Stages of Category Learning in Monkeys (*Macaca mulatta*) and Humans (*Homo sapiens*)

J. David Smith, **William P. Chapman**, and **Joshua S. Redford**
Department of Psychology, University at Buffalo, the State University of New York.

Abstract

Smith and Minda (1998) and Blair and Homa (2001) studied the time course of category learning in humans. They distinguished an early, abstraction-based stage of category learning from a later stage that incorporated a capacity for categorizing exceptional category members. The present authors asked whether similar processing stages characterize the category learning of nonhuman primates. Humans and monkeys participated in category-learning tasks that extended Blair and Homa's paradigm comparatively. Early in learning, both species improved on typical items more than on exception items, indicating an initial mastery of the categories' general structure. Later in learning, both species selectively improved their exception-item performance, indicating exception-item resolution or exemplar memorization. An initial stage of abstraction-based category learning may characterize categorization across a substantial range of the order *Primates*. This default strategy may have an adaptive resonance with the family-resemblance organization of many natural-kind categories.

Keywords

category learning; prototypes; stages of learning; abstraction; primate cognition

---

Learning and using categories is a basic cognitive function for animals and humans. Thus, categorization has been a focus in animal research (Cerella, 1979; Chase & Heinemann, 2001; D'Amato & van Sant, 1988; Herrnstein, Loveland, & Cable, 1976; Jitsumori, 1994; Lea & Ryan, 1990; Pearce, 1988; Roberts & Mazmanian, 1988; Smith, Minda, & Washburn, 2004; Thompson & Oden, 2000; Vauclair, 2002; Wasserman, Kiedinger, & Bhatt, 1988) and human research (Ashby & Maddox, 2005; Brooks, 1978; Knowlton & Squire, 1993; Kruschke, 1992; Minda & Smith, 2002; Murphy, 2003; Nosofsky, 1987; Rosch & Mervis, 1975).

Some early categorization theories assumed that organisms apply a unitary category-learning system to all category problems. Different descriptions were offered for this system. Some hypothesized that learners average their exemplar experience into the category prototype and compare new items to this in judging category membership (e.g.,

---

Correspondence concerning this article should be addressed to J. David Smith, Park Hall, University at Buffalo, the State University of New York, Buffalo, NY, 14260, or to psysmith@buffalo.edu.

Reed, 1972). Others hypothesized that learners store category exemplars as separate memory representations and compare new items to these (e. g., Medin & Schaffer, 1978).

However, categorization is probably not so simple and unitary. In fact, from an adaptive/ fitness standpoint, categorization is likely an important enough capacity to deserve (and receive) redundant and varied expression in cognition. There is now strong evidence that something like exemplar memory dominates categorization under some circumstances (Medin & Schwanenflugel, 1981; Minda & Smith, 2001). There is also strong evidence that animals and humans sometimes use prototype representations in categorization (Aydin & Pearce, 1994; Huber & Lenz, 1993; Jitsumori, 1996; Knowlton & Squire, 1993; Smith, Redford, & Haas, 2008a; Reber, Stark, & Squire, 1998a,b; Reed, 1972; von Fersen & Lea, 1990; White, Alsop, & Williams, 1993).

As a result, a multiple-systems theoretical perspective has become an important part of the human categorization literature (Ashby & Ell, 2001; Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Erickson & Kruschke, 1998; Homa, Sterling, & Trepel, 1981; Love, Medin, & Gureckis, 2004; Minda & Smith, 2001, Rosseel, 2002; Smith & Minda, 1998), based on the idea that organisms have multiple categorization utilities that learn different statistical features of the repeating and differentiating environment. For example, the dominant process in categorization depends on the number of exemplars in categories (Homa et al., 1981; Minda & Smith, 2001). As categories contain few or many exemplars, respectively, participants modally encode either separate, individuated exemplar traces (in a process akin to exemplar memorization) or the central tendency of the exemplars (in a process akin to prototype abstraction). The dominant process in categorization also depends on the perceptual coherence of categories (Blair & Homa, 2003; Smith & Minda, 1998). As category members have a weak family resemblance (barely resembling each other more than they do opposing category members) or strong family resemblance (with the central tendency of the exemplars providing a clear and useful categorization signal), exemplar processes or abstraction processes, respectively, become more prominent. All in all, the multiple-systems perspective has profoundly enriched the human categorization literature. However, this new perspective has not been extended fully to the comparative categorization literature. Furthering this extension is an important goal of the present research.

This article explores another central issue in categorization science. An intuitive view of categories is that category members form a cloud of exemplars that are overall similar to one another as family members of the category. This exemplar cloud is a central image in the categorization literature. These family-resemblance categories do not have exceptional members with a strong pull to membership in an opposing category. These categories are linearly separable, because one of these exemplar clouds can be cleanly separated from another using a linear discriminant function. In fact, these categories can be effectively learned and used just by abstracting the central tendency of the exemplars. No special process is needed for exception items.

It is a fundamental question whether humans and animals learn linearly separable categories most easily and naturally. The answer bears on how organisms learn and represent categories. For example, if organisms abstract prototypes, they will show a linear

separability constraint and they might be unable to learn exceptional category members. Yet, in seminal human research, Medin and Schwanenflugel (1981) found equal learning for both linearly separable (LS) and non linearly separable (NLS) categories. Medin and Schwanenflugel carefully constrained their interpretations. The small categories used in their studies could have encouraged learning through exemplar memorization for the reasons already described. In fact, Blair and Homa (2003) showed that such categories are sometimes learned, not as coherent families of related exemplars, but rather as a series of individual stimulus-response pairings. In addition, participants learned poorly the matched LS and NLS category tasks in the original research. Smith, Murray, & Minda (1997) showed that these tasks sampled only one small region—an inherently difficult region—of the whole space of category tasks. Showing that very difficult LS and NLS categories are equally minimally learnable is different from showing that LS and NLS categories—generally—are equally learnable.

Taking another approach to this problem, Smith and Minda (1998) showed that different processes dominate categorization at different stages of category learning. They found that humans often pass through an extended stage of category learning that is consistent with an LS assumption about categories. Early in category learning, humans performed highly accurately on typical category members—indicating an initial mastery of the categories' general structure—but below chance on exceptional category members. During this stage, a formal prototype model fit participants' data closely. Later in category learning, humans improved selectively and sharply in their performance on exception items, indicating exception-item resolution or memorization. At that point, a formal exemplar model fit participants' data closely. Smith and Minda suggested that secondary exemplar processes finally supplemented category-level knowledge. Evaluating whether the same stages of category learning characterize category learning by nonhuman primates is a principal empirical goal of the present research.

Blair and Homa (2001) extended the findings of Smith and Minda (1998) to test humans with the dot-distortion categorization paradigm that has been so influential (Posner, Goldsmith, & Welton, 1967; Knowlton & Squire, 1993; Homa et al., 1981; Smith & Minda, 2002). Their work suggested the paradigm used here to test the monkeys. In one experiment, Blair and Homa (2001) assayed participants' performance on a four-category learning task in which each category had six typical category members and three exceptional category members. The exceptions in each category were actually typical members of the three opposing categories, but they were assigned and reinforced across categories according to their roles as exceptions. Blair and Homa asked whether participants could learn to override the perceptual resemblance of these exceptions to opposing prototypes and treat them particularly and individually, or whether they were constrained to respond in accordance with prototype similarity by incorrectly placing the exceptions in their prototype-congruent category. A great majority of participants responded poorly on the exception items because they responded obediently to prototype similarity. Only a small minority finally learned late in the experiment to treat the exception items idiosyncratically and thus categorize them correctly.

Experiment 1A in the present article replicates this finding by Blair and Homa (2001) with humans, providing additional evidence for the LS assumption in human category learning. Then, Experiment 1B demonstrates that the LS assumption persists within a simplified paradigm. Experiment 2 extends Experiment 1B's paradigm comparatively, analyzing the performance of three monkeys (*Macaca mulatta*).This may provide evidence that the LS assumption in early category learning has some breadth across the order *Primates*.

To be fair, Smith and Minda (1998) and Blair and Homa (2001) confirmed an LS assumption in category learning, not an irremediable LS constraint. These researchers focused on natural predispositions and initial approaches to category tasks. In contrast, Medin and Schwanenflugel (1981) focused on ultimate limits and learnability. These researchers showed that, after hundreds of trials, exception items are learned by many participants. The focus on ultimate limits is important, and it is feasible because categorizing undergraduates can err and learn indefinitely with no cost or risk.

Animals cannot. Moreover, in some situations, the LS assumption could itself sharply limit learnability. Suppose that in the natural ecology, organisms encountered NLS categories, made the LS assumption, and systematically erred on exception items—eating the exceptional toadstool that resembles a mushroom; grazing next to the exceptional predator (the wolf in sheep's family resemblance). In these cases, the LS assumption would constrain learnability, because a coping organism is not allowed multiple poisonings or multiple predator-recognition mistakes. The LS assumption could effectively render ecological NLS categories unlearnable, or too risky to engage in a learning process.

This example makes plain that ecological and foraging/survival considerations are part of our overall theoretical perspective. What kind of categories have organisms been prepared by evolution to find most learnable? What kind of categories do they typically encounter in the natural ecology? How well is their overall categorization capacity tuned for what the world sends their way? In our view, these questions, and the ecological perspective behind them, are critical to a full understanding of the ancestral primate category system from which that of humans emerged. They also provide a useful context for understanding the behavior of humans and monkeys in the present category tasks. This ecological perspective seems not to have been acknowledged sufficiently in the human categorization literature, and this is an area in which the comparative categorization literature has a distinctive contribution to make.

Experiment 1A: Humans

Following Smith and Minda (1998) and Blair and Homa (2001), Experiment 1A evaluated— in humans—the relative strengths of the LS assumption and immediate feedback signals on the categorization of exception items in a category-learning task. We expected that the LS assumption would dominate early on, as abstraction processes placed exception items into opposing categories despite the task's feedback signals. We expected that feedback signals would dominate late in category learning, as exemplar-specific processes finally resolved the exceptions and let them be treated individually and categorized correctly.

Method

Participants—Participants were 40 undergraduates from the University at Buffalo, the State University of New York, who participated in a session lasting about an hour to fulfill a course requirement. Our participant pool contained slightly more women than men. Participants were in their late teens or early twenties with apparently normal or corrected-to-normal visual acuity. The approximate racial mix of our participant pool was 69% Caucasian, 17% Asian, 9% African-American, and 5% Other. As a tool to increase performance motivation, the top scorers were awarded cash prizes in the experiment.

Dot-pattern stimuli—The dot-distortion stimuli were created with a common method that creates families of dot patterns from prototypes. Nine random points—the prototype—were selected from within the central 30×30 area of a 50×50 grid. The distortions (the family members) were produced by applying probabilities that govern whether each dot would keep the position it had in the prototype or how far it would be moved. These distortions were generally perceptually similar to one another and to their originating prototype.

