Rules and Exemplars in Categorization: A Computational Exploration

Duncan P. Brumby (Brumby@cs.drexel.edu)
Department of Computer Science, Drexel University
Philadelphia, PA 19104 USA

Ulrike Hahn (HahnU@cardiff.ac.uk)
School of Psychology, Cardiff University
Cardiff, CF10 3AT UK

Abstract

Studies have found that human categorization judgments are affected by exemplar similarity, even when a simple, perfectly predictive rule is provided and paying attention to instance similarity is harmful to performance. These data provide an interesting challenge for recent hybrid rule-plus-exemplar models of category learning. We report the results of a modeling effort with a pre-existing hybrid model developed in the ACT-R cognitive architecture. A search of the model's parameter space revealed that increasing use of an exemplar route improved the fit of the model to the data, because it resulted in faster categorization judgments for high-similarity items compared to low-similarity items. However, use of the exemplar route carried no adaptive value for the model because it necessarily lead to more categorization errors than simply basing judgments on the categorization rule alone. The fact that people's categorization judgments juxtapose rule application with instance-similarity while maintaining very low error rates presents a non-trivial problem for current hybrid models of category learning.

Keywords: categorization; rules; exemplars; similarity; computational modeling; ACT-R

Introduction

Over the last two decades, categorization research has seen a steady rise in interest in hybrid accounts; particularly rule-plus-exemplar accounts that assume human categorization judgments are formed through some mix of exemplar- and rule-based processes (e.g., Erickson & Kruschke, 1998; Palmeri, 1997; for a different hybrid approach, see Ashby et al, 1998). This interest, for which there is good theoretical reason (e.g., Hahn & Chater, 1998), has fueled experimental tests as well as a range of computational models of varying scope and specificity.

Allen and Brooks (1991) provided a seminal experimental demonstration of the joint effects of rules and exemplars in categorization. Participants in the study were given a simple rule to classify both old and novel items. Even though the rule was perfectively predictive there was evidence for systematic effects of exemplar similarity on categorization.

Allen and Brooks' results are somewhat less surprising when one takes into account the specific nature of the rule used in the study. Specifically, the rule described an m-of-n concept ("an object is a digger if it has at least 3 of the following 5 features…"). This type of rule description is functionally equivalent to a prototype-plus-similarity threshold account. Consequently, for Allen and Brooks' materials similarity is correlated with the rule's applicability. Attending to similarity is therefore adaptive in this task in a way it need not be in general.

In response to the critique of the rule description used in Allen and Brooks' (1991) experiments, Hahn, Prat-Sala and Pothos (2002) sought to test whether exemplar similarity effects would arise in a rule-based task in which category membership was entirely uncorrelated with similarity. Hahn et al. found effects of exemplar similarity on error patterns and reaction times, even under conditions where attending to similarity interfered with performance on the rule application task. At the same time they also found very low error rates, which suggests that the rule was in fact used.

Hahn et al.'s findings are of interest because hybrid rule-plus-exemplar models of categorization would generally predict that (1) categorization errors should be associated with exemplar similarity effects and (2) any reduction in error rates should be associated with diminished similarity effects. That people's categorization judgments juxtapose rule application with instance-similarity, while maintaining very low error rates seems at odds with the basic predictions that can be derived from these models of categorization.

In this paper, we first describe the data from two experiments (Hahn et al., 2002; Hahn et al., submitted) that investigate the effect of similarity on the application of a simple, perfectively predictive rule. These data suggest that combining rule application with instance-similarity occurs even under conditions where paying attention to instance similarity is harmful to performance. A reimplementation of Anderson and Betz's (2001) ACT-R model follows, along with a computational exploration of the parameter space of the model, in order to find the best-fitting model for the first data set. Based on these best-fitting parameter values, a comparison between the performance predictions of the model and the second data set is presented.

