Journal of Experimental Psychology:
Learning, Memory, and Cognition
1994, Vol. 20, No. 6, 1264-1282

0278-7393/94/$3.00

Models of the Effects of Prior Knowledge on Category Learning

Evan Heit

When people learn about a new category, they are influenced by prior knowledge of other categories. In 5 experiments, subjects made categorization judgments after observing descriptions of persons from a location referred to as *City W.* In these experiments, prior knowledge as well as observations within City W were manipulated. The *integration, weighting,* and *distortion models* of categorization explain prior knowledge effects in different ways. The integration model, which assumes that categorization is influenced by prior examples from other categories, predicted the results of the experiments. It was found that the effect of *prior knowledge* was independent of the observed proportion of category membership in City W, that the prior knowledge effect was diminished with more observations, and that learning about City W affected subjects' judgments about the general population. The weighting and distortion models could not account for all of the results.

When people learn about a new category, they are influenced not only by observed members of this category but also by prior knowledge of other, related categories. For example, consider someone who moves to Chicago after having lived in another midwestern state. This person would be learning about the members of novel categories, such as *Chicago roads, Chicago stores,* and *Chicago restaurants.* As this person crosses into Chicago for the first time, strictly speaking this person would have no prior direct experience with members of these categories. Of course, the person would not have much trouble on Chicago roads or in Chicago stores because prior knowledge about established categories, such as *Michigan roads,* could be used to guide learning and inferences about the new categories. However, beliefs about the new categories would also be sensitive to observations, so that, for example, after months of experience on Chicago roads, what is known about this category reflects real experiences with roads in Chicago and not simply what had been expected about this category. In addition, learning about categories in this new context may transfer back to what is believed about other categories. For example, after the newcomer to Chicago experiences members of the category *Chicago restaurants,* this person may revise beliefs about the generic category *restaurants* and even revise what is believed about other context-specific categories such as *Michigan restaurants.*

Learning about categories in the context of a new location is a special case of the much more general phenomenon of category learning when prior knowledge is available. A complete psychological account of category learning should be able to address three issues. First, category learning, even category learning in the context of a psychology experiment, is influenced by prior knowledge of other categories. Second, category learning is sensitive to observations of category members. Third, learning about a new category may lead to revisions in what is believed about other categories. This third issue has not received much attention in research on category learning, although the phenomenon of general knowledge being updated after specific experiences has been studied in the domain of memory (e.g., Anderson & Ross, 1980; Brown & Siegler, 1993; Dosher & Rosedale, 1991; Potts, St. John, & Kirson, 1989). However, the first two issues have been addressed in much categorization research.

Previous research has addressed the first issue, that prior knowledge affects category learning: by argument and by experimental demonstration. Murphy and Medin (1985) argued that psychological accounts of categorization that ignore the influence of prior knowledge are incomplete because the categories people form cannot be predicted only from what people observe. Many recent studies have demonstrated the influence of prior knowledge on category learning (Barrett, Abdi, Murphy, & McCarthy Gallagher, 1993; Hayes & Taplin, 1992; Murphy & Wisniewski, 1989; Pazzani, 1991; Wattenmaker, Dewey, Murphy, & Medin, 1986; Wisniewski & Medin, 1991, 1994; see Murphy, 1993, for a review). For example, Wattenmaker et al. showed that prior knowledge of occupations helped subjects learn about novel categories of occupations. Together, these studies have made it clear that it is easier to learn new categories that are consistent with prior knowledge than categories that are inconsistent with prior knowledge. In addition, beliefs about these newly formed categories reflect knowledge from outside of these categories.

Much research on categorization has addressed the second issue, that category learning reflects what is observed. However, research on this second issue typically has not simultaneously addressed the first issue. Psychological models of categorization have been applied mainly to studies of category learning in isolated contexts (J. R. Anderson, 1990, 1991; Ashby & Gott, 1988; Estes, 1986; Gluck & Bower, 1988; Heit,

This research was supported by a National Institute of Mental Health Individual Postdoctoral Fellowship to Evan Heit and by National Science Foundation Grant 91-10245 to Douglas Medin. Part of this research was presented at the 1993 meeting of the Psychonomic Society in Washington, DC. I am grateful to Dorrit Billman, James Hampton, Caren Jones, Douglas Medin, Lance Rips, and Edward Wisniewski for comments on this research and to Seema Shastri for assistance in developing the experimental materials.

Correspondence concerning this article should be addressed to Evan Heit, Department of Psychology, Northwestern University, 2029 Sheridan Road, Evanston, Illinois 60208. Electronic mail may be sent to heit@nwu.edu.

1992; Kruschke, 1992; Medin & Schaffer, 1978; Nosofsky, 1988). Typically in these studies, subjects learned categories that were intended by the experimenter to be as unrelated as possible to prior knowledge (e.g., categories of geometric figures or fictional diseases). Although these modeling efforts have been quite successful, they have not focused on the effects of prior knowledge on category learning. The present research has the goal of developing these models further to account for the learning of new categories that make contact with old categories.

Three Theories of Prior Knowledge Effects

Illustration of the different ways that category learning might be influenced by prior knowledge can be done by considering the learning of the category *people in Chicago who jog regularly.* Assume that in previous contexts, *owns expensive running shoes* has served as a reliable cue for categorizing people as joggers or nonjoggers. A newcomer to Chicago meets several people who together do not embody the expected relation between expensive running shoes and jogging, as shown in the first section of Table 1. The next time the newcomer meets someone with expensive running shoes in Chicago, the newcomer wants to infer whether this person is a jogger. The newcomer must now rely on two sources of knowledge: what was expected of people in Chicago with expensive running shoes, and what has been directly observed of people with this description. It is likely that the newcomer will infer that a person with expensive running shoes is a jogger, despite observations in this context that half the people with expensive running shoes are joggers and half are not joggers.

Expectations might influence categorization in at least three distinct ways (Heit, 1993). First, according to *integration theory,* a categorization judgment depends on retrieving two kinds of examples from memory: observed members of the category as well as *prior examples,* members of related categories that serve as expectations for the new category. As illustrated in Table 1, the newcomer might retrieve additional descriptions of joggers and nonjoggers from outside of the context, for example, remembering three joggers from Michigan who own expensive running shoes. The newcomer would categorize someone with expensive running shoes as a jogger because most of the retrieved joggers fit this description. More generally, prior examples affect categorization in the same manner as observed examples, but the prior examples come from other sources such as observed members of other categories. In addition, the prior examples could be salient fictional examples or idealizations, such as what is remembered from stories and media sources. In the case of joggers, the prior examples could be remembered from advertisements for running shoes. The claims of integration theory are supported by research on source monitoring and reality monitoring (Johnson, Hashtroudi, & Lindsay, 1993; Johnson & Raye, 1981). This research has shown that people's judgments about what they remember are influenced not only by real observations but also by imagined examples and observations from other contexts. Integration theory is also related to information integration theory (N. H. Anderson, 1991; Busemeyer, 1991), which has investigated ways that expectations are combined with new observations to form a judgment. It is also related to Bayesian approaches to revising estimates of proportions (see the General Discussion section).

The second way that prior knowledge might influence category learning is referred to as *weighting theory.* According to this theory, prior knowledge facilitates the learning of category members that meet expectations and filters away unexpected observations so that observed category members that are congruent with prior knowledge have a greater effect on categorization than incongruent category members. As illustrated in Table 1, the congruent observations (joggers who own expensive running shoes) could have three times the weight of incongruent observations (joggers without expensive running shoes). Thus the newcomer will categorize persons in Chicago with expensive running shoes as joggers. Weighting theory is intended to represent the outcome of a range of possible sources of selective weighting, for example, processes at encoding or retrieval. Weighting theory is a component of schema theory (Alba & Hasher, 1983), which has described several ways that knowledge may influence memory and judgment, including leading people to ignore or reject anomalous data (Chinn & Brewer, 1993). Also consistent with weighting theory are results showing confirmation biases, that people seek out information consistent with what they expect (Mynatt, Doherty, & Tweney, 1977). Finally, other proposed accounts of the effects of prior knowledge on category learning (Murphy & Medin, 1985; Murphy & Wisniewski, 1989) have suggested that one of the effects of background knowledge is to lead people to focus on co-occurrences of features that are expected to co-occur.

Third, according to *distortion theory,* what is observed is distorted to make it more congruent with prior knowledge. As illustrated in Table 1, some incongruent observations of joggers (those without expensive running shoes) could be misremembered as being congruent (owning expensive running shoes). Thus the newcomer will be likely to categorize someone with expensive running shoes as a jogger because the newcomer remembers joggers from this context as having expensive running shoes. Distortion processes might take a variety of forms, including processes that occur during encoding or retrieval. Distortion theory is a component of schema

Table 1
Illustration of Models

Description	No. of joggers	No. of nonjoggers
Observed examples		
Expensive running shoes	2	2
No expensive running shoes	2	2
Integration theory		
Expensive running shoes	3 + 2 = 5	2
No expensive running shoes	2	3 + 2 = 5
Weighting theory		
Expensive running shoes	3 × 2 = 6	2
No expensive running shoes	2	3 × 2 = 6
Distortion theory		
Expensive running shoes	2 + 1 = 3	2 − 1 = 1
No expensive running shoes	2 − 1 = 1	2 + 1 = 3

theory (e.g., Taylor & Crocker, 1978), which has suggested that background knowledge may have the effect of leading us to rely on default information when observations are poorly remembered. Likewise, schematic knowledge may be used to reinterpret anomalous data (Chinn & Brewer, 1993). In addition, distortion theory is related to Asch's (1946) change of meaning hypothesis, which claimed that the interpretation of new information depends on what is already known. Finally, distortion processes have already been proposed as an account of some of the effects of prior knowledge on category learning (Wisniewski & Medin, 1991, 1994), particularly for cases of ambiguous features being interpreted as congruent with expectations.

Integration, weighting, and distortion theories are fairly broad accounts for explaining prior knowledge effects on category learning; each of these theories may have several possible instantiations at the processing level. Although these accounts are intended to be applicable to many situations, the present research focuses on cases in which the three general theories make distinct predictions. For example, if integration theory correctly predicts the results of a set of experiments for which weighting and distortion theories make incorrect predictions, then such results would suggest that an integration process is taking place and would cast doubts on the occurrence of the other kinds of processes. The predictions of the three theories are inferred by implementing them as variants of a mathematical model of categorization.

