Training in Statistical Reasoning Inhibits the Formation of Erroneous Group Stereotypes

Mark Schaller
University of British Columbia

Charles H. Asp
Michelle Ceynar Rosell
University of Montana

Stephen J. Heim
Northwestern University

Erroneous group stereotypes can result from people's failure to engage in sufficiently sophisticated reasoning strategies. Three experiments tested the hypothesis that training in statistical reasoning inhibits the formation of these stereotypes. In Study 1, 60 students were assigned randomly to a control condition or one of two training conditions in which they received training in the logic of analysis of covariance. Approximately 1 week later, they were presented with a group impression formation task. Control participants formed erroneous stereotypes, but those who received statistical training formed more accurate group impressions. Study 2 (N = 82) replicated these results, addressed concerns with experimental demand, and provided preliminary evidence concerning possible moderating effects of motivation. Study 3 (N = 44) tested a different alternative explanation and provided further clues about the inferential processes through which statistical training influences group impressions.

Functional though group stereotypes may be for their users, they have harmful consequences for members of stereotyped groups. This is especially so when, as is often the case, the stereotype inaccurately depicts the group and its members. Concern with these consequences has resulted in campaigns to control the negative attitudes and behaviors that are linked to group stereotypes (see Aronson, Stephan, Sikes, Blaney, & Snapp, 1978; Cook, 1985; Hewstone & Brown, 1986; Katz & Taylor, 1988; Miller & Brewer, 1984). The results indicate that although certain interventions might alter the expression of prejudice, underlying negative stereotypes may be largely unchanged (Devine, 1989; McConahay, 1986; Sigall & Page, 1971). More recently, there have been renewed efforts to focus on stereotypes directly (Rothbart & Lewis, 1988; Weber & Crocker, 1983). It has become abundantly clear that stereotypes are extremely hardy. Once formed, a group stereotype is difficult to eliminate from the individual mind or from the collective repositories of a culture.

It may be worthwhile to consider additional weapons with which to combat stereotypes. Like diseases, stereotypes might be fought proactively as well as reactively. Within the medical community, there has been an increased emphasis on primary prevention—interventions designed to reduce the likelihood of contracting disease in the first place. Might it be possible to develop analogous interventions that impede the initial development of certain erroneous group stereotypes?

We suggest that it is possible. Our hypothesis follows from recent research revealing the relation between statistical reasoning and stereotype formation.

STATISTICAL REASONING AND STEREOTYPE FORMATION

People are intuitive statisticians (Gigerenzer & Murray, 1987; Nisbett, Krantz, Jepson, & Kunda, 1983; Peterson &

Authors' Note: The research reported in this article was supported by a National Science Foundation EPSCoR program and by the Montana Science and Technology Alliance. We thank Laura Taylor, Rachel Brown, Pat Hagerty, Todd Havelka, Melaney Swenson, Ivan Wodarz, and Glen Zeller for assistance in data collection. Thanks also to Monica Biernat and Chris Crandall for helpful comments on an earlier draft. Address correspondence to Mark Schaller, Department of Psychology, University of British Columbia, West Mall 2136, Vancouver, B.C., Canada V6T 1Z4.

PSPB, Vol. 22 No. 8, August 1996 829-844
© 1996 by the Society for Personality and Social Psychology, Inc.

Beach, 1967). Stereotype formation offers merely one example of the intuitive statistics that enter into everyday social inference. At a descriptive level, stereotypes may be represented cognitively as the perception of covariation between groups and dispositions (Hamilton & Gifford, 1976; Schaller, 1994). Indeed, intergroup comparison is implicit—if not explicit—in virtually all group stereotypes. For example, to stereotype a group as stupid and lazy is to imply that the group is less intelligent and less hard-working than another group. Any group stereotype implies some perceived covariation.

At an explanatory level, statistical reasoning also plays a role in the development of certain group stereotypes. It appears that some erroneous stereotypes may result in part from application of overly simple models of data integration and inference (Schaller, 1994; Schaller & O'Brien, 1992). This occurs because there exist group differences in behavior and performance that are spurious, the result of other variables that are confounded with group and with behavior. For example, the observed racial differences in achievement test scores appear to result not from inherent racial differences in intelligence or motivation but in part from differential constraints imposed on people of different races (e.g., differences in quality of early education)—constraints that directly influence achievement test scores (Fairchild, 1984). Similar constraints may account in part for observed differences in the behavior of men and women (Eagly, 1987). To form accurate inferences about group differences, scientists must statistically control for such confounded third variables (Fairchild, 1991). At a more intuitive level, laypersons must do the same thing.

Clearly this is not an easy task. It is not surprising that people often fail to engage sufficiently sophisticated inference strategies and consequently form inaccurate stereotypes of various disadvantaged social groups (Eagly, 1987; Fairchild, 1984; Schaller & O'Brien, 1992). But the news is not all bad. People do sometimes engage in an impressively complex statistical reasoning strategy that—like a statistical analysis of covariance—takes confounded third variables into account (Schaller, 1994). Consequently, people can form very accurate group impressions. This is not to say that complex statistical reasoning eliminates the formation of stereotypes. But the content of these overall impressions is, at least, more accurate.

Recent research has revealed that the underlying reasoning process is responsive to a number of variables. For instance, this process is affected by the amount of information available, the order in which that information is encountered, and the cognitive load experienced by the perceiver (Schaller, 1992a; Schaller, 1992b; Schaller & O'Brien, 1992). The process is influenced also by one's expectations, motives, and goals (Schaller, 1992b). And there is evidence that individual differences in cognitive style also exert an influence on statistical reasoning and group impression formation (Schaller, Boyd, Yohannes, & O'Brien, 1995). For instance, people who like a well-structured, predictable environment (see Neuberg & Newsom, 1993) are most likely to reason according to simplistic statistical models and to form erroneous stereotypes.

These results are encouraging for two reasons. First, they reveal that although people typically may think simplistically, they do have the potential to engage in complex statistical reasoning. Thus, under the right circumstances, they form highly accurate group impressions. Second, the results remind us that stereotype formation is, to paraphrase Allport (1954), a reflection of a person's whole habit of thinking about the world he or she lives in. This suggests that by changing a person's habit of thinking, one may prevent that person from forming certain types of erroneous group stereotypes.

PROPHYLACTIC EFFECTS OF STATISTICAL TRAINING?

These habits of thinking are likely to respond to social learning processes (Mischel, 1973). Therefore, education in complex statistical reasoning may be one possible means of proactively influencing the way people draw inferences from group-relevant information. Training in the logic of statistical analysis of covariance (ANCOVA) might enhance the accuracy of the group impressions that people form.

Why might training in ANCOVA reasoning have an impact? There are at least three reasons, each related to a different step in the logical inference process. First, training might enhance awareness that seemingly obvious inferences might be wrong. In a sense, training may make people more cautious, more skeptical of their own initial impressions. Second, as a result of this caution, training might enhance vigilance or sensitivity to third variables of possible importance. Training may enhance the likelihood of actually detecting confounding variables that must then be accounted for. Third, training in ANCOVA reasoning may increase one's ability to adjust appropriately for any confound that has been detected, to "partial it out" when arriving at group impressions.

Though optimistic, these speculations are not unrealistic. Research has shown that people who receive training in logical/statistical thinking are more likely later to apply these statistical concepts when drawing inferences and making judgments (Agnoli, 1991; Crandall & Greenfield, 1986; Fong & Nisbett, 1991; Nisbett, Fong, Lehman, & Cheng, 1987). Most of this research has focused on probability theory and is not directly relevant to the process considered here. But some studies are at least indirectly relevant, revealing that broad training in statistical methods influences the extent to which people recognize confounding variables in everyday inference situations (Lehman, Lempert, & Nisbett, 1988).

No research has specifically examined training in the logic of statistical ANCOVA, nor has any research explored the possibility that such training might improve the accuracy of group impression formation. The purpose of the experiments reported here was to do just that.

STUDY 1

To test the hypothesis that statistical training enhances the accuracy of group impression formation, we recruited college students to participate in two ostensibly unrelated experiments. The first of these experiments was a training session. During this session, some of the students received procedures designed to teach them the logic of statistical analysis of covariance. One of these training procedures presented the statistical logic within the framework of sports scenarios only. The other training procedure included an additional scenario that applied this logic specifically to group stereotyping. A third group of participants received no statistical training at all.

After completing this first session, the students signed up to participate in the second experiment at a later date. This second experiment involved procedures previously used to study statistical reasoning and the formation of erroneous group stereotypes (e.g., Schaller, 1992b; Schaller et al., 1995; Schaller & O'Brien, 1992).

Given these methods, we were able to test the hypothesis that training in the logic of statistical ANCOVA can inhibit erroneous stereotype formation. We were also able to test for differences between different training approaches. It is of some interest to consider whether training in ANCOVA must be specifically relevant to group judgment tasks in order to exert a positive effect. Past research on instruction in logic and reasoning reveals a difference of opinion on what process or processes might be responsible for its long-term effects. Research by Fong and Nisbett (1991; Nisbett et al., 1987) suggests that instruction aids the development of abstract rules of inference and need not be domain relevant to influence inference and judgments. However, research on analogical transfer (e.g., Ross, 1987) suggests that superficial aspects of instruction importantly affect the generality of its influence.

