EXPERIMENTER EFFECTS IN BEHAVIORAL RESEARCH AND THEIR IMPLICATIONS FOR RESEARCH ON NONVERBAL COMMUNICATION

ROBERT ROSENTHAL

For twenty-three years now I've been studying the behavior of people who study behavior, working basically on the social psychology of the psychological experiment. My special interest in the psychological experiment has been in the operation of experimenter expectancy effects. That's the phenomenon whereby experimenters tend to get the results they expect to get, not because they're so terribly clever about anticipating what nature is really like, but simply because they do expect it.

I've also been very much interested in the operation of teacher expectancy effects—the phenomenon whereby teachers get the results they expect from their pupils simply because they expect those performances from their pupils. My colleague, Don Rubin, who has the Chair of the Department of Statistics at the Educational Testing Service in Princeton, and I have recently finished a quantitative review of 345 experiments of interpersonal expectancy effects for the new journal, *The Behavioral and Brain Sciences*.

What we found was that the estimated average size of the effect of teacher expectations and experimenter expectations was very substantial in magnitude, about .7 of a standard deviation. A standard deviation is a unit of dispersion or variability to give you a handle. Most of you are familiar with IQ tests, so if you haven't heard of standard deviation, you intuit that 100 is the average IQ and 115 or 120 is one standard deviation above the mean. So in terms of IQ units, two-thirds of a standard deviation would be maybe 10 IQ points or maybe thirteen IQ points, depending on the particular IQ test. But that's just to give you a feel for it, what we call the magnitude of the effect. Gene Glass, incidentally, found in his review of the psychotherapy literature an effect size somewhat similar to the one that we had found in our review of the interpersonal expectancy literature.

Now, you may fairly ask, what does this stuff on effect sizes have to do with parapsychology? Quite a lot, actually, as I rediscovered from a very instructive luncheon chat with Monte Ullman. I think that lots of failures to replicate psi phenomena are not failures to "get the effect," but basically failure to employ sufficient statistical power. Lots of new studies have been proposed here. I'm very dissatisfied with the use that has already been made of all the studies that have already been done. That's not a failing of parapsychology per se; I'm just as unhappy with all the experiments that have been done in the hundreds of areas of psychology and sociology as well. We have conducted thousands of experiments and keep coming to the conclusion that we don't know anything. There's a growing literature in the area of research methodology which teaches something about how to combine and how to organize the results of experiments that have been done. I think the employment of some of these techniques would be very salutory to parapsychology, as it would be to the other behavioral sciences. I do have to confess that one of my own recent research interests has been how to combine probabilities and how to summarize domains of research.

Back to the data. Of those 345 experiments, a very sizeable subset was designed and conducted in such a way that it would be very hard to figure out how experimenters communicated with their research subjects as to what they wanted them to do, or how teachers communicated to their pupils in some standard micro-teaching situations what their expectations had been for these pupils' performance, without invoking some kind of unusual subtle communication process—whether you want to call it ESP or nonverbal cuing it doesn't matter—but it was, in any case, a fairly subtle phenomenon. Let me quickly review some of the evidence that led us to conclude that it was a fairly subtle phenomenon.

One of the tasks that we have often used in our own experiments on interpersonal self-fulfilling prophesies has been to show photographs of faces to research subjects. These faces are held up for the research subjects by experimenters whom we have employed. Half the time these experimenters think these faces they're showing are faces of people who have experienced success and half the time they think they're faces of people who have experienced failure. However, they're always the same photographs. And it turns out that if the experimenters think that they are showing photographs of successful people, they tend to get those photographs judged to be of more successful people by their research subjects than if they believe that these photographs are of unsuccessful people. Or, in a Rorschach situation, if they have

been led to expect that their patients or their clients or their testees are people who tend to see a lot of animal movement patterns in the inkblots, they will get more animal movement responses from their research subjects than will those who don't. This is the kind of experimental evidence that goes to make up the 345 experiments that I've been talking about.

In the photo rating task, for example, every experimenter reads the same instructions to every subject, so it's hard to know how he communicates the expectancy unless it be by nonverbal or ESP type cues. If you put screens between experimenters and subjects, you do tend somewhat to decrease the magnitude of the phenomenon of expectancy effect, but it doesn't go away, so there's some visual effect.

In Winnipeg, in the laboratory of John Adair in collaboration with Joyce Epstein, they did the following brilliant, extraordinary experiment.

They conducted the basic photo rating study. Experimenters were led to expect that their research subjects were going to rate the photos as being of successful people. Other experimenters were led to expect failure ratings from their subjects. The results they obtained were in line with those that had been obtained before. That was Stage I.

