Journal of Personality and Social Psychology
1983, Vol. 44, No. 1, 20-33


0022-3514/83/4401-0020$00.75

# A Hypothesis-Confirming Bias in Labeling Effects

John M. Darley and Paget H. Gross
Princeton University

The present study examines the process leading to the confirmation of a perceiver's expectancies about another when the social label that created the expectancy provides poor or tentative evidence about another's true dispositions or capabilities. One group of subjects was led to believe that a child came from a high socioeconomic background; the other group, that the child came from a low socioeconomic background. Nothing in the socioeconomic data conveyed information directly relevant to the child's ability level, and when asked, both groups of subjects reluctantly rated the child's ability level to be approximately at her grade level. Two other groups received the social-class information and then witnessed a videotape of the child taking an academic test. Although the videotaped performance series was identical for all subjects, those who had information that the child came from a high socioeconomic background rated her abilities well above grade level, whereas those for whom the child was identified as coming from a lower class background rated her abilites as below grade level. Both groups cited evidence from the ability test to support their conflicting conclusions. We interpret these findings as suggesting that some "stereotype" information (e.g., socioeconomic class information) creates not certainties but hypotheses about the stereotyped individual. However, these hypotheses are often tested in a biased fashion that leads to their false confirmation.

The expectancy-confirmation process is an important link in the chain leading from social perception to social action (Darley & Fazio, 1980; Rosenthal & Jacobson, 1968; Snyder & Swann, 1978a). As research has demonstrated, two processes leading to the confirmation of a perceiver's beliefs about another can be identified. The first, called a "behavioral confirmation effect" (Snyder & Swann, 1978b), is consistent with Merton's (1948) description of the "self-fulfilling prophecy." In this process, perceiver's behaviors toward the individual for whom they hold an expectancy channel the course of the interaction such that expectancy-confirming behaviors are elicited from the other individual (Rosenthal, 1974; Snyder, Tanke, & Berscheid, 1977). The second process leads to what we may call a "cognitive confirmation effect." We use this term to refer to expectancy-confirmation effects that occur in the absence of any interaction between the perceiver and the target person. In these cases, perceivers simply selectively interpret, attribute, or recall aspects of the target person's actions in ways that are consistent with their expectations (Duncan, 1976; Kelley, 1950; Langer & Abelson, 1974). Thus, perceivers with different expectancies about another may witness an identical action sequence and still emerge with their divergent expectancies "confirmed."

The focus of the present article is on the mediation of cognitive confirmation effects. We suggest that there are at least two different processes that bring about the cognitive confirmation of expectancies. The key to separating these processes lies in recognizing that people distinguish between the kinds of information that create conceptions of other people. Perceivers may define a continuum, one end of which involves information that is seen as a valid and sufficient basis for judgments about another; at the other end is evidence that is seen as a weak or invalid basis for those judgments.

The authors are grateful for the insightful comments of Nancy Cantor, Ron Comer, Joel Cooper, E. E. Jones, Charles Lord, Mark Zanna, and the members of the Princeton Social Psychology Research Seminar. Robin Akert, Kristin Boggiano, Paul Bree, Kay Ferdinandsen, Hannah McChesney, and Frederick Rhodewalt ably assisted in creating the stimulus materials.

Requests for reprints should be sent to John M. Darley, Department of Psychology, Princeton University, Princeton, New Jersey 08544.

As an example of valid information, consider a teacher who receives the results of a standardized test indicating that a particular pupil has high ability. The expectancies this information creates about the child are assumed to reflect the child's actual capabilities and are probably quite automatically applied. At the other end of the continuum, and of primary interest to this article, is expectancy-creating information that most perceivers would regard as incomplete with respect to an individual's abilities or dispositions. Many of our social stereotypes fall into this category. For example, racial or social-class categories are regarded by most of us as an insufficient evidential basis for conclusive judgments of another's dispositions or capabilities. In this case, we suspect that perceivers are highly resistant to automatically applying their expectancies to a target person. A teacher, for example, would be extremely hesitant to conclude that a black child had low ability unless that child supplied direct behavioral evidence validating the application of the label.

The end of the continuum defining information that is seen as insufficient evidence for social judgments is of interest because we find what appears to be a paradox in the literature dealing with social stereotypes. Some recent investigations of the influence of stereotypes on social judgments have demonstrated a "fading" of stereotypic attributions (e.g., Karlins, Coffman, & Walters, 1969; Locksley, Borgida, Brekke, & Hepburn, 1980). For example, investigators have noted participants' increasing unwillingness to make stereotypic trait ascriptions (Brigham, 1971). Moreover, Quattrone and Jones (1980) demonstrate that although people may make stereotype-based judgments about a social group, they are unwilling to use category-based information to predict the behavior of any one member of that group.

Given this resistance to the utilization of expectancies when the social labels establishing them are not seen as valid guides for judgments, one might expect an elimination of the expectancy-confirmation bias. That is, perceivers would not unjustly assume the truth of a stereotype; they would instead require that evidence substantiating the accuracy of that stereotype be provided. This leads to the prediction that, ultimately, judgments about the target person will reflect the actual evidence produced by his or her behavior, unbiased by the perceivers' initial expectancies. Unfortunately, this conclusion stands in contradiction to the bulk of the self-fulfilling prophecy literature in which one finds that confirmation effects are often produced when racial, ethnic, or other negative social labels are implicated—exactly those cases in which one expects perceivers to refrain from using category-based information (e.g., Foster, Schmidt, & Sabatino, 1976; Rist, 1970; Rosenhan, 1973; Word, Zanna, & Cooper, 1974). We suggest that this apparent contradiction can be resolved if the following two-stage expectancy-confirmation process is assumed: Initially, when perceivers have reason to suspect that the information that establishes an expectancy is not diagnostically valid for determining certain of the target person's dispositions or capabilities, they will refrain from using that information to come to diagnostic conclusions. The expectancies function not as truths about the target person but rather as hypotheses about the likely dispositions of that person. If perceivers were asked for judgments at this point in the process, without any behavioral evidence to confirm their predictions, they would not report evaluations based on their expectancies. They would instead report that either they did not have sufficient information or they would make judgments consistent with normative expectations about the general population.

