JOURNAL OF EXPERIMENTAL SOCIAL PSYCHOLOGY **32,** 309–336 (1996)
ARTICLE NO. 0015

The Retrieval of Typical and Atypical Category Members

MYRON ROTHBART

University of Oregon

N. SRIRAM

National University of Singapore

AND

CARENE DAVIS-STITT

University of Oregon

Received: August 16, 1995; revised: January 10, 1996; accepted January 16, 1996

Four studies tested the hypothesis that category labels are more likely to prime retrieval of typical than atypical category members, an assumption critical to Rothbart and John's (1985) view of stereotype change. The first two experiments examined the use of a category name to retrieve category members in both a laboratory and field setting. In Experiment 1, subjects first learned to associate the names of stimulus persons to one of four social categories. They then learned new information about each group member that was either consistent with, inconsistent with, or irrelevant to the group's stereotype. It was predicted and found that a category's ability to retrieve an exemplar was proportional to the exemplar's goodness-of-fit to the stereotype. In Experiment 2, members of one fraternity recalled the members of a rival fraternity and afterward rated all members on typicality, familiarity, and liking. Consistent with prediction, the typical members of the target fraternity were recalled more often than the atypical members of the fraternity with familiarity and liking controlled. The second two experiments examined the retrievability of information associated with typical and atypical category members. In Experiment 3, subjects learned stereotypically neutral information about men and women who were either teachers or scientists and, when given a gender or occupational retrieval cue, found it

This research was supported in part by National Institutes of Mental Health Grant MH40662 to the first author. We gratefully acknowledge the comments, help, and advice of Mara Cadinu, Deborah Day, Robert Mauro, and Bernadette Park. Mary Rothbart provided invaluable help on editorial and substantive issues. Correspondence and reprint requests should be addressed to Myron Rothbart, Department of Psychology, 1227 University of Oregon, Eugene, OR 97403-1227.

0022-1031/96 $18.00

easier to retrieve information associated with typical (male scientists, female teachers) than atypical (male teachers, female scientists) category members. A fourth experiment replicated the third experiment with increased cognitive load, and found similar results. These findings have paradoxical implications for stereotype change (Rothbart & John, 1985; Rothbart & Lewis, 1988). Although information is most likely to become part of the stereotype when attached to a typical group member, the effect of pairing disconfirming information with a category member is to reduce that member's typicality.

INTRODUCTION

A half century ago, Lewin and Grabbe (1945) published an article, "Conduct, Knowledge, and the Acceptance of New Values, "dealing with the broad question of how entrenched behavior, beliefs, and values can be changed. Discussing prejudice based on age and race, they wrote:

> [Such examples] . . . indicate that favorable experiences with members of another group, even if they are frequent, do not necessarily diminish prejudices toward that group. . . . Only if a psychological linkage is made between the image of specific individuals and the stereotype of a certain group, only when individuals can be perceived as "typical representatives" of that group, is the experience with individuals likely to affect the stereotype. (p. 58)

This paper is concerned with the role of typicality in establishing the psychological linkage between the individual and the group, and is based on Rothbart and John's (1985) analysis of the dynamic and complex relation that exists between the category as a whole and the individual members (exemplars) who make up the category. A brief summary of the model follows.

Two lines of research suggest a disparity between our images of a group, and our images of the individual members who make up the group. First, Sears (1983) found that people express greater favorability toward the individual members of the group than to the group as a whole, which he labeled the person-positivity bias. Second, research on the contact hypothesis often shows that experiences with individual members of a group fail to generalize to the group as a whole, as suggested in the previous quote from Lewin and Grabbe (cf. Amir, 1976; Cook, 1984; Kelman, 1992). One possible explanation for this "semi-autonomous" relation between the group and its members is based on a distinction between the logical and psychological members of a group, parallel to the distinction made between the classical and graded views of category membership (Smith & Medin, 1981; Rosch, 1978). A college male who is a legal member of a campus fraternity, but who is sensitive, shy, poetic, scholarly, and a member of a radical environmental group may not be psychologically thought of as a fraternity member. The classic or Aristotelian view regards this fraternity member as having the same status as any other logical member of the group, but the graded or prototypic view of category membership, which defines category membership in terms of similarity to a characteristic representation, would view this member as a poor fit to the category, and not much of a "real" fraternity member at all.

When we say an exemplar is not a "real" member of the category, we specifically mean that the individual member does not "come to mind" (implicitly or explicitly) when thinking about the category as a whole. A *psychological* member is one whose identity is likely to be activated and retrieved by the category name.[1] The issue of which exemplars do or do not "come to mind" is critical, because exemplar models of judgment (Hintzman, 1986; Smith & Zarate, 1990, 1992) have shown that our judgments about the attributes of a category are based to a significant extent on the attributes of the activated category members (also, see Bodenhausen, Schwarz, Bless, & Wanke, 1995).

The classical and graded models of category membership make different assumptions about the nature of category change. The contact hypothesis, which focuses specifically on the conditions under which the attributes of individual exemplars who disconfirm the stereotype generalize to the category as a whole, is based, however implicitly, on a classical view of category membership. Since generalization of the attributes of "atypical" members is required for contact-induced change, it is assumed that the degree of generalization is independent of the exemplar's "goodness of fit" to the category—that generalization occurs for typical and atypical members alike.

In contrast, the graded view of category membership predicts that generalization will not be the same for all exemplars, and that generalization from an exemplar is proportional to goodness of fit to the category (cf., Rothbart & Lewis, 1988, Experiment 3). However, a paradox emerges when we apply the graded view of category membership to the problem of stereotype change. Contact-induced stereotype change should occur when exemplars possessing disconfirming information are retrieved by the category as a whole. But the same disconfirming information that has the potential to modify the general stereotype has the effect of making the exemplars with whom it is associated become a poorer fit to the category. To state this another way, disconfirming information associated with exemplars may have two contradictory effects on stereotype change. First, as the disconfirming nature of the information increases, the *potential* to modify the stereotype also increases. Second, as the magnitude of disconfirmation increases, the associated exemplar decreases in its goodness of fit to the stereotype, reducing the likelihood of that exemplar being activated or retrieved by the stereotype label. The paradox, then, is that the information most likely to disconfirm the stereotype is also most likely to "isolate" the exemplar from the category by decreasing the exemplar's typicality.

One important implication of the exemplar model proposed by Rothbart and John (1985) is that the retrieval of an exemplar by a category is a function, not only of the endogenous fit between the features of the category and the exemplar, but also of exogenous factors, such as social context, that are likely to link the

[1] Throughout this paper we talk about the activation or nonactivation of a category member as if activation were all or none. Category activation clearly is continuous rather than all-or-none, and we refer to "activation" rather than "probability of activation" or "degree of activation" only to simplify the argument.

category and exemplar. The "link" or degree of similarity between a category and exemplar, as Smith and Zarate (1992) have convincingly argued, is not fixed, but depends on a number of traditional social psychological variables (social context, social roles, perceiver's motivation, etc.) in addition to the intrinsic similarity between the target stimulus and the stored exemplars. The strength of the relation between category and exemplar is variable rather than fixed, and one important source of variation is the accrual of new information. During the initial stages of acquaintanceship, for example, our knowledge of an individual may be limited to category membership, such as gender, race, occupation, etc., but it is commonly assumed that as we acquire individuating information about the person, the importance of categorical knowledge decreases (e.g., Brewer & Miller, 1984; Fiske & Neuberg, 1990). Although research on the dilution effect (Nisbett, Zukier, & Lemley, 1981), and Brewer's dual process model (Brewer, 1988) suggests that new information should increase the discrepancy between category and exemplar, an exemplar model would predict different effects for information that is consistent rather than inconsistent with the stereotype. Both Smith and Zarate (1992) and Rothbart and John predict that it is the nature of the new, individuating information that determines whether the link between exemplar and category is weakened or strengthened. Individuating information can strengthen category membership if that new information is consistent with the stereotype associated with category membership, and weaken it only if the information is inconsistent with the stereotype.

A number of findings are consistent with predictions generated from the preceding analysis. To test the assumption that judgments about the attributes of a category are disproportionately influenced by the "good-fitting" or typical members of the category, Rothbart and Lewis (1988) presented subjects with "good" and "poor" examples of rectangles, triangles, ellipses, and pentagons, with one color associated with the good (typical) and a different color associated with the poor (atypical) shapes within each category. When subjects were later asked to estimate the frequency of colors associated with, for example, triangles, they estimated the color associated with the good-fitting triangles to be more common than the color associated with the poor-fitting triangles, even though both colors were presented equally often. The use of geometric shapes was considered a stringent test of the idea that good-fitting members are given disproportionate weight in judging category attributes, since mathematical categories, unlike social categories, possess defining rather than characteristic features.