Stage II was the following: Unbeknownst to experimenters in Stage I, all of their readings of the instructions (and they all read the same instructions to those subjects of Experiment I) had been tape-recorded. Experiment II involved a whole new group of research subjects who were brought in to be "run." These subjects were brought in, but they were not "run" by experimenters at all. Indeed, there were no experimenters. All there were were spoors and traces of experimenters, namely, the tape-recordings that had been made in Study I of them reading instructions to those other subjects. You have new subjects coming in. The secretary in the psychology department tells them to walk down the hall; there are signs in the hall that tell them "turn left," and they turn left—"open door"—they open door—"proceed to tape-recorder," they proceed to tape recorder. "Push a button" (there's a sign that says "button") so they push the button and each of the thirty subjects hears what one of the thirty subjects in the first experiment heard. Precisely that and nothing else! So now what we have is the audio portion of Experiment I, which is now conducting Experiment II.

What Adair predicted was a loss of statistical significance of the expectancy phenomenon, but maybe still a trend in the right direction, because, after all, the tone of voice would be important. And although you have lost all the video non-verbal cues, at least the audio is still

there. The results were quite shocking. Adair and Epstein found they had just as large an effect of experimenter expectations when there was *no* experimenter, i.e., just the spoors and traces. The audio tape had produced just as significant an effect for this new bunch of research subjects, as the experimenters' expectations had for the original group of subjects. They had a kind of bottled interpersonal self-fulfilling prophecy on these audio tapes. It was all there. How they read the instructions was what turned the trick. Other experiments have been done to show the importance of nonverbal communication in this process of mediating interpersonal self-fulfilling prophecies.

Not only human subjects have been found effective recipients of these interpersonal self-fulfilling prophecies, but animals as well. Bertrand Russell, in 1927, intuited something about how animal subjects in laboratories might be affected by the caretakers, the laboratory directors, the psychologists and the ethologists with whom they come in contact. Russell said, "Animals studied by Americans rush about frantically with an incredible display of hustle and pep; whereas animals observed by Germans sit still and think, and at last evolve the solution out of their inner consciousness."

We did some experiments with animal subjects as well. We arbitrarily labeled half of our animals as "maze-bright," and half of them as "maze-dull." We found to our surprise that experimenters expecting rats to perform well got smarter rats, even though the same rats had been assigned at random to those expecting good and poor performance. The performance that they expected was the performance that they got.

Throughout much of this research on interpersonal self-fulfilling prophecies, we made movies. Some of this work was in the late fifties and early sixties, long before video-tape was in general use, so we started with film.

As experimenters interacted with their research subjects, say in the photo-rating task, we made sound/motion pictures of the interaction and then we had observers code the behavior, making relative judgments of what was going on in the interaction between the experimenters and subjects. We found, for example, that male experimenters conducted a more friendly experiment than did female experimenters, and that female subjects elicited different behavior from their experimenters than did male subjects. For example, only twelve percent of experimenters ever smile at a male subject when they conduct these perfectly standard psychological experiments. However, seventy percent of the experimenters smile at female subjects.

How long does it take to read these perfectly standard instructions?

We found that it mattered whether you were male or female. For example, if you were a male reading instructions to a female subject, it took you longer to read them than if you were reading them to another male. Or if you were a female experimenter reading instructions to a male, it took you longer to read them than if you were reading them to another female (all the experimenters' subjects were about the same age), thus, opposite sex dyads lasted longer in the conduct of these "perfectly standard" psychological experiments.

One of the nice things, from the point of view of nonverbal communication research, that you can do when you have the sound/motion pictures is that you can split off the channels of information—audio from video. You get one group of listeners just to listen to the sound track. They never see what happened. You get another group that watches the silent film and never hears what happened. Then you get a third group that has both the sound and the video portion. Ratings based on these different channels taught us a number of things.

For instance, male experimenters were friendlier to female subjects than to male subjects, both in tone of voice and in movement patterns. Not a big surprise. Female experimenters were more interesting. When they were interacting with female subjects, they were judged as quite friendly in movement patterns. They decreased interpersonal distance, they leaned over close. They were judged not very friendly, though, in tone of voice. When the same female experimenters were conducting experiments with male subjects, they were judged not so friendly in the video portion—that is, they didn't decrease interpersonal distance particularly—but were judged very friendly in tone of voice. So it appears that the sexes may have different things to say to each other in different channels of communication.

Nonverbal communication is very subtle. Much of the work that we have done has been in laboratories. But some of the work has been in schools, in clinics and in hospitals. One of our earlier ventures into the hospital setting was an attempt to understand something about the outcome of treatment of alcoholics as a function of nonverbal cues that doctors might give the alcoholic patients when they first come in contact with them.