The second stage occurs when perceivers are given the opportunity to observe the actions of the labeled other. They then can test their hypotheses against relevant behavioral evidence. The initiation of a hypothesis-testing process would seem to be an unbiased approach for deriving a valid basis for judgments about another. If, however, individuals test their hypothesis using a "confirming strategy"—as has often been demonstrated—a tendency to find evidence supporting the hypothesis being tested would be expected (Snyder & Cantor, 1979; Snyder & Swann, 1978b). A number of mechanisms operating in the service of a hypothesis-confirming strategy may contribute to this result. First, the search for evidence may involve selective attention to information that is consistent

with expectations and a consequent tendency to recall expectancy-consistent information when making final evaluations (Zadny & Gerard, 1974). Second, a hypothesis-confirming strategy may affect how information attended to during a performance will be weighted. Typically, expectancy-consistent information has inferential impact, whereas inconsistent information has insufficient influence in social-decision tasks (Nisbett & Ross, 1980). In fact, a recent study by Lord, Ross, and Lepper (1979) indicates that even when expectancy-inconsistent information is brought to the attention of the perceiver, it may be regarded as flawed evidence and therefore given minimal weight in the evaluation process. Third, it is also possible for inconsistent actions to be attributed to situational factors and thereby be attributionally discounted (Regan, Strauss, & Fazio, 1974). Finally, apparently inconsistent behavior may be reinterpreted as a manifestation of dispositions that are consistent with the initial expectancy (Hayden & Mischel, 1976).

Given the operation of all of these biasing mechanisms, an expectancy-confirmation effect could arise even when the target person's behavior does not objectively confirm the perceiver's expectancies. Nonetheless, the opportunity to observe the diagnostically relevant information is critical to the process because it provides what perceivers consider to be valid evidence, and thus, they can feel that they have made an "unbiased" judgment.

In the present study, we attempted to find evidence of this two-stage expectancy-confirmation process. To do this, perceivers were given information that would induce them to categorize an elementary school child as belonging to a high- or low-socioeconomic-status (SES) class (cf. Cooper, Baron, & Lowe, 1975). Consistent with the two-stage model, we predicted that perceivers given only this demographic information about the child would be reluctant to provide label-consistent ability evaluations. Another group of evaluators were given the identical demographic information about the child (high or low SES) and were then shown a performance sequence that provided ability-relevant information about the child. Owing to the hypothesis-confirming bias, it was predicted that these individuals would find evidence in the identical performance sequence to support their opposing hypotheses and would thus report widely different judgments of the child's ability. Moreover, we expected these perceivers to mislocate the source of their evidence from their own expectancies to the "objective" evidence provided by the performance sequence.

## Method

As part of a study on "teacher evaluation methods," students viewed a videotape of a fourth-grade female child and were asked to evaluate her academic capabilities. Variation in the videotape determined the four experimental conditions. The first segment provided demographic information about the child and was used to establish either positive or negative expectations for the child's academic potential. Half of the participants viewed a sequence that depicted the child in an urban, low-income area (negative expectancy); the other half were shown the same child in a middle-class, suburban setting (positive expectancy).

Orthogonal to this manipulation was the performance variable. Half of the participants from each expectancy condition were shown a second tape segment in which the child responded to achievement-test problems (performance). The tape was constructed to be inconsistent and relatively uninformative about the child's abilities. The remaining participants were not shown this segment (no performance).

The design was thus a $2 \times 2$ factorial one, with two levels of expectancy (positive and negative) and two levels of performance (performance and no performance). In addition, a fifth group of participants viewed the performance tape but were not given prior information about the child's background (performance only). Their evaluations were used to determine if the performance tape was, as intended, an ambiguous display of the child's academic capabilities.

All viewers then completed an evaluation form on which they rated the child's overall achievement and academic skill level. Additional questions about the child's performance and manipulation checks were included. After completing their evaluation form, participants were given a questionnaire designed to probe their suspiciousness about the experiment. Finally, participants were debriefed, thanked, and paid.

### *Subjects*

Seventy (30 male and 40 female) Princeton University undergraduates volunteered for a study on "teacher evaluation and referral" for which they were paid $2.50 for a 1-hour session. Participants were randomly assigned to one of five (four experimental and one control) conditions, with an attempt made to have an equal number of men and women in each condition. None of the students in the study reported having any formal teacher training; two students had informal teaching experience, both at the high school level. Only three of the original

subjects were eliminated from the study because of suspiciousness about the experimental procedures.

## *Instructions*

The experimenter introduced herself as a research assistant for a federal agency interested in testing new educational procedures. Students were told that their participation would be useful for determining the reliability of a new evaluation form teachers would use when referring pupils to special programs (these included remedial classes and programs for gifted students). To test the completeness and scorability of the evaluation form, subjects, acting as teachers, were asked to provide an academic evaluation of a selected child on this specially designed form. The experimenter emphasized that all evaluations would be anonymous and confidential and asked participants not to place their names anywhere on the form. She also requested that they replace the form in its envelope and seal it when they were finished. Each participant was further admonished to be as accurate and objective as possible when evaluating their selected pupil.

The reseach assistant then went on to explain that a videotape file of elementary school children had been prepared for a previous study (numerous videotape reels were on shelves in front of the subject). Participants would be selecting one child from this sample to observe and evaluate. It was made clear that this "randomly selected sample of children includes some who perform well above their grade level, some who would benefit from remedial programs, and some at all levels between these extremes." To select a child from this file, participants drew a number corresponding to a videotape reel. The experimenter, who had been blind to condition until this point, placed the tape on a television monitor and gave the participant a fact sheet appropriate to the child they selected.

The participant actually selected one of five prepared tapes (corresponding to the four experimental and one control conditions). In all conditions, the child observed was a nine-year-old female Caucasian named Hannah, who was a fourth grader attending a public elementary school. The information about the child's name, grade, and so forth appeared on the fact sheet and was reiterated in the narration of the tape.

## *Demographic Expectancy Manipulation*

To establish either positive or negative expectancies about Hannah's ability, participants viewed a tape of Hannah that contained environmental cues indicating either a high or low socioeconomic background. Each tape included 4 minutes devoted to scenes of Hannah playing in a playground (filmed at a distance to prevent clear perception of her physical attractiveness) and 2 minutes devoted to scenes of her neighborhood and school. The tapes were filmed at two different locations.

In the negative-expectancy condition, subjects viewed Hannah playing in a stark fenced-in school yard. The scences from her neighborhood showed an urban setting of run-down two-family homes. The school she attended was depicted as a three-story brick structure set close to the street, with an adjacent asphalt school yard. The fact sheet given to participants included the following information about Hannah's parents: Both parents had only a high school education; her father was employed as a meat packer; her mother was a seamstress who worked at home.

