GENDER-BASED EXPECTANCIES AND OBSERVER JUDGMENTS OF SMILING

Nancy J. Briton and Judith A. Hall

ABSTRACT: Beliefs about gender differences in smiling were measured by asking college students to rate how much they believed hypothetical women and men smile. Women were believed to smile more than men. Individual differences in this belief did not affect subsequent scoring of smiles, whether scored by counting the number of smiles exhibited by videotaped male and female targets or by rating the amount of smiling exhibited. An expectation about gender differences in smiling was experimentally induced, either that women smile more than men or that there is no gender difference in smiling. This expectation did not affect subsequent scoring of smiles, regardless of scoring method and regardless of whether the expectation was induced as a casual aside or in more formal instructions. In all conditions female targets were observed to smile more than male targets. Rating produced larger target gender effects than counting, but this could have been due to the nature of the rating process rather than observer bias.

The study of behavioral gender differences is a topic of widespread interest. Gender differences are found for a wide range of behaviors, including nonverbal behaviors such as smiling. Numerous studies have found that women smile more than men (Hall, 1984). In fact, smiling shows more consistent and larger gender differences than do most social psychological variables (Eagly, 1987; Hall, 1984). Additionally, women are *believed* to smile more than men (Briton & Hall, in press). Smiling, like most aspects of social behavior, is not normally measured by any precise physiological means, but by observers, who record the amount of smiling they perceive as they view taped or live nonverbal communication. Observers, unfortunately, have been found to be susceptible to the effects of bias. One form of observer bias occurs when observers view behavior

The authors gratefully acknowledge the assistance of Thomas Leahy and Joe Pieri in data collection, and of Cliff Brown and an anonymous reviewer for their comments on an earlier version of this paper. Miles Patterson served as Action Editor for this article.

Inquiries and requests for reprints should be addressed to Judith A. Hall, Psychology Department, Northeastern University, 125 NI, 360 Huntington Avenue, Boston MA, 02115.

Journal of Nonverbal Behavior 19(1), Spring 1995

about which they have an expectation, then rate the behavior in line with that expectation.

A behavior or trait that has a corresponding gender-role stereotype may create an expectation in observers that affects their scoring of that behavior or trait. One such trait is aggressiveness, a stereotypically masculine characteristic. Observer bias for rating aggressiveness was found by Lyons and Serbin (1986), who asked observers to rate the aggressiveness of female and male children represented by simple line drawings. Observers rated drawings of children they believed were boys as acting more aggressively than they rated identical drawings of children they believed were girls. The authors suggested that because observers expected to see more aggression in boys, their scores were biased in the direction of that expectation.

Gender-based beliefs exist not only for traits such as aggressiveness, but also for nonverbal behaviors and skills. Women are believed to be more facially and vocally expressive than men (Briton & Hall, in press; Zuckerman & Larrance, 1979), better at recognizing faces and reading others' nonverbal cues (Briton & Hall, in press), and more affectionate, charming, and kind than men (Archer & Lloyd, 1985). With regard to specific nonverbal behaviors, women are believed to smile and laugh more than men, while men are believed to speak more loudly than women, interrupt others more, use more speech disfluencies such as "um" and "ah," and have more restless feet and legs than women (Briton & Hall, in press; Kramer, 1977).

In behavioral research where nonverbal behaviors of both women and men are quantified, many of these gender-specific beliefs are confirmed by observation. For example, women are *believed* to smile more than men (Briton & Hall, in press; Kramer, 1977) and women are also *observed* to smile more than men (Hall, 1984). Men are believed to talk louder and interrupt others more than women (Briton & Hall, in press; Kramer, 1977), and men are also observed to talk louder and interrupt others more than women (Hall, 1984).

No published study could be found that examined observer bias for the scoring of gender-stereotyped nonverbal behaviors, although expectancy-based observer errors unrelated to gender have been found for the scoring of gaze. Three studies (Martin & Rovira, 1981; Stephenson & Rutter, 1970; White, Hegarty, & Beasley, 1970) found that observers tended to make errors in line with their expectations when scoring eye contact in dyads, although the validity of these laboratory studies has been challenged (Argyle, 1970; Patterson, 1975). Nieman, Roberts, and Kantner (1983) similarly found expectancy-based observer bias in judgments of eye movements from a single target.