In the experiment on geometric shapes, subjects were presented with all members of the category and asked to summarize the category's attributes. In another experiment, using a social category, subjects were given information about a single category member, who varied in degree of prototypicality, and then asked to estimate the likelihood that the category as a whole possessed the same attribute as the individual category member. Subjects read about a single fraternity member, who was described in a way that indicated a poor, moderate, or good fit to the fraternity stereotype and who had voted for either Mondale or Reagan in

the presidential election. Subjects' inference that the fraternity as a whole voted in the same manner as the individual fraternity member was proportional to the prototypicality of that group member.

A recent experiment by Bodenhausen et al. (1995) provided evidence for the general assumption underlying exemplar models, that judgments about a group are influenced by the attributes of salient (activated) exemplars. They found that favorable judgments about Blacks (as a group) was increased by exposure to favorable exemplars (e.g., Michael Jordan or Oprah Winfrey) in comparison to either neutral exemplars (e.g., Spike Lee) or no exemplars at all. Interestingly, when subjects were asked to rate the typicality of the favorable exemplars (who were generally viewed as atypical of Blacks), these exemplars no longer influenced judgments of the group as a whole.

The findings described above provide enough encouragement about the potential strength of the model to justify its further examination. The model places central importance on the effectiveness of category retrieval cues to differentially access typical and atypical category members and the information associated with those members. By assuming that judgments about a category are based on the attributes of the exemplars most likely to be retrieved by the category and that, in turn, the exemplars most likely to be retrieved are those who already fit the category best, the role of typicality in the retrieval process becomes a critical component of the model. The goal of the present research is to directly examine this central assumption by examining the use of category retrieval cues in (a) accessing typical and atypical category members, and (b) accessing the information associated with those members.

Four studies are presented in this paper. The first is a laboratory experiment examining how the acquisition of new information about an exemplar can change the ability of the category to activate or retrieve that exemplar. The second study examines the relation between retrieval and typicality, but in a more naturalistic setting. Whereas the first two studies focus on the ability of a category name to activate typical and atypical exemplars, the third and fourth experiments focus more specifically on the ability to retrieve neutral (nonstereotypic) information that has been associated with typical or atypical members of a category.

EXPERIMENT 1

Overview

Subjects learned the category membership of 12 male stimulus persons (SPs), with three SPs each associated with four categories: Asian, Black, Gay, and Fraternity. In Phase 1, subjects learned category membership to criterion after which subjects' reaction times to category-exemplar pairings were assessed. In Phase 2, new information about each category member was presented. Within each of the four categories, one SP each was associated with information that either confirmed, disconfirmed, or was irrelevant to category membership. Following that new information, RTs were assessed again, and the changes between the first and second RT testing were compared, as was Ss' accuracy in verifying

category-exemplar relations. It was predicted that the changes in RTs and error rates from Phase 1 to Phase 2 should be linearly related to the disconfirming nature of the new information: subjects should show the greatest decrease in RT (and reduction in errors) when the new information confirmed the stereotype, the smallest decrease in RT (and reduction in errors) when the new information disconfirmed the stereotype, with the irrelevant information condition showing intermediate effects.

Subjects

Seventy-five students from introductory psychology courses participated in this experiment, in partial fulfillment of a course research requirement. Due to an error in procedure, sex of subject was not recorded. Typically, the ratio of females to males in our research is approximately 2:1.

Experimental Materials

The categories Asian, Black, Gay, and Fraternity were chosen on the basis of other research showing good interjudge agreement regarding the four stereotypes, as well as distinct differences between the groups. Twelve common male names were chosen, and 12 behaviors were selected on the basis of their confirming, disconfirming, or irrelevant relation to the stereotype for each of the four groups.

One-paragraph descriptions, designed to be confirming, disconfirming, or irrelevant, were generated for each of the SPs to be presented in phase 2. The behavioral items used as confirming, disconfirming, or irrelevant to the stereotype (for both the original paragraph and the experimental manipulation) were chosen from a larger pool of items rated by a pretest sample of 32 subjects. Subjects rated each behavior on a 9-point scale, ranging from "very uncharacteristic of X" as 1 (where X is one of the four social categories) to "very characteristic of X" as 9, and "neither characteristic nor uncharacteristic of X" as 5. Each behavior was rated for each category. Items with mean scale values of greater than 6 were considered stereotypic, less than 4 were counter-stereotypic, and items between 4.5 and 5.5 were considered irrelevant to the category. Items in the appropriate region, with the smallest standard deviations were chosen for use in the experiment.

For Asian SPs, the confirming description indicated a reserved person working for a computer firm, the disconfirming description an extraverted, boisterous person, and the neutral description a family man who enjoyed hiking. For Black SPs, the confirming description involved a highly athletic person, the disconfirming description a president of a consulting firm who listens to classical music, and the neutral description a person who enjoyed gardening and fishing. For the Fraternity SPs, the confirming description described a fashion-conscious party-goer, the disconfirming description a reflective, thoughtful English major, and the neutral description a chemistry major with a GPA of 3.0. For the Gay SPs, the confirming description indicated a hairdresser who wore brightly colored clothing, the disconfirming description a construction worker who is an avid sports fan, and the neutral description a cautious appliance store manager.

Procedure

The presentation of experimental materials and collection of responses were implemented on Apple II computers using the Applepsych system developed at the University of Oregon (Osgood, 1985, 1986). Subjects were given an overview of the procedure in groups of 2 to 5. They were familiarized with the computer keyboard and the response board used for measuring reaction time and were asked to learn the name of the group associated with various individuals. Subjects first practiced on the response board, matching targets to primes (e.g., "EUGENE" to "OREGON") and the importance of quick and accurate responses was emphasized.

The main study was divided into 2 phases. Each phase consisted of a learning sequence followed by a test sequence. In Phase 1, subjects were first required to associate 12 male names with their membership group. Subjects were presented with the first names of 12 men and their group membership, and then were tested for their learning of the pairings. On every learning trial, a name (e.g., "CARL") appeared on the center of the screen, with the names of the four groups below the male name. Subjects matched the proper name to one of the groups by positioning the cursor on one of 4 group labels (e.g., "Black") at the bottom of the screen. Feedback was presented after every trial (e.g., "CORRECT! CARL IS BLACK"). These trials were randomly permuted in blocks of 12 presentations. The learning sequence in Phase 1 was terminated on the completion of one error-free block. Subjects took an average of 4.8 blocks to reach this criterion.

The test sequence in Phase 1 was composed of 2 consecutive blocks of 24 trials each. On each test trial, Ss fixated on a "+" sign presented on the center of the screen for 1 second. This was followed by a group label (e.g., "FRAT") that was presented for 400 ms. After a delay of 300 ms, the category name went off and a person's name (e.g., "MARK") appeared on screen. Subjects had to decide whether the person was a member of the presented category by pressing the "yes" or "no" key on the response board. The inter-stimulus interval was 3 seconds. In each block of 24 trials, each individual name appeared twice—once as a "true" target and once as a "false" lure. In each block, the trials were randomly permuted with the constraint that the same category or target name not appear on consecutive trials.

In Phase 2, subjects learned additional information about the 12 stimulus persons. Subjects read the new paragraph associated with each of the 12 SPs prior to the learning sequence. On each learning trial, a stimulus person's central attribute from the paragraph was described on screen (e.g., "XXXX HAS AN ATHLETIC BUILD"), and the subject had to choose the correct name for the "XXXX." Four individual names were present at the bottom of the screen, one being the correct name, with three other incorrect names chosen to represent (a) each of the three groups, and (b) each of the three levels of congruency with the stereotype. Subjects positioned the cursor to choose the one correct name from the four names presented. Feedback was presented after every trial (e.g., "CORRECT! ERIC HAS AN ATHLETIC BUILD"). The learning trials were randomly

TABLE 1
REACTION TIMES AND ERROR RATES FOR PHASE 1 AND PHASE 2 AS A FUNCTION OF TYPE OF INFORMATION

	Type of information		
	Consistent	Irrelevant	Inconsistent
1. *Reaction times*			
Phase 1	1127	1137	1078
	(327)	(324)	(290)
Phase 2	959	1010	1022
	(294)	(347)	(348)
Phase 1–Phase 2	169	127	56
	(270)	(299)	(262)
2. *Error rates*			
Phase 1	.19	.16	.15
	(.16)	(.15)	(.13)
Phase 2	.11	.16	.18
	(.14)	(.17)	(.19)
Phase 1–Phase 2	.08	.00	−.03
	(.18)	(.15)	(.19)

Note. SDs are given in parentheses.

permuted in blocks of 12 presentations. Subjects were given an opportunity to review the paragraph information at the end of each block. The learning sequence in Phase 2 was terminated on the completion of two error-free blocks. Despite this relatively strict criterion, subjects only took 3.4 blocks to complete the learning sequence. Finally, in a test sequence identical to the one described earlier, subjects' RTs to category label-individual name pairings were assessed. Most subjects took under 50 minutes to complete the procedure. On completion, subjects were debriefed and thanked for their participation.