Empirical Data

Hahn et al. (2002), in an experiment which we will refer to as Experiment 1, constructed a set of items governed by a simple, perfectly predictive rule that specified three necessary and sufficient features for category membership (e.g., "is an A if it has an upside-down triangle at the sides, a cross in the centre, and a curly line at the top"). Participants were told this rule at the beginning of the experiment, and were then given a series of positive exemplars as illustration. At test, participants were given 96 novel items, distributed over four blocks. Participants did not receive feedback regarding the accuracy of their categorization judgments. Half of the test items complied with

the rule, and half violated it. At the same time, half of the items were high in similarity to one of the initial training exemplars (as determined by the amount of overlap in the non-rule features of the objects) and half were low in similarity to the training exemplars. Manipulations of similarity were orthogonal to category membership, such that exactly half the rule-compliant items were high in similarity to the training items and half were not, and likewise for the non-compliant items. In other words, while using the rule would lead to perfect performance, basing categorization on exemplar similarity would lead to chance performance.

The crucial question addressed by Experiment 1 was whether effects of exemplar similarity would arise even under circumstances where attending to similarity had no adaptive value. Analyzing data from 42 participants, significant effects of exemplar similarity were found on error rates and reaction time (RT). From a total of (96x42) 4032 responses, only 7.56% were errors (where the participant pressed the YES key when the NO key was expected or the other way round). In addition, there were significantly more errors on low-similarity compliant items than on high-similarity compliant items (61 vs. 46, respectively). However, there was no significant difference between response errors to low- and high-similarity non-compliant items. Analysis of reaction time data only considered responses to compliant items across each block in the experiment, excluding all trials where an incorrect response was made. Responses for high-similarity compliant items were found to be significantly faster than responses for low-similarity compliant items (1005 ms vs. 1070 ms, respectively). Average RT also speed-up over consecutive blocks of trials (1221 ms, 1048 ms, 955 ms, 920 ms, respectively). Importantly, there was not a significant interaction between similarity and block, suggesting that the effect of exemplar-similarity did not diminish with practice.

Participants did not receive any feedback on the accuracy of their categorization judgments in Experiment 1. Feedback was introduced in a follow-up study, which we shall refer to as Experiment 2 (for full details, see Hahn, Prat-Sala, Pothos, & Brumby, submitted). If participants in this study responded using the correct key, then the message "CORRECT!" appeared on the screen. If they responded using the incorrect key, then the message "WRONG!" appeared on the screen and a short beep alerted the participant to the mistake. In all other respect the two experiments were identical.

As expected the inclusion of feedback in Experiment 2 reduced the total number of categorization errors: From a total of (96x40) 3840, 4.87% were errors. As before, there were significantly more errors on low-similarity compliant items than on high-similarity compliant items (36 vs. 21, respectively). There was no significant difference between response errors to low- and high-similarity non-compliant items (60 vs. 70, respectively). Moreover, the decrease in overall error rates in the second experiment carried an associated time cost: Average RTs for Experiment 2 were elevated in comparison to Experiment 1 (1347 ms vs. 1036 ms, respectively). Regardless of this increase, the overall pattern for RT data was robust across experiments: Responses for high-similarity compliant items were significantly faster than responses for low-similarity compliant items (1290 ms vs. 1405 ms, respectively), and RT also decreased over consecutive blocks of trials (1776 ms, 1314 ms, 1205 ms, 1095 ms, respectively). There was no interaction between similarity and block.

In summary, then, these data provide robust evidence of exemplar similarity effects, even under conditions where attending to similarity *interferes* with performance on the rule application task. At the same time, the consistently low error rates demonstrate that the rule *was* used. We next describe a hybrid model of categorization and derive predictions for these data.

Hybrid Model of Categorization

Current hybrid models differ substantially in the way rules and exemplars are related. The first set of models assumes two independent routes; a route is chosen on a trial-by-trial basis depending on a number of factors such as simplicity or reliability (e.g., Anderson & Betz, 2001; Ashby et al, 1998). The second set of models assumes a parallel competition or race between the two components (e.g., Palmeri, 1997), with the fastest route governing the response. The third set again assumes that both routes operate in parallel and that their respective outputs are blended into an overall response (e.g., Erickson & Kruschke, 1998).

At present the model that is most explicitly defined in all respects is that of Anderson and Betz (2001); this computational model fully implements the Exemplar-Based Random Walk (EBRW) model of Nosofsky and Palmeri (1997) and the rule-plus-exception (RULEX) model of Nosofsky, Palmeri, and McKinley (1994) in the ACT-R cognitive architecture (Anderson et al., 2004). ACT-R is a cognitive architecture that consists of multiple modules that are integrated through a central production system to simulate cognition. As a first step, we reimplemented Anderson and Betz's original model to run in the most recent version of the ACT-R software (ACT-R 6, Anderson et al., 2004). A general strength of this modeling approach is that the model of category learning is embedded within a broader theory of human memory and perceptual/motor processing. As a consequence, predictions of both categorization judgments and reaction times are simulated.