In the positive-expectancy condition, Hannah was seen playing in a tree-lined park. The scenes from her neighborhood showed a suburban setting of five- and six-bedroom homes set on landscaped grounds. Her school was depicted as a sprawling modern structure, with adjacent playing fields and a shaded playground. Further, Hannah's fact sheet indicated that both her parents were college graduates. Her father's occupation was listed as an attorney, her mother's as a free-lance writer.

## *The Performance Manipulation*

Two groups were asked to evaluate Hannah's academic ability immediately after viewing one or the other expectancy tape (no performance); two other groups were given the opportunity to observe Hannah in a test situation (performance).

Subjects in the performance conditions observed a second 12-minute tape sequence in which Hannah responded to 25 achievement-test problems. This portion of the tape was identical for both performance groups. The problems were modified versions of items selected from an achievement-test battery and included problems from the mathematics computation, mathematics concepts, reading, science, and social studies subtests. The grade level for the problems ranged from the second to the sixth grade. Participants were told that the test included "easy, moderate, and difficult problems." The problems were given orally to Hannah by a male tester who held up the possible solutions on cards. The sequence was filmed from behind the child so the viewer was able to see the cards held by the tester but not Hannah's face.

Hannah's performance was prearranged to present an inconsistent picture of her abilities. She answered both easy and difficult questions correctly as well as incorrectly. She appeared to be fairly verbal, motivated, and attentive on some portions of the tape and unresponsive and distracted on other portions of the tape. The tester provided little feedback about Hannah's performance. After each problem, he recorded Hannah's response and went on to the next problem.

To determine what information the tape provided about Hannah's ability in the absence of a priori expectancies, a group of participants, given the same cover story as subjects in the other conditions, were shown only the performance tape. These subjects were given no information about the child other than her name, age, grade, address, and the school she attended.

## *Dependent Measures*

After reviewing the tape, participants were given an evaluation form to complete. The form contained the following sections:

*Ability measures.* Nine curriculum areas forming three broad categories were listed. Included in this section were reading (reading comprehension, reading ability, writing, language ability), mathematics (mathemat-

ical concepts, mathematical computation), and liberal arts (science, general knowledge, social studies). Each curriculum area was followed by a scale extending from kindergarten to the sixth-grade, ninth-month grade level, with points labeled at 3-month intervals. Subjects were instructed to indicate the grade level that represented the child's ability in each of these areas. For subsequent analyses, mean ratings of items within these three categories were used, and grade levels were converted to a scale with months represented as fractions of a year (i.e., third grade, sixth month would equal 3.5).

*Performance measures.* Participants in the performance conditions were asked to estimate the number of easy, moderate, and difficult problems the child answered correctly and to report the overall grade level of the test administered to the child. In an open-ended question, participants were asked to report the "information they found most useful in determining the child's capabilities."

*Supplementary academic measures.* Twenty traits or skills, followed by exemplars of classroom behaviors characterizing both the positive and negative ends of each of these traits, were listed. Subjects were asked to check the point on a 9-point scale that would best characterize the child on the dimension. Next to each scale, a box labeled "insufficient information" was also provided. Subjects were instructed to check this box rather than a scale value if they felt they had not been given sufficient information to rate the child on a given dimension.

These 20 items were selected to form five clusters: work habits (organization, task orientation, dependability, attention, thoroughness), motivation (involvement, motivation, achievement orientation), sociability (popularity, verbal behavior, cooperation), emotional maturity (confidence, maturity, mood, disposition), and cognitive skill (articulation, creativity, learning capability, logical reasoning). Mean ratings of items within these five categories were used in subsequent analysis.

*Manipulation checks.* In the last part of the booklet, subjects were asked to rate the child's "attractiveness" and the "usefulness of socioeconomic information as an indicator of a child's academic ability." The final open-ended question asked subjects to report the child's socioeconomic level.

*Suspiciousness probe.* Finally, participants filled out a questionnaire assessing, for the agency, "how they had been treated during the experimental session." This was designed to probe their suspiciousness about the experimental procedures and purpose of the study. Following this, participants were thoroughly debriefed and paid.

## Results

### *Ability-Level Ratings*

Our primary hypothesis was that expectancy-confirmation effects occur only when perceivers feel they have definitive evidence relevant to their expectations. Specifically, we predicted that subjects who viewed only the positive- or negative-expectancy tape segment (no performance) would show little, if any, signs of expectancy confirmation in their ratings of the child's ability level, whereas subjects who viewed both the expectancy segment and the test segment (performance) would show considerable signs of expectancy confirmation. As a test of this hypothesis, a 2 (positive vs. negative) × 2 (performance vs. no performance) analysis of variance (ANOVA) was performed on ability-level ratings.

As shown in Figure 1, the results support our predictions. The ANOVA interaction term was significant for each index: liberal arts, $F(1, 56) = 6.67$, $p < .02$; reading $F(1, 56) = 5.73$, $p < .03$; and mathematics $F(1, 56) = 9.87$, $p < .01$. Although a main effect for expectancy emerged for each of the three indexes—liberal arts, $F(1, 56) = 19.24$, $p < .01$;

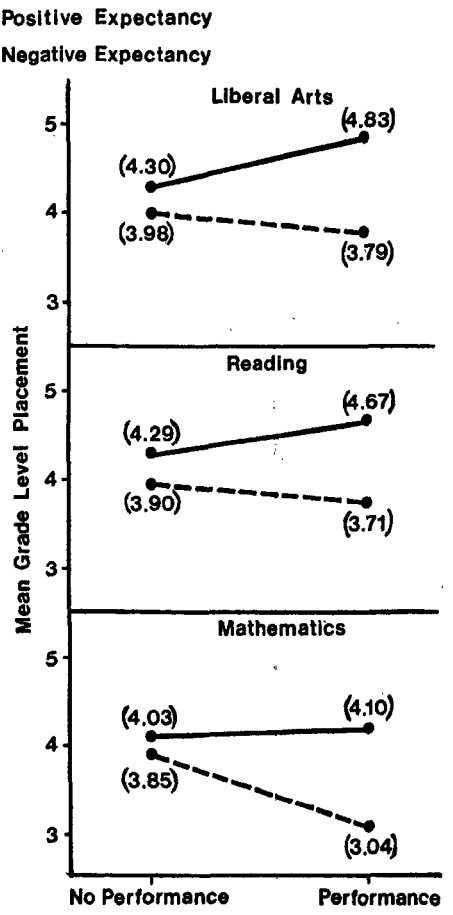


*Figure 1.* Mean grade-level placements on the liberal arts, reading, and mathematics indexes for experimental conditions.