Results

Reaction times for correct responses and error rates were analyzed. Reaction times faster than 350 ms and slower than 3000 ms (3% of total RTs) were not included in the reaction time measure, and were not included in the analysis of error rates. The RTs and error rates after Phase 1 represent the baseline measures for speed and accuracy of retrieval. The RTs and error rates after Phase 2 represent the same measures after subjects learned additional confirming, irrelevant, or disconfirming information about individual group members. Thus, the change between Phase 1 and Phase 2 measures constitutes an index of the effects of the three types of information (confirming, irrelevant, and disconfirming) on a category's tendency to retrieve an exemplar. Since we are subtracting the Phase 2 measures from the Phase 1 measures, positive difference scores reflect "improvement," that is, an increase in speed (reduction in RT) and a decrease in error rates. Table 1 contains the data on RTs and error rates for Phase 1 and Phase 2, and the

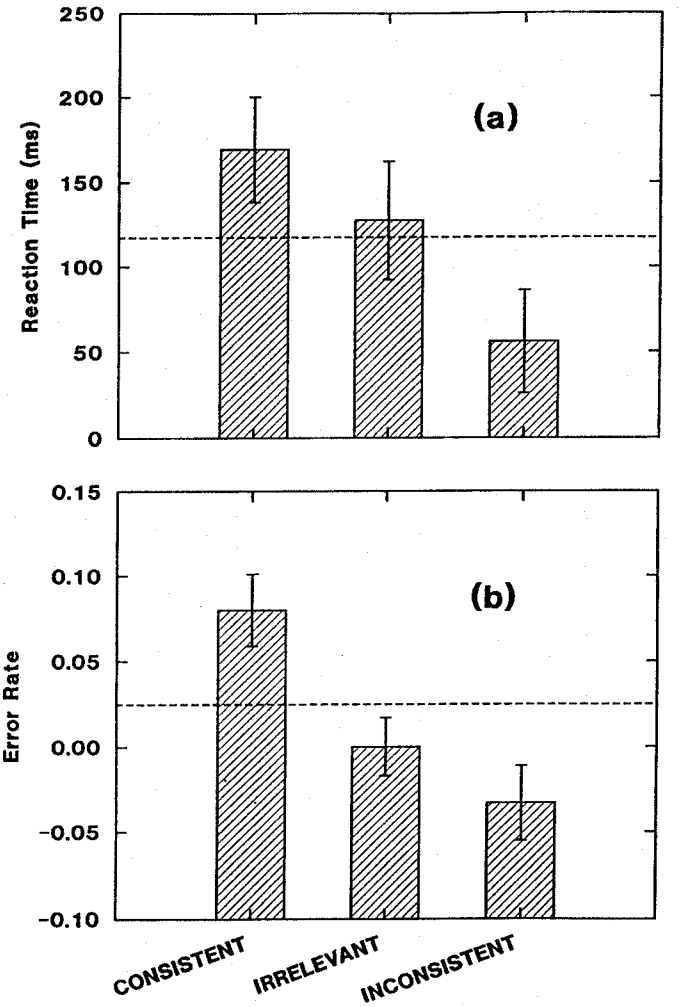

FIG. 1. Reaction time and error rates for differences between Phase 1 and Phase 2 as a function of type of information associated with exemplar (Experiment 1).

differences between Phase 1 and 2 are plotted in Figure 1.[2] It was hypothesized that the change in RTs and error rates from Phase 1 to Phase 2 should be linearly related to the disconfirming nature of the new information. That is, improvements in RTs and error rates should be greatest with confirming information, moderate with irrelevant information, and least with disconfirming information. Reaction times were first subject to log transformations to better meet the assumptions

[2] The dotted line is the mean change for all three conditions combined, and the error bars represent the standard error of the mean. For these within subject comparisons, if any pair of means is greater than two standard errors, the difference is statistically significant.

underlying analysis of variance.[3] We then tested for linear effects in two ways. First, a 2 × 3 (Phase × Consistency) within subjects ANOVA was conducted, testing for a linear trend in the Phase × Consistency interaction. There was a strong Phase effect ($F(1, 74) = 28.63$, $p < .001$), no Consistency effect ($F(2, 148) = .98$, n.s.), and a significant Phase × Consistency interaction ($F(2, 148) = 3.29$, $p < .05$). Most importantly, the Phase × Consistency interaction showed a significant linear trend ($F(1, 74) = 5.44$, $p < .03$), but no significant quadradic trend ($F(1, 74) = 0.19$, n.s.). A second test of the linearity of change was based on a regression analysis in which the linear component of Phase 2 RTs was tested for significance, after the linear component of Phase 1 RTs had been statistically removed. For each subject, Phase 2 RTs were converted to a single linear index by multiplying the RTs in the consistent condition by −1, the RTs in the irrelevant condition by 0, and the RTs in the inconsistent condition by +1. A comparable linear index was then computed for the Phase 1 data, and the Phase 2 linear index regressed on the Phase 1 linear index. The test for the linear trend in Phase 2, having removed the predicted effects from Phase 1, is the significance of the constant tested against 0. This constant was significantly different 0 ($t = 1.99$, $df = 74$, $p < .06$). A comparable test of a quadratic trend, using −1, 2, and −1 as weights, did not show a constant significantly different from 0 ($t = 0.40$, $df = 74$, n.s.).

Parallel analyses were also conducted on error rates, where a linear effect was also expected. Again the first analysis was a 2 × 3 (Phase × Consistency) within subjects ANOVA, testing for a linear trend in the Phase × Consistency interaction. There was no Phase effect ($F(1, 74) = 1.21$, n.s.), no Consistency effect ($F(2, 148) = .38$, n.s.), and a significant Phase × Consistency interaction ($F(2, 148) = 11.15$, $p < .001$). Most importantly, the Phase × Consistency interaction showed a significant linear trend ($F(1, 74) = 16.19$, $p < .001$), but no significant quadratic trend ($F(1, 74) = 0.03$, n.s.). The second test was to analyze the linearity of the Phase 2 error rates with the linear effects of Phase 1 error rates removed. For each subject Phase 2 error rates were converted to a single linear index by multiplying the error rates in the consistent condition by −1, the error rates in the irrelevant condition by 0, and the error rates in the inconsistent condition by +1. A comparable linear index was then computed for the Phase 1 data, and the Phase 2 linear index regressed on the Phase 1 linear index. The test for the linear trend in Phase 2, having removed the predicted effects from Phase 1, is the significance of the constant tested against 0. This constant was significantly different from 0 ($t = 3.46$, $df = 74$, $p < .001$). A comparable test of a quadratic trend, using −1, 2, and −1 as weights, did not show a constant significantly different from 0 ($t = 1.02$, $df = 74$, n.s.).

The data show clear, linear effects for both RT and error rates. The more

[3] Analyses were carried out on both raw scores and log transformed scores. In every case, for all RT analyses presented in this paper, the predicted effects were more significant on the raw scores. Log-transformed scores yielded more conservative results and are the ones presented here.

onsistent the newly acquired information is with the stereotype, the more rapidly and accurately the category will retrieve the exemplar.

Discussion

The RT data for correct responses support the view that the effect of new information on category-exemplar relations depends very much on the nature of that information. Information that increased goodness of fit facilitated the category's retrieval of an exemplar, and the same pattern was supported by the error data.

The results are consistent with the analysis of category-exemplar dynamics outlined by Rothbart and John. Although disconfirming information associated with category exemplars should in principle have a strong effect in modifying category attributes, the prediction that disconfirming information also has the effect of distancing the exemplar from the category is supported by the present findings. By weakening the link between exemplar and category, disconfirming information makes it less likely that the associated exemplar will be retrieved when making judgments about the attributes of the category.

Whether or not new information about a stimulus person leads to increased individuation—and decreased categorical perception—depends very much on whether the new information is consistent or inconsistent with stereotypic expectations. These findings may suggest one possible mechanism for understanding the dilution effect, in which the addition of non-diagnostic information weakens the perceived predictive power of a category (Nisbett et al., 1981). Nisbett et al. found that "an alcoholic" was judged as more likely to engage in child abuse than was "an alcoholic" who was "employed as a shoe salesman." In this case, being "employed as a shoe salesman" was nondiagnostic of child abuse, yet it weakened the predictive power of being "an alcoholic" which was diagnostic of child abuse. The findings of Experiment 1 suggest that the nondiagnostic information may have had its effects by weakening the association between category and exemplar. That is, "an alcoholic who holds a job" may be viewed as a less severe alcoholic than someone simply described as "an alcoholic," presumably because the extreme image of the latter precludes reliable employment. If this explanation is correct, it should be possible to construct nondiagnostic information that strengthens, as well as weakens, the predictive power of diagnostic categories. For example, an item such as "has trouble remembering recent events" would probably be judged as irrelevant to child abuse (and therefore nondiagnostic), but could strengthen the predictive power of the alcoholic label if the item is viewed as congruent with alcoholic amnesia.