Method

PARTICIPANTS

Participants were undergraduate students enrolled in introductory psychology courses at the University of Montana. Eighty-five students participated in the first experimental session in exchange for credit toward a course requirement. During this first session, participants were randomly assigned either to a control (no training) condition or to one of two training conditions in which they received training in the logic of statistical ANCOVA. From this initial set of students, 60 students (38 women and 22 men) participated in the second session for additional credit or for $5. An average of 7.82 days elapsed between sessions.

SESSION 1 (STATISTICAL TRAINING SESSION)

Participants were met by a male experimenter who explained that the purpose of the experiment was to evaluate a new technique for teaching statistics. They were told that they would be given a lecture using this new format and would be asked to evaluate this teaching technique.

Before proceeding, participants completed two questionnaires: the 12-item Personal Need for Structure Scale (PNS; Neuberg & Newsom, 1993) and the 28-item Attributional Complexity Scale (AC; Fletcher, Danilovacs, Fernandez, Peterson, & Reeder, 1986). Both measures have well-established reliability and validity, and both have been shown to relate to "intuitive analysis of covariance" in a group perception task (Schaller et al., 1995).

Participants in the two training conditions proceeded to receive a "lecture/workshop on analysis of covariance."

Tennis-only training. The "lecture" included a packet of written materials consisting of three separate judgment scenarios, all of which involved making judgments about tennis abilities. All three scenarios demanded that participants engage in the logic of ANCOVA to provide an accurate answer. The scenarios became progressively more complex.

To begin the "lecture," the experimenter presented participants with a written scenario describing a person who claimed to be a better tennis player than John McEnroe because "I've won every tennis match I've played this year; but he's lost quite a few of his." Participants were asked whether they believed this braggart and why. After participants responded orally, they read an explanation concerning the logic of taking into account the different ability levels of two tennis players' different opponents. These materials introduced participants to statistical terminology (e.g., covariation, controlling for confounded variables).

The second scenario presented summary data concerning the performance of two fictitious tennis players, Fred and Barney, who played matches in two different leagues. The summary data presented overall win-loss records revealing that Fred had won 40 of 100 matches while Barney had won 60 of 100. These data also presented win-loss records within each league, revealing that the overall records were misleading because (a) the leagues were apparently quite unequal in level of competition and (b) Fred and Barney had played vastly different numbers of matches in each league. Thus, despite his lower overall winning percentage, Fred had a higher winning percentage than Barney within each league when considered separately (see Schaller, 1992a, for details). Participants were asked who they thought was the better player and why. They were then given written materials that explained the logic of analysis of covariance as a means of controlling for confounded variables. These materials discussed how the results of this statistical analysis would reveal that despite his lower overall winning percentage, Fred was most likely to be the better player.

The third scenario described a scientist testing the hypothesis that the likelihood of winning a tennis match is negatively related to opponent's ability level. To test this hypothesis, the scientist gathered data by observing three target players—a particularly bad high school player, a moderately good college varsity player, and a particularly good professional player. In observing these players, the scientist recorded the ability level of the opponent and also recorded whether the target player won or lost. The scientist then computed the correlation between those two variables and found an effect opposite to the predicted negative correlation. To demonstrate this, the written materials presented a scatterplot representing a positive correlation between likelihood of winning and opponent's ability. Participants were asked to explain this unexpected and counterintuitive finding. They were then given a detailed explanation that discussed how the sampling procedure (choosing three target players of very different ability levels) may have introduced a variable (the target player's own ability) that was confounded with the two variables of interest. Participants were shown three scatterplots demonstrating the expected negative correlations between winning likelihood and opponent's ability among (a) high school players, (b) college varsity players, and (c) professional players. They were shown how collapsing these three separate scatterplots into a single aggregate scatterplot could create an overall correlation that is misleadingly reversed.

The lecture closed with a brief summary of the statistical logic involved in drawing inferences about covariation, emphasizing the value of statistical ANCOVA and three-dimensional statistical thinking in general.

The written training module was 2,166 words in length. A reading ease analysis (using the Correct Grammar software) revealed a Flesch-Kincaid grade level score of 9.6. Reading time and response time were not controlled; the experimenter allowed participants as much time as needed to understand the information. The experimenter orally supplemented the written explanations if the subject needed additional assistance to understand the logical and statistical principles. The experimenter ended the session when he felt sure the subject fully understood the material presented.

Tennis + stereotype training. Participants assigned to this condition received statistical training procedures identical to those in the tennis-only condition, with the addition of a fourth judgment scenario. This fourth scenario, presented after the three tennis scenarios, was designed to show participants the application of the logic of statistical ANCOVA to the perception of erroneous group stereotypes. The scenario presented summary data concerning the performance of Black and White high school students on achievement tests. These data revealed that, overall, 35% of Black students and 65% of White students passed the achievement test. But separate summaries of performance among students at rich schools and poor schools revealed that the overall results were entirely spurious—resulting from the fact that the type of school was highly correlated with race and with achievement test performance. Within each type of school, there was no racial difference in the percentage of students passing the test. Participants were asked to explain orally their conclusions concerning the relation between race and academic abilities. They were then given a written explanation that discussed how the apparent racial differences are revealed to be illusory when controlling for the confounded variable.

With the addition of this fourth scenario, the written training module was 2,712 words in length. A reading ease analysis revealed a Flesch-Kincaid grade level score of 9.7.

Control condition. After completing the PNS and AC questionnaires, participants in the control condition were told that they were in a control condition and would not be receiving the statistics lecture. So, for them, the experiment was over.

Questionnaire and recruitment for Session 2. At the completion of the experiment, a final questionnaire was administered to all participants. Three questions assessed (a) the extent to which participants liked math, (b) whether they had taken any statistics courses while in college, and (c) their gradepoint average (GPA). Five additional questions (not asked in the control condition) assessed participants' evaluations of the teaching methods.

Finally, participants were thanked and given credit for participating in the experiment. The experimenter then asked whether they would like to participate in a different experiment to be run approximately 1 week later. The experimenter professed ignorance of the purpose or nature of this other experiment (to enhance the perception that the two experimental sessions were un-

TABLE 1: Breakdown of Stimulus Information According to Group Membership, Anagram Length, and Performance Outcome, Studies 1 and 2

	Five-Letter Anagrams		*Seven-Letter Anagrams*		*Aggregate*	
Membership	*Success*	*Failure*	*Success*	*Failure*	*Success*	*Failure*
Group A	5	0	5	15	10	15
Group B	15	5	0	5	15	10

related). The experimenter told participants that they could earn either additional experimental credit or $5 for participation in the study. Twelve students who participated in Session 1 declined the opportunity to sign up for Session 2; an additional 13 students signed up but failed to attend. (Among these students who did not participate in Session 2, 12 were from the control condition, 5 from the tennis-only condition, and 8 from the tennis + stereotype condition.)

SESSION 2 (GROUP IMPRESSION SESSION)

The second session was conducted in a different laboratory. Students participated individually or in pairs. They were greeted by a female experimenter who was unaware of their training condition. Participants were given an introduction indicating that the purpose of the experiment was to examine the way people process information when forming impressions of others. The introduction also provided an overview of the procedures. Participants were told that they would be presented with representative information about the performance of members of two groups (called simply Group A and Group B) on anagram tasks. They were told that these people had been presented with specific anagrams as part of a separate experiment and had been given a certain amount of time to solve each anagram. Their task would be to judge the relative intelligence of the two groups on the basis of group members' ability to solve the anagrams (see Schaller & O'Brien, 1992, for details).

Composition of the group-relevant information. Stimuli consisted of 50 statements, each describing an attempt by a particular person to solve a particular anagram. Each statement revealed four pieces of information: (a) the group membership of the person, (b) whether the person solved or failed to solve the anagram, (c) the actual anagram (some anagrams were five letters long and some were seven letters long), and (d) the correct solution to the anagram. The stimuli were carefully constructed so that, across all 50 statements, Group A members correctly solved 10 of 25 anagrams and Group B members correctly solved 15 of 25 anagrams. These overall ratios were misleading, however, because anagram length covaried with both group membership and success: Group A attempted many more seven-letter anagrams, whereas Group B attempted many more five-letter anagrams; and five-letter anagrams were much more easily solved than seven-letter anagrams. In fact, on both five- and seven-letter anagrams, Group A enjoyed a higher success rate than Group B (see Table 1 for a summary). A simple assessment of how many total anagrams each group successfully solved would suggest that Group B is more intelligent. However, if anagram difficulty were taken into account (by a statistical ANCOVA, for instance), Group A would appear more intelligent.