We had available a series of nine interviews with doctors, each of whom had been in charge of the emergency admissions ward of a large, urban, high-quality hospital. A lot of alcoholics go through those doors in the emergency ward. They had a new treatment center for alcoholics which had been established right there in the hospital. For each of the nine doctors, we had a success rate—what proportion of the alcoholic

patients could he get to go into treatment? Some of the doctors were very good at getting their alcoholics to go to treatment and others weren't very good at all.

From each of the nine interviews, we clipped out a one-minute taped portion of their answer to the question: "What's been your experience with alcoholics?" We took this one-minute tape clip, and ran it through a low pass band filter that removed the frequencies above 500 hertz. What's left sounds like conversation heard through a closed door or a heavy curtain. You can certainly tell it's human speech. Though you can make out an occasional fraction of a word, you can't really understand what's being said. But you can certainly get the tone of voice. Content-filtered tapes were made from the original one-minute tape clips where you can hear what was said. They were then played for undergraduate students who made ratings of the degree of hostility in the doctor's voice. There's a point six correlation here: the greater the degree of hostility in the doctor's tone of voice in talking about alcoholics, the more successful the doctor had been in getting alcoholics to go into treatment. Point six is a very, very substantial correlation in behavioral science research. Not very high in physics or engineering or even in economics, but it's very high for most of the behavioral sciences. That was a very dramatic result, and it's one of the factors that led us to develop the PONS test—the Profile of Nonverbal Sensitivity.

The Profile of Nonverbal Sensitivity is a test that measures sensitivity to cues in eleven different channels of nonverbal communication (see Figure 1). For background, let me go back to the doctor's voice study because the two things that will be new to you are "CF" and "RS"—Content Filtered and Randomized Spliced Speech, so before I go into details of the PONS test, I'd better tell you something about both of those.

In Content Filtered Speech you just knock out the high frequenices. I'm going to play you some excerpts of some ordinary English statements that were made in four different affect quadrants. There will be a statement by a female and by a male made in a positive-submissive way; another in a positive-dominant way; another in a negative-submissive way and another in a negative-dominant way. So there are these four kinds of affects. Now, I'm going to play the content-filtered part and see if you can understand the words and then see if you can easily detect what the affect is. Then, after I've played the content-filtered speech for you, I'll play you the same four excerpts randomly spliced. Randomized splicing is an alternative way of getting rid of the content if you're a nonverbalist and want to study tone of