These findings also seem inconsistent with our reading of Brewer's dual process model of impression formation (Brewer, 1988), in which new information has the effect of decreasing the strength of association between the initial category and the exemplar. Clearly the findings of Experiment 1 suggest that category membership can, in some instances, be strengthened by individuating information.

The first study addressed the relation of typicality and retrieval using an experimental paradigm that effectively manipulated the degree of typicality between exemplar and category and then, using reaction time measures, assessed the retrievability of the exemplar following category activation. The experimental manipulation of typicality required that subjects learn about the attributes of multiple, contrived stimulus persons in ways that are notably artificial. Although there is reason to believe that in this experimental paradigm reaction time provides a more sensitive index of associative strength between category and exemplar than does free recall, it could be argued that experimental control was achieved at the expense of "ecological validity" and that under natural conditions we learn about the members of a group, and retrieve group members from memory, in ways that are not adequately captured in the first experiment.

EXPERIMENT 2

Study 2 attempts to replicate the findings of Experiment 1 in a more natural setting by examining whether the more typical members of a group are more easily retrieved from memory. Although the concern about contrived laboratory experiments is legitimate, the problems with correlational studies conducted in natural settings are not negligible. The general form of study 2 is to ask subjects to recall the individual members of a group that is well-known to the subjects and of modest size. We can then determine whether the typical members of the group are retrieved earlier, and/or more frequently, than members atypical of the category. One problem, of course, is that subjects are not equally familiar with all members of a group, and familiarity should be a powerful determinant of recall. Moreover, familiarity itself should be strongly affected by liking for the individual members of the group. When recalling the members of favorable groups, in which the most typical members are also the most likable, there is total confounding of likability, familiarity, and typicality. This problem can be solved to some degree by having subjects recall members of disliked groups, where the most typical group members are the least liked, but then there is the additional problem of avoidance of disliked group members, with the attendant lack of information about disliked and/or typical group members. The present study asks members of one campus fraternity to recall the members of another, rival fraternity. After the members are recalled, with order of recall noted, subjects are given a list of all members of the target fraternity and asked to rate all known members on liking, degree of familiarity, and typicality. It is predicted that typical members are more likely to be recalled (and recalled earlier) than atypical members, even when familiarity and liking are controlled.

Subjects

Eighteen members of one campus fraternity agreed to participate in a 20-min psychology experiment for payment of $3 each.

TABLE 2
CORRELATIONS AMONG RECALL, LIKING, FAMILIARITY, AND TYPICALITY (ACROSS TARGET PERSONS)

	Amount of recall	Order of recall	Liking	Familiarity
Order of recall	−.43*	—		
Liking	−.05	.32	—	
Familiarity	.04	−.05	.61*	—
Typicality	.49*	−.47*	−.24	.34

Note. (1) Number of cases = number of target persons = 22. Number of judges = 18. (2) Liking, familiarity, and typicality measures are average ratings across 18 judges.

* $p < .05$, one-tailed test.

Method and Procedure

Members of one campus fraternity, known to be in an antagonistic or rivalrous relationship with another nearby target fraternity, were recruited as subjects in the present study. Subjects were given a questionnaire and first asked to briefly describe, in no more than two sentences, the attributes of the target fraternity. Shortly after giving this description, the subjects were given a page with numbered lines on it and asked to recall from memory as many of the members of the target fraternity as possible. When the recall task was completed, subjects were then given a list of all of the 44 members of the target fraternity, in alphabetical order, and asked to make three judgments about each member "that was at all known to them." The three judgments were: (a) degree of liking, on a 9-point scale anchored by dislike very much at one end (1) and like very much at the other end (9), (b) degree of familiarity, on a 9-point scale anchored by not at all familiar at one end (1), and very familiar at the other end (9), and (c) how typical the member was of the group as a whole, on a 9-point scale anchored at one end as very atypical (1) and at the other end as very typical (9). Subjects were told to rate all members of the target group except for those members "they did not know at all."

Results

The original goal was to analyze the data within subject, correlating recall (present/absent) with rated typicality, partialling out liking and familiarity. This analysis was predicated on a number of assumptions, including the critical assumption that there would be a reasonable number of non-recalled target persons for whom typicality ratings would be available. For most subjects, the target persons who were not recalled were also not rated, presumably because they were not well enough known to the subjects.

It was possible, however, to conduct an analysis aggregated over target persons. That is, a list of target persons recalled by at least one subject was created. Only 22 out of the 44 possible target persons were recalled by at least one subject. For these 22 target persons, the following measures were available: (a) total number of judges recalling that person, (b) the average recall position in the list (the lower the number, the earlier the person was recalled, (c) the average typicality rating, (d) the average familiarity rating, and (e) the average likability rating. Table 2 contains the correlations among the five variables across the 22 target persons. Since the number of subjects (18) is less than the number of cases (22), we used df based on the former rather than the latter in tests of significance. First, the choice

of an unfavorable target group is somewhat vindicated since there is a negative (albeit nonsignificant) correlation between typicality and liking ($r = -.24$, n.s.). However, there is also a positive correlation (also nonsignificant) between typicality and familiarity ($r = .34$). The critical correlations between average typicality and recall were in the predicted direction and significant. There was a positive correlation between average typicality and number of judges who recalled the target person ($r = .49, p < .05$), and a negative correlation between average typicality and average position of recall ($r = -.47, p < .05$) (the lower the recall position, the earlier the person was recalled in the list). Since familiarity and liking correlate well ($r = .61, p < .01$, two tailed test), they were averaged into a single variable for use in computing partial correlations. When the combined familiarity/liking measure is partialled out of the correlation between typicality and recall, the original correlations are vitually unchanged ($r = .49$ and $-.48$, respectively). Thus, there is evidence that the more typical members of the target group are more likely to be recalled, and more likely to be recalled earlier in the list.[4]

Discussion

Although the present study is limited by the small number of subjects, and the problems of obtaining ratings on nonrecalled target persons, the data are nonetheless supportive of the experimental findings presented earlier in this paper: Typical group members are more readily retrieved than atypical members. The difficulty in testing the typicality-retrieval hypothesis in natural settings should also be clear from the present study, however. Subjects are not equally familiar with all members of the target group, and this creates more of a problem as the size of the target group increases. Indeed, for this target group, which is one of the smallest fraternities on campus, only 22 of the 44 members were recalled by at least one subject, and 12 out of the 18 judges recalled no more than 4 members (out of 44 possible). This is particularly problematic since we chose this pair of fraternities specifically because of the extensive contact and mutual knowledge between the groups. For very large target groups, judges may only be familiar with the highly typical members of the category, rendering a test of the hypothesis somewhat questionable. The virtue of the experimental procedures is that the experimenter can control the amount of information associated with typical and atypical category members, insuring a more stringent test of the relation between typicality and retrieval.

The first experiment examined how the acquisition of new information about a

[4] Caution is recommended in interpreting the order of recall. The number of target persons recalled varies across subjects, and for some target persons data from only one or two subjects determines their order of recall. Target persons who are recalled by many subjects will, to some extent, receive lower order values (i.e., earlier recall), because many of the subjects are remembering only one or two targets. In fact, there was a correlation between the number of subjects recalling a target and the target's order ($r = -.43$, $p < .05$, one-tailed test). This is a case in which correlation between typicality and order computed within-subject would have been superior, since recall and order would not be confounded.

category exemplar modified the category's strength in activating the exemplar. In that experiment, all SPs were initially "good" members of the category, but then became associated with information that strengthened, weakened, or did not alter the goodness of fit with the category. Both studies dealt with a category's tendency to activate typical and atypical members of the group. The next experiments, while using category cues to activate typical and atypical exemplars, examine directly the likelihood of associating neutral (nonstereotypic) exemplar attributes with the category labels. Since exemplar based models assume that the content of the stereotype will be some function of the attributes associated with the activated exemplars, it is important to show more directly the link between the category and the exemplar attributes. It is predicted that neutral, nonstereotypic attributes will be retrieved more readily when associated with typical than atypical category members. Subjects first learn about the behavior of SPs whose occupation is either typical or atypical of their gender category, so that the typicality/atypicality of the stimulus persons is available during initial encoding. Later, subjects are primed with either a gender or occupational category followed by the presentation of a SP's behavior (which is unrelated to gender or occupation), and subjects are asked to verify whether the behavior was engaged in by a person in the primed category. If the primed category differentially activates the typical over the atypical members of the category, subjects should be faster to verify the correct behavior of a typical than an atypical member of the category.