Presentation of group-relevant information. These stimuli were presented to participants in the form of a timed "slide show" on IBM-compatible microcomputers. Participants were asked to watch the slide show and to form impressions of which group they felt was more intelligent. Each slide presented one of the 50 stimulus statements. Each statement was shown for 8 s before it was replaced by the next slide. All participants saw the same 50 slides but were randomly assigned to view the slides in one of three different orders.

Dependent measures. After viewing the slide show, participants completed an evaluation questionnaire to assess their impressions of the groups' relative intelligence. On 9-point Likert-type scales, the participants rated which group they thought would do better if they were to attempt anagram tasks (we shall refer to this judgment as *anagram ability*) and which group was more intelligent (*Intelligence I*). Responses were scored on a scale ranging from –4 to +4, where negative values indicated judgments in favor of Group B and positive values indicated judgments in favor of Group A. On another 9-point scale, participants indicated how confident they were in their judgments. (Confidence ratings were not related to impression accuracy or influenced by the training manipulation and will not be discussed further.) Participants then rated the two groups' intelligence separately on 9-point scales. By subtracting the rating of Group B from the rating of Group A, we created a second measure of relative intelligence (*Intelligence II*), ranging from –4 to +4.

In addition to these measures of stereotyping, participants completed three measures designed to offer direct and face-valid assessments of inferential reasoning. They were asked to describe, in their own words, "What information did you consider in order to make your decision about which group had the better anagram-solvers and the more intelligent people?" After completing this written self-report, participants ranked six types of information according to "how important each type of information was in the way you mentally organized the information." Of particular interest was the ranking given to "length of anagram." Participants were then presented with a list of three possible information-organizing strategies. These varied in complexity, the third representing something akin to the logic of ANCOVA. Participants indicated which of the three strategies most closely represented their own method of organizing the information when forming impressions (for details, see Schaller & O'Brien, 1992).

Participants also completed two procedures designed to offer subtle and indirect indicators of reasoning. For one of these measures, participants were given a blank sheet of paper and asked to list any solution words that they remembered from the presentation. This list was used to generate indexes representing the way information was clustered in recall. Though prone to error variance, indexes of categorical clustering can offer subtle indicators of reasoning (see Schaller, 1992b; Schaller & O'Brien, 1992). In addition, participants were asked to estimate the number of anagrams they had seen that belonged to each of eight categories determined by the group, outcome, and anagram-length dimensions (e.g., "How many 7-letter anagrams did Group A successfully solve?"). Neither of these indirect measures revealed differences between conditions, and they will not be discussed further.

Debriefing. After completing the dependent measures, participants were fully debriefed concerning the purpose of the experiment, were asked not to discuss the experiment with other students, and were remunerated for their participation. Also as part of the debriefing, the experimenter noted whether participants volunteered a belief that the two sessions were part of the same experiment. No participant indicated any such belief.

Results

Effects of training on group impressions. To test the hypothesis that statistical training would influence relative impressions of Group A and Group B, we entered participants' three judgments of anagram-solving ability and intelligence in a multivariate analysis of variance (MANOVA). The main effect for training condition was significant, $F(6,110) = 2.39$, $p < .04$.

To explore the effect further, we conducted a series of planned contrasts. First, we tested the specific hypothesis that judgments in the two training sessions would differ from those in the control condition. Planned contrasts were conducted separately on the three impression measures, using one-tailed tests of significance because of the clear directional hypothesis. Results revealed a significant effect on Intelligence I, $t(57) = 2.31$, $p < .02$, and on anagram-solving ability, $t(57) = 1.76$, $p < .05$. The effect on Intelligence II did not quite attain conventional levels of significance, $t(57) = 1.52$, $p < .07$. In addition, we created a composite impression measure by averaging responses on the three individual measures. The planned contrast on this composite measure also indicated a significant difference between the control and the training conditions, $t(57) = 2.01$, $p < .03$ (one-tailed).

The means for the judgment of relative anagram-solving ability and the two judgments of relative intelligence are presented in Figure 1. This figure clearly reveals the effects of statistical training on the group impressions. Among participants in the control condition, mean scores on all three measures were less than zero—indicating that these participants judged Group A to be less intelligent than Group B. Control participants apparently failed to take the differential constraints into account and so formed erroneous group impressions. Participants who received the tennis-only and tennis-stereotype training did not form this erroneous stereotype. Mean judgments in both conditions were greater than zero, indicating that participants generally believed Group A to be more intelligent than Group B—the inference that would emerge from a statistical ANCOVA or partial correlation analysis.

Another set of planned contrasts, orthogonal to the first, tested whether mean impressions differed between the tennis-only and tennis + stereotype training conditions. No significant effects emerged, $ts < 1.15$. Mean differences on the measures of relative intelligence (but not the measure of relative anagram-solving ability) indicated that participants in the tennis + stereotype condition were most favorable toward Group A, but these judgments were not reliably different from those in the tennis-only condition.[1]

Effects of training on inferential reasoning. Of the 60 participants, 56 offered written reports of their reasoning processes ($ns = 16$, 20, and 20 in the control, tennis-only, and tennis + stereotyping conditions, respectively). Two judges, blind to experimental conditions, independently coded these self-reports according to three questions. The two judges agreed on 90% of their yes/no judgments across the three questions and resolved any disagreements through discussion.

One question addressed by coders was "Did the subject indicate that seven-letter anagrams were more difficult to solve than five-letter anagrams?" "Yes" ratings were recorded for 13%, 30%, and 35% of participants in the control, tennis-only, and tennis + stereotype conditions, respectively. A second question was "Did the subject indicate that the anagrams attempted by Group A were generally more difficult than the anagrams attempted by Group B?" "Yes" ratings were recorded for 6%, 40%, and 50% of participants in the same three conditions, respectively. Finally, coders addressed a third question: "Did the subject indicate that s/he considered success rates separately for the five- and seven-letter ana-

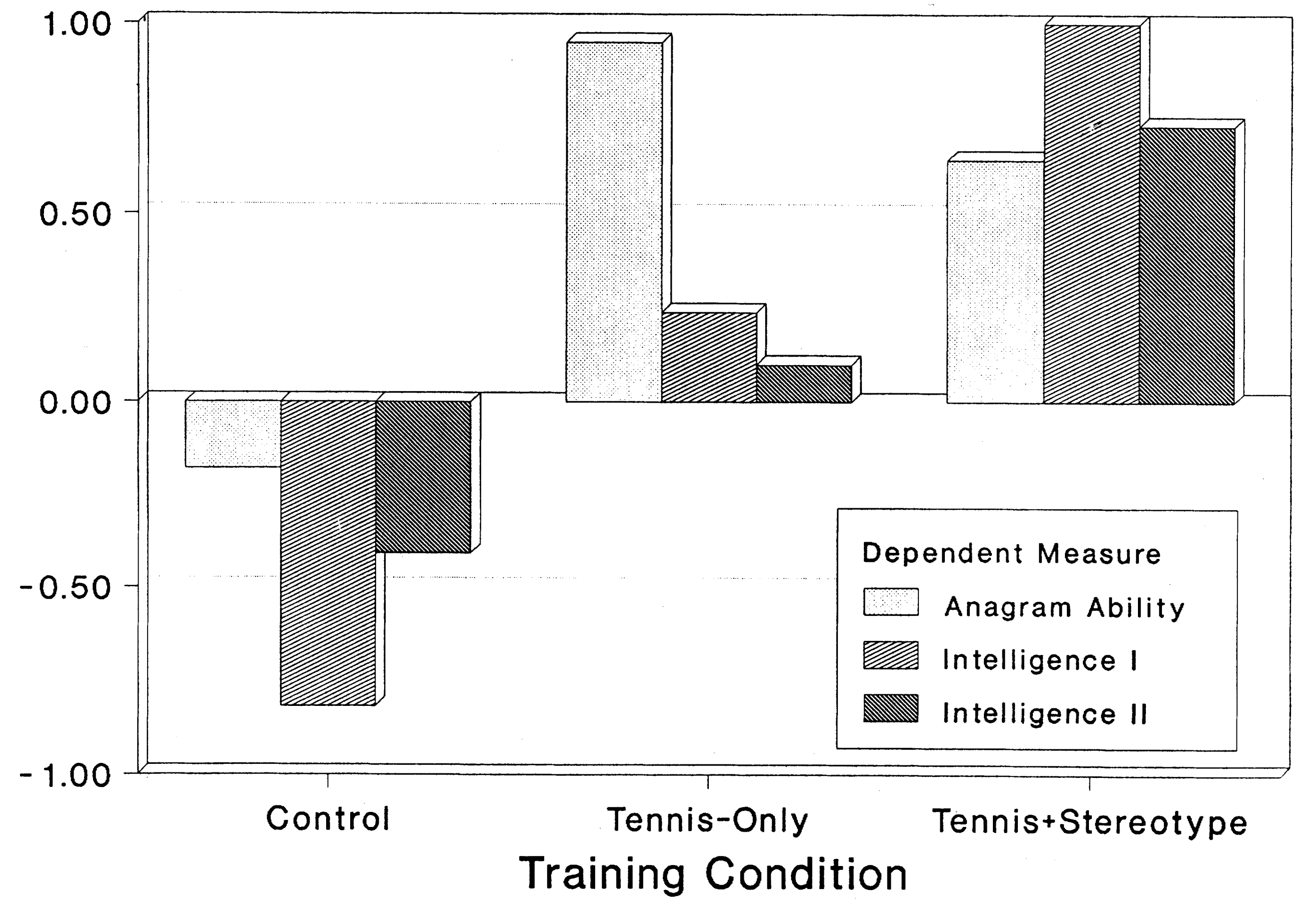

Figure 1 **Effects of ANCOVA training on the measures assessing relative group impressions, Study 1. (Scales ranged from –4 to +4; positive values indicate normatively accurate judgments favoring Group A.)**

grams?" "Yes" ratings were recorded for 25%, 45%, and 40% of participants in the three conditions, respectively.[2]