Specifically, distortions were built from prototypes by probabilistically moving each dot into one of five areas that covered the 20×20 grid of pixels that surrounded it. For Area 1, the dot kept its original position. For Area 2, the dot moved to one of the 8 pixel positions in the first ring of pixels immediately around its original position. For Area 3, the dot moved to one of the 16 pixel positions in the second concentric ring of pixels around its position. For Area 4, the dot was moved into one of the 75 pixel positions in the third, fourth, and fifth concentric rings of surrounding pixels. For Area 5, the dot was moved into one of the remaining 300 pixel positions in the surrounding 20×20 pixel grid (i.e., to the 5th, 6th, 7th, 8th, 9th, or 10th concentric ring of pixels around the dot's original position). All dot-distortion stimuli used in Experiment 1A were Level 5 distortions (5 bits/dot of uncertainty in the original formulation of Posner et al., 1967), or what are called low-level distortions. For these distortions, the probability that dots would move into each of the five areas was, respectively, .20, .30, .40, .05, and 05.

With the dot positions chosen for a pattern, the DrawPoly procedure in Turbo Pascal 7.0 connected successive dots by yellow lines. This followed one common practice for presenting dot-distortion stimuli (Homa, Rhoads, & Chambliss, 1979; Homa et al., 1981). In addition, the shape constellation was magnified to be more visible and to approximate the size traditionally used for these shapes. To do so, each pixel position in the distortion algorithm was mapped to a 3×3 pixel square on the screen, and the dot was placed in the center of the appropriate 9-pixel cell. In this way, the stimulus patterns were magnified threefold from being drawn in a virtual 50×50 coordinate space to being shown on an actual 150×150 pixel space on the screen (about 7.7 cm at maximum, viewed from a distance of about 45 cm and so with a visual angle of about 9.8°).

Category prototypes—Four prototypes were used in the construction of each four-category task. Eight sets of four prototypes were randomly constructed. One of these eight sets was used for the task of each participant as randomly determined by their sequential participant number. The use of multiple prototype sets made the experiment a more

generalizeable statement about exception-resolution processes in the categorization of dot distortions.

Category structures—The stimuli were members of four categories. All stimuli were low-level distortions of one of the four prototypes in a prototype set. Nine of these distortions were produced from each prototype. Six of these distortions were kept together to constitute a prototype-based family of exemplars. To these six were added one distortion produced from each of the three opposing prototypes. This produced four nine-exemplar categories containing six typical instances and three exceptional instances that could have been a typical exemplar of some other category. An example of one of these stimulus sets can be seen in Figure 1. In each category group of exemplars, the first two rows contain typical items, and the third row contains the exceptions that reflect their derivation from the prototypes of the other categories.

The stimuli chosen for each category were constrained in their distance relationship to their originating prototype. Following the psychophysics of dot-pattern perception established in Posner et al. (1967) and Smith and Minda (2001, 2002), our operating measure of interstimulus distance was based on the average Pythagorean distance that corresponding dots were moved between dot-distortion patterns. We set inter-stimulus distance equal to ln (1+average Pythagorean distance). This measure is now standard in studies of dot-distortion perception and categorization, and it has been shown to sensitively predict both humans' and animals' responses to dot-distortion patterns (Smith et al., 2008a,b). All exemplars were constrained to be 1.09+- 0.05 distance units from their originating prototype, and thus they were all precisely normal low-level distortions (1.09 is the mean distance of low-level distortions from their originating prototype).

Dot-distortion trials—Each trial consisted of one shape presented in the center of the top half of a computer screen against a black background. Below the shape were the four response icons—A, B, C, and D—arranged clockwise from West to South in 90° intervals. Responses were made by moving the cursor to one of the response icons using one press of the appropriate arrow keys (A-Left; B-Up; C-Right; D-Down. The key press moved the cursor on the screen to the appropriate response icon, acknowledging the participant's response. Following a correct response, participants heard an 0.5 s computer-generated reward whoop and their score (shown in a text box in the upper left-hand corner of the screen) was incremented by one. Following an error, they received a 1.0 s computer-generated buzzing sound and their score was decremented by 1. In addition, they were prompted about the correct response on that trial. That is, all the incorrect response icons disappeared and they were forced to complete the trial by making the correct response. They received no feedback sound or point adjustment for this correction procedure. Following feedback and/or the correction procedure, the next trial was presented.

Instructions—Participants received the following instructions: In this experiment you will see yellow patterns which belong to one of four categories of shapes. Move the cursor to the "A" if you think the shape belongs to category A. Move the cursor to the "B" if you think the shape belongs to category B. Move the cursor to the "C" if you think the shape belongs to category C. Move the cursor to the "D" if you think the shape belongs to category D. You

will hear a whoop sound and GAIN A POINT for each correct response. You will hear a buzz sound and LOSE A POINT for each incorrect response. If you are incorrect, the computer will prompt you to make the response that would have been correct. This week, a &dollar;10 prize goes to the two people who earn the most points.

Procedure—Each participant was tested individually, after being randomly assigned to a set of prototypes. Stimuli were presented on a computer screen in blocks of 36 trials—each block was a random permutation of the 36 stimuli in the task (9 stimuli in 4 categories). Successive blocks were presented without any break that could have been apparent to participants. Participants completed 900 trials (25 blocks of 36 stimuli).

Results

The aggregate performance of all participants is shown in Figure 2A. The Typical (T) curve and Exception (E) curve, respectively, represent the proportion of typical items (6 per category) and exception items (3 per category) that were placed into the correct category. The LS curve is discussed momentarily.

A two-way analysis of variance (ANOVA), with trial block (1–25) and stimulus type (typical, exception) as within-subject factors, found a significant main effect for both trial block, $F (24,936) = 37.323$, $p < .001$ and stimulus type, $F (1,39) = 135.017$, $p < .001$. Participants improved their performance over trial blocks and performed better on typical items than on exception items. The interaction between trial block and stimulus type, $F (24,936) = 6.081$, $p < .001$, indicated that participants improved differentially on typical and exception items. A dramatic performance improvement on the typical items occurred over the first half of the experiment (from 35% in the first trial block to 73% in the 13th block). During this time there was little change in exception performance (from 18% in the first trial block to 22% in the 13th block), with exception performance still below chance level (25%) halfway through the experiment. Over the second half of the experiment, from blocks 13–25, typical-item performance improved (from 73% to 80% correct) but by less than did exception-item performance (from 22% to 33% correct).

Figure 2A provides an additional perspective on the data. The errors made on exception items were not random. Curve LS shows the proportion of exception items on which participants ignored ongoing feedback in the task and instead placed the exception items with the other exemplars that had the same originating prototype. By this reassignment—toward linearly separable categories and away from the task's feedback signals—participants confirmed that they were making an LS assumption. Curve LS shows that for about 7 36-trial blocks, or 252 trials, participants placed the exception items and the typical items equivalently strongly into the category of their proper family resemblance. For 252 trials, feedback had essentially no effect on category learning. In fact, the LS assumption remained strong at the experiment's end. Participants were still making incorrect and systematically penalized LS responses to 50% of the exception items. Consequently, participants were only 33% correct on the exception items, barely above chance after 900 trials.

In all, the 40 participants in Experiment 1A made 6,388 errors that were obedient to their LS assumption. This level of errors would be disastrous in a foraging/survival categorization situation. Remember also that each of these errors was accompanied by an additional focal experience with the correct stimulus-response pairing during the correction procedure. This means that LS-based errors persisted through 12,776 exemplar attempts and corrections. A specific-exemplar process that is so resistant to feedback and correction is empirically striking and theoretically illuminating, and we return to this issue in the General Discussion.

Participant subgroups—The LS assumption is borne out by examining group differences in the task. Figure 2B shows the aggregate results from a subgroup of 20 participants. These participants never overcame their LS assumption. They never reduced below 70% their tendency to classify the exception items incorrectly based on family resemblance. They never reached chance performance on exception items. In contrast, Figure 2C shows the aggregate results from a subgroup of 8 participants. These participants overcame their LS assumption. Their tendency to respond wrongly based on family resemblance weakened after Block 7. Their ability to categorize exception items correctly— in accordance with the task's feedback—rose after Block 7. Thus, some humans overcame their LS assumption in category tasks, as Medin and Schwanenflugel (1981) asserted, and perhaps all participants would have eventually. However, notice that even these participants made a strong LS assumption for the first 7 blocks (252 trials) of the task, just as all participants did. Their response to feedback signals did not outweigh their response to family resemblance until Block 16 in the task, after 540 trials.

Experiment 1B Humans

Experiment 1A recreated the categorization phenomena found in humans by Blair and Homa (2001) and, with other category stimuli and tasks, by Smith and Minda (1998). We were concerned that a four-category task would cause training and testing difficulty for the monkeys, given the memory load and the complexity of the response mappings. Therefore, we sought for the monkeys a simplified version of the previous experiment that would produce the same phenomena. That simplified paradigm was evaluated with humans in Experiment 1B.

Method

Participants—Forty UB undergraduates participated from the participant pool already described.

Dot-pattern stimuli—The method already described for creating dot-distortion stimuli was used again.

Category prototypes—Two prototypes were used in the construction of each two-category task. One of eight prototype pairs was used for the task of each participant, based on their sequential participant number.

Category structures—The stimuli were members of two categories. All stimuli were low-level distortions of one of the two prototypes in a prototype set. Eight of these

distortions were produced from each prototype. Six of these distortions were kept together to constitute a prototype-based family of exemplars. To these six were added two distortions that had been derived from the opposing prototype. This produced two eight-exemplar categories containing six typical instances and two exceptions that could have been a typical exemplar of the other category. An example of one of these stimulus sets can be seen in Figure 3. In each category group of exemplars, the first two rows contain typical items, and the third row contains the exceptions that reflect their derivation from the opposing prototype. The exemplars chosen were constrained in their distance relationship to their originating prototype as already described.

Dot-distortion trials—The trial stimuli were presented as already described. Now below the stimulus were two response icons—A and B—arranged to the Left and Right of the central cursor. Responses were made by moving the cursor to one of the response icons using the Left or Right arrow key. All aspects of feedback and response correction were as already described. Following feedback and/or the correction procedure, the next trial was presented.

Instructions—Participants received the same instructions as in Experiment 1A, except that a two-category task with two responses was indicated to them.

Procedure—Each participant was tested individually, after being randomly assigned to a set of prototypes. Stimuli were presented on a computer screen in blocks of 16 trials—each block was a random permutation of the 16 stimuli in the task (8 stimuli in 2 categories). Participants completed 800 trials (50 blocks of 16 stimuli).

Results

A two-way analysis of variance (ANOVA), with trial block (1–50) and stimulus type (typical, exception) as within-subject factors, found a significant main effect for both trial block, $F (49,1911) = 25.629$, $p < .001$ and stimulus type, $F (1,39) = 84.570$, $p < .001$. Participants improved their performance over trial blocks and performed better on typical items than on exceptions (T and E curves in Figure 4A). The interaction between trial block and stimulus type, $F (49,1911) = 3.694$, $p < .001$, indicated that participants improved differentially on typical and exception items. From Blocks 1–15, performance improved 25% on typical items (59% to 84% correct) but only 16% on exception items (31% to 47% correct). From Blocks 15 to 50, performance improved 27% on exception items (47% to 74% correct) but only 6% on typical items (84% to 90% correct).

Figure 4A provides an additional constructive perspective on the data. Curve LS shows the proportion of exception items on which participants ignored the ongoing feedback in the task and instead placed the exception items with the other exemplars that had the same originating prototype. In this two-category task, the LS curve is the complement of the E curve. Participants again made a strong LS assumption early in the task, in opposition to its feedback signals. They did not improve their exception-item performance until Block 9, after 128 trials in the task. They were unable to respond more strongly to the task's feedback signals than the task's family-resemblance signals until Block 20, after 304 trials in the task.