For the training phase, the category rule and the training items were each stored in exemplar (or declarative) memory. On the subsequent test trials, the model could choose on a trial-by-trial basis whether to use the rule- or the exemplar-route. The choice between routes is determined by a route's utility. Anderson and Betz (2001) define this as a simple trade-off function between the probability P that the route would be expected to lead to a correct judgment and the expected time C required to reach that judgment Specifically, the utility U of a route is defined as,

$$U_i = P_i G - C_i + \varepsilon \quad (1)$$

where G is a constant that reflects the value of the objective (which can be thought of as a maximum time investment to complete the goal). Utility estimates are also stochastic, with the addition of transient noise ε. On each trial, the route with the greatest utility is selected.

A route's utility estimate is updated following usage. The probability *P* that a route would be expected to lead to a correct judgment is defined as follows,

$$P = \frac{successes}{successes + failures} \quad (2)$$

where *successes* is a count representing the frequency of positively rewarded responses attributed to the route and *failures* is a count of negatively rewarded responses. Initially, both rule- and exemplar-route were equally likely to be chosen. It is worth noting that because participants in Experiment 1 did not receive explicit feedback regarding the accuracy of categorization judgments, the probability *P* of a route increased at a constant rate after each successive trial that it was selected (i.e., *P'* = (*successes* + 1) / (*successes* + *failures* + 1)). In contrast, participants in Experiment 2 received feedback for the accuracy of responses; therefore, the probability *P* of a route decreased if an incorrect response was made (i.e., *P'* = *successes* / (*successes* + *failures* + 1)). This was the only difference between models.

In the model, as in ACT-R generally, declarative knowledge is represented as *chunks*. The activation *A* of chunk *i* is defined as,

$$A_i = B_i + \sum_k \sum_j W_{kj} S_{ji} + \sum_l PM_{li} + \varepsilon \quad (3)$$

which represents a summation over the base-level activation of the chunk, spreading activation, a partial matching score, and transient noise, respectively. Both the rule route and exemplar route relied on this activation-based account of declarative memory. We next provide a detailed description of each route; specifically, how the definition of chunk activation was used to judge whether or not a test item was compliant with the categorization rule. by a route.

Rule Route

The rule-based route implemented Nosofsky, Palmeri, and McKinley's (1994) rule-plus-exception (RULEX) model. The rule route determines category membership through the retrieval of a declarative memory representation of the categorization rule, which is then systematically compared to the test item. The latency and probability of the rule's retrieval is determined by its activation in memory (Eq. 3). The spreading activation and partial matching components of the equation did not play a functional role in the rule route; that is, though implemented, these processes do not contribute to the routes performance. On a small number of trials rule retrieval will fail, and a random (guessing) response is made. When retrieved, the rule is held in working memory and each of the rule feature values are iteratively compared to the feature values of the current test item. This differs from Anderson and Betz's (2001) original model, where all features were exhaustively compared. This change was forced by evidence that rule-complexity (i.e., the number of rule-features) affects reaction time (Hahn et al., 2002); thus, only rule relevant features were evaluated.

Exemplar Route

The exemplar-based route implemented Nosofsky and Palmeri's (1997) Exemplar-Based Random Walk (EBRW) model. The exemplar route determines category membership by recalling declarative memory representations of rule-compliant items leant at training. The latency and probability of retrieving an item is determined by its activation (Eq. 3), which is a summation of the chunks base-level activation, a partial matching score, and a transient noise. (As before, spreading activation did not play a functional role in determining chunk activation.) We unpack each in turn.

First, the partial matching score provided a definition for the degree of similarity between the current test item and training exemplars in memory. The matching score is a sum computed over all six dimensions of the object. The match scale *P* reflects the amount of weight given to a dimension; by default this is a constant across all dimensions. The match similarities M_{li} determine the similarity between the feature in the retrieval specification and the corresponding dimension of exemplars in memory. Matches received a value of 1.0 and mismatches received a value of -1.0, so that partial matching scores varied between 6.0 and -6.0. The net result of this is that a training exemplar becomes more likely to be retrieved, as the similarity between it and the current test item increases.