reading, $F(1, 56) = 32.98$, $p < .001$; and mathematics, $F(1, 56) = 19.78$, $p < .001$—Newman-Keuls tests revealed that the subjects in the no-performance conditions did not rate the child's ability level as differing much in either direction from her known school grade. On only one of the indexes (liberal arts) did the no-performance–positive-expectancy subjects rate the child significantly higher than the negative-expectancy subjects ($p < .05$). In the two performance conditions, however, positive-expectancy subjects made reliably higher ratings on all three indexes ($p < .05$ in all cases). The fan-shaped interaction of Figure 1 is consistent with the hypothesized two-stage confirmation process in which subjects first reserve judgment—if that judgment is based on only demographic indicators—but then allow their judgment about an ability to be biased in the direction of hypothesis confirmation.[1]

### *Manipulation Checks*

The manipulation checks indicate that the above results were not artifactually produced. First, the expectancy manipulation was as successful for subjects who viewed the child's test performance as for those who did not. Without exception, positive-expectancy subjects reported the child's socioeconomic status as upper middle or upper class, and negative-expectancy subjects reported the child's socioeconomic status as lower middle or lower class. Second, analyses of ratings of the child's attractiveness and the usefulness of socioeconomic information for predicting ability yielded no differences across groups. The latter result is especially important in indicating that those who had seen the child's test performance did not regard the demographic information as any more diagnostic than those who had not seen it. Thus, the greater impact of induced expectancies in the performance conditions was not attributable to greater confidence in an implicit theory of the social-class–ability relation. Moreover, mean ratings of the usefulness of socioeconomic information (for all groups) were just below the midpoint toward the "not useful" end of the scale.

Finally, as can be seen in Table 1, ability ratings of the performance-only group indicate that the performance segment was, as intended, an ambiguous display of Hannah's capabilities.

Table 1
*Mean Grade-Level Placements on Curriculum Areas by Performance Control Group*

| Index | *M* | Grade level | *SD* |
|---|---|---|---|
| Liberal arts | 4.0 | 3rd grade, 9th month | .505 |
| Reading | 3.8 | 4th grade | .581 |
| Mathematics | 3.5 | 3rd grade, 6th month | .238 |

*Note.* $n = 10$.

We hoped that evaluations of the child's ability would tend to be variable, reflecting the inconsistencies in the child's performance; however, mean estimates would be expected to fall close to the child's given grade level. As the data in Table 1 indicate, ratings on the three curriculum indexes do show considerable variability, and the perceivers do use the child's grade level as an anchor for their judgments. The mathematics ratings were somewhat lower and less variable than the others, indicating that her performance in this area may have been poorer and more consistent.[2] (This will be discussed at a later point.)

### *Judgments of the Performance*

If performance subjects were no less aware of, or impressed by, the relevance of the expectancy information, it follows that they found support for their divergent hypotheses in the child's performance. Measures from the academic evaluation form indicate several ways in which perceivers obtained support for their diverse hypotheses. (All of these

[1] We also analyzed this data by pooling across the three ability measures. As one would expect, because this increases the number of observations, the significance levels are improved, although the basic interactional pattern ($p = .002$) remains the same. Again, post hoc comparisons reveal that the two performance conditions are reliably different ($p < .01$), whereas the two no-performance groups do not differ reliably.

[2] An $F_{max}$ test of the difference between several variances indicates no difference between the variances of the liberal arts and reading indexes. The variance of the mathematics index is, however, significantly different from that of the liberal arts index, $F_{max}(9, 9) = 5.92$, $p < .05$, and shows a marginally significant difference from that of the reading index.

were measures of the subjects' perceptions of the performance and therefore were taken only from the groups that witnessed the test sequence. Recall that all of these subjects witnessed the identical performance tape.)

*Test difficulty.* Performance on a test is a joint function of the test taker's ability (and other personal factors) and the difficulty of the test. Therefore, one way of justifying a high-ability inference from an inconsistent test performance is to perceive the test as being very difficult. Conversely, one way of rationalizing a low-ability inference from the same performance would be to perceive the test as easy. This happened: Subjects in the positive-expectancy condition rated the test as significantly more difficult ($M$ grade = 4.8) than did those in the negative-expectancy condition ($M$ grade = 3.9), $t(28) = 2.69$, $p < .02$.

*Problems correct.* Subjects also estimated the number of problems the child answered correctly within each of the problem categories: very difficult, moderately difficult, and easy.

A repeated measures analysis of variance revealed a marginally reliable tendency for subjects with positive expectancies to estimate that the child correctly answered a higher percentage of problems, $F(1, 28) = 3.94$, $p < .06$. Follow-up analyses revealed that subjects with positive expectancies estimated that the child correctly answered more of the easy ($M$ = 94% vs. 79%) and moderately difficult ($M$ = 69% vs. 53%) problems than did subjects with negative expectancies, $t(28) = 2.55$ and 2.21, respectively, $p$s $< .05$. Expectancy did not affect estimates of answers to difficult problems ($M$ = 37% vs. 36%). The overall pattern suggests a bias to report more instances of expectancy-consistent than expectancy-inconsistent test responses.

*Reporting relevant behaviors.* Subjects had been asked to report, in an open-ended format, the performance information "most relevant for determining the child's capabilities." We expected that subjects anticipating a good performance would report more instances of positive behaviors than those expecting a poor performance. For each subject, we computed a positivity index by subtracting the number of negative instances from the number of positive instances. Consistent with predictions, positive-expectancy subjects reported a significantly greater number of positive behaviors relative to negative ones as being relevant in their judgments than did negative-expectancy subjects, $t(28) = 34.65$, $p < .001$.

To summarize the performance judgments, positive- and negative-expectancy subjects, although agreeing that the performance provided information that was sufficient to estimate the child's capabilities, disagreed on how difficult the test was, how many problems the child answered correctly, and how many of her test behaviors reflected either positively or negatively on her achievement level. On every measure, positive-expectancy subjects made interpretations more favorable to the child than did negative-expectancy subjects.