Categorization Models

The new categorization models presented in this article take exemplar theory as a starting point. Exemplar models assume that categorization depends on the similarity of a test stimulus to category exemplars retrieved from memory (Estes, 1986; Heit, 1992; Medin & Schaffer, 1978; Nosofsky, 1988). These models have been successful in terms of accounting for the results of many studies. (See Nosofsky, 1992, for a review, and see Brooks, 1978, 1987, for additional arguments on the psychological plausibility of exemplar theory.) Standard exemplar models are based on Equation 1, which describes classifying a stimulus, x, as a member of either Category A or Category B.

$$P(\text{classify x as A}) = \frac{\text{fam}_A(x)}{\text{fam}_A(x) + \text{fam}_B(x)} \quad (1)$$

Categorization depends on the underlying psychological construct of familiarity. The familiarity of Stimulus x is evaluated with respect to the members of the two categories. The likelihood of categorizing x in Category A increases with fam_A, the familiarity of x with respect to Category A, and decreases with fam_B, the familiarity of x with respect to Category B. In Equation 2a, the familiarity of Stimulus x with respect to Category A is the sum of the similarities of x to each member of Category A retrieved from memory. Likewise, in Equation 2b, $\text{fam}_B(x)$ is the sum of similarities of x to each member of Category B retrieved from memory. Here, the similarity between two identical stimuli, sim(x, x), is assigned the value of 1, and the similarity between x and a mismatching stimulus, y, sim(x, y), is assigned a fractional value s, where $0 < = s < 1$.[1]

$$\text{fam}_A(x) = \sum_{a \epsilon A} \text{sim}(x, a) = \text{no. of x in A} + s\,(\text{no. of y in A}) \quad (2a)$$

$$\text{fam}_B(x) = \sum_{b \epsilon B} \text{sim}(x, b) = \text{no. of x in B} + s\,(\text{no. of y in B}) \quad (2b)$$

Finally, Equations 2a and 2b may be substituted into Equation 1, leading to Equation 3, which is the basic model applied in this article.

$$P(\text{classify x as A}) = \frac{\text{no. of x in A} + s(\text{no. of y in A})}{\text{no. of x in A} + s(\text{no. of y in A}) + \text{no. of x in B} + s(\text{no. of y in B})} \quad (3)$$

As these models are commonly interpreted when applied to an experiment, categorization depends only on memory traces from the experimental context. That is, the numbers of Stimuli x and y retrieved from Categories A and B depend only on what was observed during the experiment. Next, three new models, which incorporate prior knowledge by means of integration, weighting, or distortion are presented. These models are illustrated for the case in which, according to prior knowledge, Stimulus x is congruent with Category A and Stimulus y is congruent with Category B. For example, x might correspond to *owns expensive running shoes,* and Category A might correspond to *joggers.* Feature y would correspond to *does not own expensive running shoes,* and Category B would correspond to *nonjoggers.*

First, according to the integration model, the Categories A and B are initially filled with prior examples. It is assumed here that the number of prior examples for Category A, as well as the number of prior examples for Category B, are each equal to some positive number G. The number of retrieved examples for a category is the actual number of observed examples plus the number of prior examples. Therefore, for Equations 2a and 3, the number of stimuli retrieved from Category A with Description x are the actual number of observations of x in Category A plus G. Likewise, for Equations 2b and 3, the number of y stimuli retrieved from Category B are the number of observed examples of y in Category B plus G. It is assumed that there are no prior examples of incongruent stimuli, so the number of y retrieved from Category A and the number of x retrieved from Category B are in each case be the number that was observed.

[1] This implementation of the sim function is a special case of the multiplicative similarity rule of Medin and Schaffer (1978) for stimuli described by a single feature. The free parameter s measures the ability of a subject to discriminate between stimuli retrieved from memory. When s is 0, memory discrimination is perfect, and mismatching stimuli have no effect on categorization. The more general version of this multiplicative similarity rule would be used for stimuli with multiple features (as in Heit, 1993).

Second, according to the weighting model, a single observation of a congruent stimulus has some multiplier, W, times the influence of a single incongruent stimulus, where $W > 1$. This claim can be implemented in Equations 2a, 2b, and 3 by assuming that the number of x descriptions retrieved from Category A, as well as the number of y descriptions retrieved from Category B, is in each case the observed number multiplied by W. For the incongruent category members, the retrieved numbers equal the observed numbers.

Third, according to the distortion model, a fixed proportion, D, of the incongruent observations are misremembered as congruent instances. A proportion, D, of the observations of y in Category A are misremembered as congruent, either as stimuli with Description y in Category B or as stimuli with Description x in Category A. Likewise, the number of retrieved members of Category B with description x are less than the observed number because a proportion of these incongruent stimuli are distorted into congruent stimuli. In addition, the number of congruent stimuli, x in Category A and y in Category B, are greater than the observed number, when Equations 2a, 2b, and 3 are applied.

The differing predictions of these models are described in detail as the experiments in this article are presented.

Overview of Experiments

Subjects were taught about familiar categories in a new context so that prior knowledge and current observations could both be manipulated. The stimuli were person descriptions in a context referred to as *City W.* In Experiments 1, 2, 3, and 4, subjects made transfer judgments categorizing additional persons from City W. For Experiments 1 and 2, the integration and weighting models made different predictions about the joint influence of prior knowledge and knowledge of City W. Experiments 3 and 4 were intended to distinguish between predictions of the integration and distortion models. In Experiments 4 and 5, subjects made categorization judgments from general knowledge rather than specifically on the basis of City W. These experiments addressed the integration model's predictions of how general knowledge about categories is updated after recent observations.

Experiment 1

In this experiment, subjects observed a set of descriptions of persons in a novel context, City W. These descriptions consisted of two features; for example, a person might be described as shy and often attending parties. In the test phase, subjects made categorization judgments from partial descriptions, such as how likely a person in City W who attends parties often would be in the category *shy.* Half of the test questions involved a description that was incongruent with a category, such as *attends parties often* and *shy,* and half had a description that was congruent with a category, such as *does not attend parties often* and *shy.* In addition, the observed category membership referred to by these test questions was systematically manipulated, for example, the observed proportion of frequent attendees of parties in City W who were shy was varied.

Method

Subjects. Thirty-eight Northwestern University undergraduates participated; they received course credit or a small payment. No subject participated in more than one study reported in this article; all subjects in these studies were recruited in the same manner.

Stimuli. Each subject saw training examples derived from the descriptive terms shown in Table 2. Each couplet of four features consists of two pairs of opposites or complements. For example, *not shy* is the complement of *shy,* and *does not attend parties often* is the complement of *attends parties often.* The first and third item in each couplet were congruent with each other (e.g., *shy* and *does not attend parties often*), likewise the second and fourth item were congruent. The first and fourth items, as well as the second and third items, were incongruent (e.g., *shy* and *attends parties often*). The stimuli were pretested on other subjects, as described in Appendix A, to validate this manipulation of prior knowledge.

Each training example was a description of a person, in terms of two features. A pairing of two features was either congruent or incongruent. In Experiment 1, the features from each of the 10 couplets in Table 2 were used for 20 person descriptions, thus there were 200 descriptions in total. Features from two different couplets never appeared in the same person description. The 10 couplets were assigned randomly for each subject to parts of the structure in Table 3. For example, for some subjects, the person descriptions referring to shyness and parties contained 0% congruent pairings and for other subjects, person descriptions referring to shyness and parties contained 20%, 50%, 80%, or 100% congruent pairings. For example, when the shyness–parties couplet was assigned to the 20% congruent condition, there were 4 congruent descriptions: two persons who were shy and did not attend parties often, and two persons who were not shy and attended parties often. The remaining 16 descriptions were incongruent: eight persons who were shy and attended parties often, and eight persons who were not shy but did attend parties often. (Note that several persons could have the same featural description.)

Table 2
Feature Couplets Presented in Experiments 1–5

Couplet	
Shy	Stubborn
Not shy	Not stubborn
Does not attend parties often	Frequently gets into arguments
Attends parties often	Does not get into arguments frequently
Jogs regularly	
Does not jog regularly	Mechanically inclined
Owns expensive running shoes	Not mechanically inclined
Does not own expensive running shoes	Fixes things as a hobby
	Does not fix things as a hobby
Travels two or more times per year	Generous
Travels less than two times per year	Not generous
Has frequent flyer number	Donates to charity
Does not have frequent flyer number	Does not donate to charity
	Usually happy
Has a college degree	Usually sad
Does not have a college degree	Smiles more than average
Higher income than average	Smiles less than average
Lower income than average	
	Watches more TV than average
Often buys expensive wine	Watches less TV than average
Does not buy expensive wine often	Reads books less than average
Often buys gourmet food	Reads books more than average
Does not buy gourmet food often	

Table 3
Stimulus Structure for Training Examples in Experiment 1

No. of couplets	Presentations per couplet	Category size	Proportion of congruent pairings (%)
2	20	10	0
2	20	10	20
2	20	10	50
2	20	10	80
2	20	10	100

The test stimuli had a two-factor design. Each test question was a conditional probability judgment, referring to the probability of one feature given another feature. The first experimental variable was whether the two features were congruent or incongruent with each other. The second variable was the conditional probability of presentation during the study phase: 0%, 20%, 50%, 80%, or 100%. Eight test questions were derived from each couplet, thus there were 80 test questions.

Procedure. Each subject saw information displayed on a computer screen. The procedure consisted of three parts. First, subjects were given the opportunity to familiarize themselves with the stimuli. The features shown in Table 2 were displayed together for 3 min, in a different random order for each subject. The subjects were simply instructed to read the stimuli. The purpose of this first phase was to make reading easier in later phases by assuring that none of the material would be unfamiliar.

Next, in the training phase, each subject saw 200 person descriptions displayed in a random order, one at a time for 3.5 s. There was a brief interval between displays (0.2 s) during which the computer screen was cleared. The person descriptions were worded as follows:

> A person from City W has this description:
> x1
> x2,

where x1 and x2 were two features. The order of these two features was determined randomly for each display. Before the study phase began, subjects were told that they would see a set of descriptions of persons living in City W, a city located in Illinois. The subjects were instructed to pay attention to these descriptions, and they were informed that they would later be tested on this information. The training phase was followed by a 2-min unfilled retention interval.

Finally, in the test phase, subjects made 80 conditional probability estimates. The test questions for each subject were presented in a random order. These questions were worded as follows:

> Consider a person from City W with the following characteristic: x
> How likely is it that this person would also have this characteristic? A,

where x and A were two features. Subjects responded by typing integers on a scale from 0% to 100%. Subjects were told to base their answers on what they inferred to be true of persons in City W after seeing descriptions of some of the citizens of City W. The entire procedure typically lasted half an hour.

Results

The results of Experiment 1, in terms of the average response for each type of test question, are shown in Figure 1a. This figure shows that subjects were influenced by the prior knowledge because the congruent judgments are greater than the incongruent judgments. In addition, subjects were influenced by what they observed of City W, as indicated by the positive slopes of both the congruent and incongruent lines. Furthermore, it appears that the effect of prior knowledge, that is, the difference between the congruent and incongruent conditions, is independent of the observed proportion of category membership.

A two-way analysis of variance (ANOVA) of the judgments confirmed these observations, showing a main effect of congruent versus incongruent test stimuli, $F(1, 37) = 20.39, p < .001$, $MS_e = 1{,}148$, and a main effect of observed proportion, $F(4, 148) = 90.75, p < .001$, $MS_e = 422$. The interaction between these two variables was not found to be significant, $F(4, 148) = 2.01$, $MS_e = 95$. Thus, the effect of prior knowledge was not found to depend on the observed proportion.

Next, the integration, weighting, and distortion models were evaluated quantitatively. Equation 3 was applied to predict the average probability judgment in each condition, from the numbers of retrieved examples according to each model. The predictions of the models are described in Table 4. For a test question involving the categorization of Stimulus x in Category A, the number of observed examples of xA is referred to as Np in the second column of the table, where N is the total number of members of Category A, and p is the proportion of members of Category A observed to have Description x. (In Experiment 1, N was always 10, as indicated in Table 3.) The proportion of members of Category A with the Description y was $(1 - p)$, so there were $N(1 - p)$ observed examples of y in Category A. Because of the balanced design, the number of x in Category B was also $N(1 - p)$ and the number of y in Category B was Np.