Second, each time an exemplar is retrieved from memory, it receives a temporary boost in base-level activation. Over the course of the experiment, exemplars that are frequently retrieved have their declarative memory representation further strengthened. These increases in base-level activation due to frequency and recency of use mean that an exemplar is more likely to be retrieved in the future, and in less time.

Finally, one also needs to define how similarities between a test item and training exemplars are translated into a category decision. Anderson and Betz's (2001) model, like the EBRW, made use of a random walk procedure. The random walk procedure makes repeated attempts to retrieve exemplars from memory. The successful retrieval of an exemplar provides positive evidence that the test item is compliant with the rule. A retrieval failure provides negative evidence to the contrary. Whether or not a training exemplar is retrieved is determined by its activation exceeding a retrieval threshold. A random walk threshold determines when 'enough' evidence is accumulated to make a decision. We explored the effect of varying the random walk threshold. The random walk threshold affected neither the actual decision nor the relative differences in reaction time between high- and low-similarity items. This is because the definition of chunk activation (Eq. 3) is itself already sensitive to similarity. Consequently, the reported model fits are based on a single step threshold, where category decisions are made on the basis of evidence from a single retrieval: If a training exemplar is retrieved in the context of a test item, then the model makes a positive response, indicating that the test item is compliant with the categorization rule. Whereas, if none of the training exemplars are retrieved (i.e., because the activation values of the training exemplars in declarative memory are less than the retrieval threshold), then a negative response is made, indicating that the test item is not compliant with the categorization rule.

Table 1: Comparison between main effects in the human data and those predicted by the model over different parameter values for Experiment 1. 'X' represents points in the parameter space where effects were consistent between model and data. Highlighted cells indicate best fitting model parameters (see test for details).

A) Main effects for error rates.

p(Rule)	Similarity effect c-items					*n.s.* effect n-items				
	Retrieval Threshold					Retrieval Threshold				
	-0.5	0.0	0.5	1.0	1.5	-0.5	0.0	0.5	1.0	1.5
0.0	-	-	X	X	X	X	X	-	-	-
0.1	-	-	X	X	X	X	-	-	-	-
0.2	-	-	X	X	X	-	-	-	-	-
0.3	-	-	X	X	X	X	X	-	-	-
0.4	-	-	X	X	X	X	X	X	-	-
0.5	-	-	X	X	X	X	X	-	-	-
0.6	-	-	X	X	X	X	X	-	-	-
0.7	-	X	X	X	X	X	X	-	-	-
0.8	-	-	X	X	X	X	X	-	-	-
0.9	-	X	X	X	X	X	X	X	-	-
1.0	-	-	-	-	-	X	X	X	X	X

B) Main effects for reaction time.

p(Rule)	Similarity effect					Block effect				
	Retrieval Threshold					Retrieval Threshold				
	-0.5	0.0	0.5	1.0	1.5	-0.5	0.0	0.5	1.0	1.5
0.0	X	X	X	X	-	X	X	X	X	-
0.1	X	X	X	X	-	X	X	X	-	X
0.2	X	X	X	X	X	X	X	X	-	-
0.3	X	X	X	X	-	X	X	X	X	-
0.4	X	X	X	X	-	X	X	X	X	-
0.5	X	X	X	X	X	X	X	X	-	-
0.6	X	X	X	X	-	X	X	X	-	-
0.7	X	X	X	X	X	X	X	X	-	-
0.8	X	X	X	X	X	X	X	X	-	-
0.9	X	X	X	X	X	X	X	X	X	-
1.0	-	-	-	-	-	X	X	X	X	X

Model Evaluation

The model initially stored the categorization rule in declarative memory, and was then presented with various training exemplars, which too, were added to memory. Model performance was then evaluated on subsequent test items. Following categorization, test items were not added to memory.