Different methods of scoring have been used in observational studies. Two of the most common are counting and rating. When counting is used, observers simply count how many times a behavior occurs within some time period (for example, Condon & Sander, 1974; Jacklin, Maccoby, & Dick, 1973; Korner, 1973). Rating is also used often, and in these studies observers rate the overall level or the amount of a behavior on a scale (for example, Binning, Zaba, & Whattam, 1986; Hechtman & Rosenthal, 1991; Jennings, 1977; Rubin, Provenzano, & Luria, 1974). Some studies use both methods of quantifying behavior (for example, Chaikin & Derlega, 1978; Fazio, Effrein, & Falender, 1981; Lyons & Serbin, 1986; Shuller & McNamara, 1976).

Rating has been found to be a more ambiguous method of quantifying behavior than counting (Brightman & Raymond, 1975; Martell & Guzzo, 1991), and has been considered more capable of reflecting judges' biases (Ritter & Langlois, 1988).

Because observer bias has been documented for gaze, and because gender-stereotyped expectancies have been found to influence scoring of gender-stereotyped traits, it may be that observer bias has also occurred in past research on gender differences in other nonverbal behaviors, for instance smiling. To test this possibility, we measured observers' expectations about gender differences in smiling, and related these expectations to the same observers' subsequent scoring of videotaped target smiles. Expectation-based observer bias would be demonstrated if the pre-existing belief scores were related to observers' scoring of smiles. In other words, observers who expected to see a large gender difference in smiling would see a larger gender difference in smiling than observers who did not expect to see such a difference.

In the current research, we also experimentally induced an expectation, first weakly and then more strongly, about gender differences in smiling, to determine if this induced expectancy would affect the scoring of smiles. It was predicted that both the pre-existing belief about gender differences in smiling and the induced expectation about gender differences in smiling would produce biased ratings from our observers, in the direction of their expectations. Additionally, the more ambiguous scoring method of rating was predicted to show greater effects of these expectancies than the more objective task of counting.

Four studies tested these predictions. Study 1 provided manipulation checks and evidence for the reliability of our measurement of participants' beliefs about female and male smiling. Study 2 provided a baseline measure of the observed levels of female and male smiling on the target videotapes. Studies 3 and 4 measured pre-existing beliefs about gender differences in smiling, and attempted two types of expectancy inductions, one

casual and one more formal. Studies 2 and 3 used both counting and rating scoring methods, while Study 4 used only rating.

Study 1

Method

The purpose of Study 1 was to demonstrate the efficacy of an experimental induction of expectancies about gender differences in smiling, and to establish the reliability of measured beliefs about smiling in women and men.

Participants

Participants (N = 30; 15 women, 15 men) were students attending Northeastern University, a large private university in Boston, MA. Participants received partial fulfillment of psychology course requirements in exchange for participation.

Procedure

Participants reported individually to the laboratory. There, they were told that they would watch a videotape of doctors interacting with their patients and rate the doctors on some characteristic. Before watching the videotape, they were given a short questionnaire asking for their names and their opinions about gender differences for a variety of nonverbal behaviors, one of which was smiling. They were asked to consider people in general, and to rate how much each gender does each behavior by indicating a number from *never* (1) to *always* (10) for each gender, and for each behavior.

After filling out the questionnaire, participants were randomly assigned to one of three conditions, with the following expectancies: a) women smile more than men; b) men smile more than women; or c) no gender difference for smiling. The expectancy induction was imbedded in the following instructions that were read to each participant:

> In this study, we are looking at smiling in relation to various physician characteristics including age, sex, experience, and status. For most of these characteristics, not much is known, but we do know that, in general, women smile considerably more than men [or men smile considerably more than women, or there is no gender difference in smil-

> ing], although we do not know if this is the case for doctors in particular.

After delivering the expectancy induction, the experimenter looked at her notes, apologized, and told each participant:

> Oh, no, I made a mistake. I gave you the wrong questionnaire. You were supposed to get the new one. You need to do *this* one (gets a different questionnaire out of a file cabinet and gives it to the participant). This one is anonymous. Just put it in the envelope on the shelf there when you're finished. I'll throw this one away.