In all, the 40 participants in Experiment 1B made 3,488 (44%) errors in category assignment on exception items, despite the fact that each of these errors was followed by a deliberate correction procedure. Once again, this level of errors would be a serious matter for a coping organism.

Participant subgroups—The idea of the LS constraint is also borne out on examining subgroup differences in the task. Figure 4B shows the performance of a group of 5 participants. This graph—showing 5 participants categorizing 4 exceptions per block—produces noisier LS and E curves. But one still sees that these participants never overcame their LS assumption. On average, they classified 70% of exceptions in accordance with the task's family-resemblance relations. Consequently, they were 30% correct on exceptions, below chance (50%). In contrast, Figure 4C shows the performance of a group of 24 participants. These participants did overcome their LS constraint in the two-category task. Even so, they did not appreciably improve their exception-item performance until Block 9, after 128 trials in the task. They were unable to respond more strongly to the task's feedback signals than the task's family-resemblance signals until Block 15, after 224 trials in the task.

Comparing Experiment 1A and 1B—Perhaps it is intuitive that the pace of exception resolution was quicker in Experiment 1B, and the break-even point of feedback responding and family-resemblance responding was earlier in Experiment 1B (Block 20, after 304 trials) than in Experiment 1A (never). Experiment 1A had more stimuli (36), more exceptions (12 or 33%), fewer stimulus repetitions (every 36 trials), and complex exception-response mappings given a four-category task. Experiment 1B had fewer stimuli (16), fewer exception items (4 or 25%), more stimulus repetitions (every 16 trials), and simpler exception-response mappings because any exception item could correctly be placed into seemingly the opposite category. Nonetheless, both experiments confirm that humans persisted through many errors and many correction procedures in making an LS assumption during category learning.

A Formal-Modeling Perspective

Formal models confirmed this description of human's performance in Experiment 1B. (Blair & Homa, 2001, extensively modeled humans' results in the paradigm of Experiment 1A. There was no need for duplicative modeling in that case.) This section describes the models used here, the procedures used for fitting models to data, and the results of our formal analyses.

The prototype model—The prototype model assumes that participants compare a to-be-categorized item to the *within* prototype in the item's own category and the *between* prototype in the opposing category. The model's comparisons incorporated the established measure of interstimulus distance (*dist*) that was described above (also Posner et al., 1967; Smith & Minda, 2001, 2002). Stimuli were 1.09 distant on average from their derivational prototypes and 2.84 distant from opposing prototypes. These values agree with those from previous research. Distances were transformed into estimates of psychological similarity using an exponential-decay function that incorporated a sensitivity parameter (*sens*) that was the model's free parameter. For example, within similarity was: $sim_{within} = e^{-sens X dist_{within}}$.

Similarities in turn were entered into the model's choice rule:

$$P(R_{within}) = \frac{sim_{within}}{sim_{within} + sim_{between}}.$$

Stronger within similarity supports correct (within) responding—typical items (similar to their within prototype) should often be categorized correctly. Stronger between-category similarity undermines correct responding—exception items (similar to the opposing prototype, dissimilar within) should often be categorized incorrectly.

The exemplar model—The exemplar model assumes that participants compare a to-be-categorized item to the 8 *within* exemplars in the item's category and the 8 *between* exemplars in the opposing category. The established distance metric already described was used again. Items compared to themselves had an interstimulus distance of 0.0. Items compared to members of the same derivational category had an interstimulus distance of 1.34. Items compared to exemplars from the opposing category had a distance of 2.88. These values differ slightly from those in previous research due to the tight controls placed on the present distortion algorithm to ensure homogeneous exemplar sets. These distances were transformed into estimates of psychological similarity using the exponential-decay function with a sensitivity parameter that was already described. Total category similarity was the sum of 8 similarities as follows:

Typical items, within similarity: 1 perfect self-similarity, 5 strong similarities, 2 weak similarities

Typical items, between similarity: 2 strong similarities, 6 weak similarities Exceptions, within similarity: 1 perfect self-similarity, 1 strong similarity, 6 weak similarities

Exceptions, between similarity: 6 strong similarities, 2 weak similarities

The similarities were entered into the choice rule already described. Typical items (mainly similar to within exemplars) and exceptions (mainly similar to between exemplars) should often be categorized correctly and incorrectly, respectively.

Measures of fit and methods of minimization—We used standard hill-climbing algorithms to maximize the fit (i.e., minimize the differential) between predicted and observed typical-exception performances. We chose a starting parameter value for sensitivity and calculated the predicted categorization probabilities for typical and exception items on that basis. The degree of fit between the predicted and observed categorization probabilities was the sum of the squared deviations (SSD) between them. Random adjustments were made to the starting value and adopted if they produced a better fit (i.e., a smaller SSD). The adjustments were small (1/10,000) and respected the parameter's bounds (0.0001 and 20). To ensure that local minima were not a problem, this fitting procedure was repeated by choosing three more starting sensitivity values and climbing from there.

Results: prototype model—Figure 5 shows the results when the prototype model was fit to the performance of all 40 participants (A), to the performance of the five participants who never overcame their LS assumption (B), and to the performance of the 24 participants who overcame their initial LS assumption (C). In all cases, the model was fit to performance of individual participants completing single blocks of trials. Then, the predicted values were obtained by averaging the individual predictions across participants within each trial block. One sees that early in performance participants combined highly accurate typical-item performance with dismal exception-item performance. This is the LS constraint that has been described. The prototype model captured this data pattern well with small fit indices during this period. Late in performance, though, the prototype model failed qualitatively for the group as a whole and for the group of participants who overcame their LS assumption. Many humans finally achieved accurate typical- and exception-item performance. The prototype model mispredicted both performance levels badly with correspondingly poor fit indices. The prototype model showed a sharp stage change during category learning by strongly succeeding then qualitatively failing.

Results: exemplar model—Figure 6 shows the results when the exemplar model was fit to the performance of all 40 participants (A), to the performance of the five participants who never overcame their LS assumption (B), and to the performance of the 24 participants who overcame their initial LS assumption (C). One sees that the exemplar captured very poorly the data pattern seen in the early trial blocks. That model could not both capture how well participants were doing on typical items and how poorly they were doing on exception items. This because the model, in predicting better typical-item performance, would raise the level of the sensitivity parameter, which it cannot do without also predicting higher exception-item performance. So, the model is pinned in the middle in a sense as the figure shows. Late in performance, though, the exemplar model fit the data pattern well for the group as a whole and for the group of participants who overcame their LS assumption. It fits comfortably, with high values for the sensitivity parameter, the data from the many humans who finally achieved accurate typical- and exception-item performance. The exemplar model showed a sharp stage change during category learning by qualitatively failing then strongly succeeding.

We stress that it is not the purpose of this article to make some one model win, or to express any view regarding the general superiority of prototype or exemplar theory or prototype or exemplar models. This article is about the psychology of category learning, and in particular it is about the different learning stages that humans and animals pass through during their acquisition of categories. Both models show the succession of stages clearly. One succeeds then fails. One fails then succeeds. In this respect, our model fitting is model- and theory-neutral. However, both patterns of fit converge to indicate strongly that the processes and representations used in categorization were sharply different early and late in acquisition. Early on, a strong LS assumption produced strong typical-item performance but terrible exception-item performance. Later on, the LS assumption eased, and some item-specific process strengthened, to the point of supporting strong typical- and exception-item performance.

This narrative would be unchanged even if one found some model that alone could fit humans' early and late performance profiles (see Experiment 2). This model would necessarily do so by selecting profoundly different parameter settings that would create the early and late performance profiles. Yet this would leave unanswered the psychological reasons why humans entered a radically different place in parameter space in performing during the two stages of acquisition. That question would still need a theoretical answer that the mathematical model could not provide. It is important to realize that the mathematics in formal models is often psychologically silent in this way. The idea of the changing balance between the initial LS assumption and the later item-specific processes are psychologically illuminating, and they explain the changing performance patterns in a way that models may not.

Experiment 2: Monkeys

Experiment 2 generalizes the human phenomena to the category learning of nonhuman primates, by examining the behavior of three rhesus monkeys (*Macaca mulatta*) in the paradigm of Experiment 1B.

Method

Participants—Hank (22 years old), Lou (14 years old), and Murph (14 years old) were tested. They had been trained, using procedures described elsewhere (Rumbaugh, Richardson, Washburn, Savage-Rumbaugh, & Hopkins, 1989; Washburn & Rumbaugh, 1992), to respond to computer-graphic stimuli by manipulating a joystick. They had been tested in prior studies on a variety of computer tasks. Lou and Murph had had experience with a related categorization paradigm; Hank had not. The monkeys were tested in their home cages at the Language Research Center of Georgia State University, with *ad lib* access to the test apparatus, working or resting as they chose during long sessions. The animals were neither food deprived nor weight reduced for the purposes of testing and they had continuous access to water.

Apparatus—The monkeys were tested using the Language Research Center's Computerized Test System—LRC-CTS (described in Rumbaugh et al., 1989; Washburn & Rumbaugh, 1992)—comprising a Compaq DeskPro computer, a digital joystick, a color monitor, and a pellet dispenser. Monkeys manipulated the joystick through the mesh of their home cages, producing isomorphic movements of a computer-graphic cursor on the screen. Contacting appropriate computer-generated stimuli with the cursor brought them a 94-mg fruit-flavored chow pellet (Bio-Serve, Frenchtown, NJ) using a Gerbrands 5120 dispenser interfaced to the computer through a relay box and output board (PIO-12 and ERA-01; Keithley Instruments, Cleveland, OH). Correct responses were accompanied by a computer-generated whooping sound that bridged the animals to their reward. On incorrect responses, the screen froze with the wrong response visible, and there was a computer-generated buzzing sound and a timeout that generally lasted 10 s.

Procedure

The stimuli, prototypes, category structures, and trial displays were those already described in Experiment 1B. The monkeys made their categorization decision using an analog joystick to move a cursor to touch the response icon of their choice. Stimuli were presented on a computer screen in blocks of 16 trials—each block was a random permutation of the 16 stimuli in the task (8 stimuli in 2 categories). The monkeys were presented again with the same trial following an error, and for them completing this trial had the same consequences for correct and incorrect responses. This seemed a more transparent way of providing correction on a trial without the complexity of the qualitatively different correction procedure used in the human studies.

Data analysis—Monkeys completed six successive category tasks, using six of the category pairs from Experiment 1B. Tasks 2–6 of this series were analyzed. For analysis, the data were organized into 64-trial blocks containing four successive runs through the 16 stimuli. Each monkey completed a different number of trials in each task. The precise trial counts were dependent on many variables including each animal's own decisions about how many trials to complete in a given day. To create a balanced design, we analyzed for each of a monkey's five tasks the minimum number of 64-trial blocks that he completed across the 5 tasks. For Hank, Lou, and Murph, respectively, these trial-block counts were 194, 164, and 158. This means that data is reported for 62,080, 52,480, and 50,560 trials for Hank, Lou, and Murph, respectively.