To provide a statistically sensitive measure of self-reported reasoning, a summary index was created by summing the number of "yes" responses to the three coding questions. Higher values on this reasoning index (possible range 0-3) signify greater attention to the confounded "third variable" and more complex inferential reasoning. To test for effects of training on this reasoning index, we performed two hypothesis-driven planned contrasts. A contrast (one-tailed because of the directional hypothesis) comparing the two training conditions against the control condition was significant, $t(53) = 2.24$, $p < .02$. A second contrast, orthogonal to the first, found no significant differences between the two training conditions, $t < 1$. As revealed in Figure 2, the mean reasoning score in the control condition was fairly low ($M = 0.44$), whereas the reasoning scores in the tennis-only and tennis + stereotype conditions were considerably higher ($Ms = 1.15$ and 1.25).

Similar mean trends were observed on the two other direct measures of inferential reasoning, and the same hypothesis-driven planned contrasts were performed on each measure. Contrasts testing the hypothesis that the training conditions would differ from the control condition approached but did not meet conventional levels of significance ($ps = .07$ and $.11$, one-tailed; see Table 2). Contrasts testing for differences between the two training conditions revealed no differences whatsoever.[3]

Discussion

Can training in ANCOVA reasoning proactively prevent people from forming erroneous stereotypes? Our results indicate that it can. Participants who received no training failed to take into account the differential situational constraints that influenced the anagram-solving performance of Group A and Group B and consequently drew the erroneous inference that Group A was less intelligent than Group B. In contrast, participants who earlier had received statistical training were more likely

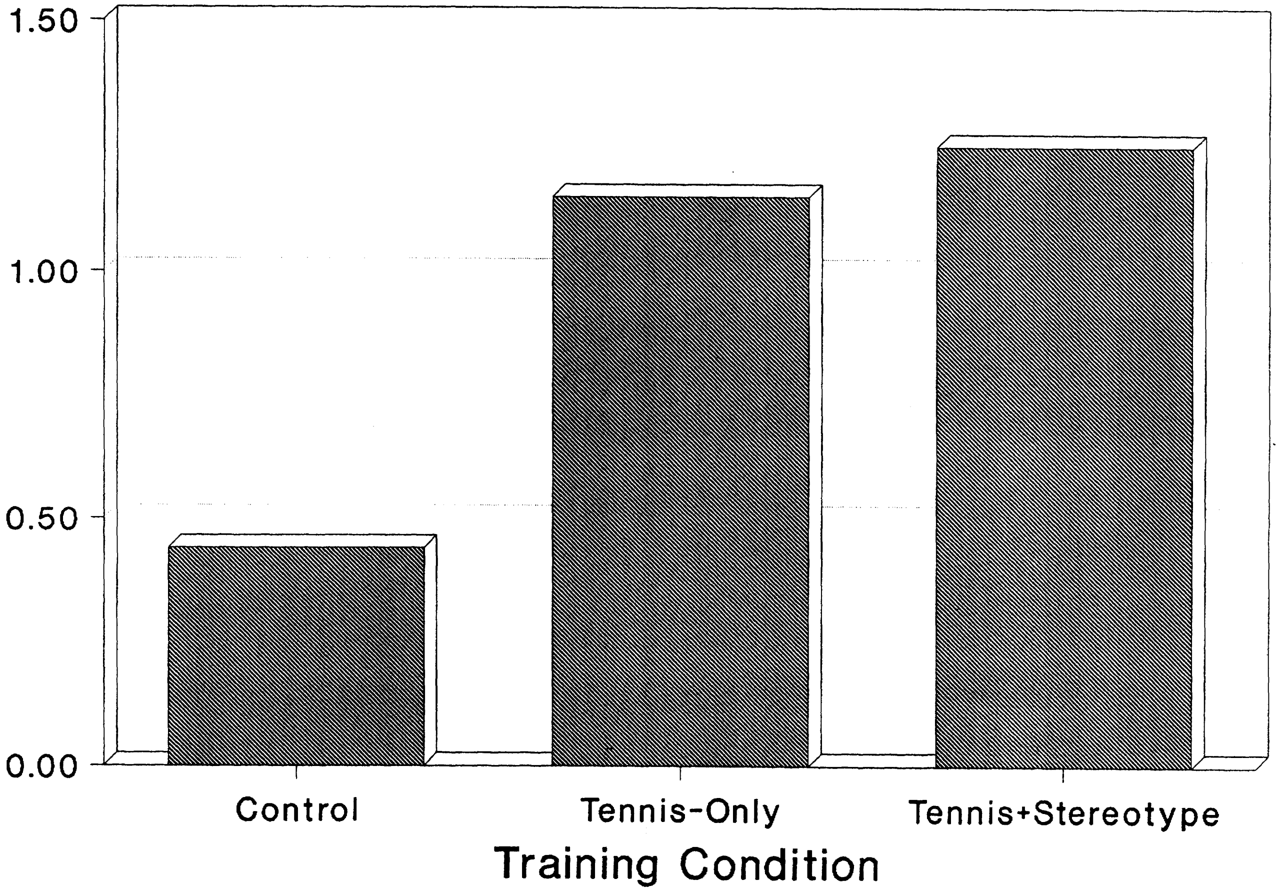

Figure 2 **Effects of ANCOVA training on the complexity of self-reported inferential reasoning, Study 1. (Scale ranged from 0 to 3; higher values indicate more complex, ANCOVA-like reasoning.)**

to identify the confound inherent in the group performance data, to adjust their inferences according to the differential constraints on performance, and to draw the normatively accurate inference that Group A was more intelligent.

There were no reliable differences between the two training conditions. This result has implications for our understanding of the process underlying the positive effects of ANCOVA training and is relevant to a current controversy over the underlying mechanisms in the transfer of statistical training in general. We will discuss these implications in our General Discussion, below. But first we must consider several issues that potentially weaken the inferential appeal of Study 1.

Recall that although participants were randomly assigned to training conditions during Session 1, not all of these participants signed up to participate in Session 2. This raises the possibility that some form of self-selection artifact might have contaminated the sample on which our analyses were conducted. Comfortingly, available evidence suggests otherwise. There were no reliable differences between conditions on personal need for structure, attributional complexity, grade point average, liking for mathematics, experience in statistics, or number of days between sessions (all *p*s > .10). In fact, to the extent that there were nonsignificant mean differences on any of these variables, the direction of these difference was generally such that they cannot account for the obtained results (e.g., participants in the control condition reported an average GPA slightly higher than that reported in either training condition). In sum, there is no evidence supporting a self-selection explanation for our results.

Perhaps a greater concern is the possibility that the differences between control and training conditions resulted not from the learning of ANCOVA logic but from demand characteristics associated with the procedures (see Orne, 1962). Despite our efforts to ensure that the two sessions would be perceived to be two entirely separate and unrelated experiments, we might not have been successful in doing so. And despite our careful attempt to disguise the purpose of the training session, participants in the training conditions might have later realized this purpose and deliberately responded in a hypothesis-

TABLE 2: **Effects of Training on the Primary Dependent Variables, Study 1**

	Experimental Condition			*Planned Contrast* p *Values*	
Dependent Variable	*Control*	*Tennis Only*	*Tennis + Stereotype*	*2 Training Conditions Versus Control*[a]	*Tennis Only Versus Tennis + Stereotype*[b]
Intelligence I	–0.82	0.24	1.00	.02	>.25
Intelligence II	–0.41	0.10	0.73	.07	>.25
Anagram-solving ability	–0.18	0.95	0.64	.05	>.25
Composite impression score	–0.47	0.43	0.79	.03	>.25
Self-report reasoning index	0.44	1.15	1.25	.02	>.25
Importance of length	3.47	4.48	4.09	.07	>.25
Organizing strategy	2.41	2.71	2.64	.11	>.25

NOTE: For all measures, higher and positive numbers indicate greater use of ANCOVA-like reasoning or normatively accurate judgments favoring Group A.

a. One-tailed test of significance.

b. Two-tailed test of significance.

confirming manner during Session 2. Although no participants indicated that they thought the two sessions to be related, we did not directly question them regarding this possibility. Study 2 further explored this possibility.