The experimenter then tossed the first questionnaire in the trash, to be later retrieved and compared with the second questionnaire. After completing the second questionnaire, participants placed it in an envelope containing several other (dummy) questionnaires and informed the experimenter they were finished. The experimenter then told participants that the experiment was complete, and debriefed each participant. It was explained in the debriefing that the experiment had been performed to determine whether their beliefs about gender differences in smiling could be altered by a simple suggestion on the part of the experimenter. All of the participants stated that they did not recognize the experimenter's comments as an experimental manipulation.

Results

Reliability of Smiling Beliefs

Test-retest reliability of the measurement of smiling beliefs was calculated by correlating the item "smiles at others" between the first (pre-induction) questionnaire and the second (post-induction) questionnaire, separately for male targets and female targets, and separately for the three expectancy conditions. Analyses were performed separately for the three expectancy conditions because a pooled analysis would confound changes due to the manipulations with lack of reliability (i.e., random errors). Table 1 presents these reliability coefficients.

As Table 1 shows, reliability was adequate for the women smile more and the no gender difference conditions. When given the expectation that men smile more than women, reliability was not adequate. The expectation that men smile more than women was not used in further experimental manipulations.

TABLE 1

Test-Retest Reliability of Smiling Belief Ratings

	Reliability	
Expectancy condition	Male targets	Female targets
Women smile more	.87	.82
No gender difference	.92	.96
Men smile more	.32	.62

Note. Reliability coefficients are correlations between pre-manipulation smiling beliefs and post-manipulation smiling beliefs.

Manipulation Check

A 2 (participant gender) × 3 (expectancy condition: women smile more than men, men smile more than women, no difference) × 2 (target gender) analysis of variance (ANOVA) was conducted on the post-manipulation item "smiles at others." Participant gender and expectancy condition were between-subjects factors, and target gender was a within-subjects (repeated measures) factor. The interaction of expectancy condition and target gender was significant, $F(2, 24) = 18.48$, $p < .0001$ (Table 2).

The pattern shows that participants' reported beliefs about smiling were affected by the information presented to them in the expectancy induction. Those told that women smile more than men reported a belief that

TABLE 2

Manipulation Check, Study 1

	Mean ratings of smiling (beliefs)		
Expectancy condition	Female targets	Male targets	Difference
Women smile more	7.50	4.00	3.50**
No gender difference	8.70	6.10	2.60*
Men smile more	6.60	7.40	−.80*

Note. Hypothetical targets were rated from *never* (1) to *always* (10).
*$p < .05$ **$p < .001$

women smile more, and those told that men smile more than women reported the reverse, with the "no gender difference" group's beliefs falling in between. Note that participants in the "no gender difference" condition still believed that women smile more than men, but less so than participants in the "women smile more" condition, and more so than in the "men smile more" condition.

Study 2

Method

The purpose of Study 2 was to determine a baseline level of target smiling on the stimulus tapes, to serve as a comparison for the remaining two studies.

Participants

Participants (N = 138; 60 women, 78 men) were students recruited as in Study 1. They participated either singly or in groups of two. When in groups of two, participants were separated by a screen so that they could not see one another while they scored smiles.

Stimuli

The stimulus tape of smiles was created using videotapes of physicians talking to patients in routine office visits at a large Boston-area hospital. The physicians had been recruited for an unrelated study of doctor-patient interaction. Informed consent had been obtained from physicians and patients, and all knew they were being taped. Neither physicians nor patients were aware of the nature of the current research or that the tape would be scored for smiling.

Segments for each of 50 target physicians were one minute in length. Segments were separated by 10-second intervals, and segments were randomly ordered within the tape. The tape was balanced with 25 male and 25 female target physicians. Participants were randomly assigned to view one half of the tape (Half 1 or Half 2). Each half contained 25 one-minute segments of communication, with 12 female and 13 male targets on the first half, and 13 female and 12 male targets on the second half. No participant saw the same target physician more than once; therefore, each par-

ticipant viewed 25 segments, either 12 female and 13 male targets or 13 female and 12 male targets. The videotape was presented without audio. Only the physicians were visible on the screen.