Results

Hank—Hank's aggregate performance across five tasks is shown in Figure 7A. The T, E, and LS curves are those already described. A two-way analysis of variance (ANOVA), with 64-trial block (1–194) and stimulus type (typical, exception) as within-subject factors, did not find a significant main effect for trial block, $F(193,772) = 1.111$, ns. Hank improved on typical items early in the experiment but there were long periods during which performance on typical and exception items was static. In addition, his performance loss on exception items somewhat canceled his performance gain on typical items early in the experiment. We did find a significant main effect for stimulus type, $F(1,772) = 3175.340$, $p < .001$. Hank performed better on typical items than on exceptions. The lack of an interaction between trial block and stimulus type, $F(193,772) = 0.999$, ns, indicated that the relation of typical and exception items remained fairly constant through the experiment. Hank reached 90% on typical items midway through the experiment. But he barely improved on the exception items. Indeed, he was still slightly below chance on them at Block 194, after 12,352 trials experienced in each task. Across his five tasks, Hank received about 15,500 exception trials, and 8,800 additional correction trials for errors made on those trials, all to no avail because performance remained below chance.

Lou—Lou's aggregate performance across five tasks is shown in Figure 7B. A two-way analysis of variance (ANOVA), with 64-trial block (1–164) and stimulus type (typical, exception) as within-subject factors, found a significant main effect for both trial block, $F(163,652) = 2.298$, $p < .001$ and stimulus type, $F(1,652) = 2913.406$, $p < .001$. Lou improved his performance over trial blocks and performed better on typical items than on

exceptions. The interaction between trial block and stimulus type, F (163,652) = 1.527, $p < .001$, indicated that Lou improved differentially on typical and exception items. In his early blocks, as he improved about 25% on typical items, his exception-item performance actually fell about 20%. These reciprocal changes in early performance are the clearest possible evidence that Lou was operating under a strong LS assumption during this phase of the task. Lou improved to 90% correct on typical items while he was still only about 35% correct on exception items. The LS curve also confirms that Lou was making a strong LS assumption. He was unable to respond more strongly to the task's feedback signals than the task's family-resemblance signals until about Block 110, after 6,976 trials in the task. Lou never fully overcame his LS assumption, though he did improve substantially on the exception items late in the task.

Murph—Murph's aggregate performance across five tasks is shown in Figure 7C. A two-way analysis of variance (ANOVA), with 64-trial block (1–158) and stimulus type (typical, exception) as within-subject factors, found a significant main effect for both trial block, F (157,628) = 8.833, $p < .001$ and stimulus type, F (1,628) = 3544.709, $p < .001$. Murph improved his performance over trial blocks and performed better on typical items than on exceptions. The interaction between trial block and stimulus type, F (157,628) = 3.927, $p < .001$, indicated that Murph improved differentially on typical and exception items. From Block 1–23, he improved 36% on typical items (57% to 93% correct), but not at all on exception items (30% correct). That is, he improved beyond 90% correct on typical items while remaining far below chance (50%) on exception items. The LS curve also confirms that he was making a strong LS assumption early in the task, in opposition to the task's feedback signals. Murph was unable to respond more strongly to the task's feedback signals than the task's family-resemblance signals until about Block 45, after 2,816 trials in the task. But Murph did overcome his LS assumption—he finally performed the exception items at a high level.

A Formal-Modeling Perspective

The formal models already described in Experiment 1B confirmed the descriptions of animals' performances.

Results: prototype model—Figure 8 (A–C) shows the results when the prototype model was fit to the performance of Hank, Lou, and Murph in the 3rd of 5 analyzed tasks. Early in performance, animals paired highly accurate typical-item performance with dismal exception-item performance. This is the LS constraint that has been described. The prototype model captured this data pattern well with small fit indices during this period. Late in performance, though, the prototype model failed qualitatively. Monkeys achieved accurate typical- and exception-item performance. The model mispredicted both performance levels badly with correspondingly poor fit indices. The prototype model showed a sharp stage change during category learning by strongly succeeding then qualitatively failing.

Results: exemplar model—Figure 9 (A–C) shows the results when the exemplar model was fit to the same performances. Early in performance, the exemplar model failed

qualitatively. It could not accommodate the LS-constrained data pattern. It badly under-predicted and over-predicted typical-item and exception-item performance, respectively. It had correspondingly poor fit indices. Late in performance, though, the exemplar model captured well with small fit indices animals' accurate performance on both item types. It accommodated well the LS-transcending data pattern. The exemplar model showed a sharp stage change during category learning by qualitatively failing then strongly succeeding.

Results: gamma model—To illustrate that every modeling framework would reflect the same stage change, we also fit a gamma model to animals' performances. This model operated like the exemplar model except that the choice-rule's quantities could be raised to any power gamma:

$$P(R_{within}) = \frac{sim_{within}{}^{gamma}}{sim_{within}{}^{gamma} + sim_{between}{}^{gamma}}.$$

This model provided an alternative formal perspective toward animals' LS-constrained and LS-transcending stages of category learning.

There are concerns to bear in mind regarding the use of the gamma model. First, the gamma parameter can push estimated high performance higher and low performance lower, but for mathematical reasons that need not reflect animals' psychological processes (Smith et al., 2008a). Second, the gamma parameter counters the tendency for exemplar-based comparisons to produce flatter typicality gradients (Smith, 2002) and acts mathematically in ways that can mimic prototype-based processing (Smith & Minda, 1998). In these respects, the gamma model is not an exemplar model. Third, the gamma model in this case has two free parameters— gamma and sensitivity—which it brings to fitting two data points— typical- and exception-item performance. Therefore, the gamma model's capacity to fit the present data is guaranteed. But still, one can examine the gamma model's best-fitting parameter configurations to confirm animals LS-constrained and LS-transcending learning stages.

Figure 10 (A–C) shows the best-fitting parameter values when the gamma model fit the performances of Hank, Lou, and Murph. Early in performance, gamma-parameter estimates spiked very high while sensitivity was estimated low. The gamma model chose this parameter configuration to accommodate animals' strong LS assumption early in category learning. Late in performance, gamma-parameter estimates fell sharply. There is no psychological rationale why animals would relax their decisional use of category evidence (i.e., their gamma) by twenty orders of magnitude. Indeed, more plausibly task expertise would cause them to respond more deterministically (i.e., with higher gammas). However, the gamma model must make this counter-intuitive change to fit the data. At the same point, sensitivity was estimated higher. This increase let individual exemplar retrievals (especially perfect self-similarities) affect performance substantially and essentially granted the gamma model an exemplar-memorization process that let it accommodate improving exception-item performance. The gamma model showed a strong stage change in category learning by choosing successive best-fitting parameter configurations that were diametrically opposed in its parameter space.

The gamma model concretizes the issue raised in the modeling section of Experiment 1B. That is, even given a model that alone fits animals' early and late performance profiles by choosing profoundly different parameter configurations, one still faces the question of why —psychologically—animals entered such a radically different place in parameter space in performing during the two stages of acquisition. That question needs a psychological answer that the mathematical model cannot provide. In our view, it is most plausible, parsimonious, and illuminating to posit that there is an early stage of LS-constrained category learning that is based in family resemblance and prototype abstraction. Then, the LS assumption is augmented or replaced by item-specific processes or memory traces that allow LS-transcending performance instead.

Nonetheless, all three formal perspectives converged in showing the same succession of stages in animals' category learning that humans showed. For making this empirical demonstration, the choice of the modeling framework does not matter.

General Discussion

We extended the findings of Smith and Minda (1998) and Blair and Homa (2001) to nonhuman primates. Humans and three rhesus monkeys expressed a two-stage category-learning process. Early on, they learned the categories' general structure, allowing typical items to be categorized accurately but causing exception items to be categorized below chance. During this stage, the categorization system of both species made a strong LS assumption. Exception items were categorized in opposition to prevailing feedback and to specific-exemplar encoding. Instead, they were categorized in a way that maintained coherent, LS categories. This LS assumption lasted through hundreds of errors and trial-correction procedures. Later, some humans and monkeys overcame their LS assumption, mastering the exceptions at above-chance levels and responding more to the prevailing feedback than to family-resemblance considerations. This transition was achieved slowly by few humans in the four-category task of Experiment 1A. In that case, most humans never made the transition, and this was true for monkey Hank in Experiment 2's two-category task.

One can characterize psychologically the initial, LS-constrained phase of category learning. Humans in dot-distortion tasks unambiguously refer to-be-categorized items to a prototype-like representation near the category's center rather than to memorized training exemplars (Smith & Minda, 2001, 2002). In those studies, prototype models fit closely participants' performance in dot-distortion tasks. Matched exemplar models made severe prediction errors. In Smith et al. (2008a), prototype models fit closely monkeys' performance in dot-distortion category tasks. Matched exemplar models could not capture what the monkeys did.

The same was true here. Matched prototype and exemplar models, respectively, fit early performance brilliantly and terribly. The only limitation on this conclusion is that the gamma model (in a sense an unmatched exemplar model with more mathematical power and more free parameters than the prototype model) did fit the data, though it did so degenerately given two free parameters fitting just two data values. Of course it is highly

implausible that animals would invoke early on a 20-orders-of-magnitude response determinism and then discard it later on. There is no precedent for this in the experimental or modeling literatures. In addition, it is known that when the prototype and gamma models are contrasted most sensitively by analyzing the shapes of typicality gradients, the gamma model fails to account for dot-distortion performance by either humans or animals. Gamma is not the mechanism at work in the present data, either.

Thus, the LS-based stage of category learning is likely a stage of abstraction and prototype formation wherein the central tendencies of coherent exemplar families are derived and used as the comparative standard in making categorizations decisions. In this abstractive, averaging process, the details of particular exemplars are naturally lost in their assimilation to the prototypes. Therefore, participants naturally place exception items obeying family-resemblance considerations, not feedback considerations. Assuming an early stage of abstraction provides the most intuitive explanation for why it would be that the cognitive system of humans and monkeys blurs the specific details of the individual exemplars and learns slowly individualized responses to them.

The present results also illuminate the specific-exemplar process that was also at work within the present category tasks. The conceptual core of exemplar theory is that participants learn about and store exemplars as separate, individuated memory representations. Yet, in this case, hundreds of repetitions of exception items, with many of those repetitions augmented by trial-correction interventions, were insufficient to allow consistent feedback signals to override family resemblance. The data point to a surprising conclusion. Humans and monkeys in the present experiments expressed a specific-exemplar system that is slow learning and resistant to persistent feedback.

This conclusion illuminates a wide range of human categorization data. Medin and Schawenflugel (1981) grounded the narrative that humans are indifferent to LS or NLS category structures and to categories with different degrees of category coherence. They are not. Humans have a difficult time with NLS categories that might depend on specific-exemplar processes. For example, in the original articles that supported exemplar processes by showing that humans can ultimately learn NLS categories, 30%, 36%, 40%, 60%, 66%, and 72% of participants failed to reach criterion in various experiments (Medin & Schwanenflugel, 1981; Medin, Dewey, & Murphy, 1983; Medin & E. Smith, 1981; Medin & Schaffer, 1978). Even learners had an asymptotic performance of only about 80% correct. (The humans in Experiments 1A and 1B, respectively, had an asymptotic performance of 64% and 86% correct.) In fact, Blair and Homa (2003) showed that in some cases NLS categories are learned through an arduous memorization process that may be paired-associate learning, not categorization.