### *Supplementary Academic Measures*

*Information sufficiency.* Recall that subjects reporting on these measures were allowed to check a scale value or a box labeled insufficient information. We believed that the no-performance subjects who had only demographic-based expectancies to rely on would display a greater reluctance to evaluate the child and that this reluctance would lead to more frequent use of the insufficient information answer. A $2 \times 2$ analysis of variance on these data yielded only a main effect for performance, $F(1, 56) = 12.86$, $p < .001$, such that no-performance subjects chose this option more often ($M$ = 43% of the items) than subjects who viewed the test sequence ($M$ = 22% of the items). A one-way analysis of variance, comparing the no-performance, performance, and performance-only conditions yielded a reliable effect, $F(2, 67) = 12.41$, $p < .001$. Moreover, comparing these means (via Duncan's test), we find that the performance-witnessing groups were not significantly different from each other, whereas both were significantly different from the no-performance groups ($p < .01$). Thus, the difference found between performance and no-performance groups on the use of the insufficient information option did not seem to depend on the fact that the mere quantity of evidence provided to performance subjects

Table 2
*Mean Ratings on Trait Measures for Experimental Groups*

| Condition | Dependent measure | | | | |
|---|---|---|---|---|---|
| | Work habits | Motivation | Sociability | Maturity | Cognitive skills |
| Positive, performance | $5.21_{a}$ | $5.16_{a}$ | $5.25_{a}$ | $5.33_{a,b}$ | $4.73_{a}$ |
| Positive, no-performance | $4.92_{a,b}$ | $5.31_{a}$ | $4.82_{a,b}$ | $5.65_{b}$ | $5.55_{b}$ |
| Negative, no-performance | $5.13_{a}$ | $4.80_{a,b}$ | $4.38_{b}$ | $4.67_{a}$ | $4.83_{a}$ |
| Negative, performance | $4.36_{b}$ | $4.11_{b}$ | $4.58_{a,b}$ | $4.77_{a,b}$ | $4.12_{a}$ |

*Note.* $n = 15$ per condition. The higher the number, the more positive the evaluation. Letter subscripts indicate vertical comparisons of cell means by Duncan's multiple-range test. Means that do not share a common subscript are significantly different from each other at the .05 level.

was greater (two tape segments) than that given to no-performance subjects (one tape segment). The performance-only subjects, who also saw only one tape segment, did not differ from performance subjects on this measure. The difference is better attributed to the greater perceived diagnostic utility of the performance segment. Performance subjects apparently felt that the child's test performance provided sufficient diagnostic information on which to base their evaluations.

*Trait measures.* A 2 (expectancy) × 2 (performance) analysis of variance was performed on ratings for each of the five trait dimensions. Because participants were given the option of not checking a scale value on these measures, missing values were given a score of 5, which, on a 9-point scale, represents the neutral point.[3] These data are presented in Table 2.

Consistent with our findings for the curriculum indexes, a significant interaction emerged for the work habits index such that individuals expecting the child to perform well rated her more positively after viewing the performance tape whereas those expecting her to perform poorly rated her more negatively after viewing the performance tape, $F(1, 56) = 5.15$, $p < .03$. The predicted interaction effect was not obtained for the motivation, sociability, emotional maturity, or cognitive skill measures. For each of these measures, we found a main effect for expectancy, with the positive-expectancy groups rating the child significantly higher than the negative-expectancy groups, $F(1, 56) = 6.99$, 4.57, 5.76, and 5.84, respectively, $ps < .05$. In addition, there was a significant effect for performance on the cognitive skill index, with the performance groups showing lower ratings than the no-performance groups, $F(1, 56) = 7.73$, $p < .05$. These data indicate that certain expectancy-consistent judgments may not require a two-stage process. Although it may be necessary to provide performance information to obtain judgments of a child's ability level, judgments about other dispositional characteristics may be made without this information.

## Discussion

Unlike many previous studies demonstrating expectancy-confirmation effects, the expectancies in the present study were not created by information that most people would regard as definitively establishing their validity. They were not created by objective test results, expert judgments, or other authoritative information. Instead, the expectancies were conveyed by such cues as the child's clothes, the bleakness of the playground on which she played, or the high- or low-status character of her parents' occupations.

We suggested that perceivers would realize

[3] Analyses of these data require a decision about how to treat the responses of subjects who checked the "not enough information to rate" alternative. The means presented in Table 2 are calculated by assigning a score of 5 to missing scale values. This assumes that the nonresponding subjects would have checked the scale midpoint if forced to respond. Another way of dealing with the same issue is to insert the cell-mean score for each such subject. Using this procedure, the pattern of results is essentially unchanged. The same effects emerged as significant.

that expectancies created by this information do not form a completely valid basis for some of the evaluations they were asked to make. The results indicate that this is so: Perceivers who were given only demographic information about the child demonstrated a resistance to making expectancy-consistent attributions on the ability indexes. Their estimations of the child's ability level tended to cluster closely around the one concrete fact they had at their disposal: the child's grade in school. When given the opportunity to avoid making dispositional attributions altogether, nearly half of the time these perceivers chose that option.

In contrast, a marked expectancy-confirmation effect was evident for those perceivers who evaluated the child after witnessing an ability-relevant performance. Those who believed the child came from a high socioeconomic class reported that her performance indicated a high ability level, whereas those who believed the child came from a low socioeconomic class reported that the identical performance indicated a substantially lower level of ability.

This pattern of results suggests that when the diagnostic validity of a perceiver's expectations is suspect, expectancies function as hypotheses, and the task of evaluating an individual for whom one has an expectancy is a hypothesis-testing process. Expectancy confirmation, then, does not always result from an automatic inference process. Instead, it occurs as the end product of an active process in which perceivers examine the labeled individual's behavior for evidence relevant to their hypothesis.[4]

As is apparent from our data, the hypothesis-testing strategy that perceivers use has a bias (as Snyder & Cantor, 1979, have suggested) toward confirmation of the hypothesis being tested. The literature suggests a number of related mechanisms that can contribute to this effect (see Nisbett & Ross, 1980, for a review). We do have evidence to suggest what some of these mechanisms may have been in our study. First, there seems to be a selective recall of evidence: Perceivers who expected the child to do well reported the child as having answered more easy and more moderately difficult problems correctly than those expecting the child to do poorly. Second, there seems to be a selective weighting of the evidence such that hypothesis-consistent behaviors are regarded as more "typical" of the child's true capabilities. When people were asked to report what evidence they found most useful in determining their evaluations, they reported only those test items on which the child's performance was consistent with their initial expectations. Third, perceivers appeared to develop auxilliary hypotheses that would render apparently inconsistent behavior consistent with their hypotheses. These auxilliary hypotheses did not seem to be revised assessments of the actor but rather assessments of situational factors that could account for discrepancies in the actor's behavior. For instance, we found that persons who expected a good performance decided that the test given to the child was very difficult, a conclusion that would account for instances of otherwise inconsistent poor performance; whereas persons who expected a poor performance reported that the test was easy, which would account for inconsistent good performance. Finally, we found evidence in the open-ended reports of some participants to suggest that the meaning given to the child's behaviors was often consistent with the perceivers' initial hypotheses. For example, a low-SES Hannah was reported to have "difficulty accepting new information," whereas a high-SES Hannah was reported to have the "ability to apply what she knows to unfamiliar problems."