STUDY 2

Study 2 was essentially a replication and extension of Study 1, with two notable additions and one subtraction.

First, during debriefing, we questioned participants explicitly to probe for the possibility that they were responding to the demand characteristics of these procedures.

Second, we added another variable to the design to test for the possibility that motivation might moderate the effects of ANCOVA training on stereotype formation. Specifically, we included a group categorization manipulation (e.g., Schaller, 1992b): Before being presented with the anagram information during Session 2, participants were led to believe that they were members of either Group A or Group B. Given past research on the interplay between motivation and ability (e.g., Santioso, Kunda, & Fong, 1990), we anticipated the possibility that the positive effects of statistical training might occur only when participants are motivated to think complexly—in the condition in which participants were themselves members of Group A.

Third, we subtracted one of the training conditions. Given that Study 1 revealed no reliable differences between the two training conditions, we streamlined our design by dropping the tennis-only condition. Thus, during Session 1, participants were randomly assigned either to the control (no training) condition or to the tennis + stereotype training condition.

Method

PARTICIPANTS

Participants were undergraduate students enrolled in introductory psychology courses at the University of Montana. In Session 1, 127 students were randomly assigned either to a control (no training) condition or to the training condition. Eighty-four of these students (60 women and 24 men) participated in Session 2 for additional credit or for $5. An average of 10.05 days elapsed between sessions.

Of the 43 students who failed to participate in Session 2, 20 had been in the control condition and 23 in the training condition. We explored the possibility that self-selection might lead to different subject profiles in the two conditions. Analyses on all measured individual difference variables indicated only one that correlated with experimental condition: More participants from the training condition indicated that they had taken a statistics course ($p < .03$). However, this variable was unrelated to any of the measures of stereotype formation ($-.06 < rs < .06$). Thus there is no evidence to suggest that any self-selection artifact seriously compromises the internal validity of the study.

PROCEDURE

Session 1. The materials and procedures used during Session 1 were identical to those of Study 1 except that there was no tennis-only training condition.

Session 2. The materials and procedures used during Session 2 were identical to those of Study 1 except for the addition of procedures designed to lead participants to believe that they were themselves members of either Group A or Group B. These procedures were adapted from those described by Schaller (1992b, Study 3). Before being presented with the stimulus information, participants were told that their own membership in Group A or Group B had been determined on the basis of questionnaires they had filled out in a different experiment. At this point the experimenter gave the name of the experiment, which was, in fact, Session 1.[4] After consulting a computer printout, the experimenter wrote either "A" or "B" on an index card, according to a predetermined randomized block procedure, and gave a card to each subject indicating his or her group membership. After this group assignment manipulation, the procedures resembled those of Study 1.

Debriefing. After informing participants of the true nature of the experiment, the experimenter explicitly asked participants whether they had realized that the two experimental sessions were related in any way other than to determine participants' ostensible group membership. Only 9 of the 84 participants answered that they had thoughts along these lines—4 participants in the control condition and 5 in the training condition. None of these participants identified the actual hypotheses; rather, they tended to indicate merely that the two sessions had seemed in some way similar. Separate analyses that excluded these 9 participants revealed effects virtually identical to those reported below. Thus there was no evidence indicating that the effects of training were merely the result of demand characteristics.

Results

Effects on group impressions. Separate 2 × 2 (Training × Group Categorization) analyses of variance (ANOVAs) were performed on the three primary measures of group impression formation and on a composite group impression score computed by averaging the three scales. Significant or near-significant main effects of training emerged on all four variables: Intelligence I, $F(1, 80) = 7.18$, $p < .01$; Intelligence II, $F(1, 80) = 3.58$, $p < .06$; anagram-solving ability, $F(1, 80) = 2.45$, $p < .12$; composite group impression score, $F(1, 80) = 5.54$, $p < .02$.

Figure 3 presents the means for the three primary stereotyping measures within the control and training conditions and reveals the positive effects of ANCOVA training on group impressions. Among participants in the control condition, mean scores on all three measures were less than zero, indicating that these participants erroneously judged Group A to be less intelligent than Group B. Among participants who received training in ANCOVA logic, mean scores on all three measures were greater than zero, indicating the more accurate inference that Group A was more intelligent.

On none of the three individual stereotyping measures nor on the composite was there any main effect of group categorization, *F*s < 1. And only on Intelligence II did the interaction approach significance, $F(1, 80) = 2.46$, $p < .12$. Given a prediction that participants would be most likely to form accurate impressions if they had both the ability and the motivation to do so, the most appropriate test is a 3 versus 1 contrast, comparing the training/Group A condition against the other three conditions. Tests of this contrast revealed a significant effect for Intelligence II, $t(80) = 2.50$, $p < .01$ (one-tailed). Participants in the training/Group A condition accurately judged Group A to be more intelligent ($M = 0.39$), whereas participants in all three other conditions judged Group A to be less intelligent ($-0.70 < Ms < -0.33$). No significant effects emerged on the other two group impression measures or on the composite (*p*s > .15).

Effects on inferential reasoning. For the 72 participants who wrote interpretable self-reported reasoning strategies (*n*s = 37 and 35 in the control and training conditions, respectively), self-reports were coded by a single rater, who was blind to experimental condition, in the same manner as for Study 1. (A second rater coded a subset of the self-reports and agreed with the primary rater's codings 93% of the time.)

Results indicated that participants in the training condition were more likely than those in the control condition (a) to note that five- and seven-letter anagrams differed in difficulty (26% vs. 5%), (b) to note that anagrams attempted by Group A were more difficult than those attempted by Group B (37% vs. 16%), and (c) to attempt to calculate success rates separately for the five- and seven-letter anagrams (34% vs. 19%). As in Study 1, these yes/no ratings were positively related to participants' judgments about the groups' relative anagram abilities and intelligence, both across all participants and within training conditions. Also as in Study 1, the one exception to this pattern was that, within the control condition only, there was no relation between group impressions and explicit recognition that the anagrams differed in difficulty ($-.05 < rs < .10$ in the control condition, vs. *r*s > .42 in the training condition).

The summary reasoning index was computed in the same way as in Study 1 and was subjected to a 2 × 2 (Training × Group Categorization) ANOVA. Only one significant effect emerged, a main effect for training, $F(1, 68) = 5.61$, $p < .02$. Participants in the training condition ($M = 0.97$) indicated more complex ANCOVA-like reasoning than those in the control condition ($M = 0.41$). No effects on any other measures were significant.

Discussion

The results of Study 2 replicated the primary results of Study 1. Participants who received training in the logic of statistical ANCOVA were less likely later (a week and a half later) to form erroneous stereotypes of novel groups.

In addition, Study 2 provided some support for the hypothesis that an in-group favoritism motive might moderate the positive effect of statistical training. But this support was mixed. No strong conclusions can yet be drawn concerning this hypothesis. Conceptually, it seems reasonable to expect some interactive effects of one's ability to think complexly and one's motivation to do so. Empirically, this hypothesis awaits further testing.