The present results explain this broad-based pattern of learning difficulty in two ways. First, NLS categories poorly suit humans' and animals' initial approach toward abstraction and an LS assumption in category tasks. NLS categories often offer no robust abstractive or similarity basis on which category decisions might be based. Second, when that initial approach proves unworkable, the specific-exemplar process that steps in to support learning tends to be weak and slow. The impoverished nature of the exemplar system that serves

humans' and animals' categorization is perhaps the most important implication of the present studies.

The results in the present article also have ecological-fitness implications. Humans and animals sometimes make hundreds of errors on the way to NLS concept mastery (if mastery eventuates at all). If these were toadstool-foraging errors or predator-recognition errors, fitness would take a direct, powerful hit. If species predominantly encountered NLS ecological categories, they would have developed a sharper, faster, specific-exemplar system and they would have de-emphasized their strong LS assumption entering category tasks.

Therefore, it becomes an important possibility that there is an adaptive resonance between the LS categories that organisms seem prepared to learn by default and the natural-kind categories that species have frequently encountered during evolutionary time. In fact, if one examines the ecological categories faced daily by monkeys (e. g., vervet monkeys, Cheney & Seyfarth, 1990), one sees that all of them—eagles, snakes, leopards, bushes, trees, leaves, seeds, shoots, insects, grubs, rodents, eggs, Masai tribespeople, dominant males, estrous females, infants—are family-resemblance categories as described by Rosch and Mervis (1975). The abstractive stance, the LS assumption, and the weaker specific-exemplar system as elements in cognitive systems for category learning may reflect that cognitive evolution has taken advantage of a natural category structure that organisms have often experienced (see also Ruts, Storms, & Hampton, 2004).

If there is an adaptive resonance between abstractive approaches to initial category learning and natural-kind categories, then the psychological transition demonstrated here should be broad phylogenetically. In fact, Cook and Smith (2006) asked whether similar processing stages extend to category learning in birds. Using the approach taken by Smith and Minda (1998), they found the same psychological transition—from abstraction to exception-resolution processing—in pigeons and humans who have more than 100 million years of phylogenetic separation. An analogous finding was reported by Wasserman et al. (1988). The phylogenetic breadth of this transition strengthens the suggestion that this succession of learning stages is not accidental, but grounded in some entrainment between cognitive systems and cognitive environments.

We stress that this discussion does not preclude that humans and animals will sometimes encounter NLS category structures in nature, or that they will do so frequently enough that they need additional components within their overall categorization system. For example, considering vervet monkeys again, their social-networking system is intimately nuanced and does require detailed exemplar-by-exemplar (i. e., animal-by-animal) representations. Likewise, in the present research, some humans and monkeys ultimately showed an ability to master exceptions by treating them individually instead of assimilating them to their prototype. The modeling also showed a succession between categorization strategies that were abstraction and exemplar based. Our view, shared by many neuroscientists studying categorization, is that the cognitive systems that serve categorization are multiple and diverse. For example, there is increasing interest in trying to localize in brain the process of prototype abstraction (Aizenstein et al., 2000; Ashby & Maddox, 2005; Coutinho, Couchman, Redford, & Smith, 2008; Reber et al., 1998a,b). Likewise, there is a striatal

categorization system that learns slowly to map responses to particular regions of perceptual space, and this system would have the right character to explain later exception-item performance in our tasks (Ashby & Maddox, 2005; Ashby et al., 1998).

We also stress that we do not question the elegant results that have supported exemplar processes and theory over many years. Our difference with that work is a matter of emphasis. It focused on the ultimate learnability of category tasks, which is one criterion for evaluating whether humans and animals have an LS constraint in category learning. By that criterion, they often do not. Our work focused on humans' and monkeys' first line of defense in category learning, which is often that of abstraction, prototype formation, the blurring of fine distinctions among exemplars, and an LS assumption that trumps the task's reinforcement landscape. This assumption has not received its due from human or comparative scientists of categorization, and the primary theoretical contribution of the present article is to encourage that recognition.

We have also suggested that the LS assumption may be rooted in a long-lived affordance offered by natural kinds. This affordance has been present through all the hundreds of millions of years that animals have confronted family-resemblance category problems. There may have been in many species a gentle pressure toward abstraction and toward blurring the fine distinctions among exemplars in the service of forming coherent categories. The result would be the default category-learning system displayed by humans and animals in the present studies: a strong and persistent LS assumption paired with a weaker exemplar-resolution system. In this respect our article joins others (Shepard, 1987, 1994, 2001) in suggesting that basic principles may attend the structure and geometry of natural kinds, that principles of optimality follow from these, and that evolutionary internalization may have incorporated those principles so that they are reflected in the minds of species that experience those natural kinds.

Acknowledgments

References

Aizenstein HJ, et al. Complementary category learning systems identified using event-related functional MRI. Journal of Cognitive Neuroscience. 2000; 1:977–987. [PubMed: 11177418]

Ashby FG, Alfonso-Reese LA, Turken AU, Waldron EM. Aneuropsychological theory of multiple systems in category learning. Psychological Review. 1998; 105:442–481. [PubMed: 9697427]

Ashby FG, Ell SW. The neurobiology of human category learning. Trends in Cognitive Science. 2001; 5:204–210.

Ashby FG, Maddox WT. Human category learning. Annual Review of Psychology. 2005; 56:149–178.

Aydin A, Pearce JM. Prototype effects in categorization by pigeons. Journal of Experimental Psychology: Animal Behavior Processes. 1994; 20:264–277.

Blair M, Homa D. Expanding the search for a linear separability constraint on category learning. Memory & Cognition. 2001; 29:1153–1164. [PubMed: 11913752]

Blair M, Homa D. As easy to memorize as they are to classify: The 5–4 categories and the category advantage. Memory & Cognition. 2003; 31:1293–1301. [PubMed: 15058690]

Brooks, LR. Nonanalytic concept formation and memory for instances. In: Rosch, E.; Lloyd, BB., editors. Cognition and categorization. Hillsdale, NJ USA: Erlbaum; 1978. p. 169-211.

Cerella J. Visual classes and natural categories in the pigeon. Journal of Experimental Psychology: Human Perception and Performance. 1979; 5:68–77. [PubMed: 528931]

Chase, S.; Heinemann, EG. Cook, RG., editor. Exemplar memory and discrimination. Avian visual cognition. 2001. [On-line]. www.pigeon.psy.tufts.edu/avc/chase/

Cheney, DL.; Seyfarth, RM. How monkeys see the world: Inside the mind of another species. Chicago: University of Chicago Press; 1990.

Cook RG, Smith JD. Stages of abstraction and exemplar memorization in pigeon category learning. Psychological Science. 2006; 17:1059–1067. [PubMed: 17201788]

Coutinho MVC, Couchman JJ, Redford JS, Smith JD. Refining the Visual Cortical Priming Hypothesis in Category Learning by Humans (*Homo sapiens*) and Rhesus Monkeys (*Macaca mulatta*). Manuscript submitted for publication. 2008

D'Amato MR, Van Sant P. The person concept in monkeys (*Cebus apella*). Journal of Experimental Psychology: Animal Behavior Processes. 1988; 14:43–55.

Erickson MA, Kruschke JK. Rules and exemplars in category learning. Journal of Experimental Psychology: General. 1998; 127:107–140. [PubMed: 9622910]

Herrnstein RJ, Loveland DH, Cable C. Natural concepts in pigeons. Journal of Experimental Psychology: Animal Behavior Processes. 1976; 2:285–311. [PubMed: 978139]

Homa D, Rhoads D, Chambliss D. Evolution of conceptual structure. Journal of Experimental Psychology: Human Learning and Memory. 1979; 5:11–23.

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

Huber L, Lenz R. A test of the linear feature model of polymorphous concept discrimination with pigeons. Quarterly Journal of Experimental Psychology. 1993; 46B:1–18.

Jitsumori, M. Discrimination of artificial polymorphous categories in humans and nonhumans. In: Hayes, SC.; Hayes, LJ.; Sato, M.; Ono, K., editors. Behavior analysis of language and cognition. Washington, DC USA: APA; 1994. p. 91-106.

Jitsumori M. A prototype effect and categorization of artificial polymorphous stimuli in pigeons. Journal of Experimental Psychology: Animal Behavior Processes. 1996; 22:405–441. [PubMed: 8865609]

Knowlton BJ, Squire LR. The learning of categories: Parallel brain systems for item memory and category knowledge. Science. 1993; 262:1747–1749. [PubMed: 8259522]

Kruschke JK. ALCOVE: An exemplar-based connectionist model of category learning. Psychological Review. 1992; 99:22–44. [PubMed: 1546117]

Lea, SEG.; Ryan, CME. Unnatural concepts and the theory of concept discrimination in birds. In: Commons, ML.; Herrnstein, RJ.; Kosslyn, SM.; Mumford, DB., editors. Quantitative analyses of behavior. Vol. VIII. Hillsdale, NJ USA: Erlbaum; 1990. p. 165-185.

Love BC, Medin DL, Gureckis TM. SUSTAIN: A network model of category learning. Psychological Review. 2004; 111:309–332. [PubMed: 15065912]

Medin DL, Dewey GI, Murphy TD. Relationships between item and category learning: Evidence that categorization is not automatic. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1983; 9:607–625.

Medin DL, Schaffer MM. Context theory of classification learning. Psychological Review. 1978; 85:207–238.

Medin DL, Schwanenflugel PJ. Linear separability in classification learning. Journal of Experimental Psychology: Human Learning and Memory. 1981; 7:355–368.

Medin DL, Smith EE. Strategies and classification learning. Journal of Experimental Psychology: Human Learning and Memory. 1981; 7:241–253.

Minda JP, Smith JD. Prototypes in category learning: The effects of category size, category structure, and stimulus complexity. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2001; 27:775–799.

Minda JP, Smith JD. Comparing prototype-based and exemplar-based accounts of category learning and attentional allocation. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2002; 28:275–292.

Murphy, GL. The big book of concepts. Cambridge, MA USA: MIT Press; 2003.

Nosofsky RM. Attention and learning processes in the identification and categorization of integral stimuli. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1987; 13:87–108.

Pearce, JM. Stimulus generalization and the acquisition of categories by pigeons. In: Weiskrantz, L., editor. Thought without language. Oxford UK: Oxford University Press; 1988. p. 132-152.

Posner MI, Goldsmith R, Welton KE. Perceived distance and the classification of distorted patterns. Journal of Experimental Psychology. 1967; 73:28–38. [PubMed: 6047706]

Reber PJ, Stark CEL, Squire LR. Cortical areas supporting category learning identified using functional MRI. Proceedings of the National Academy of Science. 1998a; 95:747–750.

Reber PJ, Stark CEL, Squire LR. Contrasting cortical activity associated with category memory and recognition memory. Learning and Memory. 1998b; 5:420–428. [PubMed: 10489259]

Reed SK. Pattern recognition and categorization. Cognitive Psychology. 1972; 3:382–407.

Roberts WA, Mazmanian DS. Concept learning at different levels of abstraction by pigeons, monkeys, and people. Journal of Experimental Psychology: Animal Behavior Processes. 1988; 14:247–260.

Rosch E, Mervis CB. Family resemblances: Studies in the internal structure of categories. Cognitive Psychology. 1975; 7:573–605.

Rosseel Y. Mixture models of categorization. Journal of Mathematical Psychology. 2002; 46:178–210.