Both stimuli and participants were randomly assigned to either Half 1 or Half 2, and statistical analysis showed no differences between Half 1 and Half 2 for the effects of pre-existing belief or induced expectation. Data were subsequently pooled over both halves, and Half will not be discussed further as a variable in statistical analyses.

Procedure

The participants in Study 2 simply watched the 25 target clips and scored the amount of smiling they observed for each target. In order that participants were not sensitized to the nature of the study by stating their beliefs about gender differences in smiling, no pre-existing belief was measured and no expectation was induced.

Scoring Methods

Participants were randomly assigned either to act as counters (n = 56; 33 men, 23 women) or raters (n = 82; 45 men, 37 women).[1] Counters were provided with a scoring sheet for each segment and placed a tic mark on the sheet each time they observed a smile. Raters watched each entire clip, then rated the amount of smiling that had occurred for each clip using a scale of *did not smile* (1) to *always smiled* (10).

Results

Data were not normally distributed and were therefore normalized by a log transformation. A 2 (gender of participant) × 2 (gender of target) analysis of variance, with repeated measures on the second factor, was conducted for each scoring method. These analyses showed that female targets were observed to smile more than male targets for both scoring methods: for counting, $F(1, 54) = 27.42$, $p < .0001$, $r = .58$; for rating $F(1, 80) = 315.36$, $p < .00001$, $r = .89$. Effect sizes, expressed as r, were obtained by converting from F using the formula $\sqrt{F/(F + df \text{ error})}$ (Rosenthal & Rosnow, 1991). There were no significant differences found for participant gender, nor for the interaction of participant gender with target gender. Table 3, top row, shows the target gender results. For display purposes the raw (not log transformed) data are shown.

TABLE 3

Observed Smiling in Male and Female Targets by Different Scoring Methods, Studies 2-4

	Counting[a]			Rating[b]		
Study	Male targets	Female targets	*r*	Male targets	Female targets	*r*
Study 2	2.00	2.61	.58	2.49	4.14	.89
Study 3	1.32	1.94	.62	2.67	3.98	.88
Study 4	[c]	[c]	[c]	2.50	4.20	.92

Note. The column labeled "*r*" shows the effect size associated with the target gender effect.

[a]Mean number of smiles counted for each 60-second segment.

[b]Possible range was 1 to 10.

[c]Counting was not used in Study 4.

Differences between scoring methods were analyzed by comparing target gender effect sizes (*r*) for both scoring methods. The difference between the effect sizes for rating and counting smiles was computed by a *Z*-test (Rosenthal & Rosnow, 1991). The target gender effect size was significantly larger for the participants who rated ($r = .89$) than for those who counted ($r = .58$), $Z = 4.28$, $p < .00001$. In other words, when participants rated smiles, their ratings showed a larger target gender difference than when participants counted smiles.

Studies 3 and 4

Method

The purpose of studies 3 and 4 was to combine both sources of possible observer bias, pre-existing beliefs and induced expectations about gender differences in smiling, with a scoring task.

Participants and Procedure

Two hundred forty-three students acted as observers in Studies 3 ($N = 130$; 64 women, 66 men) and 4 ($N = 113$; 52 women, 61 men). They

were recruited as in Studies 1 and 2. Procedures for Studies 3 and 4 were identical to those for Study 2, with the following additions.

Pre-existing beliefs. Prior to viewing the tape, pre-existing beliefs about gender differences in smiling were assessed using the same questionnaire that was used as the pre-expectancy measure in Study 1. Participants in Study 3 were asked their beliefs immediately prior to scoring the tape for smiles. Participants in Study 4 were asked their beliefs several weeks before they scored the tape.

Induction of expectations. Participants were randomly assigned to one of two conditions, with the following expectancies: a) women smile more than men; b) no gender difference.

The expectation was induced in two ways, casually and formally. In Study 3, the expectation was phrased as a casual conversational aside from the experimenter to each participant. The wording of the belief induction depended on the participants' pre-existing belief about gender differences in smiling, as assessed by the questionnaire item described above. For Induction *a* (women smile more than men), the experimenter looked at each participant's just completed questionnaire and casually said to that participant, "I see you think that women smile more than men. That's true. Research shows that women *do* smile more than men"; or "I see you think that there is no gender difference in smiling [or that men smile more than women]. That's actually not true. Research shows that women smile more than men."