Implicit in this data is the conclusion that perceivers seem to be aware that witnessing a particular test performance does not give

[4] In the experimental paradigm in which expectancy effects are typically demonstrated, perceivers are always provided with the opportunity to observe or interact with the labeled target person. By using this research design, one cannot conclusively determine whether the resulting expectancy effect was due to differential perceptions of the target person, as most researchers suggest, or if subjects had simply based their evaluations on the information provided by the label and had ignored the performance. By including conditions in the present study in which some perceivers are not provided with performance information, it becomes possible to distinguish between expectancy effects arising from a nonobservationally based inference process and those arising from expectancy-guided search processes.

them automatic access to an individual's underlying ability. Many other factors, such as luck, task difficulty, or lack of motivation, may intervene (Darley & Goethals, 1980; Weiner et al., 1971). Therefore, the meaning of a person's performance is susceptible to multiple interpretations that can be consistent with, and even supportive of, opposing hypotheses about that person's ability.

Thus far, we have treated information as creating expectancies that are either valid and automatically applied to others or weak and only hypothesis generating. It is more likely that any item of information about a person generates some certainties and some hypotheses, depending on the domain to which it is applied. In the present study, the demographic information seems to have this character of creating both certainties and hypotheses. On the supplemental measures related to school achievement—specifically, on measures of motivation, sociability, and emotional maturity—a simple main effect was obtained such that people who saw the child as coming from a high socioeconomic background judged her more positively, and those who did not see the performance had as extreme ratings as those who did. (But keep in mind that individuals had the opportunity not to rate the child on these measures and that, overall, many more people from the no-performance conditions chose not to rate.) Apparently, some individuals felt that demographic data alone was sufficient evidence on which to base an evaluation of, for example, a child's likely achievement orientation. Thus, the addition of performance information was not necessary for a conclusive judgment in this area. In general, our social categories do trigger expectancies for a constellation of dispositions and behaviors, and for some of these, it may not be necessary to rely on performance evidence to feel certain that one's expectations are accurate.

### *The Validity of Demographic Evidence*

From another perspective, one could ask whether demographic information does not warrant correspondent inferences of ability. Certainly, numerous studies show correlations between social class and school performance (Dreger & Miller, 1960; Kennedy, VanDeReit, & White, 1963; Lesser, Fifer, & Clark, 1965). From this perspective, the differential judgments of people who witnessed the same test with different demographically produced expectancies was less evidence of bias than it was of an understanding of the true workings of the world. Two things can be said about this: First, part of the general argument of those concerned with self-fulfilling prophecies is that the present process is exactly how the link between social class and academic performance comes about. Second, the data from our no-performance perceivers indicate that people regard the question of what exists in the world as a separate question from that addressed in the present study. Base-rate information (i.e., estimates of the frequencies with which an attribute or capability level occurs in a social group) represents probabilistic statements about a class of individuals, which may not be applicable to every member of the class. Thus, regardless of what an individual perceives the actual base rates to be, rating any one member of the class requires a higher standard of evidence. When one child's ability is being considered, demographic information does not appear to meet the perceiver's criteria for a valid predictor; performance information, on the other hand, clearly does.

There is yet another way to pose the validity question, and that is to consider the use of demographic evidence when perceivers formulate a working hypothesis. From an information-processing perspective, hypothesis formulation serves a useful function: It allows one to make better use of subsequent evidence. The rub, of course, is that once a hypothesis is formulated, regardless of our judgments of the validity of the evidence on which it is based, our cognitive mechanisms are biased toward its eventual confirmation. Thus, when asking whether the final judgments of the perceiver accurately reflect what exists in the world, we should not obscure an important point: how those judgments come about. To clarify further, the "judgmental bias" in the present study does not refer to the indiscriminate use of category-based information, or to the (in)accuracy of final judgments, but

to the processes that determine what those judgments will be.

*An Alternative Explanation*

An alternative explanation for the general pattern of results reported here is possible. The individuals who witnessed only the demographic information may have actually made ability inferences but chose not to report them. Their failure to report their evaluations may have been due to fears that the experimenter would regard the inferences as unjustified. However, in the experiment we minimized the possible cause for this concern by demonstrating to the participants that their responses would be anonymous. The experimenter was not present while the participant filled out the dependent-measures form and did not return until he or she deposited the questionnaire in an anonymity-guaranteeing location. Furthermore, on the final questionnaire, participants were asked if they were sufficiently assured of the anonymity of their responses, and all of them replied affirmatively.

It is, of course, still possible to make a generalized version of the same point: The perceiver's resistance to using the demographic information could, at least in part, be motivated by the awareness that their behavior was under scrutiny by others. This does not necessarily diminish the interest in the phenomenon. In the real world, people who make judgments frequently know their judgments may be public. Teachers classifying students, clinicians diagnosing clients, and employers selecting new personnel are all aware that their actions may be scrutinized by others. Thus, whether this awareness is based on personal knowledge, social pressures, or internalized social desirability concerns, both the processes that bring about those judgments and the consequences in terms of judgmental bias are likely to be the same.

The present study finds results that at first glance seem contradictory to results of some other studies. One thinks particularly of the work of Locksley et al. (1980) and that of Kahneman and Tversky (1973). In the Locksley et al. study, the direction of the interaction appears to be the reverse of that obtained here. A strong stereotype effect in trait ratings is found with category-membership information (gender labels) or when nondiagnostic information accompanies the category label. This stereotype effect disappears with diagnostic information. However, consider the differences in the type of information given to perceivers in the present study and that given to perceivers in the correspondent conditions of the Locksley et al. study. The perceivers in the present study who were reluctant to apply stereotypes (the no-performance conditions) received nondiagnostic case information, as do some in the Locksley et al. study. However, perceivers in the present study observe the child they will rate and are given a fair amount of family data—information that would certainly distinguish the child from others in her social group. The diagnostic information used by Locksley et al. consists of information that could be applied to almost any person and may not have created an individuated impression of the person to be rated. The two conditions, then, are not identical, and comparing them leads us to the following possibility: Stereotype effects persist with information that does not distinguish the target from the target's social category, whereas dilution effects (nonstereotypic judgments) appear when case information successfully creates an individuated impression of the target. Recent studies by Quattrone and Jones (1980) and Locksley, Hepburn, and Oritz (in press) support this conclusion.