In Table 4, the test questions are consistently described in terms of evaluating the probability of Category A given Feature x. For congruent test questions, x is congruent with

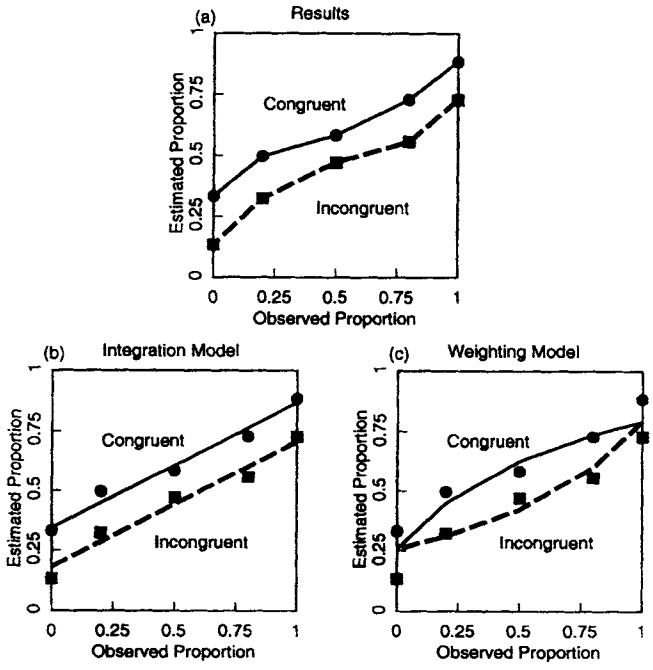

Figure 1. Results of Experiment 1 and predictions of integration and weighting models.

Table 4
Predictions of Models for Congruent and Incongruent Test Questions

Stimulus	Observed no.	Retrieved no.		
		Integration model	Weighting model	Distortion model
Congruent test item: Evaluate $P(A \mid x)$ when xA and yB were congruent				
xA	Np	$Np + G$	WNp	$N[p + D(1-p)]$
yA	$N(1-p)$	$N(1-p)$	$N(1-p)$	$N(1-D)(1-p)$
xB	$N(1-p)$	$N(1-p)$	$N(1-p)$	$N(1-D)(1-p)$
yB	Np	$Np + G$	WNp	$N[p + D(1-p)]$
Incongruent test item: Evaluate $P(A \mid x)$ when xB and yA were congruent				
xA	Np	Np	Np	$N(1-D)p$
yA	$N(1-p)$	$N(1-p) + G$	$WN(1-p)$	$N(1-p+Dp)$
xB	$N(1-p)$	$N(1-p) + G$	$WN(1-p)$	$N(1-p+Dp)$
yB	Np	Np	Np	$N(1-D)p$

Category A, and y is congruent with Category B. For example, when a subject judged the probability that a shy person does not attend parties often, x refers to *shy,* Category A refers to *does not attend parties often,* y refers to *not shy,* and Category B refers to *attends parties often.* For incongruent test questions, x is congruent with Category B and y is congruent with Category A. For the incongruent test question concerning the probability that a shy person attends parties often, x refers to *shy,* Category A refers to *attends parties often,* y refers to *not shy,* and Category B refers to *does not attend parties often.*

The third column of Table 4 shows the predicted number of retrieved examples for the integration model. The number of congruent examples retrieved is the observed number plus G, the number of prior examples. G is a free parameter of the integration model. For congruent test questions, the top of the table shows that G prior examples of xA and yB are retrieved. Likewise, for incongruent test questions, prior examples of xB and yA would be retrieved, as indicated at the bottom.[2] The fourth column of Table 4 shows the predictions for the weighting model for congruent and incongruent questions. Here the number of retrieved congruent examples is W times the number of observed congruent examples, where W is a free parameter of the weighting model. Finally, the fifth column shows the predictions of the distortion model, where D, the distortion rate, is a free parameter of this model. The number of retrieved incongruent examples is the number of observed incongruent examples decremented by the proportion D. This distortion of incongruent examples leads to an increase in the retrieved number of congruent examples, as shown.

All three models predict an average judgment of 50% for this experiment because the test questions were organized in complementary pairs. However, the average judgment was 52.4%; this result may be attributed to a slight lack of calibration by the subjects.[3] In the interest of giving each model a chance to succeed, a bias value .024 was added to each model's predicted probabilities. In addition, Equation 3 has a free parameter, s, that refers to the ability of subjects to discriminate mismatching stimuli retrieved from memory.

When the integration model was applied to the results, the parameters s and G were estimated by minimizing the squared differences between the model's predictions and the average responses by the subjects (Chandler, 1965). The best-fitting value of s was .19, and the best value of G was 3.11. Interpreted within the framework of the integration model, the average subject retrieved about three prior examples for each category. Figure 1b shows the integration model's predictions overlaid on the average responses, where the lines indicate the model and the points indicate the data. The integration model predicts the qualitative result that the effect of prior knowledge is independent of observed proportion. Quantitatively, the average discrepancy between the model and the data, measured by the root-mean-square error, is .0335.

Next, the weighting model was fitted to the results. The best-fitting value of s was .30, and the best value of W was 2.23. Thus, within the framework of the weighting model, congruent category members had about twice the influence of incongruent category members. Figure 1c shows the best predictions of the weighting model overlaid on the data; the correspondence is rather poor. The problem of the weighting model is that it predicts that the effect of prior knowledge depends on the observed proportion of category membership. Quantitatively, the weighting model also performed worse than the integration model. The root-mean-square error of the weighting model was .0651, about twice the error of the integration model.

Finally, the distortion model was applied. The best value of s was found to be .19, and the best value of D was found to be .23. In terms of the framework of the distortion model, about one fourth of the incongruent observations were distorted to become congruent examples. Of special note, the distortion model makes the same predictions as the integration model for Experiment 1. Thus Figure 1b also shows the distortion model's predictions.

[2] Note that the underlying psychological claim is the same regardless of the notation at the top or bottom. For example, subjects are always assumed to have G prior examples in memory with the description of shy and does not attend parties often, but depending on the test question, this description would be denoted in Table 4 as xA, xB, yA, or yB.

[3] This finding is not uncommon; several studies by Heit (1992), with procedures similar to the present studies, also obtained a slight degree of superadditivity (see also Wright & Whalley, 1983).

Discussion

The results of Experiment 1 clearly favored the integration model and the distortion model over the weighting model. The difference between the models is particularly striking at a qualitative level. As predicted by the integration and distortion models, the effect of prior knowledge was independent of what was observed in City W. In contrast, the weighting model made two incorrect predictions. First, the weighting model incorrectly predicted that prior knowledge would have no effect in cases of unmixed evidence, that is, when $p = 0$ or $p = 1$. Second, the weighting model incorrectly predicted that the effect of prior knowledge would be greatest when the evidence is completely mixed, that is, when $p = .5$.

The reason that the integration model predicts that the prior knowledge effect is independent of what was observed is that the number of prior examples added to the retrieved examples is the same regardless of the value of p. In contrast, the weighting model predicts that prior knowledge has no effect when $p = 0$ or $p = 1$ because in these cases the observations were completely consistent, that is, all incongruent examples or all congruent examples. The relative weight of congruent examples to incongruent examples has no impact because there is only one kind of example. These predictions of the integration and weighting models are derived in more detail in Appendix B.

It is perhaps surprising that the distortion model gave the same predictions as the integration model for Experiment 1, given that the two models embody such different psychological claims. As shown in Appendix B, the integration model and the distortion model do not always make the same predictions. However, for the design of Experiment 1, where the category size, N, is fixed, the distortion model is able to mimic the predictions of the integration model. The issue of distinguishing between the integration model and the distortion model is revisited in Experiment 3.

Experiment 2

Experiment 2 was intended to replicate Experiment 1, with a slightly different design. Experiment 2 also included an exploratory attempt to manipulate the accessibility of observed category members by varying the retention interval. This manipulation was not critical to distinguishing between integration and weighting, however.

Method

Experiment 2 was identical to Experiment 1, with the following exceptions. The training stimuli had the structure shown in Table 5. Sixteen descriptions were shown per couplet, for a total of 160 training examples. The 80 test stimuli concerned features that had been observed together 0%, 25%, 50%, 75%, or 100% of the time. Sixty-two subjects participated. Half of these subjects, in the no-delay condition, proceeded from the training phase to the test phase with no retention interval. The remaining subjects, in the delay condition, had a 20-min filled retention interval between the training phase and the test phase. During the interval, these subjects performed an unrelated task, making same-difference judgments between pairs of random letter strings.

Table 5
Stimulus Structure for Training Examples in Experiments 2 and 5

No. of couplets	Presentations per couplet	Category size	Proportion of congruent pairings (%)
2	16	8	0
2	16	8	25
2	16	8	50
2	16	8	75
2	16	8	100

Results and Discussion

A three-way ANOVA on the judgments showed a main effect of congruent versus incongruent test question, $F(1, 60) = 63.61$, $p < .001$, $MS_e = 1{,}313$, and a main effect of observed proportion, $F(4, 240) = 45.38$, $p < .001$, $MS_e = 372$. The interaction between these two variables was not found to be statistically significant, $F(4, 240) = 0.35$, $MS_e = 121$. The main effect of the delay condition versus the no-delay condition was not found to be statistically significant, $F(1, 60) = 2.14$, $MS_e = 582$. Furthermore, the delay variable did not appear significant in any of the interaction terms (all Fs < 1). Therefore, the results of this experiment were analyzed after collapsing the data across the delay and no-delay conditions. Figure 2a shows the results of Experiment 2; the pattern is similar to that of Experiment 1.

Next, the integration and weighting models were evaluated quantitatively. The average judgment was .529, thus a bias parameter of .029 was added to the predictions of each model. For the integration model, the best-fitting parameter values of s and G were .31 and 6.09, respectively. Figure 2b shows the predictions of the integration model superimposed over the data; the correspondence is outstanding. The root-mean-square error is only .0111. As in Experiment 1, the distortion model can mimic the predictions of the integration model, here by assuming $D = .43$.

For the weighting model, the best-fitting parameter values were $s = .50$ and $W = 6.15$. Figure 2c shows the weighting model's predictions; again the weighting model gives a poor account. The root-mean-square error is .0729, over six times the error of the integration model. Because of the two failures of the weighting model, it will not be considered as an account for the remaining experiments.[4]

Experiment 3

The critical difference between the integration model and the distortion model is that the integration model predicts that

[4] It is also interesting to consider a mixed model in which prior examples influence categorization, but congruent instances are also weighted more than incongruent instances. Such a model would predict a compromise between Figures 2b and 2c, with some prior knowledge effect at $p = 0$ and $p = 1$, and an increasing prior effect as p approaches .5. However, there was no evidence in Experiment 1 or Experiment 2 for an interaction between the two factors. Thus the integration model can completely account for the results without an added weighting component.