EXPERIMENT 3

Overview

Subjects acquired information about six male and six female stimulus persons (SPs), with half the SPs within each gender described as scientists and half described as teachers. Pretesting indicated that male scientists and female teachers were considered typical, while female scientists and male teachers were considered atypical. Subjects also learned the hobbies associated with SPs, where hobbies had been pretested to be equally associated with both genders and both occupations. Subjects were later asked to verify whether or not a specific hobby or activity (e.g., swimming) was associated with a given category (i.e., man, woman, scientist, or teacher). It was expected that the category name would serve as a prime to activate the members of the category, with the typical members more strongly activated than the atypical members. Thus, when asked to verify whether a behavior was engaged in by a member of the primed category, it was predicted that subjects would be faster and more accurate in correctly verifying the hobbies associated with typical (male scientists, female teachers) than atypical (male teachers, female scientists) category members.

Method

The goal of this experiment was to examine the relative ease of retrieving the attributes of typical and atypical exemplars. To achieve this goal, subjects were given information about twelve stimulus persons, and later were asked to verify

TABLE 3
MASTER LIST OF NAMES, OCCUPATIONS, AND HOBBIES OF STIMULUS PERSONS

First name	Last name	Area of work	Hobby
John	Lewis	Nuclear Energy	Tennis
Kevin	Bennett	Astrophysics	Jogging
Robert	Brady	Electronics	Wine-Tasting
Melissa	Johnson	Nursery Education	Trivial Pursuit
Rebecca	Little	Special Education	Nature Walking
Sharon	Garrison	Music Education	Movie Going
Harry	Schmidt	Ballet Instruction	PBS TV Viewing
Peter	Franklin	Grade School Education	Swimming
David	Newman	Sunday School Education	Color Photography
Helen	Ambrose	Atmospheric Science	Training Dogs
Susan	Hays	Chemistry	Downhill Skiing
Jennifer	Stuart	Computers	Bicycling

statements about SPs as true or false. More specifically, subjects were presented with a category name (MAN, WOMAN, SCIENTIST, or TEACHER) on a computer screen, and then asked whether a specific hobby was correctly associated with the presented category. "Neutral" hobbies were used as the target behavior so that none of the four category names would have an advantage or disadvantage in retrieving the target behavior. By choosing hobbies/activities that were independent of gender and occupation, none of the gender or occupation categories could retrieve the hobby by direct prior association.

Subjects

Sixty-two students, 39 women and 23 men, from introductory psychology courses participated in this experiment, in partial fulfillment of a course research requirement. Again due to an error in procedure, sex of subject information was not paired with experimental sessions and its effects could not be assessed.

Materials

In order to generate appropriate stimulus materials for this study, a pilot study was conducted to select gender-typed occupational categories as well as hobbies perceived to be independent of gender and occupation (see Table 3). Six teaching specialities perceived to be predominantly associated with women and six science specialities perceived to be predominantly associated with men were selected. Subjects estimated, on average, 72% of the people in the 6 teaching specialities to be women, and 72% of the people in the 6 scientific specialities to be men. Subjects in the pilot study also rated proportions of men, women, teachers, and scientists pursuing various hobbies. Twelve hobbies were chosen to be neutral to both the gender and occupation (teacher–scientist) dichotomies. Pretest subjects were asked to judge the likelihood that each hobby was preferred by men/women and by teachers/scientists, and only hobbies showing close to a 50/50 split on both dimensions were chosen for use in this experiment.

Based on the occupations and hobbies selected, 12 stimulus persons (SPs) were constructed. The 12 SPs consisted of 3 male teachers, 3 male scientists, 3 female teachers and 3 female scientists. For every subject, the specific names, occupations, and hobbies of these 12 SPs were constructed randomly from the master list of attributes. Table 4 contains examples of constructed SPs.

TABLE 4
EXAMPLES OF STIMULUS PERSONS PRESENTED TO SUBJECTS

Example of male scientist:
David Garrison, 49, is a scientist working in the field of chemistry.
David feels energized by swimming.
Example of male teacher:
Robert Johnson, 35, is a teacher working in the field of nursery education. Robert does regular jogging.
Example of female scientist:
Rebecca Schmidt, 32, is a scientist working in the field of electronics. Rebecca keeps fit by bicycling.
Example of female teacher:
Jennifer Newman, 38, is a teacher working in the field of special education. Jennifer is accomplished in color photography.

Procedure

The presentation of experimental materials and collection of responses were implemented on IBM PS/2 computers, with 8088 Intel chips, using the MEL system (Schneider, 1990). Two to five subjects were tested individually in sound-proof booths. They were first familiarized with the functions of keys on the computer keyboard, and then were told that they would be forming impressions of certain individuals based on short descriptions that included information about names, occupations, and hobbies.

There were three phases to the experiment. In the introductory phase, subjects were presented with a description of each of the 12 SPs, one at a time, on the computer screen (examples of the descriptions are given in Table 4). For each presented description, subjects were asked to imagine what each person would be like if they actually met her/him. The array of 12 SPs were presented twice, in different random orders. The second phase, a learning sequence, was designed to help subjects learn the correct associations of name (where first name conveys gender), occupation, and hobby. On each trial subjects were provided with two of a SPs characteristics (e.g., name and hobby) and had to select the correct third characteristic (e.g., occupation) from among four options. The four options included the one correct option and three incorrect options drawn equally from the other three cells of the design (male scientist, male teacher, female scientist, female teacher). After each trial, subjects received feedback in the form of the word CORRECT or INCORRECT presented on the screen, with the correct option indicated in the event of an error. Since each of the 12 SPs could have 3 possible questions asked about them (all combinations of three attributes taken two at a time), there were 36 trials in a block, with subjects receiving two randomized blocks of trials.[5]

The second phase, or learning sequence, was followed by an irrelevant intervening task lasting 15 min. After the intervening task, the third phase began in which subjects received the test sequence, in which subjects had to verify whether a name, occupation, or hobby was correctly paired with each of the four possible category names.

On each trial, subjects were presented with one of the four category names (MAN, WOMAN, SCIENTIST, TEACHER) followed by either (1) an occupational speciality (e.g., CHEMISTRY, NURSERY EDUCATION), 2) hobby (e.g., JOGGING, SWIMMING), or 3) proper name (e.g., JOHN LEWIS, SHARON GARRISON), and asked to judge, as quickly as possible, whether the occupation, hobby or name was correctly associated with the named category). The only data of immediate interest was the speed and accuracy of verifying category and hobby, since information about hobby was designed to be independent of category. The pairing of category and hobby was embedded within the

[5] That is, if the correct occupation was chemist, the incorrect options would be astrophysicist, ballet teacher, Sunday school teacher. The same patterns were followed when the missing information was name or hobby.

TABLE 5
REACTION TIME AND ERROR RATE AS A FUNCTION OF TYPICALITY AND RETRIEVAL CATEGORY (EXPERIMENT 3)

	Type of category serving as retrieval cue			
	Gender (Male/female)		Occupation (Scientist/teacher)	
Stimulus person:	Typical	Atypical	Typical	Atypical
Reaction time (ms)	1335 (300)	1364 (282)	1743 (383)	1828 (325)
Error rates	.13 (.15)	.14 (.17)	.35 (.17)	.40 (.24)

$N = 56$

Note. SDs are given in parentheses.

other verification tasks (category and name, category and occupation), to mask the intent of the experiment. During the test series, each of the 12 SPs were cued by each of the 4 categories, resulting in 48 trials. For each name, two of the four category cues were correct and two were false (e.g., if "MAN" was a correct pairing, "WOMAN" was an incorrect pairing). All hobbies, occupations and names appeared both as "true" targets and as "false" lures. The 48 trials of the test sequence were presented in a random order. On each trial, the prompt "Get Ready" appeared on screen for 300 ms, followed by an offset for 300 ms, followed by the category name presented for 1000 ms. Three hundred milliseconds after the offset of the category name, the target information was presented. Subjects were instructed to decide, as quickly and accurately as possible, whether the target information (hobby, occupation, or name) was correctly or incorrectly paired with the category by pressing the "Y" or "N" keys on the response board with a finger of their dominant hand. Subjects were given several practice trials to stabilize their response times. The intertrial interval was 1600 ms. Subjects typically took about 50 min to complete the procedure. Upon completion, subjects were debriefed and thanked for their participation.

Results

Reaction times for correct responses and error rates were analyzed. Reaction times faster than 350 ms and slower than 3000 ms (6% of total RTs) were not included in the reaction time measure, and were not analyzed further. A $2 \times 2 \times 2$ (Typicality of Exemplar $\times$ Type of Category [gender versus occupation] $\times$ Sex of Stimulus Person) repeated measures analysis of variance was run on log-transformed reaction times and on error rates, and the results are presented in Table 5 and Fig. 2.