For Induction *b* (no gender difference), the experimenter said, "I see you think that there is no gender difference in smiling [or that men smile more than women]. That's true. Research shows that there is no gender difference in smiling"; or "I see you think that women smile more than men. That's actually not true. Research shows that there is no gender difference in smiling."

In Study 4 the expectation was induced more strongly, as part of the formal instructions to each participant. Expectancy remarks in Study 4 were included in the following formal instructions, read by the experimenter to each participant:

> In this study, we are looking at smiling in relation to various physician characteristics including age, sex, experience, and status. For most of these characteristics, not much is known, but we do know that, in general, women smile considerably more than men [or, there is no difference in smiling between women and men], although we don't know if this is the case for doctors in particular.

Manipulation checks. As reported in Study 1, a manipulation check of the induction used in Study 4 was conducted as a separate experiment. As another check, participants in Studies 3 and 4 were asked if they recalled the experimenter telling them anything about gender differences in smiling; and if so, what. Ninety-four percent correctly recalled the information provided.

In order to determine if participants detected the goal of the research and would be at risk of responding to demand characteristics, after scoring the tape, participants were asked to write on a blank piece of paper what they believed to be the purpose of the study. Virtually all participants stated that the experiment was concerned with doctor-patient communication. One participant correctly stated that the study was concerned with expectancy effects, and her data were excluded from the study.

Scoring methods. In Study 3, participants were randomly assigned to either the counting or rating conditions described above. In Study 4, only rating was used.

Results

Overall Target Gender Effects

Table 3 presents the mean (not log-transformed) gender differences in smiling for Studies 3 and 4, for comparison with Study 2. These effects were extremely similar across studies for both scoring methods, showing that target females were observed to smile more than target males.

Pre-Existing Beliefs

To assess participants' beliefs about gender differences in smiling, the pre-existing belief data from Studies 3 and 4 were pooled. A 2 (gender of participant) × 2 (gender of target) analysis of variance, with repeated measures on the second factor, was performed on the smiling belief data. A main effect of target gender revealed that participants believed hypothetical women ($M = 7.52$) smile more than hypothetical men ($M = 5.86$), $F(1, 439) = 174.64$, $p < .0001$, $r = .53$. A detailed analysis of the belief data is available elsewhere (Briton & Hall, in press).

Participants' reported beliefs were divided into 5 categories:

1) Reverse Stereotype: Men smile more than women (6% of participants fell into this category). 2) No Difference: Men and women smile the same amount (34%). 3) Low Stereotype: Women smile more than men by

one or two points (22%). 4) Moderate Stereotype: Women smile more than men by three or four points (28%). 5) High Stereotype: Women smile more than men by five or more points (10%).

Data in Studies 3 and 4 were subjected to a log transformation which normalized the distribution of data. A 2 (gender of participant) × 5 (belief category) × 2 (gender of target) analysis of variance, with repeated measures on the last factor, was performed for the counting data (Study 3 only) and the rating data (pooled over Studies 3 and 4). Again, for both scoring methods, female targets were observed to smile more than male targets. For counting (Study 3), $F(1, 76) = 26.26$, $p < .0001$, $r = .51$; for rating (Studies 3 and 4), $F(1, 127) = 350.48$, $p < .00001$, $r = .86$.

The test of the biasing effect of pre-existing beliefs was the interaction of belief (five categories) with target gender. This effect was nonsignificant and very small: for counting (Study 3), $F(4, 76) = 0.05$; for rating (Studies 3 and 4), $F(4, 127) = 0.55$. All other effects were also nonsignificant for both counting and rating. See Table 4 for the means (not log-transformed) from this analysis.

TABLE 4

Observed Smiling in Female and Male Targets as a Function of Pre-existing Belief, Studies 3 and 4 Pooled

	Pre-existing belief category[a]				
Target gender	1	2	3	4	5
	Counting (n = 84)				
Female	[b]	1.95	1.80	1.93	2.09
Male	[b]	1.32	1.13	1.48	1.47
Difference		.63	.67	.45	.62
	Rating (n = 137)				
Female	4.06	3.99	3.97	4.02	3.91
Male	2.44	2.71	2.53	2.69	2.53
Difference	1.62	1.28	1.44	1.33	1.38

[a]Participants' belief categories: 1 = Men smile more than women, 2 = Men and women smile the same, 3 = Women smile more than men by 1-2 points, 4 = Women smile more than men by 3-4 points, 5 = Women smile more than men by 5 or more points.