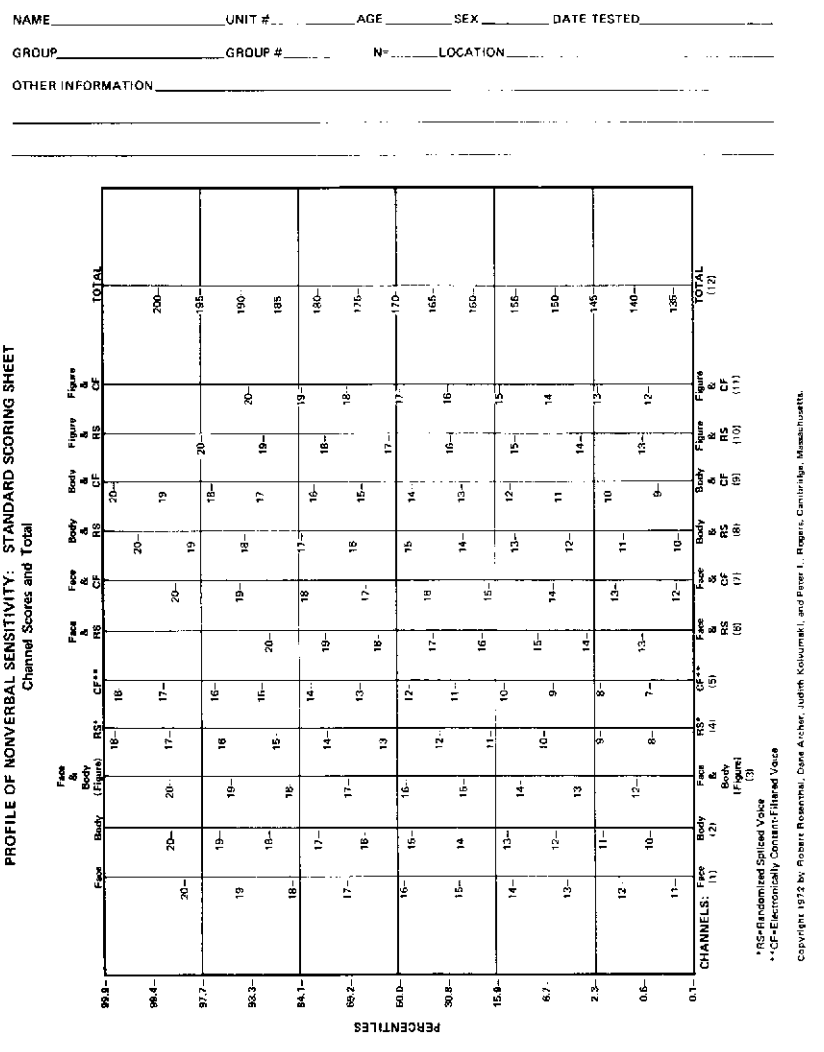

Figure 1

voice. You've got to get rid of the content for some purposes because the world doesn't come in standard speech. If you want to compare the sincerity of presidential candidates, a good way to do that is to remove the content so that you can play it for judges who won't be affected by what the person is saying, but will be listening to pure tone. The problem with the content-filtered speech is that by throwing out the high frequencies, you're losing important affective information. Then

a former student of mine, Klaus Scherer, who is now at the University of Giessen, as a Professor of Social Psychology, developed a randomized splicing procedure as an alternative way to keep all of the tone, but throw out the content. You cut the audio tape into lots of little pieces. Then you get a table of random numbers on the basis of which you randomly rearrange these little clips and splice them together. I now play you some content filtered and random spliced speech.

Before I play the original English of which the following excerpts were transformations of various kinds, did any of you understand any of the statements that were being made there? They're very hard to decipher. Some people do very well in somehow picking out the original English of content filtered speech. There are two kinds of people who do particularly well at that. One is mothers of toddlers learning how to talk. They are constantly getting content-filtered speech at home and they're pretty good at understanding that kind of speech. The other kind of people who are good at doing that are people who work with speech disordered people or with the deaf, and some speech therapists. Sometimes, but not as often, psycholinguists.

Now, I'll play you the four statements by the male and the female speaker before they were transformed. Here are the original versions:

Positive submissive: "I wonder if I could ask you a favor? Could I borrow enough money for lunch?" MALE VOICE
SAME QUESTION REPEATED—FEMALE VOICE
Positive dominant: "Well, hello. I haven't seen you for an age. We should get together more often." MALE
SAME STATEMENT REPEATED—FEMALE
Negative submissive: "Gosh, I'm sorry I borrowed the car without your permission, Dad." MALE
SAME STATEMENT REPEATED—FEMALE
Negative dominant: "Confound it, if that's how it has to be, then you can do it without me." MALE
SAME STATEMENT REPEATED—FEMALE

That's the end of the audio portion demonstration. The video is yet to come. Let me give you just a quick theoretical footnote on the development of the PONS. It was not within the context of psi research that we did all this, but in the context of social psychological thinking. We wanted to be able to do a better job of predicting the outcomes of interpersonal interactions, whether these interactions were teaching interactions or psychotherapeutic interactions or everyday chance encounters of any sort.

What we thought we might be able to do in principle, though it would be difficult, we felt, to do in fact, would be to map out for every person who might encounter any other persons, two kinds of traits having to do with nonverbal transmission and reception. That is, how good they would be in encoding nonverbal cues in different channels and how good they would be in decoding cues in different nonverbal channels. By a channel, I just mean something like seeing the face without looking at the body, or the body without looking at the face, or both together—or randomized spliced speech or content filtered speech. I'm arbitrarily calling those things channels just to make them a handy thing to talk about.

So, we have face or body, and face plus body, random spliced and content filtered speech and various combinations of those. If we knew for any one person how good he was at sending in those channels, how decodable are his messages in those nonverbal terms, and also knew for that person how good he was at decoding those particular messages, we would know something useful that might lead us to improve the predictions of outcomes of psychotherapy, of teaching and of chance encounters. We call it the Diogenes model for no good sound scholarly reason of any sort.

We haven't yet got standardized measures of encoding ability, though we have some people who have been studied intensively from the point of view of encoding ability. Most of our work on the PONS does give us confidence that we know something about how to measure decoding ability in eleven channels of nonverbal communication, and that we can predict things like the outcome of psychotherapy from knowing how good a decoder you are. Not as a patient, but as a therapist.

For example, if you are the average psychotherapist, you are not any better as a decoder of nonverbal cues than your average high school student. But if you are an above average psychotherapist, as judged by your peers and supervisors, you are a better decoder of nonverbal cues than if you are a bad psychotherapist as judged by your peers and by your supervisors. So we can make some useful distinctions with the standardized test called the Profile of Nonverbal Sensitivity.