Rumbaugh DM, Richardson WK, Washburn DA, Savage-Rumbaugh ES, Hopkins WD. Rhesus monkeys (*Macaca mulatta*), video tasks, and implications for stimulus-response spatial contiguity. Journal of Comparative Psychology. 1989; 103:32–38. [PubMed: 2924530]

Ruts W, Storms G, Hampton J. Linear separability in superordinate natural language concepts. Memory & Cognition. 2004; 32:83–95. [PubMed: 15078046]

Shepard RN. Toward a universal law of generalization for psychological science. Science. Sep; 1987 237(4820):1317–1323. [PubMed: 3629243]

Shepard RN. Perceptual-cognitive universals as reflections of the world. Psychonomic Bulletin & Review. 1994; 1:2–28. [PubMed: 24203412]

Shepard RN. Perceptual-cognitive universals as reflections of the world. Behavioral and Brain Sciences. 2001; 24:581–601. [PubMed: 12048942]

Smith JD. Exemplar theory's predicted typicality gradient can be tested and disconfirmed. Psychological Science. 2002; 13:437–442. [PubMed: 12219810]

Smith JD, Minda JP. Prototypes in the mist: The early epochs of category learning. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1998; 24:1411–1436.

Smith JD, Minda JP. Journey to the center of the category: The dissociation in amnesia between categorization and recognition. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2001; 27:984–1002.

Smith JD, Minda JP. Distinguishing prototype-based and exemplar-based processes in dot-pattern category learning. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2002; 28:800–811.

Smith JD, Minda JP, Washburn DA. Category learning in rhesus monkeys: A study of the Shepard, Hovland, and Jenkins tasks. Journal of Experimental Psychology: General. 2004; 133:398–414. [PubMed: 15355146]

Smith JD, Murray MJ, Minda JP. Straight talk about linear separability. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1997; 23:659–680.

Smith JD, Redford JS, Haas SM. Prototype abstraction by monkeys (*Macaca mulatta*). Journal of Experimental Psychology: General. 2008a; 137:390–401. [PubMed: 18473665]

Smith JD, Redford JS, Haas SM. The comparative psychology of same-different judgments by humans (*Homo sapiens*) and monkeys (*Macaca mulatta*). Journal of Experimental Psychology: Animal Behavior Processes. 2008b; 34:361–374. [PubMed: 18665719]

Thompson RKR, Oden DL. Categorical perception and conceptual judgments by nonhuman primates: The paleological monkey and the analogical ape. Cognitive Science. 2000; 24:363–396.