[b]Random assignment to counting condition produced no participants with a belief in this category.

Induced Expectancy

A 2 (gender of participant) × 2 (induced expectancy) × 2 (gender of target) analysis of variance with repeated measures on the last factor showed no effects of the induced expectation for either scoring method or type of induction. The test of this effect was the interaction of induced expectancy and target gender. This effect was not significant: for counting, casual induction (Study 3), $F(1, 61) = 0.87$; for rating, casual induction (Study 3), $F(1, 61) = 2.86$, $p = .11$; for rating, formal induction (Study 4), $F(1, 109) = 0.00$. The only significant effect was again for gender of target. For counting, casual induction (Study 3), $F(1, 61) = 37.50$, $p < .0001$, $r = .62$; for rating, casual induction (Study 3), $F(1, 61) = 200.52$, $p < .00001$, $r = .88$; for rating, formal induction (Study 4), $F(1, 109) = 583.15$, $p < .00001$, $r = .92$. The effect size for rating using the casual induction ($r = .88$) did not differ significantly from the effect size for rating using the formal induction ($r = .92$), $Z = 1.26$, $p = .20$. See Table 5 for the untransformed means from the analysis of induced expectancies.

Crossing induced expectancy with pre-existing belief in analyses yielded no effects of this interaction regardless of scoring method or type of induction.

TABLE 5

Observed Smiling in Female and Male Targets as a Function of Induced Expectancy, Studies 3 and 4

	Induced expectancy	
Target gender	Women smile more	No difference
Study 3: Casual induction, counting ($n = 65$)		
Female	2.12	1.78
Male	1.42	1.24
Difference	.70	.54
Study 3: Casual induction, rating ($n = 65$)		
Female	4.03	3.92
Male	2.59	2.76
Difference	1.44	1.16
Study 4: Formal induction, rating ($n = 113$)		
Female	4.06	4.36
Male	2.38	2.63
Difference	1.68	1.73

Scoring Methods

In all cases the target gender effect sizes for rating were significantly larger than the corresponding target gender effect sizes for counting. In the analysis of effects due to pre-existing belief, the target gender effect size for rating ($r = .86$) was significantly larger than the effect size for counting ($r = .51$), $Z = 5.19$, $p < .00001$. In the analysis of effects due to induced expectancy, the target gender effect size for rating ($r = .88$) was also significantly larger than the counting effect size ($r = .62$), $Z = 3.62$, $p < .0005$.

General Discussion

Three studies showed large gender of target effects for both counting and rating of smiling. These observed differences were not related to observers' pre-existing belief nor to an induced expectation about gender differences in smiling. This finding supports the use of observers in nonverbal research, at least for the measurement of gender differences in smiling.

The size of the target gender effect is consistent with observational studies finding that women smile more than men. Our effect sizes were larger than the average point biserial correlation of .30 between gender and smiling reported for adults in Hall's (1984) meta-analysis, due to the repeated measures approach used in the present study. When frequency of smiling in the present study is analyzed as a between-targets effect, the result is in line with other published results, $r = .36$ (Hall, Irish, Roter, Ehrlich, & Miller, 1994).

The present research does not answer the question of rating versus counting accuracy, although there was a difference in target gender effect size between the two scoring tasks. As mentioned above, regardless of the scoring task, observers saw target women smiling more than men. When rating smiles, however, observers' scores showed a larger gender difference than when counting smiles. One possible interpretation of this effect suggests that counting may be the more accurate of the two scoring methods. In this view, counting may be considered a more straightforward task than rating, as indeed it has been considered in past research. In the present research, it may have been that the larger effect sizes found for rating were indeed the result of some observer bias that our pre-existing belief questionnaire did not measure. In this vein, one must find and accurately measure the biasing condition(s) that account for this effect. If this view were supported, it would suggest that counting is the more accurate measure of smiling.

