to show that the balance between exemplar influence and rule influence might change depending on category size, number of trials and similarity between training and transfer stimuli.

In the low similarity condition, the difference between the GoodT and BadT test items was smaller but significant. What is notable about this effect is that it operates at retrieval rather

[1] One might argue that participants are using simpler rules that do "well enough" at training but lead to errors on the BadT items (see Lacroix et al. 2005; Wills et al. 2015, for discussion). This seems possible when participants must discover the rule but is less likely when they're told the rule, which is constantly in view. Furthermore, it isn't clear what such a rule could be, given people's relatively high level of accuracy. (Almost all learning conditions have error rates of .10 or under, as reported in Table 2.) If they used only two dimensions, they would be guessing much of the time. For example, if they used only the first two dimensions instead of three, Table 1 shows that five of the eight learning items give opposite answers and so would have to be guessed on. This is incompatible with the observed accuracy levels. Analyses at the end of the Results section showed that even those with low accuracy showed exemplar effects, so their performance cannot be attributed to the use of imperfect rules.

than learning—the two similarity groups learned the same categories but differed in their test items. This suggests that even the low similarity group, which showed smaller effects of exemplar use, must have encoded the exemplars well enough to allow reminding effects.

In the high similarity condition, the similarity between the training and the transfer stimuli explains the small difference between Pos and GoodT stimuli and the significant difference between Neg and BadT items. GoodT items are classified easily because rule and similarity point to the same category. In contrast, for the BadT items, similarity and the rule lead to different categories. When training stimuli are encoded in memory as distinct exemplars, the more similar they are to their transfer twins the more this will facilitate classification of the items in the same category, and the more this will interfere with the classification of items in the opposite category (BadT items).

Going beyond previous studies relying on the same paradigm (Hahn et al., 2010; Lacroix et al. 2005: Regehr & Brooks, 1993; Thibaut & Gelaes, 2006), our study shows that the balance between exemplar and rule use also depends on category structure (eight or 20 training stimuli). In this respect, it is interesting to compare category types 8-30 and 20-12, which had the same number of learning trials. There was a larger difference between training and transfer RTs in 8-30 than in 20-12, suggesting a stronger exemplar influence in 8-30. Although they had the same number of opportunities to practice rule use, the smaller number of items in 8-30 presumably led to their being encoded more strongly, leading to exemplar retrieval as a faster route to classification than rule use.

There were also significantly more errors in 8-5 than in 20-2, which elicited few errors. This is consistent with the idea that with such a high number (20) of exemplars, each presented only twice, subjects were less able to form a strong memory representation that could then enable exemplar retrieval at test. Thus, they had to rely on the rule, leading to high accuracy. The 8-5 group, with the same amount of learning, did show clearer exemplar effects. We do not argue that there is no influence of exemplars (or rules) in some conditions. No doubt some exemplars were encoded and had an effect even in large categories. However, exemplar effects were larger and more robust when categories were small and the items repeated more often.

Another new and important contribution of our experiment was that individual participants appeared to be using both rules and exemplars, even on the same trials. Participants who were accurate in their classification, which in the case of BadT items must have been accomplished through rule use (see Allen & Brooks, 1991), nonetheless showed exemplar effects in their RT data. Slower responses on BadT items that were eventually correctly classified suggest a competition between exemplar information, providing evidence for a negative response, and rule use, indicating a positive response. Thus, even high accuracy participants who responded consistently with the rule had encoded the exemplars and were influenced by them. Although learners can exert executive control to overcome the influence of a highly similar exemplar, the exemplars still affect their performance.

When we analyzed individual participants, we found more people with no errors (and fewer with many errors) in the low similarity condition. This suggests that even though exemplars exerted their influence in both similarity conditions, exemplar effects were more important with high similarity. It is important to note that calling the low similarity “low” may be somewhat misleading. Although participants judged the similarity between training and transfer items higher in the high similarity case, even in the low similarity condition, the similarity was quite high, since training and transfer items shared four out five features. Thus, exemplar effects may drop off quickly as a function of similarity.

Previous studies focused on which characteristics of stimuli gave rise to exemplar effects. As mentioned in the introduction, the most compelling evidence was obtained with holistically individuated stimuli (see Hahn et al., 2010; Lacroix et al., 2005; Regehr & Brooks, 1993), which is why we started with such holistically individuated stimuli. If the stimuli had not been so well individuated (e.g., if each feature appeared in identical form in each stimulus), the exemplar effects would likely be less, and we suspect that we wouldn’t have found exemplar use in as many conditions. Thus, one limitation of our results is that they primarily apply to cases in which exemplars are fairly distinctive, such as different medical patients, dogs, or college essays. We cannot say whether they would also be found in less distinctive categories such as squirrels or cell phones where memories of individual items may not be strongly represented.