Let's go back now to the Profile sheet (Figure 1) that I introduced you to a few moments ago. Across the top is a readout of the eleven channels: Face, Body, Face and Body, Random Splice, Content Filtered, Face plus Random Splice, Face plus Content Filtered, Body plus Random Spliced, etc. Those are the eleven basic channels, and the numbers that you see under each of those are the percentiles—for example, under Face, down around the middle of the sheet it says "50"

on the left. That's your fiftieth percentile score. If you got sixteen out of twenty face items right on our test, you would be an average face decoder. This is a standard scored profile sheet. If you got about fifteen items right on the body, you would be an average body decoder. If you got only twelve items right on content filtered; (that's a lot less than face), you would still be an average content filtered decoder. In other words, it's much more difficult to decode content filtered cues than to decode face cues or body cues. In fact, based on several thousand subjects, there's a neat kind of relationship that's linear in the logarithms that goes from tone-to-body-to-face in terms of the amount of information in this test. Not under all circumstances; the ratio is that for one unit of information. One standard deviation's worth of information in tone is what you get by adding tone to the total message. Two standard deviation's worth is what you get by adding body, and four is what you get by adding face. Those are the ratios, not the actual standard deviations. You get more in standard deviation units.

If you turn to Figure 2, you'll see summarized channels. I won't take time with the left half. The right half has four combinations: Positive-Submissive, Positive-Dominant, Negative-Submissive and Negative-Dominant. In addition to the score on eleven channels of nonverbal communication, we also get a score for how good you are at decoding negative affects, positive affects, submissive affects and dominant affects and all combinations of those two dimensions. We find that different people are differentially good at decoding different ones.

One of our most dramatic findings concerns sex differences. Women, in all the societies that we've tested, are very substantially and reliably better than men at decoding nonverbal cues across the board. That was true when we tested American college students and Mexican college students, college students in Israel and England, teachers in Singapore, Indians and Eskimos and New Guineans, Australian aborigines and immigrants. Wherever we go and test—and we've tested pretty much all over—women are better than men. They are most especially better than men at negative affects. And one of our female graduate students has put forward a kind of impression hypothesis that women needed to be better at decoding negative cues, because it was life-saving to be so in eons gone by—and maybe yesterday as well.

One of the things that surprised us when we constructed the test, was that we found that everybody was getting everything right. The original length of the video clips was five and a half seconds. So, just for psychometric reasons, to make a better test, we had to shorten the video clips to make it tougher. We wanted to have an answer sheet that had