A second possibility is that the task of counting may fail to capture all the dimensions on which smiling may be quantified. Counting only tabulates frequency; it does not reflect intensity or duration. When rating smiles, observers may use a combined evaluation of frequency, duration, and intensity of smiling. This interpretation would suggest the opposite from the preceding one, pointing to rating as being the more accurate measure.

A word needs to be said about the reliability of observers' beliefs about gender differences in smiling, assessed in Study 1. We call these beliefs "reliable" even though test-retest reliability was assessed only minutes after the first assessment. Ordinarily test-retest reliability is assessed after more time has passed. This short-term test-retest reliability is appropriate in this case, because it replicates the short time lapse between measurement of beliefs and scoring of smiles used in two of the subsequent studies in this series (Studies 3 and 4). In these studies, a similarly small time lapse occurred between the initial belief assessment and the scoring of the tape.

The reader may also note that the separate manipulation check for our expectancy induction, also in Study 1, was conducted only for the formal induction method. This decision was reached after unsuccessful attempts to construct a believable scenario incorporating both a casual induction and the ruse of giving participants the wrong questionnaire. It was determined that the combination of these two elements, both of which may be rightly considered unprofessional behavior, would be impossible to deliver without seriously challenging the ecological validity of the experiment. Aside from the acting challenge imposed on the experimenter, it was judged that participants would perceive the situation as implausible.

We were unable to induce an expectation in our observers that produced biased smiling scores, despite the fact that observers' beliefs about gender differences in smiling were affected by that expectation. Previous studies have shown effects of bias on observer scores. What was different about ours? Past studies that have shown effects of bias have tended to be studies that used rating scales or global descriptors of behavior. For example, children described as boys were observed to be more aggressive than the same children described as girls (Lyons & Serbin, 1986; Meyer & Sobieszek, 1972). However, participants in those studies did not say that the boys were observed to have performed more specific acts that constitute aggression. Bias may therefore be reduced if observers are provided a clear-cut description of a single discrete behavior to be judged. Instead of rating a child's aggressiveness, for example, observers could measure or count the discrete behaviors that constitute an operational definition of

aggression: e.g., interpersonal distance, frowns, specific verbalizations, or certain touching behaviors.

Our findings demonstrate remarkable resistance to bias for the scoring of smiling, in spite of differing pre-existing observer expectancies for female and male targets, and in spite of two interventions designed to influence expectancies. In addition, observers' gender did not affect the size of the target gender difference. This is highly encouraging for investigators who score smiling and who may have been concerned about bias based on stereotypes or gender of observer.

One possible limitation of our studies is that they were based on tapes of physicians speaking to patients. We did not sample broadly from different target groups. In choosing these stimuli, we believed that demand characteristics to conform with our expectancy inductions would be minimized due to participants' focus on the distinctive nature of the medical setting. Future studies may examine other target populations as well as other nonverbal behaviors for the possible presence of stereotype-based observer bias.

Notes

1. Other research using identical methods but different stimulus tapes demonstrated that the counting and rating methods have equivalent, excellent interobserver reliability (Payne, 1994).

References

Archer, J., & Lloyd, B. (1985). *Sex and gender*. New York: Cambridge University Press.

Argyle, M. (1970). Eye-contact and distance: A reply to Stephenson and Rutter. *British Journal of Psychology, 61*, 395-396.

Binning, J. F., Zaba, A. J., Whattam, J. C. (1986). Explaining the biasing effects of performance cues in terms of cognitive categorization. *Academy of Management Journal, 29*, 521-535.

Brightman, D., & Raymond, B. (1975). The effects of task ambiguity and expectancy control groups on the experimenter bias effect. *Journal of Social Psychology, 96*, 277-287.

Briton, N. J., & Hall, J. A. (in press). Beliefs about female and male nonverbal behavior. *Sex Roles*.

Chaikin, A. L., & Derlega, V. J. (1978). Nonverbal mediators of expectancy effects in black and white children. *Journal of Applied Social Psychology, 8*, 117-125.

Condon, W. S., & Sander, L. W. (1974). Synchrony demonstrated between movements of the neonate and adult speech. *Child Development, 45*, 456-462.