Rule to exemplar shift

Is there any behavioral evidence for a rule to exemplar (or an exemplar to rule) shift in the present studies? Recall that a shift should not be expected in the context of rule verification. A shift assumes that the exemplar influence requires a large amount of practice and doesn’t appear immediately. Along this line, Smith et al. (1998) and Ashby et al. (1998) claimed that the similarity procedure takes longer to become effective than the rule procedure. Johansen and Palmeri (2002) and Raijmakers et al. (2014) have described a shift from rules to exemplars resulting from supplementary practice with the categorization task (see also Pothos, 2005; Smith et al., 1998; Smith & Minda, 1998, 2000). Homa et al. (1981, 1987) made the opposite claim, that classifications become more prototype-based, i.e., more abstraction-based, with more expertise with the task. However, in their task, classification could not be accomplished by a rule, and the abstraction had to be learned.

Our results showed that exemplar effects were modulated by practice and by the number of training stimuli. The signs of exemplar use were generally larger with more practice in the learning phase (compare 8-5 to 20-2 and 8-30 to 20-12 in Fig. 3 and Fig. S2 in the Supplementary Materials). However, exemplar effects were still found at our lowest levels of exposure. The GoodT-BadT effect was present in 8-5—that is, after 40 learning trials Also, BadT items were significantly slower than Negative training items in 8-5, as well as in the conditions with more practice. These results suggest that exemplars exert their influence with a small number of presentations even though there is still room for performance improvement, as suggested by the vast decrease in RTs in the training phase between the 8-5 and 8-30 conditions. Interestingly, the only difference between Positive training items and GoodT items was obtained in the 8-30 case (see Supplementary Materials B for evidence). This suggests that extremely fast exemplar access requires a large number of presentations of a small number of exemplars. In sum, even with this very predictable rule, exemplar similarity could exert its influence, which indicates that participants' selective attention never became optimal (see Rehder & Hoffman, 2005).

Hybrid models of categorization

A number of hybrid models have been proposed in recent years. However, they were meant to deal with experimental situations that differ broadly from the one displayed here, where no rule is provided at the onset of the experiment (e.g., Smith & Minda, 1998). On the one hand, some models posit that the same module performs both types of computations. For example, in ACT-R (Anderson & Betz, 2001), PRAS (Vandierendonck, 1995), and SUSTAIN (Love et al., 2004), rules and exemplars are combined in the same representational system. On the other hand, in ATRIUM (Erickson & Kruschke, 1998, 2002), rule-based and exemplar-based information can be stored in distinct modules that compete. As discussed by Johansen and Palmeri (2002), a number of different architectures are compatible with the same set of results. Our design was not built to test these models. What our results show is that more experience with the rule (more training trials) and higher similarity between training and test items generally meant more exemplar influence, and that the most accurate participants still displayed exemplar influence in both accuracy and RTs.

Our results generally fit with Smith et al.'s (1998) description of categorization, linking rule-based categorization with analytic and strategic processing, differential weighting of a small subset of relevant attributes, and linking similarity-based behavior with holistic and automatic processing, equal weighting of attributes, and matching concrete information (Hahn & Chater, 1998; Norenzayan et al., 2002; Pothos, 2005; Regehr & Brooks, 1993). We have shown that these two procedures are used in parallel but that depending on the structure of the stimuli (e.g., the training-transfer similarity, stimulus distinctiveness) and the category (e.g., category size), the influence of each component might differ. However, we did not find that even highly accurate participants could "turn off" the exemplar route to classification, so any model claiming that only the more accurate path would be used would have difficulty accounting for these results. When a rule is available, other information can intrude itself into the classification decision. However, this intrusion depends on the structure of the training and the test stimuli.

Acknowledgements This research was supported by a grant from the French Community of Belgium Ministry, Convention 99/04-242, a grant from the University de Liège, "Fonds Spéciaux de Recherche," and a grant from the Conseil Régional de Bourgogne (Pari) to the first author.

We thank Maria Chuy, Anne-Sophie Detiffe, and Gwenaëlle Melchiore for their help in running the experiment.

We thank two anonymous reviewers for their comments and thoughtful suggestions.

References

Allen, S. W., & Brooks, L. R. (1991). Specializing the operation of an explicit rule. *Journal of Experimental Psychology: General*, *120*, 3–19.

Anderson, J. R. (1993). *Rules of the mind*. Mahwah: Erlbaum.

Anderson, J. R., & Betz, J. (2001). A hybrid model of categorization. *Psychonomic Bulletin & Review*, *8*, 629–647.

Anderson, J. R., Fincham, J. M., & Douglass, S. (1997). The role of exemplars and rules in the acquisition of a cognitive skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 932–945.

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

Bourne, L. E. (1970). Knowing and using concepts. *Psychological Review*, *77*, 546–556.

Brooks, L. R., Norman, G. R., & Allen, S. W. (1991). Role of specific similarity in a medical diagnostic task. *Journal of Experimental Psychology: General*, *120*, 278–287.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking.* Oxford : Wiley.