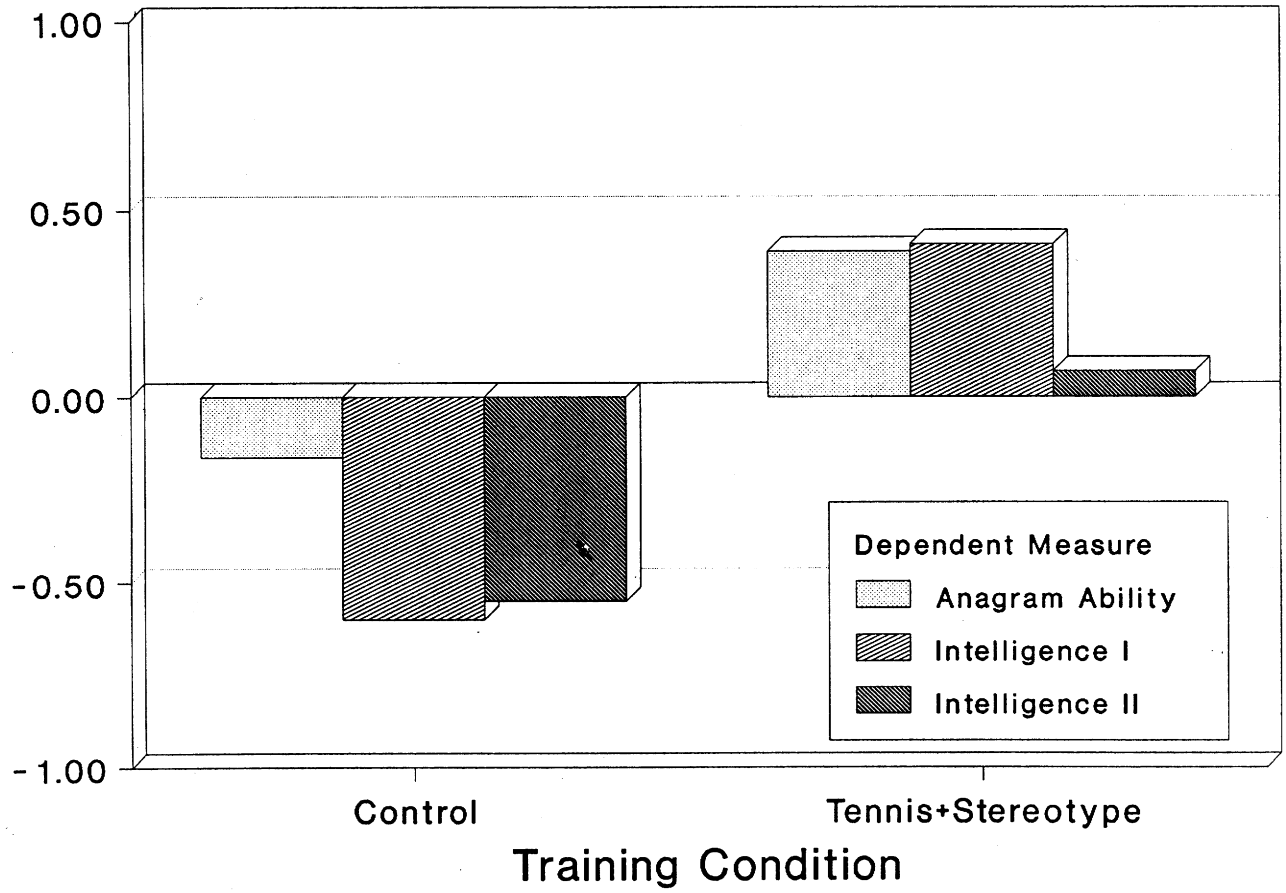

Figure 3 Effects of ANCOVA training on the measures assessing relative group impressions, Study 2. (Scales ranged from –4 to +4; positive values indicate normatively accurate judgments favoring Group A.)

Study 2 also examined more directly the possibility that the effect of the training manipulation might be due merely to experimental demand. As in Study 1, we maintained the appearance that the two sessions were not importantly related. In addition, we explicitly provided participants with a plausible but misleading link between the two, so that they would not attempt to guess the true connection even if they did perceive the two sessions to be related. We also crafted debriefing procedures to directly question participants concerning their perceptions. Responses indicated that participants were not merely responding to demand characteristics. One other result also argues against a demand interpretation. Had participants responded merely in the manner that they believed they should, it seems likely that participants in the tennis + stereotype training condition would have judged Groups A and B to be equally intelligent. This was not the nature of their responses. Instead, across both studies, these participants actually judged Group A to be more intelligent than Group B, indicating that they truly adjusted for the confounded third variable. Together, these considerations indicate that experimental demand cannot provide a compelling alternative explanation for our results.

Before turning to a more general discussion of these results, however, it is worthwhile to address an additional alternative explanation. Perhaps, rather than learning to apply more sophisticated inferential logic, participants in the training conditions were sensitized to the trickery of psychological inference tasks and adopted a simplistic "reversal" rule: Whenever some person or group appears at first to be less talented, that person or group is actually more talented. Although the training procedures never suggested any such rule, the tennis scenarios we used may have implied to participants that this rule had some merit. This alternative explanation can indeed account for the significant effects of training on ratings of relative group intelligence. Although nothing in participants' self-reports indicated the use of any sort of simple reversal rule, it may be premature to dismiss the alternative explanation without testing it directly.

STUDY 3

To test this alternative explanation, subjects received either ANCOVA training or no training and were later presented with a group inference task on which ANCOVA reasoning and a reversal rule would lead to quite different judgments. Participants applying ANCOVA reasoning would be expected to favor Group A; participants applying a reversal rule would be expected to form inaccurate impressions, favoring Group B.

Method

Participants were 44 University of Montana undergraduates who participated in exchange for course credit. Procedures were virtually identical to those used in Studies 1 and 2, except that participants completed the two sessions of the experiment on the same day. During the training session, participants were randomly assigned to a control condition (during which they simply filled out questionnaires) or to receive ANCOVA training. Those receiving training received the tennis-only training procedures described under Study 1.

At the completion of the training session, all participants were directed to a different laboratory to participate in an ostensibly unrelated experiment. During this session, they were presented with a group impression task identical in all respects but one to that of Studies 1 and 2. The one respect in which the task differed from the preceding studies was in the structure of the anagram information presented to participants. Members of Group A correctly solved 15 of 25 anagrams, whereas members of Group B correctly solved 10 of 25 anagrams. Group A's superior overall performance occurred despite the fact that Group A was presented with a higher percentage of longer (i.e., more difficult) anagrams. Group A attempted 19 seven-letter anagrams and correctly solved 9 of them; Group B attempted only 6 seven-letter anagrams and solved none of them. On five-letter anagrams, Group A's success rate (6/6) was also higher than Group B's success rate (10/19).

Given these data, both simplistic and complex reasoning favor Group A; in fact, a statistical analysis controlling for anagram length implies all the more strongly that Group A is more intelligent than Group B. Impressions of relative group intelligence were measured in a manner identical to that used in Studies 1 and 2—on 9-point scales with positive values indicating the normatively accurate inference favoring Group A. Participants also completed several of the additional measures to assess their reasoning.

This task allowed a test of the alternative explanation posed above. If, as a result of ANCOVA training, participants merely learned to distrust their first impressions and instead to record the opposite, these participants would be expected to indicate the (normatively inaccurate) impression that Group B was smarter. In addition, the task allowed us to examine whether the prophylactic effects of training generalized to situations in which the third variable acted, not as a confound, but as a "suppressor"—masking somewhat the magnitude of the relation between group and intelligence (but not altering the basic judgment about which of the two groups is more intelligent).

Results and Discussion

The results did not support the "reversal rule" alternative explanation. In both control and training conditions, participants rated Group A members as more intelligent and better at solving anagrams (*M*s on all rating measures in both conditions were significantly greater than zero, *p*s < .05).

Nor did the results offer any evidence that the effects of training generalized to tasks involving "suppressor" variables. Group A was not favored more strongly in the training condition than in the control condition; in fact, there was a trend in the reverse direction. Mean ratings of relative group intelligence were nonsignificantly lower in the training condition (*M*s were 1.33 on both intelligence measures) than in the control condition (*M*s were 1.85 and 1.80). Mean ratings of relative anagram ability were significantly lower (though still significantly positive) in the training condition (1.17) than in the control condition (2.05), $F(1, 42) = 5.74$, $p < .03$. There were no differences between the two conditions on any of the measures of reasoning.

What do these results mean? Of foremost importance to us, they provide no evidence to suggest that the training effects found in Studies 1 and 2 were the result of superficial heuristic responding. Instead, these results indicate that participants who received ANCOVA training in Studies 1 and 2 made more accurate group inferences because of some substantive learning during the training session. Just what did they learn? The results of Study 3 provide some interesting clues that are consistent with earlier speculation about the multiple possible effects of training.

Participants who received training, like those in the control condition, perceived Group A to be more intelligent than Group B. But their ratings were somewhat more moderate. This result is consistent with the speculation that one consequence of training is to enhance inferential caution. (Also consistent is a nonsignificant mean trend on confidence ratings: Participants in the training condition rated themselves as somewhat less confident in their intelligence judgments.) In this particular case, the effect on caution was apparently more powerful than any effects on statistical reasoning concerning the third variable. Why might this be? We suggest that the reason may be that caution-induced vigilance did not turn up any confound because there was no confound to detect; the third variable merely suppressed the magnitude of the relation between group and intelligence but did not change the direction of the relation. Therefore, in the absence of any confound, participants in Study 3 did not appear to apply the formal complex reasoning that might have led them to adjust their judgments in a more extreme direction.

GENERAL DISCUSSION

In two studies, participants who received ANCOVA training later made more accurate judgments about the relative intelligence of two novel groups. Together, these two studies support the hypothesis that training in ANCOVA reasoning can proactively inhibit the formation of erroneous group stereotypes. A third study addressed a possible alternative explanation and provided clues to how ANCOVA training influenced group impression formation. We now consider in more detail the possible processes through which this effect occurred.

Through What Process(es) Did Training Exert Its Effects?

One question that we must consider is whether participants in the training conditions really learned to apply a more sophisticated inferential logic or merely responded in a way that *appeared* as though they had learned that logic. Results from Studies 2 and 3 indicated that the training effects cannot be accounted for by either (a) experimental demand or (b) heuristic responding that masquerades as complex reasoning. It appears that subjects who received ANCOVA training really did learn something and that they really did apply what they learned later when forming group impressions. So just what did they learn as a result of the training?

A consideration of results across all studies suggests that ANCOVA training may influence several steps in the logical inference process. At the very least, training seems to have created some initial caution or wariness about the accuracy of inferences based on a simplistic assessment of information. This is suggested by the fact that training led to (nonsignificantly) less judgmental confidence and slightly more moderate impressions in Study 3. But the effects of training are not limited to increased caution. There was no evidence that the training manipulation decreased judgmental confidence in Studies 1 and 2. Nor were responses in the training conditions of Studies 1 and 2 more moderate than in the control conditions. Cautiousness may simply be the first of several inferential steps affected by ANCOVA training.