ACT-R makes theoretical commitments about the amount of time it takes to encoding a stimuli item. It is assumed that visual encoding entails a number of basic processes, which were represented as production rules. An initial observation is that while ACT-R provided timing estimates for these encoding processes, it was apparent that these estimates were massively greater than the RTs found in the empirical data. Specifically, in ACT-R the visual encoding of each feature of an item takes 185 ms (representing a 50 ms cognitive cycle to initiate perception and 135 ms for a shift of visual attention). Given that the exemplar route requires that all six features of an item are encoded, the model predicts that the encoding of an item should take 1,100 ms. In contrast, the empirical data show that participants RT (which not only included visual encoding, but also decision and response processes) was approximately 1,000 ms on average. The only way to account for the human RT data therefore, is to assume that stimuli features are encoded more rapidly than predicted by ACT-R's theory of visual attention. Consequently, we assume a constant time for the visual encoding all six of the stimuli features of 555 ms.

To compare model and experimental results, we simulated a population of 'model participants'. This approach was necessary because the model's behavior is stochastic. In particular, the error data could not be fit in any other way because the relevant quantity of interest was the total frequency of categorization errors made. The model was rerun over the experimental procedure, with each model run representing a single participant in the experiment. This meant that the error distributions could be fit to the data.

One important question was whether the model could fit both error data and RT data. Both error rates and RT differences are directly related to the core theoretical assumptions of the model, in that, given the nature of the test items, (systematic) errors only arise through the use of the exemplar route, as do the RT differences between high- and low-similarity exemplars. Correcting an excessive number of total errors means that the exemplar route has to have been used on proportionally fewer trials; however, reducing the relative usage of the exemplar route necessarily reduces any effect of similarity on RT data. It is clear that these two aspects of the data might not be trivial to satisfy. In addition, the exemplar route's retrieval threshold influences the distribution of errors over high- and low-similarity compliant and non-compliant items. If the retrieval threshold is very low, then a match will always be found and the exemplar route will be biased toward "yes" responses; if the retrieval threshold is very high, the exemplar route will be biased toward "no" responses. Only somewhere in between will systematic differences between high- and low-similarity items emerge.

Model Results for Experiment 1

We conducted a systematic exploration of the models performance across different proportions of rule route usage and varying retrieval thresholds. We factorially combined 5 possible retrieval threshold values (-0.5, 0, 0.5, 1, & 1.5) with increments of .1 in the probability of rule use within the range from 0 to 1. For each combination we ran 42 'model participants'. At each of these points in the parameter space,

model predictions for error and RT data were statistically evaluated and compared to the human data. Recall that the main empirical findings from the human data were: (1) Significantly fewer errors for high-similarity compliant items than low-similarity compliant items, (2) no effect of similarity on errors for non-compliant items, (3) significantly faster RTs for high-similarity compliant items than low-similarity items, and (4) a significant speed-up in RT over successive trial blocks. Table 1 summarizes comparisons between these modeling results and the human data, where "X" entries signal a match between the corresponding statistical tests.

Several things are apparent from Table 1. First, RT patterns are easier to capture than the error patterns, as can be seen in the greater number of "X" cells in the RT panels of the table (panel A vs. B). Second, the trade-off between capturing error and RT patterns goes beyond that intuitively described above. Scanning the table, one sees that the number of parameter combinations that successfully capture the error patterns increases as one moves right in the table, whereas the opposite is true for the RT patterns. In other words, the pattern of errors and RTs place conflicting demands on the retrieval threshold parameter. This suggests that collecting both error and RT data provide a far more stringent test of the model, than either kind of data alone. Finally, and most importantly, there are a number of cells where all four behavioral criteria are satisfied (i.e., cells where there is a "X" entry in all four panels). In other words, there are multiple parameter combinations that reproduce the key qualitative aspects of the participant data. In order to discriminate among these different possibilities, we consider the overall number of error responses made by the model at various points in the parameter space. This additional factor heavily constrains the model. We found that an exemplar retrieval threshold of 0.5 and probability of rule use of 0.9 gave model performance that not only satisfied the main qualitative aspects of the data, but also matched the data quantitatively as well. This is because a high probability of rule use guarantees that the overall error rate remains low, but allows the exemplar route to be selected often enough to give an overall effect of exemplar-similarity on RT.