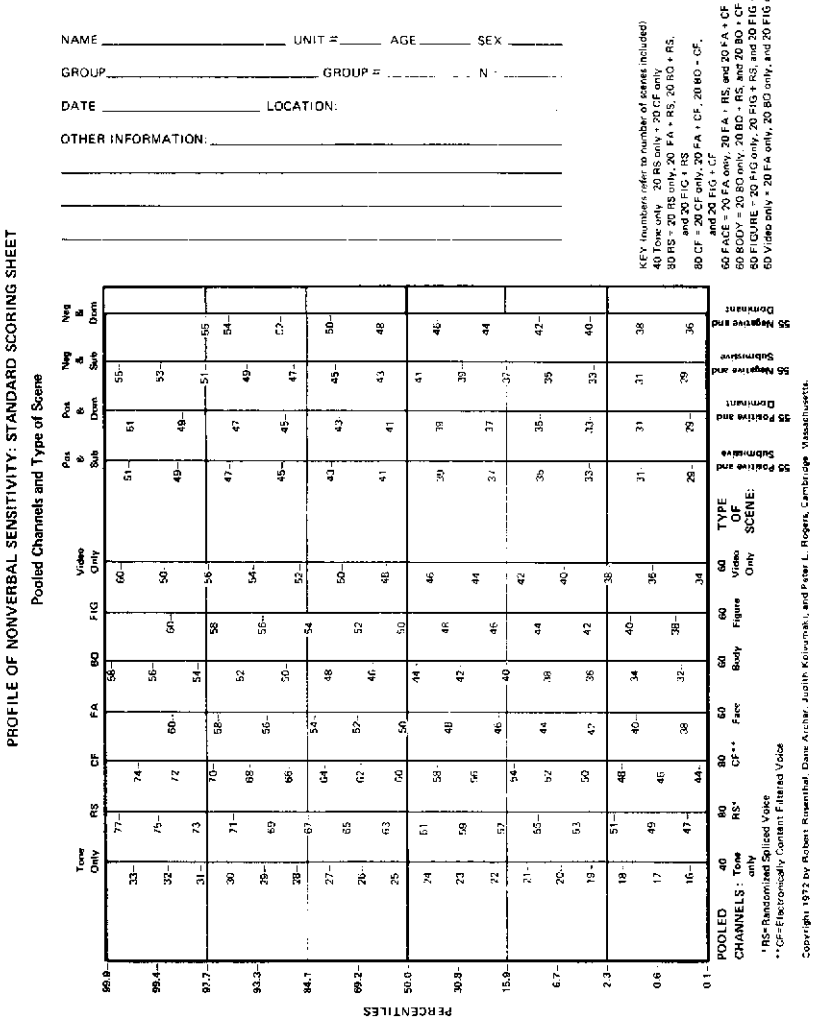

Figure 2

two answers—*A* and *B*. If you pick *A* or *B*, you have a fifty percent base rate; but you'd like to get something where the average accuracy is around seventy-five percent. If the average accuracy were fifty percent, that would just be chance, if you have only two alternatives. If the average accuracy were around one hundred percent, which we had when we first made the test, that's no good, because everybody gets everything and you can't discriminate among people. That's the

purpose of the test—to be able to score people on some trait or characteristic. So we kept cutting down and when we got to two seconds we quit. At two seconds, people were getting about seventy-five percent of the information. That's a lot of information. We wondered how far we could cut things down further. Could we go to one second? Would people still get information from the face and from the body, for example, in one second? How about a half a second? How about an eighth of a second? How about a twenty-fourth of a second, or is that so fast that the subject would think he hadn't even seen it? Figure 3 is a test form called "Nonverbal Communication BP." (BP means Brief PONS.) There is a series of film clips that vary in exposure length from one frame (that's one twenty-fourth of a second) to twenty-seven frames (that's over a second).

The score curves are quite monotonic in the sense that people do very, very well if they have twenty-seven frames; they do quite well if they have nine; they do just about as well if they have three. The big drop is when we go from three frames to one. The interesting research question is—what's the difference between one and three frames? Is it that when you go from one to three frames you introduce movement which you can't have on a single frame, or is it simply that the subject sees more of the same? The experiment that we have planned is to simply take those three frames and then scramble them and see if people will do about as well when they're scrambled. If so, it can't be the movement per se that does it, because if it's backwards, forwards, sideways, you always get about the same accuracy score. It must be just the total amount rather than the movement per se, and that's an experiment that we're planning to do.

I've already mentioned that women are better than men pretty well across the board. There's one interesting thing we call Discrepancy Sensitivity. It is a fifth index of sensitivity to nonverbal communication developed with Bella De Paulo and others at Harvard. That's a test that measures how sensitive you are to discrepancy between tone of voice and body movements or tone of voice and facial expressions. We find that that's a skill only partially related to the skill of decoding face, body and tone cues. It takes to some extent a different kind of skill to be able to detect discrepancies in tone of voice. The importance of discrepancies in psychopathology is given to us by Bateson, Jackson, Haley and others, the formulators of the Double Bind Hypothesis, that sources of psychopathology may reside in childhood, when children get different conflicting cues from their parents in different channels of communication.

Sometimes nonverbal information is not so much sent, but given off,

NONVERBAL COMMUNICATION (BP)

Name _____ Present address _____

Town and country of birth _____ Age _____ Sex _____

Primary language spoken _____ Secondary language spoken _____

Father's occupation _____ Mother's occupation _____

Field of study _____ Average grade in last year of school _____

INSTRUCTIONS: Please circle the letter (A or B) next to the label which best describes the scene you have just seen.