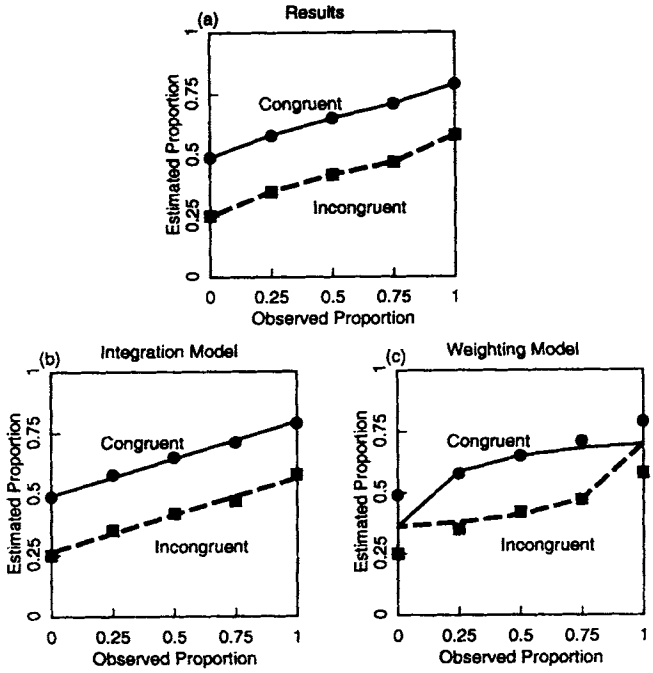

Figure 2. Results of Experiment 2 and predictions of integration and weighting models.

the effect of prior knowledge will be diminished as more category members are observed, but the distortion model predicts no decrease in the effect of prior knowledge as more category members are observed. According to the integration model, early in the course of learning the prior examples have a great influence because there are few observed examples. However, with more observations, the prior examples have relatively less influence on categorization. In contrast, according to the distortion model, a fixed proportion of examples are distorted. These predictions are derived in Appendix B. As an effort to distinguish between the two models, in Experiment 3 the observed category size was manipulated.

Method

Experiment 3 was identical to Experiment 1, with the following exceptions. The training stimuli had the structure shown in Table 6. Eight of the 10 couplets from Table 2 were used, randomly chosen for each subject. Four of these couplets were assigned to the small-category condition; these couplets described only eight citizens of City W. The other four couplets, in the large-category condition, were each used for 32 person descriptions. Thus, 160 person descriptions were presented. In the small-category condition, two couplets were arranged to produce 25% congruent pairings, and two couplets were arranged to produce 75% congruent pairings. For example, when the shyness–parties couplet was assigned to the small category, 25% congruent condition, one person was shy and did not attend parties, one person was not shy and attended parties, three persons were shy and attended parties, and three persons were not shy and did not attend parties. Likewise, in the large-category condition, two couplets were arranged to produce 25% congruent pairings, and two couplets were arranged to produce 75% congruent pairings.

Table 6
Stimulus Structure for Training Examples in Experiment 3

No. of couplets	Presentations per couplet	Category size	Proportion of congruent pairings (%)
2	8	4	25
2	8	4	75
2	32	16	25
2	32	16	75

The 64 test stimuli were organized according to a three-factor design. The first variable was whether the judgment concerned a pair of congruent or incongruent features. The second variable was whether the features referred to a small category or a large category. Finally, the third variable was whether the observed conditional probability for these two features was 25% or 75%.

Forty-three subjects participated.

Results

Figure 3 shows the results of Experiment 3 in both the small-category and large-category conditions. The prior knowledge effect diminished from the small-category condition to the large-category condition, as predicted by the integration model. A three-way ANOVA on the judgments showed a main effect of congruent versus incongruent, $F(1, 42) = 38.73, p < .001$, $MS_e = 782$, and a main effect of observed proportion, $F(1, 42) = 47.30$, $p < .001$, $MS_e = 365$. The interaction between these two variables was not found to be significant, $F(4, 240) = 0.34$, $MS_e = 143$. The main effect of small versus large category was found to be significant, $F(1, 42) = 13.29, p < .001$, $MS_e = 316$. The critical result is that the interaction between category size and congruency was significant, $F(1, 42) = 7.62, p < .01$, $MS_e = 101$, such that prior knowledge had a greater effect in the small-category condition than in the large-category condition. Otherwise, the category size variable did not appear to be significant in interaction terms (all Fs < 1).

Next, the integration and distortion models were evaluated quantitatively. In the small-category condition, the category size, N, was 4, and N was 16 in the large-category condition. The average judgment was .536, so .036 was added to the predictions of each model. For the integration model, the best-fitting parameter values of s and G were .39 and 5.04,

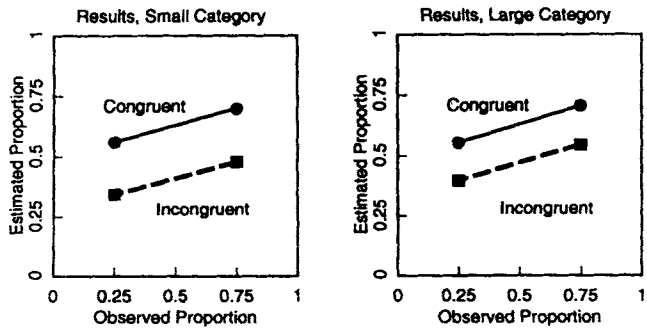

Figure 3. Results of Experiment 3.

respectively. The average discrepancy, in terms of root-mean-square error, was .0299. Figure 4a shows the predictions of the integration model superimposed over the data. Qualitatively, the model does well, in that it correctly predicts the direction of change of the prior knowledge effect. However, the integration model overestimates the impact of the category size manipulation; it predicts too large a prior knowledge effect in the small-category condition and too small an effect in the large-category condition. For the distortion model, the best-fitting parameter values were .36 and .40 for s and D, respectively. In terms of the distortion model, 40% of the incongruent memory traces were distorted into congruent memory traces. Figure 4b shows the predictions of the distortion model. The distortion model predicts no effect of the category size manipulation; in this respect, the distortion model fails to capture the results. Yet at a quantitative level, the distortion model performed somewhat better than the integration model; the average error for the distortion model was .0219.

Because the integration model correctly predicted the direction of the effect of category size but not the magnitude, it is worthwhile to use this discrepancy to help recast the integration model. Increasing the category size from 4 to 16 did not diminish the influence of prior knowledge as much as predicted. One way to examine this result is to consider, in terms of the integration model, what was the effect of increasing category size. Assume that for the small-category condition, the integration model is applied with the actual category size of $N = 4$. The N value for the large-category condition may then be treated as a free parameter, rather than assigning it the actual value of 16. The *extended integration model* was identical to the integration model, except for this additional free parameter representing the theoretical category size in the large-category condition. The best-fitting parameter values for the extended integration model were the following: $s = .36$, $G = 3.33$, and $N = 6.42$ for the large-category condition. This theoretical N value is surprisingly small compared with the actual value of 16. Figure 4c shows the predictions of the extended integration model superimposed over the data; the correspondence is now quite good. The root-mean-squared error of the model is only .0160.

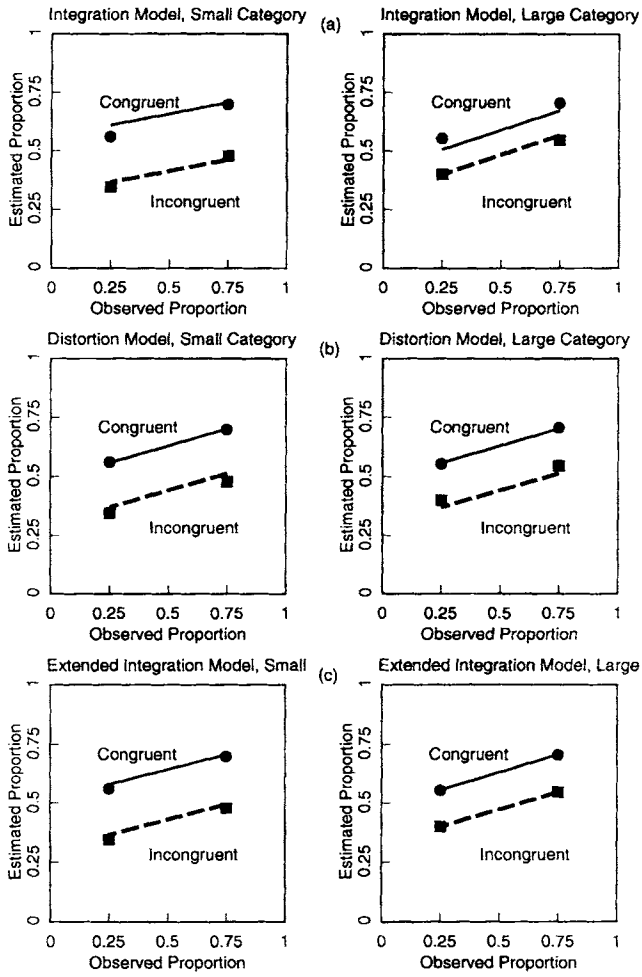

Figure 4. Predictions of integration, distortion, and extended integration models for Experiment 3.

Discussion

As predicted by the integration model, in Experiment 3 subjects showed a smaller effect of prior knowledge when more category members were observed. However, the amount of this decrease was less than what was predicted by the integration model. Subjects continued to be influenced by prior knowledge despite an abundance of observations in the context of City W. Although this result was unanticipated, it is quite interesting. This persistent influence of prior examples is reminiscent of the phenomenon of perseverance of belief in social cognition (Ross & Anderson, 1982). In addition, these results resemble conservatism effects in probability revision experiments in which subjects continue to rely on prior beliefs to a greater degree than what is prescribed by a Bayesian statistical model (Edwards, 1968; Phillips & Edwards, 1966). There are many possible explanations for the surprisingly small theoretical value of N for the large-category condition. One explanation is that repeated presentations of the same stimulus had a diminishing marginal impact on memory such as a habituation effect.

The prediction of the distortion model, that the prior knowledge effect is not influenced by category size, was not supported by the results. However, to give fair consideration to the distortion model, it should be noted that the distortion model can mimic the success of the integration model by assuming that the distortion rate, D, differs in the small- and large-category size conditions. (The distortion model can mimic the best fit of the extended integration model by assuming that $D = .45$ for the small-category condition and $D = .34$ for the large-category condition.) The plausibility of this assumption is addressed after Experiment 4.

Experiment 4

This experiment also investigated the effects of increasing category size. In addition, Experiment 4 addressed a new issue: How does category learning in a particular context affect

knowledge about categories in general? In the integration model, to make a categorization judgment for a particular context, people combine prior examples with recently observed examples. It is plausible that integration theory could be applied in the same manner to make judgments about more general categories so that both examples from general knowledge as well as examples from recent experience would influence judgments about categories in general. Thus, judgments about shy people in general would be affected by what was recently learned about shy people in City W. The integration model provides an implementation of what Rothbart (1981) has called a *bookkeeping* process in which beliefs about social categories are updated gradually as new individuals are encountered.

In contrast, the distortion model and the weighting model conceive of the relation between general knowledge and contextualized knowledge as unidirectional. These two theories attempt to describe how prior knowledge might affect learning of new categories, but they provide no account for how new experiences might in turn affect general knowledge. Thus, the weighting and distortion models could not be applied to these judgments about general knowledge.