Six subjects whose overall error rates were at chance level (>.45) were dropped from the analyses. There was again a significant effect of typicality ($F(1, 55) = 6.74, p < .02$), with faster responses to typical ($M = 1539$) than to atypical exemplars ($M = 1596$). Type of category was also highly significant ($F(1, 55) = 331.24, p < .001$), with subjects responding far more quickly to gender ($M = 1350$) than to occupation ($M = 1785$), but there was no significant interaction between typicality and type of category ($F(1, 55) = 0.85$, n.s.). There were no significant higher order interactions.

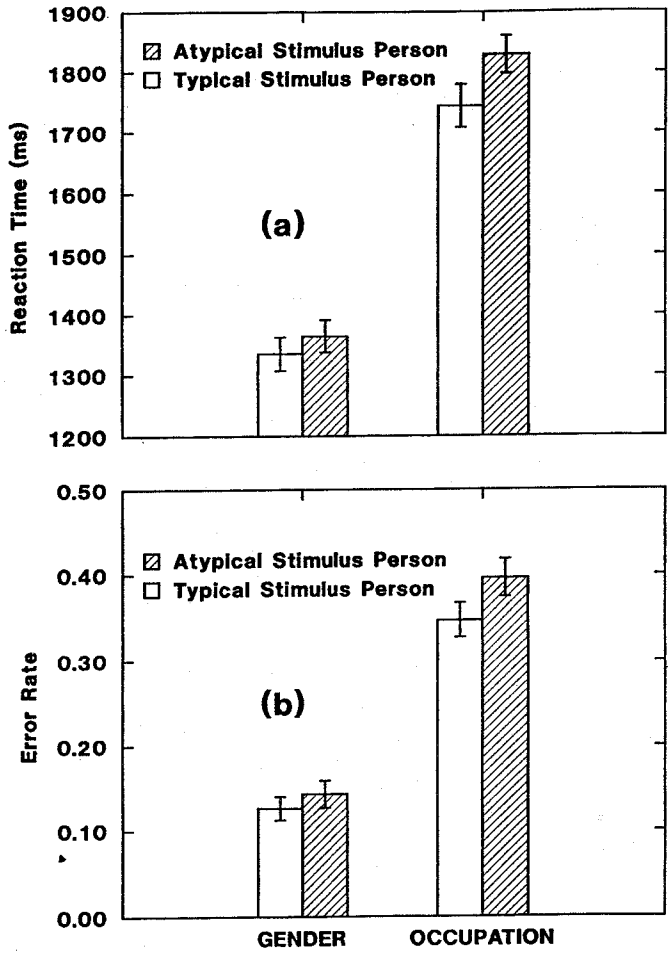

FIG. 2. Reaction time and error rates as a function of categorical retrieval cue and typicality (Experiment 3).

The error rates generally parallel the RT data. There was a marginally significant effect of typicality ($F(1, 55) = 3.57$, $p < .07$), with fewer errors for typical ($M = .24$) than atypical ($M = .27$) exemplars. There was also a strong effect of type of category ($F(1, 55) = 167.14$, $p < .001$) with fewer errors for gender ($M = .14$) than for occupation ($M = .37$). There was no significant interaction between typicality and type of category ($F(1, 55) = 1.18$, n.s.). There were no significant higher order interactions.

Discussion

Although information associated with typical stimulus persons was retrieved more accurately and more quickly than information associated with atypical SPs,

the effect was significant for reaction time, but of marginal significance for error rates. One possible reason for the lack of significance for the error rates is that the overall level of error is relatively low, particularly for gender. Accuracy based on gender is relatively high (.86), and there may be a ceiling effect that minimizes the effect of typicality.

In general, the findings are consistent with the view that when category names are used as retrieval cues for information about category members, it will be more difficult (as measured by speed and accuracy) to retrieve the information for atypical than typical category members. In a fourth experiment, we attempt to achieve a more sensitive test of the effects of typicality on error rates repeating the basic structure of Experiment 3 with more stimulus persons (to increase the error rate) and a larger sample. In Experiment 4, we also correct the procedural error to enable a test for sex of subject effects.

EXPERIMENT 4

Overview

Using the same basic paradigm as in Experiment 3, the number of stimulus persons was increased from 12 to 20, using 10 male and 10 female stimulus persons, with half of each gender described as scientists and half as teachers. It was again predicted that subjects should verify category information more quickly and more accurately for typical than atypical category members. By increasing the number of stimulus persons, the number of errors would increase, permitting a more sensitive test of the effects of typicality on gender.

Subjects

One hundred students, 66 female and 34 male, drawn from introductory psychology courses participated in this experiment, in partial fulfillment of a course research requirement.

Materials

The same materials were used as in Experiment 3, except that 8 new hobbies, pretested as neutral with respect to gender and occupation, along with 8 new occupations and names were added.

Procedure

Except for the change in number of stimulus persons, and a changeover from IBM PS/2 with 8088 chips to IBM clone computers with 80486 chips, the procedure was the same as in Experiment 3.[6] During the test trials, each of the 20 SPs were cued by each of the 4 categories, resulting in 80 trials.

Results and Discussion

A 2 × 2 × 2 × 2 (Typicality of Exemplar × Type of Category (gender versus occupation) × Sex of Stimulus Person × Sex of Subject) repeated measures analysis of variance was run on log-transformed reaction time and accuracy.

[6] The change to computers operating at faster clock speeds resulted in artifactually faster RTs in Experiment 4 than Experiment 3 by approximately 165 ms.

TABLE 6
REACTION TIME AND ERROR RATE AS A FUNCTION OF TYPICALITY AND RETRIEVAL CATEGORY (EXPERIMENT 4)

	Type of category serving as retrieval cue			
	Gender (Male/female)		Occupation (Scientist/teacher)	
Stimulus person:	Typical	Atypical	Typical	Atypical
Reaction time (ms)	1240	1270	1485	1618
	(242)	(246)	(366)	(394)
Error rates	.16	.18	.34	.44
	(.15)	(.18)	(.18)	(.20)
$N = 83$				

Note. SDs are given in parentheses.

The increased memory load from 12 to 20 stimulus persons increased the error rate substantially. Seventeen subjects (as compared to 6 from Experiment 2), 8 males and 9 females, had an overall error rate greater than 45% and were dropped from the experiment.[7] There were no significant main effects or interactions with sex of subject for either reaction time or error rates.

Log-transformed reaction times for correct responses only were analyzed (see Table 5), and the results are presented in Table 6 and Fig. 3. Extremely short (<350 ms) and long (>3000 ms) reaction times (7% of total RT responses) were excluded from the RT analysis and not analyzed further. There was again a highly significant effect of Typicality ($F(1, 82) = 21.46, p < .001$), with faster responses to typical ($M = 1362$) than to atypical ($M = 1444$) exemplars. The main effect of Type of Category was also significant, $F(1, 82) = 165.01$, $p < .001$. Subjects responded more quickly to gender ($M = 1255$) than to occupation ($M = 1551$). There was a significant interaction between typicality and type of category ($F(1, 82) = 8.25, p < .01$), showing a greater typicality effect for occupation ($M = 1485$ vs $M = 1618$) than gender ($M = 1240$ vs $M = 1270$). Planned comparisons indicated a highly significant difference between RTs for typical and atypical exemplars when verifying hobbies with occupations ($F = 32.78, p < .001$) and a non-significant difference between RTs for typical and atypical exemplars when verifying hobbies with gender ($F = 2.75$, $p < .11$). There were no significant higher order interactions.

For error rates, there was a significant main effect of Typicality ($F(1, 81) = 14.42, p < .001$), with greater errors for atypical ($M = 0.31$) than typical ($M = .25$) exemplars. There was also a significant effect of Type of Category ($F(1, 81) =$

[7] Statistical analyses were performed on the original full sample, as well as on the selected sample, and all the critical statistical decisions were the same for both samples. The presented analyses and results focus on the selected sample, since there seemed to be little justification for including subjects performing at chance level.