SAMPLE ANSWER: Scene 1. A. admiring a baby
(B) applying for a job

1. A. ordering food in a restaurant
   B. threatening someone
2. A. nagging a child
   B. criticizing someone for being late
3. A. expressing strong dislike
   B. expressing deep affection
4. A. expressing motherly love
   B. threatening someone
5. A. returning faulty item to a store
   B. helping a customer
6. A. talking about one's divorce
   B. returning faulty item to a store
7. A. saying a prayer
   B. threatening someone
8. A. admiring nature
   B. saying a prayer
9. A. talking about one's divorce
   B. trying to seduce someone
10. A. admiring nature
    B. expressing motherly love
11. A. saying a prayer
    B. talking about one's wedding
12. A. talking about the death of a friend
    B. expressing jealous anger
13. A. talking about one's wedding
    B. talking about the death of a friend
14. A. expressing strong dislike
    B. ordering food in a restaurant
15. A. criticizing someone for being late
    B. expressing gratitude
16. A. expressing deep affection
    B. nagging a child
17. A. expressing deep affection
    B. admiring nature
18. A. helping a customer
    B. asking forgiveness
19. A. talking about one's divorce
    B. leaving on a trip
20. A. saying a prayer
    B. nagging a child
21. A. leaving on a trip
    B. nagging a child
22. A. talking about one's wedding
    B. expressing gratitude
23. A. talking about one's divorce
    B. asking forgiveness
24. A. returning faulty item to a store
    B. talking about the death of a friend
25. A. asking forgiveness
    B. nagging a child
26. A. expressing motherly love
    B. asking forgiveness
27. A. returning faulty item to store
    B. expressing strong dislike
28. A. talking about one's wedding
    B. talking about one's divorce
29. A. expressing strong dislike
    B. helping a customer
30. A. threatening someone
    B. expressing strong dislike
31. A. criticizing someone for being late
    B. expressing gratitude
32. A. expressing jealous anger
    B. nagging a child
33. A. expressing motherly love
    B. talking to a lost child
34. A. leaving on a trip
    B. expressing deep affection
35. A. admiring nature
    B. helping a customer
36. A. leaving on a trip
    B. trying to seduce someone
37. A. expressing jealous anger
    B. threatening someone
38. A. nagging a child
    B. talking to a lost child
39. A. ordering food in a restaurant
    B. expressing jealous anger
40. A. trying to seduce someone
    B. talking to a lost child

Figure 3

as Goffman has said—or leaked, as Ekman has said. Face is easiest to control; we feel it leaks the least, though there may be leakage there. Body is easier to control than tone. You can't control Brief PONS because you don't know they're there. And finally, you're least able to control discrepancies among different channels. What has this all to do with the sex differences we've looked at cross-culturally? This—that although across the board women are uniformly better than men, women give up their advantage over men in accuracy of decoding more and more as you get to the leakier and leakier channels. Women are most superior to men at the face. They are next most superior at the body. They are third most superior at tone. They are fourth most superior at the brief PONS, and they are fifth most superior or probably not at all superior at the Discrepant PONS. Our interpretation is that, at least in our culture—and most of this five-level work has been done in our culture—women have adopted a kind of courteous orientation interpersonally, not only in nonverbal communication, but linguistically as well. If you read the face well you can be more assured that you're reading what the person wants you to read—the person that's sending the message. As you move down that, you're getting more and more into leakage. We have some evidence to suggest that if you're very good, for example, at micromomentary display—if you can decode the one twenty-fourth second—that may not be so good for your social adjustment. For example, although social adjustment correlates well if you're a good face/body decoder, if you are a good micromomentary display decoder, that's likely to be slightly negatively correlated with social adjustment. It may be that because of this phenomenon, you may know too much about other people. Like science fiction stories where some people come from another planet and they have a heart of gold but they also have this unfortunate gift of psi, and they know what everybody thinks and nobody likes them because people like their privacy.

Those lower channels are maybe getting closer to psi in the sense that they're more covert, they're more leaked, and therefore they may do you less good in a social sense. Women, having been taught in our culture to do the right and courteous thing, are maybe showing a special kind of courtesy in the avoidance of eavesdropping on nonverbal cues to which they are not supposed to have interpersonal access.

Let me link this to the general literature to point out that generally in the studies that have been done, women have proved to be more supportive in conversations; they laugh at others' jokes more; they listen more; they argue less; they interrupt less. They use less profane

and hostile language. They are more visually attentive. They smile more; they look more pleasant. They intrude less on others' personal space; they even use up less space controlling, for their size, than do men. That is, they shrink up more. So the pattern we get is one, overall, of having women in our culture be more polite, more courteous in their use of nonverbal as well as verbal cues.

Women in our culture, then, seem to have learned that there may be social hazards to knowing too much about others' feelings. This pattern of politeness is consistent with the traditional sex role that has been ascribed to women in our culture—the sex role that is only now beginning to change.

DISCUSSION

MORRIS: On the PONS test—what is the effect of us knowing versus not knowing the set of potential descriptors, and does that interact with the length of the film?

ROSENTHAL: Judy Hall, one of our research group who helped to develop the PONS test—she's at Johns Hopkins University—has done research to test whether you do better if you know the descriptor before you see the clip or not until after. Although most people phenomenologically feel that they will do better if they know the two alternatives (A and B) before they look at it, in fact it doesn't make any difference whether you do it before or after. Then, does it matter as a function of the length of the stimulus? That's not been done. All we've done it for is the full PONS test, which you didn't see any of, and that's two-second clips, and for that it doesn't matter. But that doesn't mean it wouldn't make a difference for these four lengths of one, three, nine, and twenty-seven frames.

MORRIS: I would expect that with the longer clips it might stand a chance of making a difference.

ROSENTHAL: It might.