Vauclair, J. Categorization and conceptional behavior in nonhuman primates. In: Bekoff, M.; Allen, C., editors. The cognitive animal: Empirical and theoretical perspectives on animal cognition. Cambridge, MA: MIT Press; 2002. p. 239-245.

von Fersen L, Lea SEG. Category discrimination by pigeons using five polymorphous features. Journal of the Experimental Analysis of Behavior. 1990; 54:69–84. [PubMed: 16812617]

Washburn DA, Rumbaugh DM. Testing primates with joystick-based automated apparatus: Lessons from the Language Research Center's Computerized Test System. Behavior Research Methods, Instruments, and Computers. 1992; 24:157–164.

Wasserman EA, Kiedinger RE, Bhatt RS. Conceptual behavior in pigeons: Categories, subcategories, and pseudocategories. Journal of Experimental Psychology: Animal Behavior Processes. 1988; 14:235–246.

White KG, Alsop B, Williams L. Prototype identification and categorization of incomplete figures by pigeons. Behavioural Processes. 1993; 30:253–258. [PubMed: 24896949]
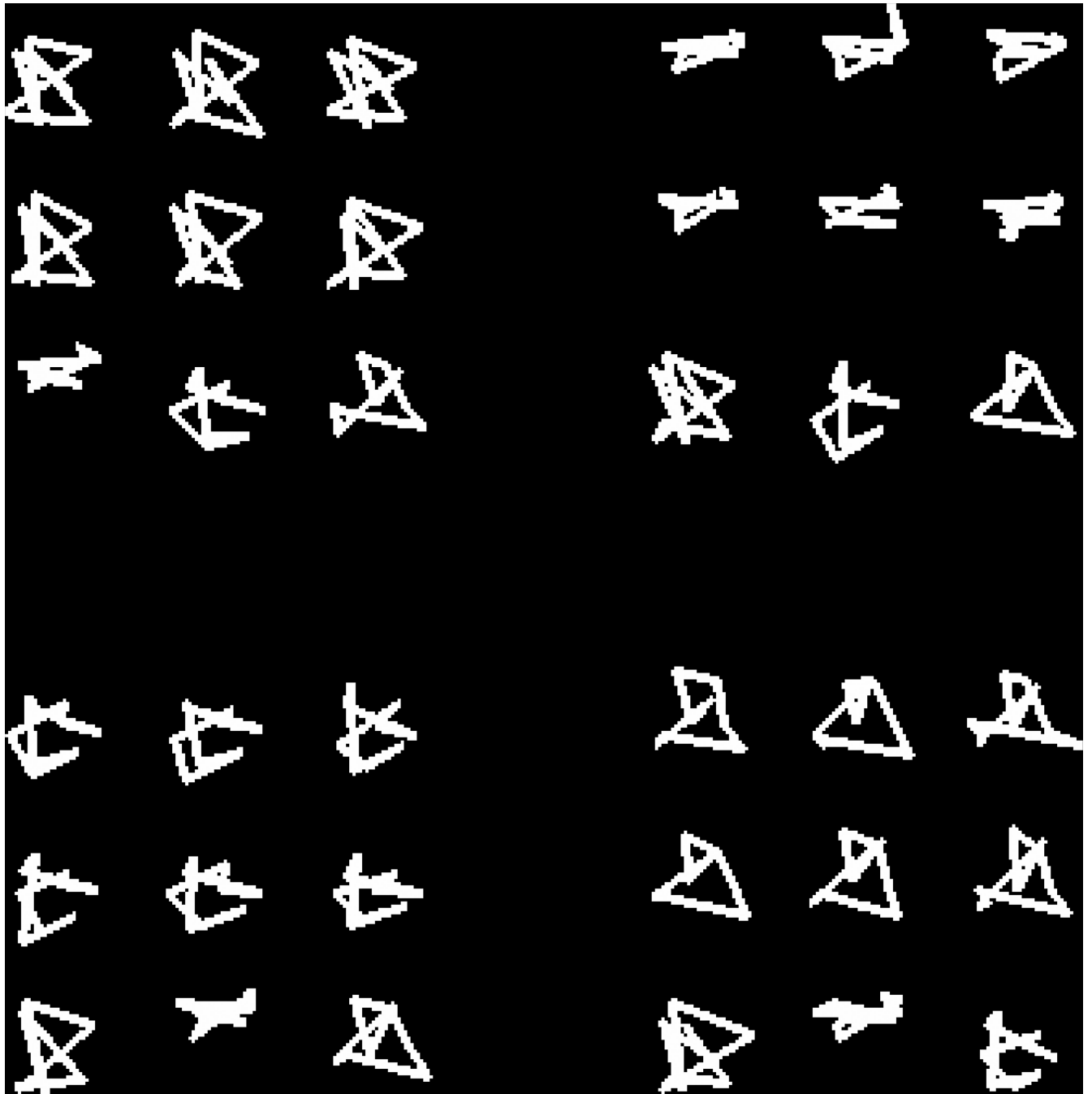
Figure 1.
Examples of Experiment 1A's categories and stimuli. Each nine-shape grouping is a
category. The six shapes in the top and middle row of each category were derived from the
same prototype. The bottom row contains exception items that were derived from the
prototypes of the three opposing categories.
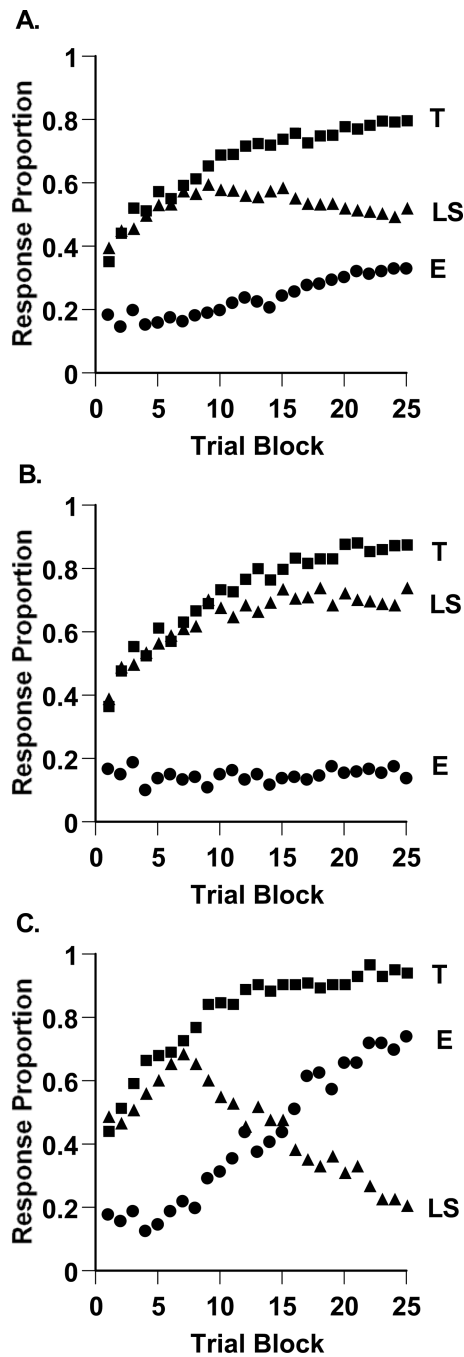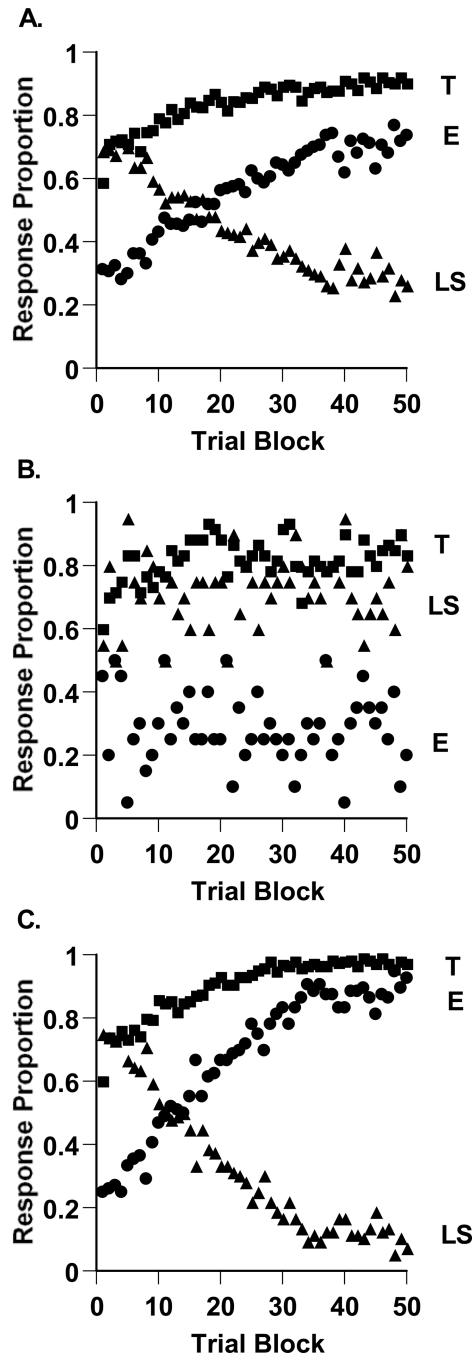
Figure 2.
A. The performance of all participants in the four-category task of Experiment 1A. Curves T and E, respectively, show the proportion of correct responses made to the six typical items and three exception items in each category. Curve LS shows the proportion of exception items placed incorrectly into the category of their originating prototype. *B*. The performance of participants in Experiment 1A whose LS assumption lasted through all 900 trials of the task, depicted in the same way. C. The performance of participants in Experiment 1A who overcame their LS assumption in the task and mastered the exception items.
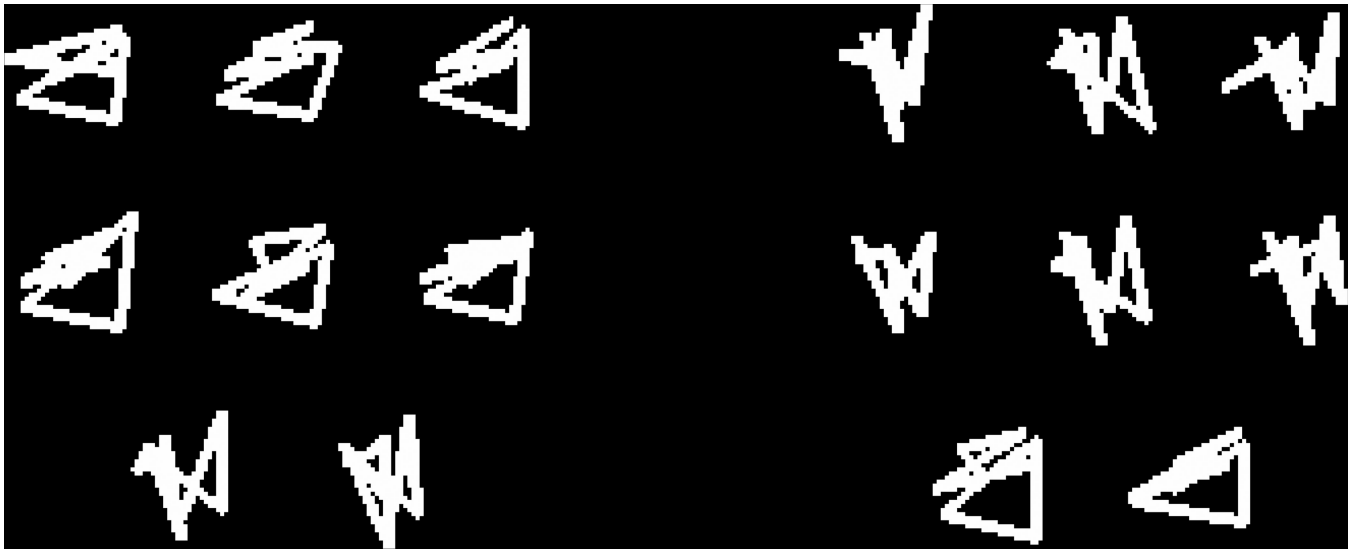
Figure 3.
Examples of the categories and stimuli used in Experiments 1B and 2. Each eight-shape grouping is a category. The six shapes in the top and middle row of each category were derived from the same prototype. The bottom row contains exception items that were derived from the prototype of the opposing category.
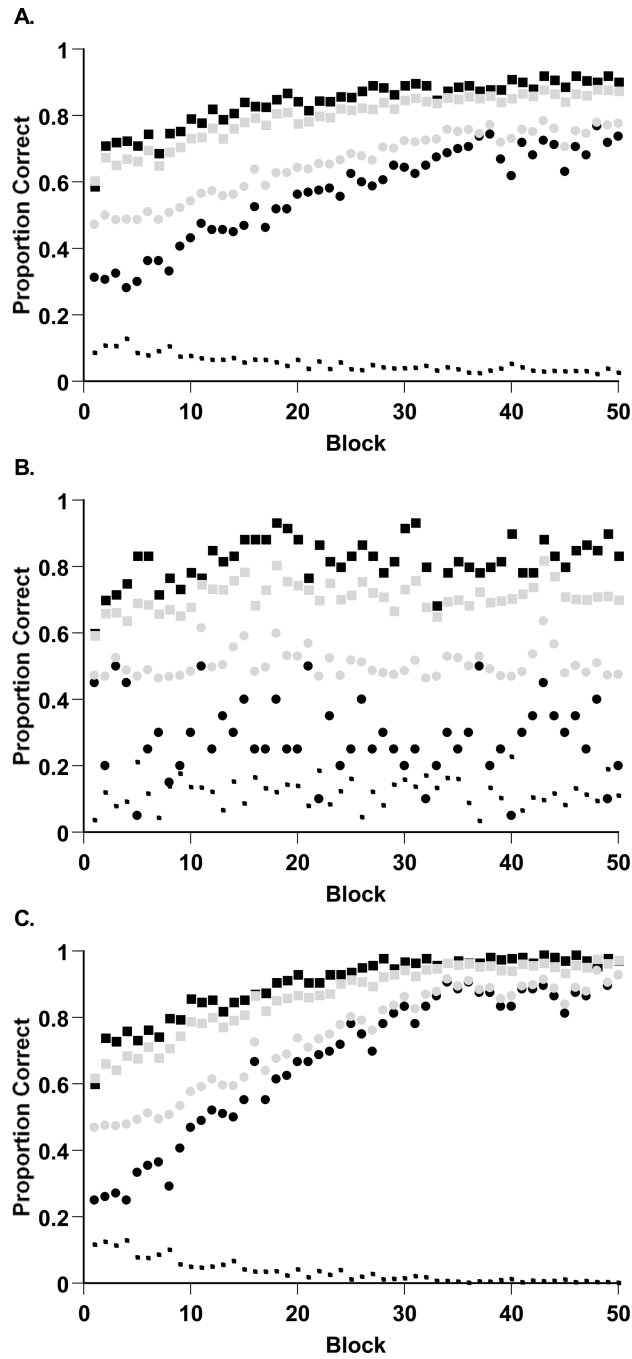
A.