Training is also likely to have led to enhanced vigilance and sensitivity to possible confounding third variables. Participants who received ANCOVA training were more likely to accord importance to anagram length (Study 1) and to mention differences in anagram length in their self-reports (Studies 1 and 2). It is interesting, though, that these effects were not replicated under conditions in which anagram length acted as a suppressor variable rather than a confound (Study 3). Our training materials apparently sensitized people to the possible presence of confounds that might qualitatively alter group judgments but did not make them wary of subtler covariates that merely influence judgmental extremity.

Several results from Studies 1 and 2 indicate that ANCOVA training affected a third step in the inference process as well—the attempt to take the confounded third variable into account in some complex computational manner. Self-reports revealed that participants in the training conditions were more likely to consider group success rates separately for anagrams of different length. In addition, only in training conditions were there significant correlations between group impressions and explicit mention that five- and seven-letter anagrams differed in difficulty. Apparently, training not only led participants to be more likely to recognize the confound as an inferential problem but provided them with some means of figuring out how to solve this problem.

These considerations imply that the potential value of ANCOVA training is not limited simply to its effects on statistical thinking per se. Just as a solid statistical education helps scientists to think like scientists in a multitude of ways (Lehman et al., 1988), ANCOVA training seems to influence a number of "prestatistical" steps in the process of drawing logic-based inferences about groups. These steps involve skepticism and informed vigilance—logical tools that trained scientists typically apply as well before subjecting their data to formal statistical analyses. Schaller (1994) speculated that to engage in a process of intuitive ANCOVA, one must shift from a "storyteller" to a "scientist" mode of thinking (cf. Epstein, Lipson, Holstein, & Huh, 1992; Zukier, 1986). Our current data suggest that this shift is unlikely to occur in the absence of (a) initial skepticism about the veracity of simplistic inferences and (b) recognition of the existence of a confounded third variable.

Must Training Be Obviously Relevant to Stereotyping to Exert Its Effects?

Just how specific to group stereotyping must the training be to inhibit the formation of erroneous group stereotypes? This is an important question when considering educational interventions and is germane to a current controversy over the underlying mechanisms in the transfer of statistical training (Fong & Nisbett, 1991; Ploger & Wilson, 1991; Reeves & Weisberg, 1993).

The controversy centers on two different explanations for the effects of instruction on everyday reasoning and problem solving. Some researchers adopt a *concretist* position, suggesting that training effects occur through a process of reminding and analogical transfer. Concrete or superficial aspects of a problem remind people of similar problems that they have solved before (e.g., during a training session). They then use their approach to the earlier problem(s) as a template to direct their approach to the new problem. Support for the concretist position can be seen in studies showing that superficial similarities between training problems and testing problems are associated with enhanced reasoning on the testing problems (e.g., Ross, 1987; Spencer & Weisberg, 1986).

Other researchers have adopted a *formalist* position, suggesting that the effects of statistical training result from the development and subsequent application of abstract rules for inference. To some extent, the development of these abstract rules is contingent on some preexisting intuitive feel for those rules within certain domains. Statistical training formalizes these intuitions, providing people with the ability to apply the logic in new domains. Prior research supporting this position has revealed that formal training in the statistical "law of large numbers" transfers across time and domain (e.g., Fong & Nisbett, 1991; Lehman et al., 1988) and that this transfer is not dependent on recall for superficial aspects of training problems (Fong & Nisbett, 1991).

Though not designed to distinguish crucially between these two positions, for several reasons Study 1 may have provided an especially stringent test of the formalist hypothesis. First, participants were led to believe that the two experimental sessions were two entirely unrelated experiments. Second, we disguised the "training" aspect of the first session behind a duplicitous façade so that participants would be unaware of the true purpose of the experiment—in fact, unaware that they were even being "trained" for any particular purpose. Third, superficial features of the group impression task were quite different from those of the training problems, both in content (in the tennis-only condition) and in form (both training conditions). Unlike the training problems, in which all relevant data were presented in summary form, the group impression task demanded that participants perceive and encode the relevant data over time.

Considering these methods, the results were consistent with the formalist hypothesis. Participants in the tennis-only condition made more accurate inferences and engaged in a higher level of reasoning than those in the control condition. Thus there was a positive effect of training even though there were only minimal superficial similarities between the training problems and the Group A/B inference task. This is not only striking because of its conceptual implications but also heartening because of its educational implications: It may not be absolutely necessary to preach and pontificate about stereotypes, prejudice, and bigotry to thwart the development of stereotypes, prejudice, and bigotry.

It is worth noting, however, that not all abstract rule training is likely to have a positive impact. Fong and Nisbett (1991) suggest that statistical training may lead to the application of abstract rules in new domains only if people already have some intuitive comprehension for the logic. We would argue that this is the case with ANCOVA logic. We suspect that our "workshop" on ANCOVA successfully inhibited stereotype formation because we began with a very simple and intuitively appealing problem and then gradually introduced the formal logic through increasingly complex problems. In general, instruction that draws on intuitions that have been culled from experience, elaborates on these intuitions, and translates them into more abstract and generalizable rules is likely to have the greatest long-term impact (cf. Bruner, 1960).

What Are the Limitations and Implications of This Research?

Over the last two decades, social psychologists adopting a cognitive perspective have contributed tremendously to our understanding of what stereotypes are, where they come from, and why they are so difficult to get rid of (Hamilton & Sherman, 1994; Stangor & Lange, 1993; Stephan, 1985). This knowledge is not merely descriptive, but prescriptive as well: Through an understanding of the processes through which stereotypes form and endure, we may develop means with which to do battle against stereotypes and their deleterious consequences. We believe that the present research constitutes an encouraging step toward developing a proactive campaign to buttress this battle.

Nevertheless, as with any single step, there are some obvious restrictions and limitations that should be noted. For instance, ANCOVA training is not likely to influence all aspects of stereotype development. People who engage in intuitive ANCOVA may still form the overgeneralized group impressions that are the hallmark of stereotypes. Although ANCOVA training may influence the accuracy of these general group impressions, it is not likely to influence the variability that people perceive within groups. Nor is this sort of training likely to influence the development of all stereotypes. Although this reasoning process may play some role in the development of stereotypes about groups whose outcomes are systematically influenced by factors outside their control (circumstances that describe a great number of groups), many other processes irrelevant to logical reasoning also contribute to stereotype formation. ANCOVA training is best viewed as but a single leg of a necessarily multipedal platform targeting erroneous stereotypes.

Just as we must keep in mind the scope of this general line of inquiry, it is also necessary to consider some of the limitations of this first empirical investigation into the effects of statistical training on stereotype formation. First, the elapsed time between training and testing sessions was just more than a week. Although this lag is greater than that in many studies on statistical training, it is short in view of the broader temporal canvas of social intervention. Additional research is necessary to assess the truly long-term consequences of statistical training for stereotype formation.

Second, consider the context in which we assessed stereotype formation. Like much contemporary research on stereotype formation, our group impression task presented participants with a cleaner and more compressed inference situation than they might typically encounter. Although many real-life inference situations do fit a similarly compressed time frame (e.g., the undergraduate admissions officer who reads through college applications providing information on ethnic identity, standardized test performance, and socioeconomic status), many do not. In addition, despite our apparently successful attempt to minimize the links between the training and testing sessions, some similarities remained and may have served as cues to help participants retrieve and apply the knowledge they gained through training. Just how well might ANCOVA training transfer to impressions formed outside the psychological laboratory? That question remains unanswered.

Third, our participants were all college students, whereas many actual stereotypes are formed at a much earlier age. Is it realistic to suppose that one can teach the logic of ANCOVA to people who are still vulnerable to the formation of new stereotypes and prejudices? Several lines of thought suggest that it is. First, it is clear that infants have some intuitive grasp of correlational concepts (Younger, 1990). Second, even young children are attentive to situational constraints on behavior and can take these constraints into account when making causal judgments (Lepper, Greene, & Nisbett, 1973; Newman & Ruble, 1992). Third, it would be a mistake to suggest that college students are no longer susceptible to the formation of new stereotypes. Research on the stability of attitudes (e.g., Krosnick & Alwin, 1989) reveals that attitudes are still quite malleable among people in their late teens and early twenties. Thus, even young adults are vulnerable to the formation of new stereotypes. Nevertheless, research on younger populations would be necessary before drawing general conclusions concerning the impact of ANCOVA training on stereotype formation.

These and other limitations remind us that it is premature to speculate wildly about interventions that might prevent stereotype formation. Clearly, further research is necessary. But it is not inappropriate to be encouraged by these initial results. If a 40-min training session prevents the formation of erroneous group stereotypes a week and a half later, it is certainly possible that more sustained educational interventions may have similarly prophylactic effects that last far longer.