MORRIS: Since we are talking about the possibility of ESP here, have you thought about trying to imbed some sort of ESP control testing just to make sure that you're not in part testing people's ESP abilities?

ROSENTHAL: Well, for some of these short tests, like the Brief PONS test, we have done what we call ESP controls. But now let me hasten to say that we don't regard those as good tests for ESP. We regard those as

question arrangement on paper problems. That is, it may be that if you give anybody an answer booklet with multiple choice items and have him go through and circle the answer, I don't care if its a third semester course in calculus or whatever it is, if it's multiple choice, you may somehow by the way you have organized the page and by the way you've formed your questions, have the average computer do better than chance. Now, the longer the test, the less likely that is. If you've assigned your alternatives to the correct answer at random and if you've assigned questions to page location at random, then in a long test it's very unlikely to happen. But the shorter the test, the greater the problem. For a forty-item test there's more of a problem than for a two hundred and twenty item test. So we did do these ESP controls. Where you'd expect fifty per cent by chance, under our ESP control conditions, you'd get about fifty-one per cent accuracy. I'm not willing to say that's a test of ESP, although it's statistically significantly better than fifty per cent for the psi sample we used. It could be that that's an effect of the particular answer sheet layout that we had. But we used that to compare accuracy above that "ESP" base rate.

MORRIS: Taking into account individual differences?

ROSENTHAL: Individual differences among items. Each item gets a special ESP correction.

NASH: Dr. Rosenthal, do you think any experimenter effects may be due to psi?

ROSENTHAL: That's a really tough question. I don't know whether any experimenter effects may be due to psi. Rhine didn't think so. I had a long talk with him fairly early in the history of the research project, in 1960 when I was visiting at Ohio State University and he was there on a visit. We discussed it, and it was his feeling that the magnitude of the effect that we were getting in the expectancy studies was inconsistent with the magnitude of the effect that he was getting or that he knew about other people's research. While he certainly would have never ruled it out as a phenomenon, he felt that it was a different order of magnitude of effect. At the time I felt pretty good about that, because I really didn't want to get involved with ESP research; thought it might be detrimental to my health as a young academic. Now I'm old and it doesn't matter. But I think, since having heard some more recent papers on ESP, that I may change my mind on that. I heard some papers at a parapsychology convention a couple of years ago at which people were giving the results of recent experiments of theirs. I did some very quick calculations on the size of the effect shown and the effects were very large, much larger than the average size of the effect

we get in standard social psychological or cognitive psychological, or developmental psychological phenomena. I guess I would hold in abeyance any judgment on it. I certainly wouldn't rule it out.

BYERS: Some years ago I was doing some experiments, when I was teaching photography, on how many frames people can see and what do they get out of it. What I discovered then was one could see very brief frames if they were evenly spaced so that they presented a rhythm that one could eventually learn to relate to. The different ones threw them off. What I want to ask you is, when you were working with the very brief one to three frame clips, were those randomly different lengths, or did you space the one frame to bring three frame clips together?

ROSENTHAL: Is your question when we had a forty-item brief PONS test, where did we put the one frames and where did we put the three frames?

BYERS: Yes—were they randomly distributed?

ROSENTHAL: They were randomly distributed with some blocking, so that a bad luck draw table of random numbers couldn't put, say, all the one framers at the end, so that if everybody did really badly on that, we wouldn't know if it was because they were one frames or if it was a fatigue effect. So we employed a procedure called "blocking," so that within each quarter or fifth of the test, we have fairly equal representation of the four time lengths. It's not one, three, nine, twenty-seven obviously. I mean, there's no fixed order; it's random, but it's blocked so that at least the first half has equally often one, three, nine and twenty-seven, and the last half has equally often one, three, nine and twenty-seven.

MORRIS: What instructions did you give your speakers on the tapes we've heard, and have you thought about varying the instructions to them? I gather that they are acting.

ROSENTHAL: These people were acting, and the young woman that you saw in the Brief PONS was also acting.

MORRIS: You might be in part studying artifacts of what kind of acting they do of their own conceptions, and I was wondering how you perceive that.

ROSENTHAL: We don't really know for any particular senders how different they themselves as senders might be when they've been asked to act or if they were in a spontaneous, everyday life situation.

MORRIS: It would be interesting to give them some of your tests and see how they do.

ROSENTHAL: You mean, how the encoders do as decoders?

MORRIS: Yes.

ROSENTHAL: Well, that's a growing area of research actually. More and more people are now beginning to look at the relationship between encoding and decoding skills, and it's very, very ambiguous. The correlation seems to be small, but positive. If I had to guess, I'd say the correlation based on some fifteen studies is maybe point two. So it's not terribly impressive. Better encoders tend to be better decoders.