Eagly, A. H. (1987). *Sex differences in social behavior: a social role interpretation*. Hillsdale, New Jersey: L. Erlbaum Associates.

Fazio, R. H., Effrein, A. E., & Falender, V. J. (1981). Self-perceptions following social interaction. *Journal of Personality and Social Psychology, 41*, 232-242.

Hall, J. A. (1984). *Nonverbal sex differences: Communication accuracy and expressive style.* Baltimore: Johns Hopkins University Press.

Hall, J. A., Irish, J. T., Roter, D. L., Ehrlich, C. M., & Miller, L. H. (1994). Gender in medical encounters: An analysis of physician and patient communication in a primary care setting. *Health Psychology, 13,* 384-392.

Hechtman, S. B., & Rosenthal, R. (1991). Teacher gender and nonverbal behavior in the teaching of gender-stereotyped materials. *Journal of Applied Social Psychology, 21,* 446-459.

Jacklin, C. N., Maccoby, E. E., & Dick, A. E. (1973). Barrier behavior and toy preference: Sex differences (and their absence) in the year-old child. *Child Development, 44,* 196-200.

Jennings, K. D. (1977). People versus object orientation in preschool children: Do sex differences really occur? *Journal of Genetic Psychology, 131,* 65-73.

Korner, A. F. (1973). Sex differences in newborns with special reference to differences in the organization of oral behavior. *Journal of Child Psychology and Psychiatry, 14,* 19-29.

Kramer, C. (1977). Perceptions of male and female speech. *Language and Speech, 20,* 151-161.

Lyons, J. A., & Serbin, L. A. (1986). Observer bias in scoring boys' and girls' aggression. *Sex Roles, 14,* 301-313.

Martell, R. F., & Guzzo, R. A. (1991). The dynamics of implicit theories of group performance: When and how do they operate? *Organizational Behavior and Human Decision Processes, 50,* 51-74.

Martin, W. W., & Rovira, M. L. (1981). An experimental analysis of discriminability and bias in eye-gaze judgment. *Journal of Nonverbal Behavior, 5,* 155-163.

Meyer, J. W., & Sobieszek, B. I. (1972). Effect of a child's sex on adult interpretations of its behavior. *Developmental Psychology, 6,* 42-48.

Nieman, C. E., Roberts, W. T. & Kantner, J. E. (1983). Theoretical biases in judging eye movements. *Journal of Nonverbal Behavior, 7,* 179-182.

Patterson, M. L. (1975). Eye contact and distance: A re-examination of measurement problems. *Personality and Social Psychology Bulletin, 1,* 600-603.

Payne, R. (1994). *Gender, status, and smiling.* Unpublished master's thesis, Northeastern University, Boston.

Ritter, J. M., & Langlois, J. H. (1988). The role of physical attractiveness in the observation of adult-child interactions: Eye of the beholder or behavioral reality? *Developmental Psychology, 24,* 254-263.

Rosenthal, R. (1976). *Experimenter effects in behavioral research,* enlarged edition. New York: Irvington.

Rosenthal, R., & Rosnow, R. L. (1985). *Contrast analysis: Focused comparisons in the analysis of variance.* New York: Cambridge University Press.

Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis* (2nd ed.). New York: McGraw-Hill.

Rubin, J. Z., Provenzano, F. J., & Luria, Z. (1974). The eye of the beholder: Parents' views on sex of newborns. *American Journal of Orthopsychiatry, 44,* 512-519.

Shuller, D. Y., & McNamara, J. R. (1976). Expectancy factors in behavioral observation. *Behavior Therapy, 7,* 519-527.

Stephenson, G. M., & Rutter, D. R. (1970). Eye-contact, distance, and affiliation: A re-evaluation. *British Journal of Psychology, 61,* 385-393.

White, J. H., Hegarty, J. R., & Beasley, N. A. (1970). Eye contact and observer bias: A research note. *British Journal of Psychology, 61,* 271-273.

Zuckerman, M., & Larrance, D. T. (1979). Individual differences in perceived encoding and decoding abilities. In R. Rosenthal (Ed.), *Skill in nonverbal communication: Individual differences.* Cambridge, MA: Oelgeschlager, Gunn & Hain.