In Experiment 4, subjects in the *background* condition observed descriptions of persons in City W but made category judgments based on more general knowledge, that is, knowledge of the general population of Illinois. In the *situational* condition, subjects received the same instructions as the previous experiments, that is, they were instructed to respond on the basis of what they had learned in the context of City W.[5]

Method

Experiment 4 was the same as Experiment 3, except for the following. First, the training stimuli had the structure shown in Table 7. The difference from Table 6 was that the large-category couplets were now used to describe 40 persons, for a category size of 20. The purpose of this increase was to strengthen the category size manipulation. There were 192 training stimuli.

The second change was that subjects made forced-choice classification judgments rather than probability estimates. The forced-choice method was introduced to provide some generalizability in the series of experiments. (The same categorization models were used to make predictions for forced choices, except that the probability in Equation 3 was now interpreted as a choice proportion, see Heit, 1992.) For each test question, subjects were presented information about one feature, then they chose which of two complementary features would be more likely to occur. For example, subjects were told that a person attended parties often, then they classified this person as either *shy* (the incongruent choice) or *not shy* (the congruent choice). The test questions were organized according to two within-subject variables: category size (small or large) and observed probability of the congruent feature (25% or 75%). There were 32 distinct test questions, but each question was asked twice, once with the congruent choice listed first and once with the incongruent choice listed first. Thus, each subject made 64 forced-choice responses.

The third change was that two sets of instructions were used in the test phase. Subjects in the situational condition received instructions comparable to the previous experiments, that is, they were given a description of a person from City W, and they were instructed to make the classification judgment on the basis of what had been learned about other persons in City W. In contrast, subjects in the background condition were instructed to base their answers on general knowledge about adults in Illinois. The test questions told these subjects to "Consider a person from Illinois with the following characteristic."

Table 7
Stimulus Structure for Training Examples in Experiment 4

No. of couplets	Presentations per couplet	Category size	Proportion of congruent pairings (%)
2	8	4	25
2	8	4	75
2	40	20	25
2	40	20	75

Ninety-two subjects participated, half assigned to the situational condition, and half assigned to the background condition. The training and test stimuli were assigned according to a yoked design. Forty-six stimulus sets were created, and each set was used for one subject in the situational condition and one subject in the background condition. The yoked design ensured that differences in results for the two conditions would not be due to differences in stimuli.

Results

Overall, subjects tended to choose congruent features over incongruent features; the mean choice proportion was .68, which was reliably greater than the chance value of .50, $t(91) = 9.16, p < .001$. A three-way ANOVA was performed on the proportion of congruent choices in each condition. This analysis indicated a main effect of the observed proportion of the congruent feature, $F(1, 90) = 95.07, p < .001, MS_e = 0.075$, indicating that congruent choices were more likely in the 75% condition than in the 25% condition. There was a main effect of category size, $F(1, 90) = 10.00, p < .01, MS_e = 0.024$, indicating that congruent choices were more likely when category size was small than when it was large. This result supports the prediction of the integration model that the prior knowledge effect is diminished in the large-category condition. The interaction between observed proportion and category size was significant, $F(1, 90) = 11.30, p < .001, MS_e = 0.023$, indicating that subjects were more sensitive to observed proportion in the large-category condition than in the small-category condition.

The main effect of the between-subjects variable of situational versus background judgment was also significant, $F(1, 90) = 5.61, p < .02, MS_e = 0.138$, indicating that subjects in the background condition were more likely to make congruent choices than subjects in the situational condition. The situational versus background variable did not show interactions with either of the other variables taken alone (both Fs < 1). Finally, the interaction between all three variables was not significant, $F(1, 90) = 2.41, MS_e = 0.023$. Figure 5 shows the results of Experiment 4.[6] These graphs indicate that prior

[5] The terms *background* and *situational* are used by analogy to the experimental paradigms of background and situational frequency judgments (Zechmeister & Nyberg, 1982).

[6] For the purpose of making Figure 5 comparable to the graphs of the previous experiments, Figure 5 shows each data point twice. For example, consider the choice between a congruent stimulus with an observed proportion of 75% and an incongruent stimulus with an

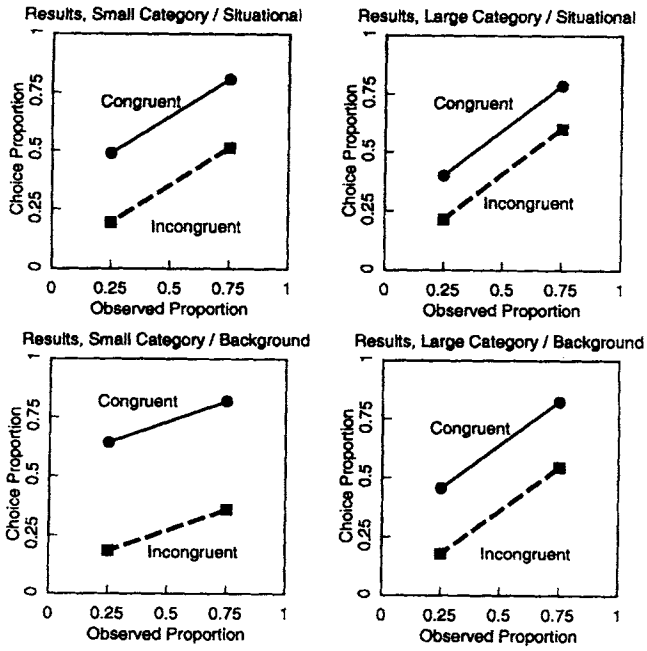

Figure 5. Results of Experiment 4.

knowledge had a greater effect in the background conditions than in the situational conditions. In addition, the observed proportion within City W affected judgments in both the situational and background conditions.

The integration model was fitted to the results of Experiment 4 with two different values of G for the situational and background conditions. A single value of s was used for both conditions because it was expected that memory for observed descriptions would be equivalent in the two conditions. The best value of s was found to be .06, the value of G in the situational condition was 2.24, and G in the background condition was 5.61, over twice the number of prior examples for the situational condition. The root-mean-square error of the model was .0375.

Figure 6 shows the predictions of the integration model superimposed over the results. In terms of predicting the greater prior knowledge effect in the background conditions compared with the situational conditions, the integration model does well. In addition, the integration model predicts that the prior knowledge effect will be diminished in the large-category conditions. However, as in Experiment 3, the integration model overestimates the impact of the category size manipulation.

Next, the extended integration model was applied; the N value for the large-category condition was estimated rather than using the actual value of 20. The best-fitting parameter values of the extended integration model were s = .04, G for situational judgments = 1.98, G for background judgments = 4.12, and N for the large-category condition = 9.67. The root-mean-square error was only .0190, about half the error of the standard integration model. The predictions of the extended integration model are not shown separately because this model's predictions are nearly indistinguishable from the data points themselves. Figure 5 gives a good representation of the extended integration model's predictions.

observed proportion of 25%. In the small category, situational condition, subjects chose the congruent stimulus about 80% of the time. This result is indicated in two places: the congruent line passes through the 80% point, and the incongruent line passes through the 20% point.

Discussion

Experiment 4 replicated Experiment 3, supporting the prediction of the integration model that the prior knowledge effect decreases as the size of the observed category increases. Again, the analysis using the extended integration model showed that the category size manipulation has a smaller impact than what the integration model predicts. The success of the extended integration model in Experiment 4, with a different design and a different response measure than Experiment 3, is suggestive of its generality.

As predicted by the integration model, category learning in a particular context led subjects to revise their judgments about the categories in general. For example, what subjects stated about shy people in the world at large was influenced by the descriptions of shy people they saw in the experiment. It may seem surprising that observations from City W influenced background judgments so much, as shown in Figure 5. Several aspects of Experiment 4 likely contributed to the fairly large effect of contextualized observations on judgments about general categories. First, the examples from City W were particularly memorable to subjects because these examples were recently observed. It would be interesting to examine in future research the longevity of this influence. Second, subjects likely did not have exact prior beliefs about, say, the proportion of shy persons in the general population who attend parties often. Thus subjects were willing to revise their estimates after seeing concrete information. Finally, the instructions for the background judgments may have encouraged the subjects to respond to the recently observed examples. Although the effect of new observations on the updating of general categories was almost surely magnified in this experimental context, it is expected that the functional form of category updating found in Experiment 4, if not the exact amount, would also describe the updating phenomenon outside of the laboratory.

As in Experiment 3, the distortion model can mimic the predictions of the integration model for situational judgments by assuming that the distortion rate decreases from the small-category condition to the large-category condition, regardless of what is observed. Yet this assumption is not consistent with the background judgments of Experiment 4, which serve as a measure of strength of belief in different conditions. Figure 5 shows that background judgments were quite sensitive to what was observed. For large categories, the judgments about the general population when the observed proportion of congruent examples was 75% were much higher

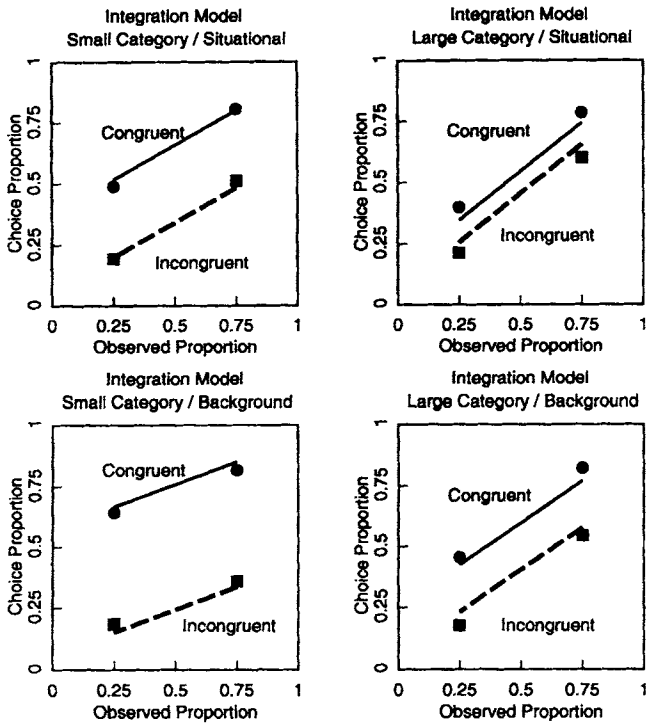

Figure 6. Predictions of integration model for Experiment 4.

than the judgments in the 25% condition. The modified distortion model must assume that both conditions correspond to the same low distortion rate, even though subjects in the 75% condition clearly had a stronger belief about category membership. The plausibility of this assumption is questionable because it requires the same decrease in the distortion rate whether expectations are confirmed or disconfirmed.

Experiment 5

In Experiment 5, the effects of category learning in a particular context on category knowledge in general were also investigated. However, the training stimuli were structured so that the observed proportion was varied over a wider range than in Experiment 4.

Method

Experiment 5 was the same as Experiment 2, including the stimulus design shown in Table 5, except for the following. For the test phase, subjects received instructions analogous to the background condition of Experiment 4. Subjects estimated conditional probabilities of features, but they did so for descriptions of persons from Illinois rather than from City W in particular. Subjects were encouraged to make these judgments on the basis of their general knowledge of people in Illinois. Because the retention interval manipulation did not measurably affect the results of Experiment 2, this manipulation was dropped. Instead, Experiment 5 had a 2-min unfilled retention interval. Thirty-eight subjects participated. They were recruited from the same population as the subjects in Experiment 2, however the two experiments were not run simultaneously.