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

Erickson, M. A., & Kruschke, J. K. (2002). Rule-based extrapolation in perceptual categorization. *Psychonomic Bulletin and Review*, *9*, 160–168.

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

Estes, W. K. (1994). *Classification and cognition*, London: Oxford University Press.

Hahn, U., & Chater, N. (1998). Similarity and rules: Distinct? Exhaustive? Empirically distinguishable. *Cognition*, *65*, 197–230.

Hahn, U., Prat-Sala, M., Pothos, E. M., & Brumby, D. P. (2010). Exemplar similarity and rule application. *Cognition*, *114*, 1–18.

Homa, D., & Vosburgh, R. (1976). Category breadth and the abstraction of prototypical information. *Journal of Experimental Psychology: Human Learning and Memory*, *2*, 322–330.

Homa, D., Sterling, S., & Trepel, L. (1981). Limitations of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Human Learning and Memory*, *7*, 418–439.

Homa, D., Burruel, L., & Field, D. (1987). The changing composition of abstracted categories under manipulations of decisional change, choice difficulty, and category size. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*, 401–412.

Johansen, M. K., & Palmeri, T. J. (2002). Are there representational shifts during category learning? *Cognitive Psychology*, *45*, 482–553.

Lacroix, G. L., Giguère, G., & Larochelle, S. (2005). The origin of exemplar effects in rule-driven categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 272–288.

Lamberts, K., Brockdorff, N., & Heit, E. (2003). Feature-sampling and random-walk models of individual-stimulus recognition. *Journal of Experimental Psychology: General*, *132*, 351.

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

Logan, G. D., & Etherton, J. L. (1994). What is learned during automatization? The role of attention in constructing an instance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 1022–1050.

Love, B. C., Medin, D. L., & Gureckis T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *11*, 309–332.

Medin, D. L., & Ross, B.H. (1989). The specific character of abstract thought: Categorization, problem solving, and induction. In R. J. Sternberg (Ed.), *Advance in the psychology of human intelligence* (pp. 189–223). Hillsdale: Lawrence Erlbaum Associates.

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

Murphy, G. L., Bosch, D. A., & Kim, S-W. (2017). Do Americans have a preference for rule-based classification? *Cognitive Science, 41*, 2026–2052. https://doi.org/10.1111/cogs.12463

Norenzayan, A., Smith, E. E., Kim, B. J., & Nisbett, R. E. (2002). Cultural preferences for formal versus intuitive reasoning. *Cognitive Science*, *26*, 653–684.

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

Nosofsky, R. M. (1988). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 54–65.

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

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*, 53–79.

Nosofsky, R. M., & Zaki, S. R. (2002). Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 924–940.

Pinker, S. (1999). *Words and rules: The ingredients of language*. NY: Basic Books.

Pothos, E. M. (2005). The rules versus similarity distinction. *Behavioral and Brain Sciences*, *28*, 1–14.

Raijmakers, M. E., Schmittmann, V. D., & Visser, I. (2014). Costs and benefits of automatization in category learning of ill-defined rules. *Cognitive Psychology*, *69*, 1–24.

Regehr, G., & Brooks, L. R. (1993). Perceptual manifestations of an analytic structure: The priority of holistic individuation. *Journal of Experimental Psychology: General*, *122*, 92–114.

Rehder, B., & Hoffman, A. B. (2005). Eyetracking and selective attention in category learning. *Cognitive psychology*, *51*, 1–41.

Rips, L. J. (1989). Similarity, typicality, and categorization. In S. Vosnadiou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 21–59). New York: Cambridge University Press.

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

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*, 3–22.

Smith, E. E., Patalano, A. L., & Jonides, J. (1998). Alternative strategies of categorization. *Cognition*, *65*, 167–196.

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

Smith, J. D., & Minda, J. P. (2000). Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 3–27.

Smith, J. D., Murray, M. J., Jr., & Minda, J. P. (1997). Straight talk about linear separability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 659–680.

Thibaut, J.P., Dupont, M., & Anselme, P. (2002). Dissociations between categorization and similarity judgments as a result of learning feature distributions. *Memory & Cognition*, *30*, 647–656.

Thibaut, J. P., & Gelaes, S. (2006). Exemplar effects in the context of a categorization rule: Featural and holistic influences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 1403–1415.

Vandierendonck, A. (1995). A parallel rule activation and rule synthesis model for generalization in category learning. *Psychonomic Bulletin & Review*, *2*, 442–459.

Ward, T. B., & Scott, J. (1987). Analytic and holistic modes of learning family-resemblance concepts. *Memory & Cognition*, *15*, 42–54.

Wills, A. J., Inkster, A. B., & Milton, F. (2015). Combination or differentiation? Two theories of processing order in classification. *Cognitive Psychology*, *80*, 1–33.