We provide detailed analysis of the model's performance using the best fitting parameter values. The model made fewer errors for high-similarity compliant items than low-similarity items, $t\,(41) = 3.27$, $p < 0.005$, at a rate comparable to the human data (i.e., 54 vs. 97 for the model compared to 41 vs. 61 for data). For the non-compliant items the model did not predict a difference in errors for high- and low-similarity items, $t\,(41) = 1.08$, $p = 0.29$; again these error rates were comparable to the human data (127 vs. 107 for model compared to 92 vs. 107 for data). Figure 1 shows the RT fits for the model compared to data. Although the absolute magnitude of speed-up in RT over block is under-predicted, the model demonstrates a reliable effect of block, $F\,(3, 123) = 57.61$, $p < 0.001$. Model responses for high-similarity compliant items were significantly faster than responses to low-similarity complaint items, $F\,(1, 41) = 42.38$, $p < 0.001$. Moreover, the interaction was non-significant, $F\,(3, 123) = 2.05$, $p = 0.11$.

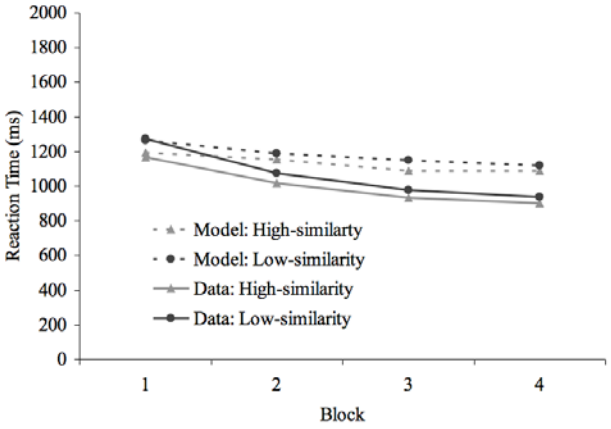

Figure 1. Data and model fits of reaction time across similarity manipulations and trial block for Experiment 1.

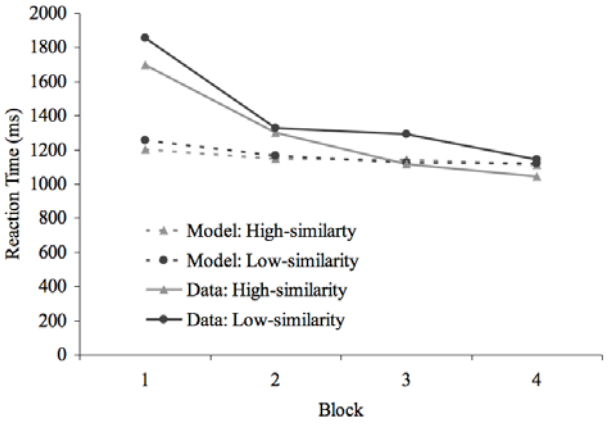

Figure 2. Data and model predictions of reaction time across similarity manipulations and trial block for Experiment 2.

Model Results for Experiment 2

Given the best-fitting parameters for the first data set, we ran the model on a supervised version of the experimental materials where feedback was given about the accuracy of each categorization judgment. For each trial a route was selected, and if it resulted in a correct judgment, then the probability that the route was selected on future trials was increased. Whereas, if a route resulted in an incorrect judgment, then the probability that it was selected on a future trial was reduced (see Eq. 2).

Figure 2 shows the RT predictions for the model compared to the human data. It is clear that in contrast to the human data, the model does not predict the presence of instance-similarity effects on RT, $F\,(1, 39) = 3.70$, $p = 0.06$. However, the model does predict a significant similarity x block interaction, $F\,(3, 117) = 3.80$, $p < 0.05$. That is, the model started out predicting a reliable effect of similarity on RT, but only for the first block of trials — for all subsequent blocks there were no predicted differences in RT across similarity manipulations.

The model demonstrated a reliable speed-up effect over successive blocks of trials, $F\,(3, 123) = 43.97$, $p < 0.001$; however, as before the absolute magnitude of this speed-up was under predicted. This is particularly pertinent in the

current data set because the human data shows elevated RTs compared to Experiment 1. The model did not predict this increase in RT, which presumably reflects changes to the participant's speed/accuracy trade-off in the context of explicit feedback.

Finally, the model's predictions for the frequency of error rates across different conditions for Experiment 2 was also inconsistent with the empirical data. The model predicted more errors than were observed in the empirical data (283 vs. 187). Furthermore, the model predicted an effect of similarity for non-compliant items (92 vs. 70, for high- and low-similarity conditions), $t(39) = 2.04$, $p < 0.05$, but no such effect for compliant items (62 vs. 59, for high- and low-similarity conditions), $t(39) = 0.26$, $p = 0.80$. These predictions are inconsistent with the human data, which found effects of similarity on error rates for compliant but not for non-compliant items.