In comparing other conditions of the Locksley et al. (1980) and the present study, differences in the type of information given to perceivers produces discrepant results. The diagnostic information given to perceivers in the Locksley et al. study consists of a single behavioral exemplar that confirms or disconfirms a gender-based trait expectancy. A dilution effect is found only with a disconfirming behavior sample. In contrast, the diagnostic test sequence in the present study contains both confirming and disconfirming behavioral evidence. Furthermore, we know from supplementary measures that perceivers found the expectancy-consistent portions more diagnostically informative than the in-

consistent portions. Therefore, with a source of multiple information—with many information elements that serve to confirm expectancies—a confirmation effect is not surprising. Had the performance tape in the present study provided only compelling disconfirming evidence, we suspect a dilution effect might have been found here as well.

Discrepancies between this work and that of Kahneman and Tversky (1973) can be addressed as well. In Kahneman and Tversky's studies, individuals are asked to predict a target person's occupational-category membership from a brief personality description. Predictions are overwhelmingly based on the degree to which the personality information "fits" with an occupational stereotype (i.e., a representativeness effect). This appears inconsistent with the stereotype-resisting judgments of the perceivers in our study who received no performance information. However, the demographic information given to our no-performance perceivers, although it does allow for a judgment of fit to a social category, does not provide information for a judgment on an ability dimension. The condition, then, is similar to Kahneman and Tversky's Experiment 3 in which the personality description is uninformative with regard to the target person's profession (i.e., it contains no occupation-relevant personality traits). In their study, occupational-category predictions were essentially random. That is, they were based neither on prior probabilities nor on similarity. This is essentially the same effect we find for no-performance perceivers in the present study. Apparently, the representativeness effect (or an expectancy effect) depends on the provision of information that allows for a similarity match to the categories perceivers are asked to judge.

Further, the no-performance conditions in the present study are not identical to Kahneman and Tversky's (1973) null-description condition. In that condition, subjects are given no information whatsoever about the target—neither individuating information nor category-relevant information. Here a strong base-rate effect emerges. Although this might cause one to predict a strong stereotype effect in the present no-performance conditions, our earlier point about individuating information may explain why it is not obtained. No-performance perceivers may lack relevant case information, but they do have a significant amount of individuating target information; apparently, this significantly alters the framework for prediction.

We might summarize as follows: Representativeness and expectancy effects are found when relevant case data are provided so that individuals can determine the target person's fit to a category. Base-rate effects (and non-observationally based stereotypic judgments) are found when neither case data nor individuating information is given. Finally, assume that three conditions are met: individuating information is given, information about base rates is withheld, and a priori expectancies are not relied on because they may not be applicable to a particular target. Then, without relevant case data, a judgment of fit is precluded, attenuating a representativeness or a biased confirmation effect. In these circumstances, judgments are made at the scale midpoint or the chance level. We find this latter effect in both Kahneman and Tversky's (1973) uninformed condition and in the no-performance conditions of the present study.

A final point is relevant to both of the studies reviewed above. Predicting ability from social-class information may not be equivalent to predicting personality traits from gender labels or occupational membership. The nature of the prediction required (ability rather than personality characteristics) may cause individuals to regard social-class information as at the invalid end of the continuum we have defined. But an individual's gender or occupation, on the other hand, may be regarded as valid information on which to base an inference about personality. Related to this point is that the standards of evidence required for different stereotype-confirming judgments may be different. Automatic assumptions about personality may be made from occupation or gender label, and thus, stereotype effects are obtained with this information alone, or with minimal additional information. To make judgments of a low-SES child's ability, perceivers require more information and, specifically, criterion-relevant information. Thus, stereotypic judg-

ments are not found with only category or nondiagnostic information but are found only when a sufficient amount of apparently confirming diagnostic information is provided.

*Limits to the Confirmation Process*

The present study finds results consistent with those of many other studies. For instance, Swann and Snyder (1980) found that target individuals labeled as dull witted were still seen as dull witted even after the perceivers had witnessed a sequence in which these target individuals outperformed those labeled as bright, a situation in which a cognitive confirmation effect triumphed over apparently strongly disconfirming evidence. Nonetheless, we suspect that there are limits to the cognitive confirmation process.

We can suggest several variables, some of which we have mentioned, that may determine whether a confirmation or a disconfirmation effect is found. First, there is the clarity of the disconfirming evidence. In the domain of abilities, in spite of the above example, a sustained high-level performance is compelling evidence for high ability. I may perceive another as a slow runner, but if I see him or her do several successive 4-minute miles, my expectancy must change. When this occurs, it is possible that a contrast effect will take place in which the significance of the disconfirming behavior will be exaggerated and the initial expectancy reversed. Intuitively, no such unambiguous evidence exists in the personality realm, where even compelling positive behavior can be attributed to negative underlying motives or dispositions. Second, the strength with which the initial expectancy or hypothesis is held may produce conflicting effects. "Strength of expectancy" is an ambiguous phrase. It may refer to one's degree of commitment to an expectancy of a fixed level, or it may refer to the extremity of the expectancy. In the first instance, the stronger the commitment to the expectancy, the more resistant it would be to disconfirmation. However, the more extreme the expectancy, the more evidence there is that potentially disconfirms it. Finally, the perceiver's motivation may play a role. Under certain circumstances, an individual may prefer to see his or her expectancy confirmed; in other situations he or she may have a preference for the disconfirmation of the same expectancy. All of these suggestions, of course, require empirical testing.