NOTES

[1] Any reader inclined to speculate on the basis of the mean (differences between the two training conditions should recognize that it is impossible to determine whether these nonsignificant differences resulted from the addition of a fourth training scenario specifically relevant to stereotypes or simply from the addition of a fourth training scenario.

[2] It is worth pointing out that, across all participants, these yes/no ratings were positively related to participants' judgments about the groups' relative anagram abilities and intelligence (.35 < *r*s < .58). Participants were more likely to make accurate group inferences if they (a) noted that five- and seven-letter anagrams differed in difficulty, (b) noted that Group A attempted anagrams that were more difficult than those attempted by Group B, and (c) attempted to calculate group success ratios separately for the two types of anagrams. These correlations were consistently positive within each of the three training conditions as well, with one notable exception: Within the control condition only, participants indicating that five- and seven-letter anagrams differed in difficulty were *not* more likely to judge Group A to be more intelligent (–.17 < *r*s < –.10). This lack of correlation stands in contrast to the strong positive correlations found in the two training conditions (.31 < *r*s < .63).

[3] Ancillary analyses examined the correlations between individual differences measured during Session 1 and stereotype formation assessed during Session 2. Only two results emerged worth noting. Consistent with previous research (Schaller, Boyd, Yohannes, & O'Brien, 1995), personal need for structure (PNS) was negatively related to measures of group impression formation, although these correlations were weak (–.25 < *r*s < –.13). Self-reported grade point average was weakly positively related to measures of group impression formation (.15 < *r*s < .25).

[4] These procedures not only created a plausible basis for categorization into groups but also provided a plausible reason that the sign-up procedure for Session 2 occurred at the end of Session 1. By providing this plausible but deceptive reason for the tenuous linkage between experimental sessions, we expected to decrease further the likelihood that participants might develop their own correct theories about that linkage and thus respond to demand characteristics.

REFERENCES

Agnoli, F. (1991). Development of judgmental heuristics and logical reasoning: Training counteracts the representativeness heuristic. *Cognitive Development, 6,* 195-217.

Allport, G. W. (1954). *The nature of prejudice.* Reading, MA: Addison-Wesley.

Aronson, E., Stephan, C., Sikes, J., Blaney, N., & Snapp, M. (1978). *The jigsaw classroom.* Beverly Hills, CA: Sage.

Bruner, J. S. (1960). *The process of education.* Cambridge, MA: Harvard University Press.

Cook, S. W. (1985). Experimenting on social issues: The case of school desegregation. *American Psychologist, 40,* 452-466.

Crandall, C. S., & Greenfield, B. S. (1986). Understanding the conjunction fallacy: A conjunction of effects? *Social Cognition, 4,* 408-419.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology, 56,* 5-18.

Eagly, A. H. (1987). *Sex differences in social behavior: A social-role interpretation.* Hillsdale, NJ: Lawrence Erlbaum.

Epstein, S., Lipson, A., Holstein, C., & Huh, E. (1992). Irrational reactions to negative outcomes: Evidence for two conceptual systems. *Journal of Personality and Social Psychology, 62,* 328-339.

Fairchild, H. H. (1984). School size, per-pupil expenditures, and academic achievement. *Review of Public Data Use, 12,* 221-229.

Fairchild, H. H. (1991). Scientific racism: The cloak of objectivity. *Journal of Social Issues, 47*(3), 101-115.

Fletcher, G.J.O., Danilovacs, P., Fernandez, G., Peterson, D., & Reeder, G. D. (1986). Attributional complexity: An individual differences measure. *Journal of Personality and Social Psychology, 51,* 875-884.

Fong, G. T., & Nisbett, R. E. (1991). Immediate and delayed transfer of training effects in statistical reasoning. *Journal of Experimental Psychology: General, 120,* 34-45.

Gigerenzer, G., & Murray, D. J. (1987). *Cognition as intuitive statistics.* Hillsdale, NJ: Lawrence Erlbaum.

Hamilton, D. L., & Gifford, R. K. (1976). Illusory correlation in interpersonal perception: A cognitive basis for stereotypic judgments. *Journal of Experimental Social Psychology, 12,* 392-407.

Hamilton, D. L., & Sherman, J. W. (1994). Stereotypes. In R. S. Wyer, Jr., & T. K. Srull (Eds.), *Handbook of social cognition* (2nd. Ed., Vol. 2, pp. 1-68). Hillsdale, NJ: Lawrence Erlbaum.

Hewstone, M., & Brown, R. J. (1986). *Contact and conflict in intergroup encounters.* Oxford, UK: Basil Blackwell.

Katz, P. A., & Taylor, D. A. (1988). *Eliminating racism: Profiles in controversy.* New York: Plenum.

Krosnick, J. A., & Alwin, D. F. (1989). Aging and susceptibility to attitude change. *Journal of Personality and Social Psychology, 57,* 416-425.

Lehman, D. R., Lempert, R. O., & Nisbett, R. E. (1988). The effects of graduate training on reasoning: Formal discipline and thinking about everyday-life events. *American Psychologist, 43,* 431-442.

Lepper, M. R., Greene, D., & Nisbett, R. E. (1973). Undermining children's intrinsic interest with extrinsic rewards: A test of the "overjustification" hypothesis. *Journal of Personality and Social Psychology, 28,* 129-137.

McConahay, J. B. (1986). Modern racism, ambivalence, and the Modern Racism Scale. In J. F. Dovidio & S. L. Gaertner (Eds.), *Prejudice, discrimination, and racism* (pp. 91-125). Orlando, FL: Academic Press.

Miller, N., & Brewer, M. B. (1984). *Groups in contact: The psychology of desegregation.* New York: Academic Press.

Mischel, W. (1973). Toward a cognitive social learning reconceptualization of personality. *Psychological Review, 80,* 252-283.

Neuberg, S. L., & Newsom, J. T. (1993). Personal need for structure: Individual differences in the desire for simple structure. *Journal of Personality and Social Psychology, 65,* 113-131.

Newman, L. S., & Ruble, D. N. (1992). Do young children use the discounting principle? *Journal of Experimental Social Psychology, 28,* 572-593.

Nisbett, R. E., Fong, G. T., Lehman, D. R., & Cheng, P. W. (1987). Teaching reasoning. *Science, 238,* 625-631.

Nisbett, R. E., Krantz, D. H., Jepson, C., & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive thinking. *Psychological Review, 90,* 339-363.

Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist, 17,* 776-783.

Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin, 68,* 29-46.

Ploger, D., & Wilson, M. (1991). Statistical reasoning: What is the role of inferential rule training? Comment on Fong and Nisbett. *Journal of Experimental Psychology: General, 120,* 213-214.

Reeves, L. M., & Weisberg, R. W. (1993). Abstract versus concrete information as the basis for transfer of problem solving. Comment on Fong and Nisbett (1991). *Journal of Experimental Psychology: General, 122,* 125-128.

Ross, B. H. (1987). This is like that: The use of earlier problems and the separation of similarity effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 629-639.

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.

Santioso, R., Kunda, Z., & Fong, G. T. (1990). Motivated recruitment of autobiographical memories. *Journal of Personality and Social Psychology, 59,* 229-241.

Schaller, M. (1992a). Sample size, aggregation, and statistical reasoning in social inference. *Journal of Experimental Social Psychology, 28,* 65-85.

Schaller, M. (1992b). Ingroup favoritism and statistical reasoning in social inference: Implications for formation and maintenance of group stereotypes. *Journal of Personality and Social Psychology, 63,* 61-74.

Schaller, M. `(1994). The role of statistical reasoning in the formation, preservation, and prevention of group stereotypes. *British Journal of Social Psychology, 33,* 47-61.

Schaller, M., Boyd, C., Yohannes, J., & O'Brien, M. (1995). The prejudiced personality revisited: Personal need for structure and the formation of erroneous group stereotypes. *Journal of Personality and Social Psychology, 68,* 544-555.

Schaller, M., & O'Brien, M. (1992). "Intuitive analysis of covariance" and group stereotype formation. *Personality and Social Psychology Bulletin, 18,* 776-785.

Sigall, H., & Page, R. (1971). Current stereotypes: A little fading, a little faking. *Journal of Personality and Social Psychology, 18,* 247-255.

Spencer, R. M., & Weisberg, R W. (1986). Context-dependent effects on analogical transfer during problem solving. *Memory & Cognition, 14,* 442-449.

Stangor, C., & Lange, J. (1993). Mental representations of social groups: Advances in understanding stereotypes and stereotyping. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 26, pp. 357-416). San Diego, CA: Academic Press.

Stephan, W. G. (1985). Intergroup relations. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology* (3rd ed., Vol. 2, pp. 599-658). New York: Random House.

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

Younger, B. (1990). Infants' detection of correlations among feature categories. *Child Development, 61,* 614-620.

Zukier, H. (1986). The paradigmatic and narrative modes in goal-guided inference. In R. M. Sorrentino & E. T. Higgins (Eds.), *Handbook of motivation and cognition: Foundations of social behavior* (Vol. 1, pp. 465-502). New York: Guilford.

Received November 22, 1994
Revision accepted August 13, 1995