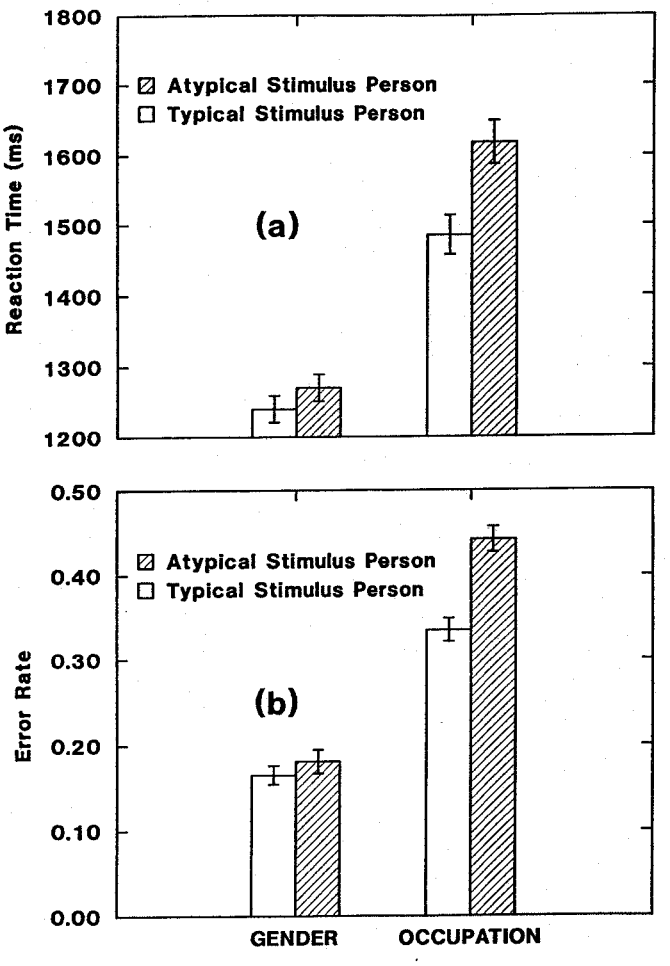

FIG. 3. Reaction time and error rates as a function of categorical retrieval cue and typicality (Experiment 4).

183.14, $p < .001$), with greater errors for occupational categories ($M = .39$) than for gender categories ($M = .17$). The interaction of Typicality and Category was also significant ($F(1, 81) = 7.76$, $p < .01$), with a much smaller typicality effect for gender (.16 vs .18) than for occupation (.34 vs .44). Planned comparisons indicated a highly significant difference in error rate for typical and atypical exemplars when verifying hobbies with occupations ($F = 22.25$, $p < .001$) and a nonsignificant difference in error rate for typical and atypical exemplars when verifying hobbies with gender ($F = 0.52$, n.s.).

The results of Experiment 4 consistently replicate the general findings of Experiment 3. By increasing the memory load of subjects in Experiment 4, we increased the overall error rate only slightly over those of subjects in Experiment

3. The goal of increasing the error rate was to determine whether typicality differences in error rate would emerge. Although the typicality effect for both RT and error rate was strong, the typicality effect was now significantly stronger for verification based on occupational categories than for verification based on gender categories. Planned comparisons showed very large effects for typicality based on occupational categories, for both RT and error rates, but for gender categories showed only a small effect for RT and no effect for error rate.

Most of the typicality effect, for both Experiments 3 and 4, seems to be carried by the occupational categories. That is, attempts to retrieve information about teachers who are male, or scientists who are female is much more difficult than retrieving the same information when associated with teachers who are female and scientists who are male. We are not sure why the typicality effect associated with gender, while tending in the predicted direction, is so much weaker than that based on occupation.

One possible interpretation is that encoding by gender, even with the slightly increased error rate, is still so efficient that it is relatively uninfluenced by typicality. Clearly, however, gender information is having a powerful effect on typicality, but it is having its effects through occupation. That is, a teacher who is male, or a scientist who is female, is much more difficult to retrieve than the same occupations when paired with female and male, respectively.

One issue of concern is whether the typicality effects in Experiments 3 and 4 are due to differential learning of the typical and atypical stimulus persons, or to differential retrieval. Although subjects in Experiment 1 learned the stimulus persons' attributes to a criterion, subjects in Experiments 3 and 4 did not learn the information to criterion, but were presented with two complete blocks of learning trials. For Experiment 4 it was possible to determine whether the difference in retrieval strength of the typical and atypical exemplars was associated with degree of initial learning.[8] For each subject, the number of errors made during the learning sequence was correlated with the strength of the typicality effect, as measured by the difference in reaction time to typical minus atypical stimulus persons. There was no significant correlation between initial learning and the strength of the typicality effect ($r = .02$, n.s.).

GENERAL DISCUSSION

The original model proposed by Rothbart and John attempted to specify the relation between the attributes of a category and the attributes of the exemplars that make up the category. It explained the often-observed disparity between a category and its members by using ideas implicit in a graded model of social category membership in the context of an exemplar model of judgment (Hintzman, 1986). It assumed that the judgments about the attributes of a category are based on the attributes of those exemplars most likely to be activated by the

[8] The same analysis was not possible for Experiment 3 where number of errors made during the initial learning sequence was inadvertently not recorded.

category, and that the exemplars most likely to be activated by the category are the ones who already fit the category best—that is, are most typical and/or most similar to the category. It is the circularity of this last assumption that functionally isolates "poor-fitting" group members, allowing stereotypes to remain insulated from disconfirming information, and permits a continuing disparity between social perception and social reality (Judd & Park, 1993). The two key components of the original model are thus (a) the general assumption common to exemplar models that the attributes of exemplars most similar to a target influence judgment, and (b) the more specific assumption that similarity is based upon the goodness of fit between the attributes of the group stereotype and the group member.

The present experiments, then, have general implications for exemplar models of judgment (e.g., Smith & Zarate, 1992), as well as more specific implications for thinking about the complexity of stereotype change. Experiment 1 provided strong support both for the importance of typicality as the basis for category activation of an exemplar, and for the idea that the link between category and exemplar is dynamic and not static. In that experiment new information acquired about the exemplar changed the link between category and exemplar, suggesting a possible mechanism for explaining changes in the diagnosticity of information (Nisbett et al., 1981). Experiment 2 showed the amount and order of recall of real category members in a more natural setting was also related to the exemplar's goodness of fit to the group stereotype. Experiments 3 and 4 also demonstrated the importance of an exemplar's typicality in retrieving the specific attributes of category members, even when the attributes were unrelated to the category retrieval cue. These data suggest that it is the similarity of the retrieval cue to the exemplar, rather than to the attribute itself, that guides judgment.

A recent paper by Maurer, Park, and Rothbart (1995) also illustrates the importance of typicality in judging a group's attributes, as well as the mutable nature of category-exemplar relations. All subjects were presented with identical information about a group's members. In one condition, subjects judged each member's goodness of fit to the group as a whole, while another condition was asked to sort the members into subcategories that minimized differences within subcategories, and maximized differences between subcategories. In a baseline control condition, subject were given no organizing instructions. Subjects in the "goodness of fit" (or subtyping) condition viewed the group in the most stereotypic and most homogeneous way, while "subcategorizing" subjects viewed the group in the least stereotypic and most heterogeneous way. Subjects appeared to implicitly define group boundaries in different ways, with the "goodness of fit" subjects functionally excluding the disconfirming exemplars from the superordinate category, while the "subcategorizing" subjects included the disconfirming exemplars within the superordinate category, and viewed the group in an appropriately complex manner. One measure of "psychological exclusion" is the difference between the typicality ratings for the good and poor fitting group members, which was found to be much larger in the "goodness of fit" condition than in the

"subcategorizing" condition. One important implication of the study is that typicality itself is not invariant, and can be influenced by the subject's implicit or explicit organizing strategies. In this way, the results are consistent with the previously cited research by Bodenhausen et al. (1995) which showed that awareness of the atypicality of the exemplar inhibited generalization from exemplar to category.

Although the model was developed to account for the isolation of stereotype-disconfirming information, it does predict change under limited conditions. In particular, moderately disconfirming exemplars should have more disconfirming impact on the stereotype than would strongly disconfirming exemplars (Weber & Crocker, 1983; Hewstone et al., 1994; Kunda & Oleson, 1995). We would expect that as moderately disconfirming information is accommodated into the stereotype, the central tendency of the stereotype would begin to change and thus would begin to activate a different set of exemplars—ones that are more discrepant from the original stereotype. This line of reasoning illustrates the bidirectional and dynamic fashion in which stereotype change may occur. The attributes of the existing category select the exemplars, but new, good-fitting exemplars (or new attributes of old exemplars) can also modify the attributes of the category in a recursive process.

Lord and his colleagues have also shown the importance of typicality in the process of stereotype change and have identified a number of interesting factors that mediate change. Ramsey, Lord, Wallace, and Pugh (in press) noted that the functional definition of the superordinate category (e.g., "mentally ill") may or may not match the attributes of the exemplar (e.g., as "depressed" or "paranoid"). Attitudes toward the category and exemplar were more consistent when the exemplar was considered a member of superordinate category. Lord, Desforges, Fein, Pugh, and Lepper (1994) found that for proponents of some social issues (e.g., the death penalty), application of a social policy was more consistently applied for typical than atypical members of the relevant category (e.g., murderers), presumably because the policy is to some extent based on typical or idealized exemplars.

To return to the issue raised by Lewin and Grabbe, how do we come to view an individual group member as typical of the category? Rothbart and John made the implicit assumption that the activation of an exemplar by a category at the time of judgment was directly related to the exemplar's original association with a category during encoding. That is, if our hypothetical fraternity member was though of as a "brain" rather than as a "frat" member, it is unlikely that he would later be thought of as a fraternity member. It was argued, however, that the goodness of fit during encoding is dependent on a number of contextual factors that could influence the category associated with the exemplar. That is, if we learned that our poetic, scholarly fraternity member was also a terrific quarterback, beer swiller, and party-goer, the goodness of fit to the stereotype of fraternity would be increased substantially. More generally, as the salience of a particular category was increased, either endogenously by the attributes of the

exemplar, or exogenously by social context, the likelihood of the link between exemplar and category increased.