Results and Discussion

Figure 7a shows the results of Experiment 5. Judgments about categories in general were influenced by what was observed in City W. In addition, this figure shows a large effect of prior knowledge. This prior knowledge effect appears to be greater than the prior knowledge effect for Experiment 2 (in Figure 2a) when subjects made situational judgments rather than background judgments. This comparison is consistent with the comparison of background and situational conditions within Experiment 4.

A two-way ANOVA on the results of Experiment 5 showed a main effect of congruent versus incongruent feature, $F(1, 37) = 137.47, p < .001, MS_e = 467$, and a main effect of observed proportion, $F(4, 148) = 22.59, p < .001, MS_e = 266$. The interaction between these two variables was also found to be statistically significant, $F(4, 148) = 3.21, p < .05, MS_e = 92$. This slight interaction appears to be attributable to the incongruent judgment condition in which the observed proportion is zero; no explanation is offered for this result.

Next, the integration model was fitted to the results. The mean judgment in this experiment was .55, thus a value of .05 was added to each prediction of the model. The best parameter values were $s = .31$ and $G = 11.05$. This value of G is about twice the value of G found for Experiment 2 (6.15), suggesting that prior examples had twice as much influence on background judgments compared with situational judgments. The root-mean-square error for the account of the integration model is .0206. Figure 7b shows the predictions of the integration model superimposed over the results. (The extended integration model was equivalent to the standard integration model because category size was not varied.)

The results of Experiment 5 are consistent with the results of the background condition of Experiment 4. Observation of category members in the context of City W influenced subject's judgments about categories in general, in a manner closely matching the predictions of the integration model. In both experiments, the integration model accounts for the results by assuming that the subjects mainly respond according to a large number of prior examples, but they are also influenced somewhat by recently observed examples.

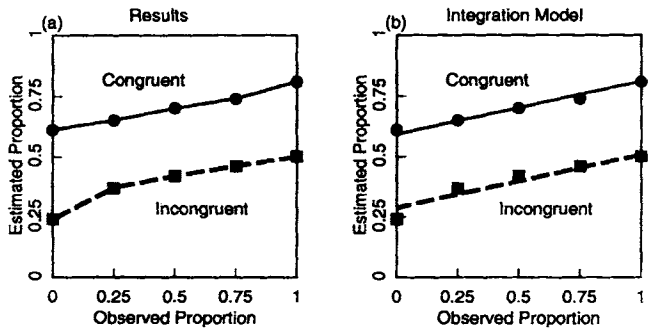

Figure 7. Results of Experiment 5 and predictions of integration model.

General Results

These interpretations of models fitted to averaged data depend on the assumption that average responses are representative of what an individual subject might produce. Otherwise, the predictions of the model would correspond to pattern of results that might never be obtained from a subject. (See Pavel, Gluck, and Henkle, 1988, for a discussion of this issue.) In these experiments it was found that average judgments in the various conditions were influenced by both prior knowledge and observed category membership. Yet it could be possible that no individual subject used both prior knowledge and observations, even though all three models predict that both sources of information will influence an individual subject's judgments.

In Experiments 1, 2, 3, and 5, each subject's expectation effect was computed as the difference between the subject's average probability judgment on incongruent test questions and the subject's average judgment on incongruent test questions. A positive difference means that the subject tended to give higher ratings when questions were congruent with prior knowledge. Likewise, the observation effect was computed as the difference between the subject's average judgment on test questions in which the observed proportion was greater than .5 and the subject's average judgment on test questions in which the observed proportion was less than .5. For Experiment 4, each subject's expectation effect was the proportion of congruent choices; a positive expectation effect is signified by a value greater than one half. Likewise, a subject's observation effect in this experiment was the proportion of choices of a $p = .75$ stimulus over a $p = .25$ stimulus.

Figure 8 shows scatter plots of individual subjects' expectation effects and observation effects. For each experiment, the proportion of subjects showing both a positive expectation effect and a positive observation effect is approximately two thirds or greater. The proportions for Experiments 1, 2, 3, 4, and 5, respectively, are 74%, 68%, 74%, 65%, and 82%. If individual subjects only attended to either prior knowledge or to observations, then the proportion of data points expected by chance to fall in the first quadrant would be only 50%. For each experiment, the proportion of subjects showing both positive effects was greater than chance by a sign test ($p < .01$).

This analysis supports the conclusion that a significant proportion of subjects made use of both expectations and observations. Thus, the averaged data are representative of individual data in terms of showing an influence of both factors. However, the scatter plots do indicate that individuals tended to respond more to either prior knowledge or to observations. The correlations between the expectation effect and the observation effect across subjects were negative, ranging from −.361 in Experiment 2 to −.685 in Experiment 5. In further analyses, the integration model was able to successfully account for two separate sets of results for each experiment, after a median split of subjects in terms of expectation effect. For example, the 62 subjects in Experiment 2 were assigned to two groups, depending on whether their expectation effects were below or above the median. With application of the integration model to subjects who showed high expectation effects, the G parameter was found to be 19.7, and the root-mean-square error of the model was only .0134. Likewise, with application of the integration model to subjects who showed low expectation effects, the G parameter was .8, and the root-mean-square error of the model was .0224. Thus the integration model provides an account of individual differences in terms of differing numbers of prior examples.

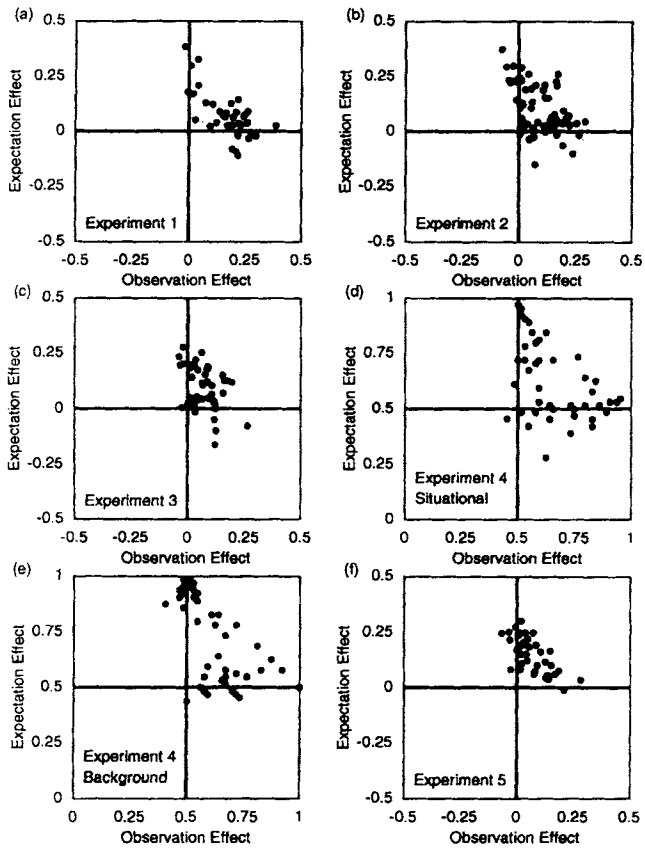

Figure 8. Individual subjects' expectation effects and observation effects.

General Discussion

Summary

Experiments 1, 2, 3, and 4 showed that when people learn about categories in a particular context, they are influenced both by what they observe in that context and by their prior knowledge. The integration model, which assumes that people are influenced both by prior examples and by observed examples, provided an excellent qualitative and quantitative account of the results. The weighting model, which assumes that prior knowledge serves to facilitate memory for congruent examples, performed poorly in accounting for Experiments 1 and 2. The weighting model incorrectly predicted for these experiments that the effect of prior knowledge depends on the consistency of new observations.

The integration model predicts that the effect of prior knowledge will be attenuated as more examples are observed. In contrast, the distortion model, which assumes that incongruent observations are distorted to be congruent with prior knowledge, does not predict this attenuation. The results of Experiments 3 and 4 favored the prediction of the integration model, although the degree of attenuation was smaller than predicted.

Finally, Experiments 4 and 5 showed that people's general knowledge about categories was influenced by the learning of new categories. Thus, there was not simply a unidirectional flow of information from prior knowledge to newly learned categories. Again, the integration model provided an account of these findings, which could not be addressed by the other models.

Assumptions of Approach

These conclusions depend, of course, on the assumptions underlying this research. First, the experimental stimuli were somewhat different from what has been used in many experiments on category learning. For example, subjects were asked to judge both the likelihood of a shy person attending parties often and the likelihood that a person who attends parties often is shy. Thus, the features *shy* and *attends parties often* were each treated as category labels on different questions. In much categorization research, some features are always treated as category labels, and the remaining features are treated only as cues for inferring the category label. However, recent research (Heit, 1992) has shown that people are more flexible in their inferences; they can treat different features as category labels to answer different questions. (See J. R. Anderson, 1991, and Billman and Heit, 1988, for further arguments on the importance of inference of features.) In addition, the person descriptions contained only two pieces of information. The rationale of this approach was to find the simplest cases that lead to distinct predictions of the models. With other stimulus designs, the models might make practically indistinguishable predictions. For the present experiments, when the weighting and distortion models failed, it was clear why they failed. Now that the integration model has been supported over the alternatives, future research can evaluate the integration model as applied to more complex stimuli.

The conclusions about integration, weighting, and distortion processes also depend on assumptions about the categorization models that were implemented. Interpretation of the equations depends on the framework of exemplar theory. For example, interpreting the parameter G as referring to some number of memory traces of category members is natural within this framework. These prior examples were not directly observable, but their existence was inferred by fitting the integration model for varying values of G. (See Busemeyer, Dewey, and Medin, 1984, and Nosofsky, 1986, for related applications of this technique.)

In addition, the conclusions favoring integration processes over weighting and distortion processes are, strictly speaking, limited to the framework of exemplar models. These studies were not intended to distinguish between the general framework of exemplar models and other general modeling frameworks, such as abstraction models or connectionist models (see Barsalou, 1990). It seems plausible, however, that comparable conclusions would be obtained by implementing the three processes within other modeling frameworks. For example, the prediction for the weighting process that prior knowledge will have the greatest effect on mixed observations is likely to be a general prediction. So far, these three processes have not been compared in other frameworks. However, Choi, McDaniel, and Busemeyer (1993) have recently incorporated prior knowledge into connectionist networks with pretraining on prior examples. These models, which may be considered as a connectionist implementations of an integration process, gave promising accounts of people's biases in learning about categories defined by logical rules.

Applications to Memory

An additional benefit of the framework of exemplar models is that these models of categorization are quite similar to multiple-trace models of memory (Estes, 1994; Gillund & Shiffrin, 1984; Hintzman, 1988). Underlying exemplar models of categorization and multiple-trace models of memory is the psychological construct of familiarity, which is the total similarity of a test stimulus to items retrieved from memory (Jones & Heit, 1993). In Equation 1, familiarity is used to make a categorization judgment, but familiarity can also be used with other response rules to make recognition judgments or frequency judgments. Theoretical developments in categorization models have the potential to contribute to the development of memory models, and vice versa. Indeed, the integration, weighting, and distortion models presented in this article were originally developed to account for the effects of prior knowledge on recognition memory (Heit, 1993).