## A Final Comment

The self-fulfilling prophecy and the expectancy-confirmation effect have been of interest to psychologists partially because of the social policy implications of the research. However, in many of the research studies that document the effect, the specific and limited character of the material that creates the expectancies is lost, and we talk as if any material that creates expectancies is automatically accepted as valid by the perceiver. The image of the perceiver that emerges is one of an individual who takes his or her stereotypes and prejudices for granted and indiscriminantly applies them to members of the class he or she has stereotyped without any consideration of the unjustness of such a proceeding. The present study suggests that this is an oversimplification that in turn does some injustice to the perceiver. There are times when perceivers resist regarding their expectancies as truths and instead treat them as hypotheses to be confirmed or disconfirmed by relevant evidence. Perceivers in the present study did not make the error of reporting stereotypic judgments without sufficient evidence to warrant their conclusions. They engaged in an extremely rational strategy of evaluating the behavioral evidence when it was available and refraining from judgment when it was not. It was the strategy perceivers employed to analyze the evidence that led them to regard their hypotheses as confirmed even when the objective evidence did not warrant that conclusion. The error the perceivers make, then, is in assuming that the behavioral evidence they have derived is valid and unbiased. Future research could profitably address the question of the conditions under which this general confirmation strategy can be reversed or eliminated. In the meantime, however, the image of the perceiver as a hypothesis tester is certainly more appealing than that of a stereotype-ap-

plying bigot, even though the end result of both processes, sadly enough, may be quite similar.

## References


Brigham, J. C. Ethnic stereotypes. *Psychological Bulletin*, 1971, *76*, 15–38.

Cooper, H. M., Baron, R. M., & Lowe, C. A. The importance of race and social class information in the formation of expectancies about academic performance. *Journal of Educational Psychology*, 1975, *67*, 312–319.

Darley, J. M., & Fazio, R. H. Expectancy confirmation processes arising in the social interaction sequence. *American Psychologist*, 1980, *35*, 867–881.

Darley, J. M., & Goethals, G. R. People's analyses of the causes of ability-linked performances. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 13). New York: Academic Press, 1980.

Dreger, R. M., & Miller, S. K. Comparative psychological studies of Negroes and whites in the United States. *Psychological Bulletin*, 1960, *57*, 361–402.

Duncan, B. L. Differential social perception and attribution of intergroup violence: Testing the lower limits of stereotyping of blacks. *Journal of Personality and Social Psychology*, 1976, *34*, 590–598.

Foster, G., Schmidt, C., & Sabatino, D. Teacher expectancies and the label "learning disabilities." *Journal of Learning Disabilities*, 1976, *9*, 111–114.

Hayden, T., & Mischel, W. Maintaining trait consistency in the resolution of behavioral inconsistency: The wolf in sheep's clothing? *Journal of Personality*, 1976, *44*, 109–132.

Kahneman, D., & Tversky, A. On the psychology of prediction. *Psychological Review*, 1973, *80*, 237–251.

Karlins, M., Coffman, T. L., & Walters, G. On the fading of social stereotypes: Studies in three generations of college students. *Journal of Personality and Social Psychology*, 1969, *13*, 1–16.

Kelley, H. H. The warm–cold variable in first impressions of persons. *Journal of Personality*, 1950, *18*, 431–439.

Kennedy, W. A., VanDeReit, V., & White, J. C. A normative sample of intelligence and achievement of Negro elementary school children in the southeastern United States. *Monographs of the Society for Research in Child Development*, 1963, *28*, 13–112.

Langer, E. J., & Abelson, R. P. A patient by any other name . . . : Clinician group differences in labeling bias. *Journal of Consulting and Clinical Psychology*, 1974, *42*, 4–9.

Lesser, G. S., Fifer, G., & Clark, D. H. Mental abilities of children from different social class and cultural groups. *Monographs of the Society for Research in Child Development*, 1965, *30*, 1–115.

Locksley, A., Borgida, E., Brekke, N., & Hepburn, C. Sex stereotypes and social judgment. *Journal of Personality and Social Psychology*, 1980, *39*, 821–831.

Locksley, A., Hepburn, C., & Ortiz, V. Social stereotypes and judgments of individuals: An instance of the base-rate fallacy. *Journal of Experimental Social Psychology*, in press.

Lord, C., Ross, L., & Lepper, M. E. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, 1979, *37*, 2098–2109.

Merton, R. K. The self-fulfilling prophecy. *Antioch Review*, 1948, *8*, 193–210.

Nisbett, R., & Ross, L. *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, N.J.: Prentice-Hall, 1980.

Quattrone, G. A., & Jones, E. E. The perception of variability within in-groups and out-groups: Implications for the law of small numbers. *Journal of Personality and Social Psychology*, 1980, *38*, 141–152.

Regan, D. T., Strauss, E., & Fazio, R. Liking and the attribution process. *Journal of Experimental Social Psychology*, 1974, *10*, 385–397.

Rist, R. C. Student social class and teacher expectations: The self-fulfilling prophecy in ghetto education. *Harvard Educational Review*, 1970, *40*, 411–451.

Rosenhan, D. L. On being sane in insane places. *Science*, 1973, *179*, 250–258.

Rosenthal, R. *On the social psychology of self-fulfilling prophecy: Further evidence for Pygmalion effects and their mediating mechanisms*. New York: MSS Modular Publications, Module 53, 1974.

Rosenthal, R., & Jacobson, L. *Pygmalion in the classroom*. New York: Holt, Rinehart & Winston, 1968.

Snyder, M., & Cantor, N. Testing hypotheses about other people: The use of historical knowledge. *Journal of Experimental Social Psychology*, 1979, *15*, 330–342.

Snyder, M., & Swann, W. B. Behavioral confirmation in social interaction: From social perception to social reality. *Journal of Experimental Social Psychology*, 1978, *14*, 148–162. (a)

Snyder, M., & Swann, W. B. Hypothesis-testing processes in social interaction. *Journal of Personality and Social Psychology*, 1978, *36*, 1202–1212. (b)

Snyder, M., Tanke, E. D., & Berscheid, E. Social perception and interpersonal behavior: On the self-fulfilling nature of social stereotypes. *Journal of Personality and Social Psychology*, 1977, *35*, 656–666.

Swann, W. B., & Snyder, M. On translating beliefs into action: Theories of ability and their application in an instructional setting. *Journal of Personality and Social Psychology*, 1980, *6*, 879–888.

Weiner, B., et al. *Perceiving the causes of success and failure*. Morristown, N.J.: General Learning Press, 1971.

Word, C. O., Zanna, M. P., & Cooper, J. The nonverbal mediation of self-fulfilling prophecies in interracial interaction. *Journal of Experimental Social Psychology*, 1974, *10*, 109–120.

Zadny, J., & Gerard, H. B. Attributed intentions and informational selectivity. *Journal of Experimental Social Psychology*, 1974, *10*, 34–52.


Received August 31, 1981
Revision received April 2, 1982 ■