It is very possible, however, that the link between exemplar and category during encoding and the link between category and exemplar during judgment is not symmetrical. For example, if white subjects were presented with video clips of a speech by Martin Luther King, Jr., there is little doubt that the category "black" would be strongly activated by the speech. But there is also little doubt that Martin Luther King, Jr., is unlikely to be activated as an exemplar when whites think of the category "blacks" since King is considered quite atypical of that category. It may be that the category activated by an exemplar during encoding is a necessary but not a sufficient condition for that exemplar being activated by the category at a later time. More generally, the question of what category is used to encode a stimulus person, and how that relates to later judgments about the stereotype is an important issue and needs to be addressed in future research.

The present experiments also have some relevance for the continuing debate on retrieval of confirming and disconfirming information (cf. Stangor & McMillan, 1992). To the extent that category retrieval cues are more likely to access typical than atypical exemplars, as suggested by all of the present experiments, as well as by Experiments 1 and 2 in Rothbart and Lewis (1988), it suggests superior memory for instances that are consistent, rather than inconsistent, with stereotypic expectancies. More generally, it should be noted that research on prototypes, going back to the early work of Rosch (1977, 1978), established that a category name is more likely to activate good than poor fitting members of the category. The current findings on memory for congruent and incongruent information, for both groups and individuals, are complex, however, and do not yield to simple conclusions. There is evidence that when the target stimulus is an individual or a highly cohesive group, there is generally superior memory for incongruent over congruent information. However, for groups characterized by low cohesion, there is evidence of superior memory for congruent over incongruent items (Rothbart, Evans, & Fulero, 1979; Srull, Lichenstein, & Rothbart, 1985). Srull et al. argued that the key to the pattern of findings lay in the asymmetry in activation of congruent and incongruent items. Incongruent items were more likely to be retrieved because they were cued or activated by both congruent and incongruent items, whereas congruent items were mostly activated only by other congruent items. It may be useful to consider two processes to be involved: (a) one process involves the activating links between items, as described by Srull, and (b) the other involves the activating link between category and exemplar, as described by Rosch. The first process seems to be driven by an attempt to form a coherent or unitary impression from an array of heterogeneous information, which directs attention to contradictory or discrepant information (e.g., Hastie, 1984; Hamilton & Sherman, 1996). Since we expect coherence from individuals and cohesive groups, more effort is spent integrating incongruent information. For groups with low cohesion, or for whom there are only weak expectancies, there is less need to

integrate the discrepant individuals, since observers do not expect a high level of coherence. For the latter process, the category name is the potent retrieval cue, and that name is more likely to activate a good rather than a poor member of the category (e.g., furniture is more likely to activate chair than fan). In any given context, both processes may be implicated, with the exact mixture of the two depending on a number of factors, including processing goals, strength of expectations, and intrinsic coherence of the material. Since many stereotypes involve relatively weak expectations, allowing many "exceptions-to-the-rule," we would expect the process based on category-exemplar relations to be paramount when thinking about the relation between a group and its members.

REFERENCES

Amir, Y. (1976). The role of intergroup contact in change of prejudice and ethnic relations. In P. Katz (Ed.), *Towards the elimination of racism.* New York: Pergammon Press.

Bodenhausen, G. V., Schwarz, N., Bless, H., & Wanke, M. (1995). Effects of atypical exemplars on racial beliefs: Enlightened racism or generalized appraisals? *Journal of Experimental Social Psychology,* **31,** 48–63.

Brewer, M. B. (1988). A dual process model of impression formation. In T. K. Srull & R. S. Wyer, Jr. (Eds.), *Advances in social cognition* (pp. 1–36). Hillsdale, NJ: Erlbaum.

Brewer, M. B., & Miller, N. (1984). Beyond the contact hypothesis: Theoretical perspectives on desegregation. In M. B. Brewer & N. Miller (Eds.), *Groups in contact: The psychology of desegregation.* New York: Academic Press.

Cook, S. W. (1984). Cooperative interaction in multiethnic contexts. In N. Miller & M. B. Brewer (Eds.), *Groups in contact: The psychology of desegregation* (pp. 155–185). New York: Academic Press.

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In *Advances in experimental social psychology* (Vol. 23). New York: Academic Press.

Hamilton, D. L., & Sherman, S. J. (1996). Perceiving persons and groups. *Psychological Review,* **103,** 336–355.

Hastie, R. (1984). Causes and effects of causal attribution. *Journal of Personality & Social Psychology,* **46,** 44–56.

Hewstone, M., Macrae, C. N., Griffiths, R., Milne, A. B., & Brown, R. (1994). Cognitive models of stereotype change: (5) Measurement, development and consequences of subtyping. *Journal of Experimental Social Psychology,* **30,** 505–526.

Hintzman, D. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review,* **93,** 411–428.

Judd, C. M., & Park, B. (1993). Definition and assessment of accuracy in social stereotypes. *Psychological Review,* **100,** 109–128.

Kelman, H. C. (1992). Coalitions across conflict lines: The interplay of conflicts within and between the Israeli and Palestinian communities. In S. Worchel & J. A. Simpson (Eds.), *Conflict between people and groups* (pp. 236–258, 293–294). Chicago: Nelson-Hall.

Kunda, Z., & Oleson, K. C. (1995). Maintaining stereotypes in the face of disconfirmation: Constructing grounds for subtyping deviants. *Journal of Personality and Social Psychology,* **68,** 565–580.

Lewin, K., & Grabbe, P. (1945). Conduct, knowledge, and acceptance of new values. *Journal of Social Issues,* 53–64.

Lord, C. G., Desforges, D. M., Fein, S., Pugh, M. A., & Lepper, M. R. (1994). Typicality effects in attitudes toward social policies: A concept-mapping approach. *Journal of Personality and Social Psychology,* **66,** 658–673.

Maurer, K. L., Park, B., & Rothbart, M. (1995). Subtyping versus subgrouping processes in stereotype representation. *Journal of Personality and Social Psychology,* **69,** 812–824.

Nisbett, R. E., Zukier, H., & Lemley, R. E. (1981). The dilution effect: Nondiagnostic information weakens the implications of diagnostic information. *Cognitive Psychology,* **13,** 248–277.

Osgood, G. W. (1985, August). *Software program development guide* (2nd ed.). Eugene, OR: University of Oregon Psychology Department.

Osgood, G. W. (1986, April). *Hardware implementation and utility manual* (2nd ed.). Eugene, OR: University of Oregon Psychology Department.

Ramsey, S. L., Lord, C. G., Wallace, D. S., & Pugh, M. A. (in press). The role of subtypes in attitudes towards superordinate social categories. *British Journal of Social Psychology.*

Rosch, E. (1977). Human categorization. In N. Warren (Ed.), *Advances in cross-cultural psychology* (Vol. 1). London: Academic Press.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization.* Hillsdale, NJ: Erlbaum.

Rothbart, M., Evans, M., & Fulero, S. (1979). Recall for confirming events: Memory processes and the maintenance of social stereotypes. *Journal of Experimental Social Psychology,* **15,** 343–355.

Rothbart, M., & John, O. P. (1985). Social categorization and behavioral episodes: A cognitive analysis of the effects of intergroup contact. *Journal of Social Issues,* **41,** 81–104.

Rothbart, M., & Lewis, S. (1988). Inferring category attributes from exemplar attributes: Geometric shapes and social categories. *Journal of Personality and Social Psychology,* **55,** 861–872.

Schneider, W. (1990). MEL user's guide: Computer techniques for real time psychological experimentation. Pittsburgh, PA: Learning Research Development Center, University of Pittsburgh.

Sears, D. O. (1983). The person-positivity bias. *Journal of Personality and Social Psychology* **44,** 233–250.

Smith, E. E., & Medin, D. L. (1981). *Categories and concepts.* Cambridge, MA: Harvard University Press.

Smith, E. R., & Zarate, M. A. (1990). Exemplar and prototype use in social categorization. *Social Cognition,* **8,** 243–262.

Smith, E. R., & Zarate, M. A. (1992). Exemplar-based model of social judgement. *Psychological Review,* **99,** 3–21.

Srull, T., Lichtenstein, M., & Rothbart, M. (1985). Associative storage and retrieval processes in person memory. *Journal of Experimental Psychology: Learning, Memory and Cognition,* **11,** 316–345.

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.

Weber, R., & Crocker, J. (1983). Cognitive processes in the revision of stereotypic beliefs. *Journal of Personality and Social Psychology,* **45,** 961–977.