Many studies have demonstrated the effects of prior knowledge on recognition, such as the effects of social stereotypes on recognition of person descriptions (see Stangor & McMillan, 1992, for a review). People show a relatively high hit rate for stimuli that are congruent with prior knowledge. For example, imagine that a subject has seen descriptions of two lawyers, one congruent with a lawyer stereotype and one incongruent with a lawyer stereotype. The congruent description is more likely to be recognized later. However, people show poor recognition memory for congruent stimuli in terms of discriminability, or d'. For example, not only would the hit rate be high for a stereotype-congruent description of a lawyer, but the false-alarm rate for new, congruent descriptions would also be high.

This pattern of results for hit rates and d' is sufficient to distinguish between the processes of integration, weighting, and distortion. Heit (1993) implemented these processes as three versions of a multiple-trace model of memory. The integration model, which assumed that memory contained prior examples that were consistent with a stereotype, correctly predicted that the hit rate will be higher for congruent stimuli and that d' will be higher for incongruent stimuli. The

weighting and distortion models also predicted a higher hit rate for congruent stimuli, but the predictions of these models for d' were incorrect. The weighting model predicted no difference in d' for incongruent and congruent stimuli, and the distortion model predicted that d' would be lower for incongruent stimuli. Thus, the recognition results provide converging evidence for the conclusions in this article favoring integration over weighting and distortion.[7] Unlike the categorization studies described here, in the recognition paradigm the distortion model could not mimic the predictions of the integration model. Indeed, the distortion model gave the worst account of the three models for recognition.

Relations to Bayesian Statistics

The learning problem faced by subjects in the present experiments may be conceived of as revising a prior estimate of a proportion, such as the proportion of shy people who do not attend parties often, after observation of additional cases. Say that a subject's prior estimate of the proportion of times that stimulus x occurs in Category A is q, and that the subject then observes N additional cases of x. The proportion of these N additional cases of x that fall into Category A is p. Then, according to Bayesian statistical theory (Edwards, Lindman, & Savage, 1963; Raiffa & Schlaifer, 1961), the posterior estimate of the proportion is shown in Equation 4:

$$P(\text{x is an A}) = \frac{Np + Gq}{N + G}, \tag{4}$$

where G is a measure of the strength of the prior belief. In effect, the posterior estimate for the proportion is a weighted mean of the prior estimate of the proportion, q, and the observed proportion, p.

The integration model closely resembles this Bayesian formula. Consider the case in which memory discrimination is perfect, that is, the parameter s is zero. The integration model, after substituting the predictions from Table 4 into Equation 3, may be rewritten as Equation 5.

$$P(\text{classify x as A}) = \frac{Np + G}{N + G} \tag{5}$$

Here, the parameter G refers to the number of prior examples of x in Category A. This special case of the integration model is equivalent to the Bayesian model when the prior estimate of the proportion, q, is 1.[8]

The integration model is consistent with the Bayesian formula for revising proportions, yet the integration model is part of a broader framework of descriptive models for memory and categorization. In contrast, it is not clear how Equation 4 would be applied to recognition memory tasks (as in Heit, 1993) or to categorization of multidimensional stimuli.[9]

Conclusion

The present research has clear implications for future research on category learning. The development and testing of three categorization models that incorporate effects of prior knowledge demonstrate that research on category learning in knowledge-rich domains can build on previous research on models of category learning in less meaningful domains. It is certainly possible to falsify these previous models by showing that extra-experimental knowledge does have an influence, but the present research undertakes the task of developing more complete models rather than falsifying the incomplete models.

As described so far, the integration, weighting, and distortion models represent only initial steps in the development of a better account of prior knowledge effects on category learning. Yet the results of these five experiments provide constraints for future accounts of category learning. First, these results cannot be explained by a weighting process in which people pay more attention to category members that are congruent with prior knowledge. This result is perhaps the most surprising finding of this article because many researchers have argued for weighting processes of various forms. Yet these results do not lead to the conclusion that selective weighting never occurs, only that a weighting process has no explanatory power for these experiments. Future models of category learning cannot explain prior knowledge effects in terms of weighting alone.

The weighting model presented in this article addresses how prior knowledge might affect the weighting of whole exemplars, but other work on categorization models has emphasized selective attention on features (Aha & Goldstone, 1992; Billman & Heit, 1988; Kruschke, 1992; Nosofsky, 1986; Smith & Zaraté, 1992). Selective weighting of features has also been proposed as one of the processes by which prior knowledge may affect category learning (Keil, 1989; Murphy & Medin, 1985; Ward, 1993). This alternate form of selective weighting cannot explain the results of present experiments because both the congruent and incongruent examples were composed of the same features. (For example, the congruent and incongruent descriptions had the same proportion of the feature *shy*.) However, it is plausible that when stimuli are described in terms of many features, prior knowledge would lead to feature weighting. Compare the task of learning to categorize persons

[7] Like memory judgments and categorization judgments, judgments of contingency or causality are also influenced by both prior knowledge and observations (see Alloy & Tabachnik, 1984; Shanks, 1991, for reviews). So far, the integration, weighting, and distortion models have not explicitly been compared as accounts for the numerous studies showing prior knowledge effects on causality judgments.

[8] The assumption of the integration model as implemented so far has been that all prior examples are congruent category members. However, future versions of the integration model could also incorporate a q parameter that would allow for some proportion of prior examples to be congruent and the rest to be incongruent. (For the present experiments, adding such a q parameter did not appreciably improve the account of the integration model.)

[9] Of course, Equation 4 is part of a different broad framework, the set of Bayesian statistical models, and it is possible to use models in this framework to address a variety of results from memory and categorization (J. R. Anderson, 1990).

as good or bad tennis players with the task of learning to categorize good or bad conversationalists. The features that a learner would consider for categorizing tennis players probably differ from the features that a learner would consider for categorizing conversationalists (Heit & Rubinstein, 1994).

Second, future models of categorization cannot explain prior knowledge effects only in terms of a simple distortion process that alters incongruent category members to make them congruent. The basic distortion model did not predict the result that prior knowledge has a diminishing effect as more category members are observed. However, the distortion model can fit this result by assuming that the distortion rate decreases as category size increases. Thus, future models could possibly succeed with a distortion process, if this effect of prior knowledge were itself modified by experience. Yet the distortion model must make a questionable assumption, that the distortion rate decreases equally either after confirming evidence or disconfirming evidence. A distortion process also cannot account for Experiments 4 and 5 in which category learning in context affected general knowledge about categories. Finally, a distortion process cannot account for the effects of prior knowledge on recognition memory, as described by Heit (1993).

Third, the process of integration has received strong support as a candidate for inclusion in future models of category learning in meaningful contexts. Yet the integration model is not yet a complete process account. For example, this model describes how people use prior examples and observed examples to perform categorization, but the model does not account for the origin of the prior examples. In some cases, such as learning about *Chicago joggers,* it is plausible that the prior examples would simply be memories of joggers from other places. However, in other cases the prior examples could be the end product of sophisticated reasoning processes, including conceptual combination. Consider the situation of meeting and learning about a group of Republican social workers. People have expectations about Republican social workers that are not directly attributable to known examples of Republicans and social workers (Hastie, Schroeder, & Weber, 1990; see also Heit & Barsalou, 1993, and Murphy, 1988). Thus the prior examples for this category would not only consist of known Republicans and known social workers. In this case, the retrieval of prior knowledge would require additional reasoning processes (for some proposals, see Ward, in press). The integration model presented in this article will serve as a starting point for future developments in modeling the effects of prior knowledge on category learning.

References

Aha, D. W., & Goldstone, R. L. (1992). Concept learning and flexible weighting. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Erlbaum.

Alba, J. W., & Hasher, L. (1983). Is memory schematic? *Psychological Bulletin, 93,* 203–231.

Alloy, L. B., & Tabachnik, N. (1984). Assessment of covariation by humans and animals: The joint influence of prior expectations and current situational information. *Psychological Review, 91,* 112–149.

Anderson, J. R. (1990). *The adaptive character of thought.* Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98,* 409–429.

Anderson, J. R., & Ross, B. H. (1980). Evidence against a semantic-episodic distinction. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 441–466.

Anderson, N. H. (1991). Stereotype theory. In N. H. Anderson (Ed.), *Contributions to information integration theory, Volume II: Social* (pp. 183–240). Hillsdale, NJ: Erlbaum.

Asch, S. E. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology, 41,* 258–290.

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 33–53.

Barrett, S. E., Abdi, H., Murphy, G. L., & McCarthy Gallagher, J. (1993). Theory-based correlations and their role in children's concepts. *Child Development, 64,* 1595–1616.

Barsalou, L. W. (1990). On the indistinguishability of exemplar memory and abstraction in memory representation. In T. K. Srull & R. S. Wyer (Eds.), *Advances in social cognition* (pp. 61–88). Hillsdale, NJ: Erlbaum.

Billman, D., & Heit, E. (1988). Observational learning without feedback: A simulation of an adaptive method. *Cognitive Science, 12,* 587–625.

Brooks, L. R. (1978). Nonanalytic concept formation and memory for instances. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 169–211). Hillsdale, NJ: Erlbaum.

Brooks, L. R. (1987). Decentralized control of categorization: The role of prior processing episodes. In U. Neisser (Ed.), *Concepts and conceptual development: Ecological and intellectual factors in categorization* (pp. 141–174). Cambridge, England: Cambridge University Press.

Brown, N. R., & Siegler, R. S. (1993). Metrics and mappings: A framework for understanding real-world quantitative estimation. *Psychological Review, 100,* 511–534.

Busemeyer, J. R. (1991). Intuitive statistical estimation. In N. H. Anderson (Ed.), *Contributions to information integration theory, Volume I: Cognition* (pp. 187–215). Hillsdale, NJ: Erlbaum.

Busemeyer, J. R., Dewey, G. I., & Medin, D. L. (1984). Evaluation of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 638–648.

Chandler, J. P. (1965). STEPIT: Finds local minima of a smooth function of several parameters. *Behavioral Science, 14,* 81–82.

Chinn, C. A., & Brewer, W. F. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. *Review of Educational Research, 63,* 1–49.

Choi, S., McDaniel, M. A., & Busemeyer, J. R. (1993). Incorporating prior biases in network models of conceptual rule learning. *Memory & Cognition, 21,* 413–423.

Dosher, B. A., & Rosedale, G. (1991). Judgments of semantic and episodic relatedness: Common time-course and failure of segregation. *Journal of Memory & Language, 30,* 125–160.

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.), *Formal representation of human judgment* (pp. 17–52). New York: Wiley.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70,* 193–242.

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

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

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91,* 1–67.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General, 117,* 227–247.

Hastie, R., Schroeder, C., & Weber, R. (1990). Creating complex social conjunction categories from simple categories. *Bulletin of the Psychonomic Society, 28,* 242–247.

Hayes, B. K., & Taplin, J. E. (1992). Developmental changes in categorization processes: Knowledge and similarity-based models of categorization. *Journal of Experimental Child Psychology, 54,* 188–212.

Heit, E. (1992). Categorization using chains of examples. *Cognitive Psychology, 24,* 341–380.