B.

C.

Figure 4.
A. The performance of all participants in the two-category task of Experiment 1B. Curves T and E, respectively, show the proportion of correct responses made to the six typical items and two exception items in each category. Curve LS shows the proportion of exception items placed incorrectly into the category of their originating prototype. *B*. The performance of participants in Experiment 1B whose LS assumption lasted through all 800 trials of the task, depicted in the same way. C. The performance of participants in Experiment 1B who overcame their LS assumption in the task and mastered the exception items.

A.

B.

C.
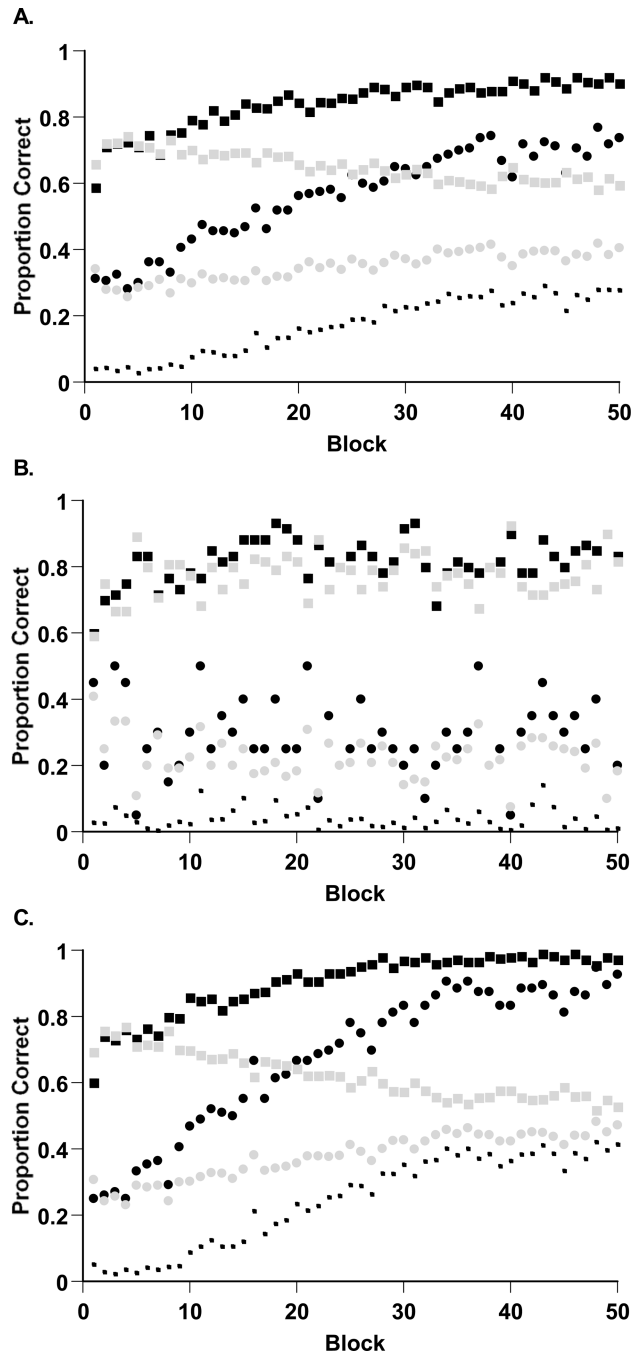
Figure 5.
Results from Experiment 1B when performance data from all humans (A), LS-constrained humans (B), and LS-transcending humans (C) were fit with a formal prototype model: observed typical-item performance (black squares); observed exception-item performance (black circles); model-estimated typical-item performance (gray squares); model-estimated exception-item performance (gray circles); fit index (pepper sprinkles).
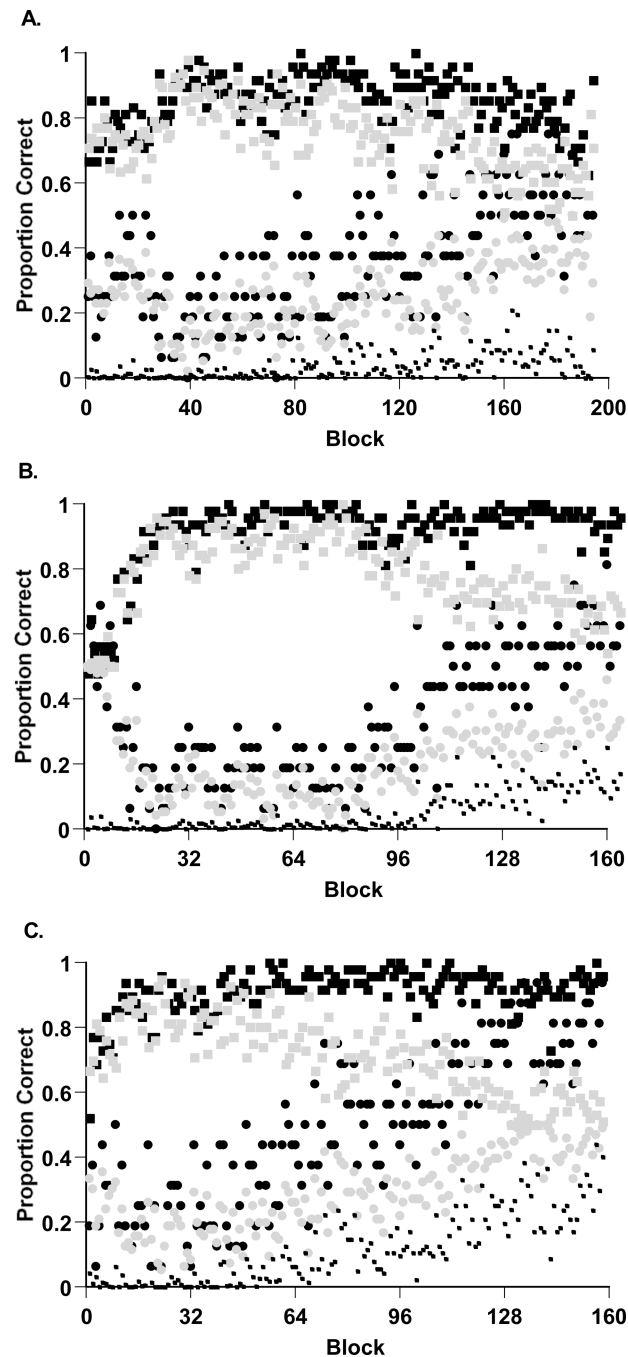
Figure 6.
Results from Experiment 1B when performance data from all humans (A), LS-constrained humans (B), and LS-transcending humans (C) were fit with a formal exemplar model: observed typical-item performance (black squares); observed exception-item performance (black circles); model-estimated typical-item performance (gray squares); model-estimated exception-item performance (gray circles); fit index (pepper sprinkles).
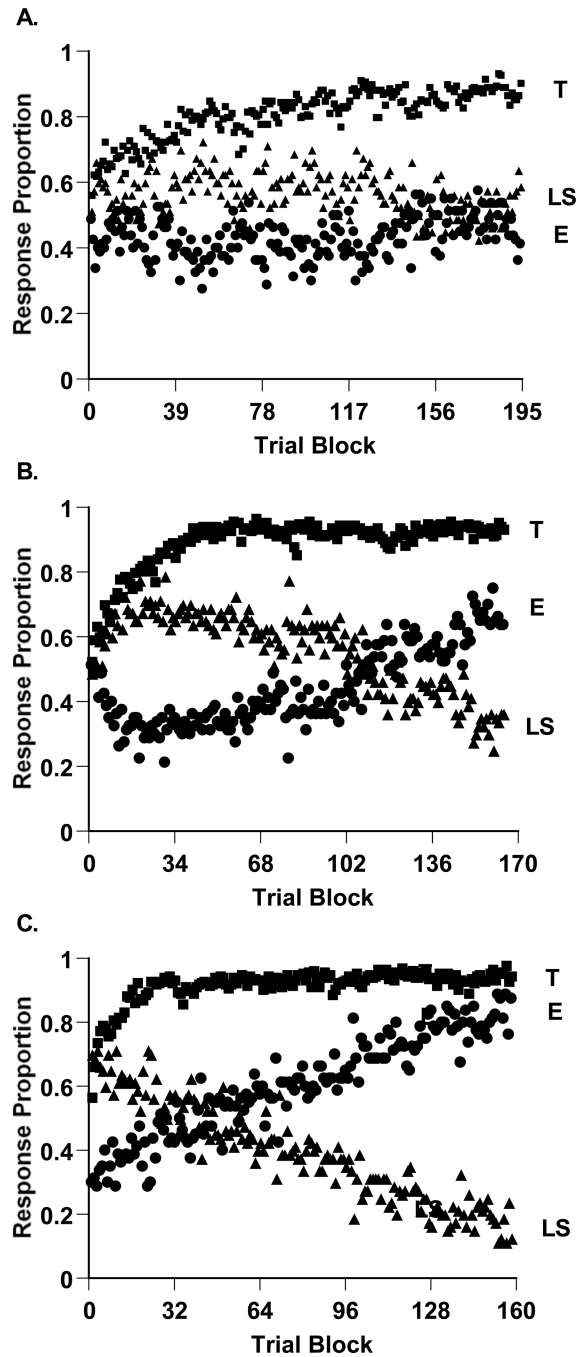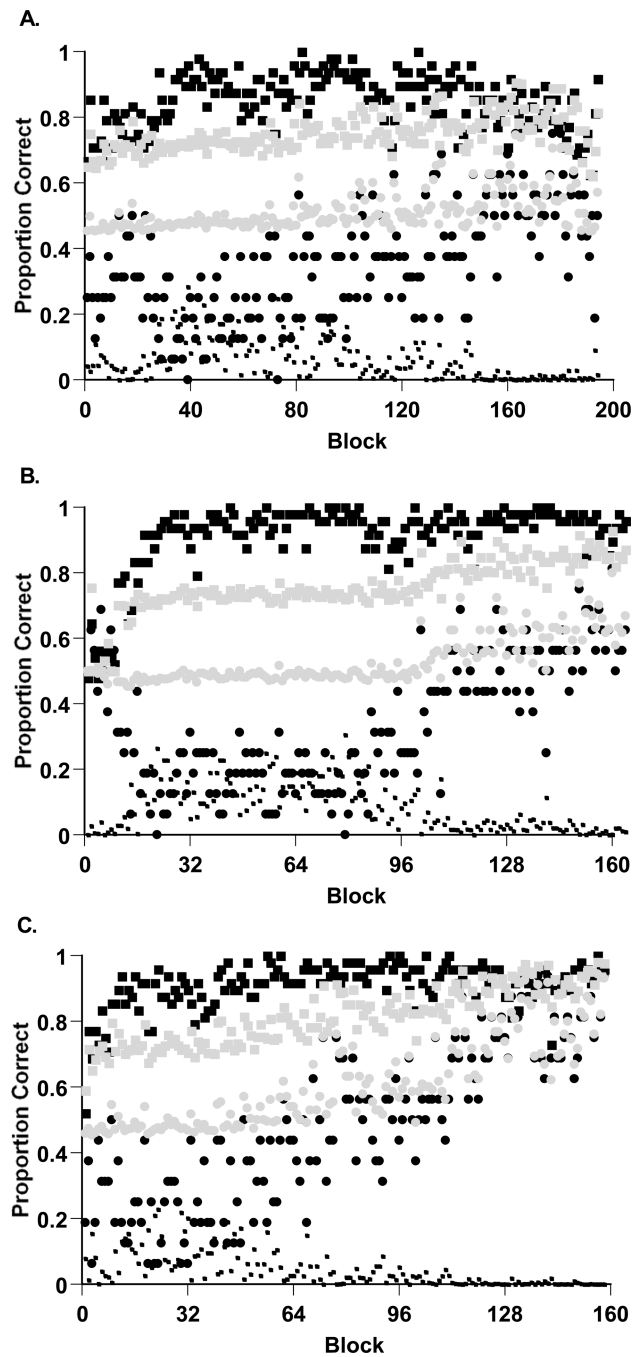
Figure 7.

A. The performance of monkey Hank in the two-category task of Experiment 2. Curves T and E, respectively, show the proportion of correct responses made to the six typical items and two exception items in each category. Curve LS shows the proportion of exception items placed incorrectly into the category of their originating prototype. *B*. The performance of monkey Lou, depicted in the same way. *C*. The performance of monkey Murph, depicted in the same way.
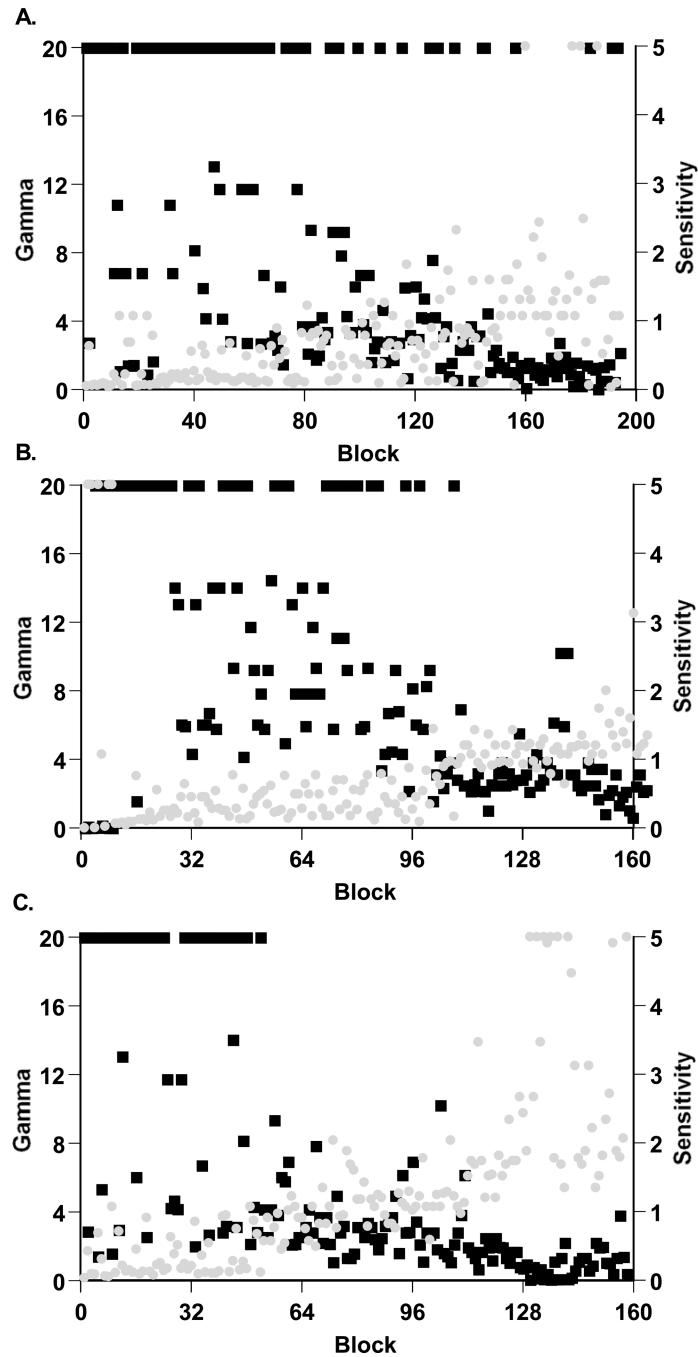
A.

B.

C.

Figure 8.
Results when Hank's (A), Lou's (B), and Murph's (C) performance data from task 4 were fit with a formal prototype model: observed typical-item performance (black squares); observed exception-item performance (black circles); model-estimated typical-item performance (gray squares); model-estimated exception-item performance (gray circles); fit index (pepper sprinkles).

Figure 9.
Results when Hank's (A), Lou's (B), and Murph's (C) performance data from task 4 were fit with a formal exemplar model--observed typical-item performance (black squares); observed exception-item performance (black circles); model-estimated typical-item performance (gray squares); model-estimated exception-item performance (gray circles); fit index (pepper sprinkles).

Figure 10.
Estimates across trial blocks of the gamma parameter (black squares) and sensitivity parameter (gray circles) when Hank's (A), Lou's (B), and Murph's (C) performance data from task 4 were fit with a formal gamma model. To scale the range of the sensitivity parameter for clarity and visibility, 4, 5, and 7 instances of sensitivity estimated > 5.0 were plotted at 5.0 in the three graphs, respectively.