General Discussion

The juxtaposition of low error rates and similarity effects reported in Hahn et al.'s (2002) study sets an interesting benchmark for hybrid theories of categorization (whether that be accounts assuming independent routes, a parallel competition between routes, or a blending of routes) because sensitivity to exemplar-similarity in this task should necessarily result in categorization errors. In this paper we focused on Anderson and Betz's (2001) hybrid model of categorization. This model was chosen for evaluation because it is currently one of the most explicitly defined computational models in the categorization literature, and it is fully implemented within a general framework of the human cognitive architecture. The model was constrained by theoretical commitments about the cognitive architecture and used largely default parameter values. A search of the model's parameter space revealed that the fit between the performance of the model and the empirical data was improved by increasing the use of an exemplar-based route, relative to a rule-based route. This is because use of the exemplar route resulted in faster categorization judgments for high-similarity items compared to low-similarity items; however, its use carried no adaptive value because it necessarily lead to more errors than simply basing judgments on the categorization rule alone. That Anderson and Betz's model gave qualitative as well as quantitative fits with the empirical data was somewhat surprising and demonstrates that the model is robust enough to capture data that seems intuitively outside of its range of behavior.

However, there were at least two important limitations of Anderson and Betz's (2001) model. First, while ACT-R provided timing estimates for the visual encoding of stimuli features, it was apparent that these estimates were much greater than those observed in the human data. Throughout, we assumed a constant time for the visual encoding of 555 ms. We speculate that rule-relevant features are directly encoded based on evidence that reaction time is strongly associated with rule-complexity (see, Hahn et al., 2002). However, exemplar similarity can only be determined when all of the stimuli features are encoded; therefore, we tentatively propose that participants in Hahn et al's experiment may have encoded rule-irrelevant features parafoveally at the same time that rule-relevant features were encoded; thus, incurring no additional time cost. Eye-tracking data would be useful to evaluate this proposal.

Providing feedback about the accuracy of categorization judgments (Experiment 2) revealed a critical weakness of Anderson and Betz's (2001) model. The model predicted that similarity effects should diminish over time as feedback demonstrates that paying attention to exemplar-similarity is harmful to performance. The reasons why the model predicts that the effect of instance-similarity diminishes over successive trials is quite clear: Exemplar-similarity effects are brought about through the use of the exemplar route. However, because the exemplar route leads to frequent categorization errors, its utility is strategically lowered, which results in it being chosen less frequently. In contrast, the rule route, which does not convey any effect of instance-similarity, has its utility strategically increased following use because its use rarely leads to an incorrect judgment. This strategic account of choosing between routes was not supported by the empirical data. In fact, the empirical data suggests that the coupling of rule application with instance similarity might be mandatory in the formation of human categorization judgments. We speculate that these data would also be problematic for other hybrid models of categorization in the literature.

Acknowledgments

This work was supported by European Commission grant 51652 (NEST).

References

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

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

Anderson, J.R., Bothell, D., Byrne, M.D., Douglass, S., Lebiere, C., and Qin, Y. (2004). An integrated theory of mind. *Psychological Review, 111*, 1036-1060.

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

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

Hahn, U., Prat-Sala, M., Pothos, E. & Brumby, D. (submitted). How Exemplar Similarity Influences Rule Application.

Hahn, U and Chater, N. (1998). Similarity and Rules: Distinct? Exhaustive? Empirically Distinguishable? *Cognition, 65*, 197-203

Hahn, U., Prat-Sala, M., & Pothos, E. (2002). How similarity affects the ease of rule application. In *Proceedings of the 24th Annual Meeting of the Cognitive Science Society.* Mahwah, NJ: Erlbaum.

Nosofsky, R.M., & Palmeri, T.J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review, 101*, 53-79.

Nosofsky, R.M., Palmeri, T.J., & McKinley, S. (1994). Rule-plus-exception model of classification learning. *Psychological Review, 104*, 266-300.

Palmeri, T.J. (1997). Exemplar similarity and the development of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 324-354.