Heit, E. (1993). Modeling the effects of expectations on recognition memory. *Psychological Science, 4,* 244–252.

Heit, E., & Barsalou, L. W. (1993). *An instantiation model of typicality in complex categories.* Manuscript submitted for publication.

Heit, E., & Rubinstein, J. (1994). Similarity and property effects in inductive reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 411–422.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review, 95,* 528–551.

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin, 114,* 3–28.

Johnson, M. K., & Raye, C. L. (1981). Reality monitoring. *Psychological Review, 88,* 67–85.

Jones, C. M., & Heit, E. (1993). An evaluation of the total similarity principle: Effects of similarity on frequency judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 799–812.

Keil, F. C. (1989). *Concepts, kinds, and cognitive development.* Cambridge, MA: MIT Press.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99,* 22–44.

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

Murphy, G. L. (1988). Comprehending complex concepts. *Cognitive Science, 12,* 529–562.

Murphy, G. L. (1993). Theories and concept formation. In I. V. Mechelen, J. Hampton, R. Michalski, & P. Theuns (Eds.), *Categories and concepts: Theoretical views and inductive data analysis* (pp. 173–200). San Diego, CA: Academic Press.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92,* 289–316.

Murphy, G. L., & Wisniewski, E. J. (1989). Feature correlations in conceptual representations. In G. Tiberghein (Ed.), *Advances in cognitive science* (pp. 23–45). Chichester, England: Ellis Horwood.

Mynatt, C. R., Doherty, M. E., & Tweney, R. D. (1977). Confirmation bias in a simulated research environment. *Quarterly Journal of Experimental Psychology, 29,* 85–95.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General, 115,* 39–57.

Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 700–708.

Nosofsky, R. M. (1992). Exemplars, prototypes, and similarity rules. In A. M. Healy, S. F. Kosslyn, & R. M. Shiffrin (Eds.), *From learning theory to connectionist theory: Essays in honor of William K. Estes* (pp. 149–167). Hillsdale, NJ: Erlbaum.

Pavel, M., Gluck, M. A., & Henkle, V. (1988). Generalization by humans and multi-layer adaptive networks. In *Proceedings of the 10th Annual Conference of the Cognitive Science Society* (pp. 680–687). Hillsdale, NJ: Erlbaum.

Pazzani, M. J. (1991). Influence of prior knowledge on concept acquisition: Experimental and computational results. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 416–432.

Phillips, L. D., & Edwards, W. (1966). Conservatism in simple probability inference tasks. *Journal of Experimental Psychology, 72,* 346–357.

Potts, G. R., St. John, M. F., & Kirson, D. (1989). Incorporating new information into existing world knowledge. *Cognitive Psychology, 21,* 303–333.

Raiffa, H., & Schlaifer, R. (1961). *Applied statistical decision theory.* Cambridge, MA: Harvard University.

Ross, L., & Anderson, C. A. (1982). Shortcomings in the attribution process: On the origins and maintenance of erroneous social assessments. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 129–152). Cambridge, England: Cambridge University Press.

Rothbart, M. (1981). Memory processes and social beliefs. In D. L. Hamilton (Ed.), *Cognitive processes in stereotyping and intergroup behavior* (pp. 145–181). Hillsdale, NJ: Erlbaum.

Shanks, D. R. (1991). On similarities between causal judgments in experienced and described situations. *Psychological Science, 2,* 341–350.

Smith, E. R., & Zaraté, M. A. (1992). Exemplar-based models of social judgment. *Psychological Review, 99,* 3–21.

Stangor, C., & McMillan, D. (1992). Memory for expectancy-congruent and expectancy-incongruent information: A review of the social and social developmental literatures. *Psychological Bulletin, 111,* 42–61.

Taylor, S. E., & Crocker, J. (1978). Schematic bases of social information processing. In E. T. Higgins, C. P. Herman, & M. P. Zanna (Eds.), *Social cognition: The Ontario symposium* (pp. 89–134). Hillsdale, NJ: Erlbaum.

Ward, T. B. (1993). Processing biases, knowledge, and context in category formation. In G. V. Nakamura, R. Taraban, & D. L. Medin (Eds.), *Categorization by humans and machines* (pp. 259–282). San Diego, CA: Academic Press.

Ward, T. B. (in press). Structured imagination: The role of category structure in exemplar generation. *Cognitive Psychology.*

Wattenmaker, W. D., Dewey, G. I., Murphy, T. D., & Medin, D. L. (1986). Linear separability and concept learning: Context, relational properties, and concept naturalness. *Cognitive Psychology, 18,* 158–194.

Wisniewski, E. J., & Medin, D. L. (1991). Harpoons and long sticks: The interaction of theory and similarity in rule induction. In D. H. Fisher, M. J. Pazzani, & P. Langley (Eds.), *Concept formation: Knowledge and experience in unsupervised learning.* San Mateo, CA: Morgan Kaufman.

Wisniewski, E. J., & Medin, D. L. (1994). On the interaction of theory and data in concept learning. *Cognitive Science, 18,* 221–282.

Wright, G., & Whalley, P. (1983). The supra-additivity of subjective probability. In B. P. Stigum & F. Wenstop (Eds.), *Foundations of utility and risk theory with applications* (pp. 233–244). Dordrecht, Holland: Reidel.

Zechmeister, E. B., & Nyberg, S. E. (1982). *Human memory.* Pacific Grove, CA: Brooks/Cole.

Appendix A

Pretest

Method

Stimuli. The test stimuli were derived from 20 couplets of four features. (Table 2 shows the 10 couplets that were eventually selected.) Each couplet was used for 8 conditional probability test items. These test questions included all possible pairs of features, in both directions, except for pairing of the first and second feature together and the third and fourth feature together. The whole test consisted of 160 conditional probability questions, but to minimize fatigue, each subject was asked only 80 questions. A subject was only given one item, randomly selected, from a pair of complementary questions, for example, a subject would rate either P(shy | attends parties often) or P(not shy | attends parties often). Each subject saw questions in a different random order.

Subjects and procedure. Twenty-four Northwestern University undergraduates participated. The subjects were instructed to answer the questions on the basis of their general knowledge of adults in the state of Illinois. The 80 questions appeared one at a time on a computer screen. Each question was presented in the following manner.

> Consider 100 people with the following characteristic: (a feature) How many of them would also have this characteristic? (another feature from the couplet)

Then the subject typed an integer estimate from 0 to 100. The entire procedure typically lasted 15 min.

Results

The purpose of the pretest was to find stimuli that met two criteria: that congruent probabilities are greater than incongruent probabilities, and that complementary probabilities add up to about 100%. The 10 best couplets, shown in Table 2, were chosen. These couplets had the following characteristics. The mean congruent conditional probability of a couplet was 76%, in which a couplet's congruent probability was the average of four ratings: the two conditional probabilities relating Features 1 and 3, and the two conditional probabilities relating Features 2 and 4. The range of congruent probabilities for couplets was 69% to 80%. The mean incongruent conditional probability of a couplet was 27%, with a range from 21% to 35%. The mean sum of complementary probabilities for a couplet was 102%, in which each couplet's score was determined from the average of four sums of complementary probabilities. Across the 10 couplets, the range of means was 94% to 108%.

Appendix B

Predictions of Models

The Integration Model

When Stimulus x is congruent with Category A, after substitution of the appropriate terms from Table 4 into Equation 3, the integration model may be stated as Equation B1.

$$P(\text{classify x as A}) = \frac{Np + G + sN(1-p)}{Np + G + sN(1-p) + N(1-p) + s(Np + G)} \tag{B1}$$

When Stimulus x is incongruent with in Category A, the integration model may be stated as Equation B2.

$$P(\text{classify x as A}) = \frac{Np + s[N(1-p) + G]}{Np + s[N(1-p) + G] + N(1-p) + G + sNp} \tag{B2}$$

Note that when $G = 0$, that is, the number of prior examples is zero, Equation B1 is identical to Equation B2. The prior knowledge effect predicted by the integration model is found by subtracting Equation B2 from Equation B1 to yield Equation B3.

$$\left(\frac{G}{G+N}\right)\left(\frac{1-s}{1+s}\right) \tag{B3}$$

The integration model predicts that the prior knowledge effect is independent of the observed proportion, p. Also, Equation B3 predicts that as N, the number of observed category members, increases, the prior knowledge effect will decrease.

The Weighting Model

When Stimulus x is congruent with Category A, the predictions of the weighting model are described by Equation B4, and Equation B5 shows the predictions for incongruent test questions.

$$P(\text{classify x as A}) = \frac{WNp + sN(1-p)}{WNp + sN(1-p) + N(1-p) + sWNp} \tag{B4}$$

$$P(\text{classify x as A}) = \frac{Np + sWN(1-p)}{Np + sWN(1-p) + WN(1-p) + sNp} \tag{B5}$$

(*Appendix B continues on next page*)

When $W = 1$, the relative weight of congruent observations is the same as that of incongruent observations, so Equations B4 and B5 are equivalent. However, when $W > 1$, the predicted prior knowledge effect of the weighting model is shown by Equation B6.

$$(p)(1-p)\left[\frac{W^2-1}{(p+W-Wp)(1-p+Wp)}\right]\left(\frac{1-s}{1+s}\right) \quad \text{(B6)}$$

Although this equation is not as simple as Equation B3, several implications are clear. First, unlike the integration model, according to the weighting model the prior knowledge effect depends on the observed proportion, p. This prior knowledge effect will be zero when $p = 0$ or $p = 1$. Furthermore, it can be shown (by taking the derivative of Equation B6 with respect to p) that the prior knowledge effect is maximized when $p = .5$.

The Distortion Model

When Stimulus x is congruent with Category A, the predictions of the distortion model are described by Equation B7, and Equation B8 corresponds to the predictions for incongruent test questions.

$$P(\text{classify x as A}) = \frac{N[p+D(1-p)]+sN(1-D)(1-p)}{N[p+D(1-p)]+sN(1-D)(1-p)+N(1-D)(1-p)+sN[p+D(1-p)]} \quad \text{(B7)}$$

$$P(\text{classify x as A}) = \frac{N(1-D)p+sN(1-p+Dp)}{N(1-D)p+sN(1-p+Dp)+N(1-p+Dp)+sN(1-D)p} \quad \text{(B8)}$$

The prior knowledge effect, the difference between Equations B7 and B8, has a simple form, as shown in Equation B9:

$$D\left(\frac{1-s}{1+s}\right). \quad \text{(B9)}$$

As in the integration model, the distortion model predicts a constant effect of prior knowledge regardless of the proportion p. Furthermore, the prior knowledge effect will increase in size as D, the proportion of distorted instances, increases. Finally, the distortion model predicts that the prior knowledge effect is not affected by the category size N.

Comparing Equation B3 and Equation B9 suggests that the distortion model can mimic the predictions of the integration model for a single level of N. In particular, the distortion rate, D, can be made a function of G, the number of prior examples, and N, the category size: $D = G/(G + N)$. With this substitution, the distortion model predicts the same prior knowledge effect as the integration model. In addition, with this substitution into the distortion model's Equations B7 and B8, Equations B1 and B2 of the integration model may be obtained.

Received August 6, 1993
Revision received January 6, 1994
Accepted January 24, 1994 ■