DOCUMENT RESUME

ED 317 551 TM 014 556

AUTHOR Rosenthal, Robert
TITLE Experimenter Expectancy, Covert Communication, and Meta-Analytic Methods.
SPONS AGENCY Center for Advanced Study in the Behavioral Sciences, Stanford, Calif.; National Science Foundation, Washington, D.C.
PUB DATE Aug 89
NOTE 43p.; Paper presented at the Annual Meeting of the American Psychological Association (97th, New Orleans, LA, August 11-15, 1989).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Academic Achievement; *Affective Behavior; *Effect Size; *Meta Analysis; Researchers; *Research Methodology; Statistical Analysis; Teacher Attitudes; *Teacher Expectations of Students; Teacher Influence
IDENTIFIERS *Affect Effect Theory; Communication Channels; *Covert Communication; Research Replication

ABSTRACT

The affect/effect theory of the mediation of teacher expectation effects is presented, and a research agenda is suggested for the investigation of this theory. In addition, some methodological issues in psychology are reviewed. The affect/effect theory states that a change in the level of expectations held by a teacher for the intellectual performance of a student is translated into a change in the affect shown by the teacher toward that student and the degree of effort shown by that teacher in teaching that student. The affect/effect theory is consistent with much experience and research. The following aspects of the theory are considered: (1) dynamic features; (2) communication channels; (3) molar versus molecular variables; (4) redundancy versus specificity; (5) channel discrepancy; (6) interactional synchrony; and (7) direct intervention. Future efforts should extend the affect/effect theory to domains other than education. The methodological features of psychology reviewed include: a discussion of effect size; a consideration of what constitutes successful replication of research; and a review of the benefits of meta-analysis, both obvious and less apparent. A 52-item list of references is included. (SLD)

EXPERIMENTER EXPECTANCY, COVERT COMMUNICATION, AND META-ANALYTIC METHODS

Robert Rosenthal

Harvard University

Experimenter Expectancy, Covert Communication, and Meta-Analytic Methods

Robert Rosenthal

Harvard University

Exactly thirty years ago I presented my first paper at APA. Both presentations, that one in Cincinnati in 1959 and this one in New Orleans in 1989, would have been quite impossible without Donald Campbell.

Donald Campbell is not only a brilliant scholar of the social and behavioral sciences, he is an inspired and inspiring teacher as well; one who has affected the intellectual lives of scientists and scholars of all kinds. His impact on me was enormous. In addition to his intellectual inspiration, he provided me with great personal support thirty years ago when I was engaged in very controversial research on the unintended effects of psychological researchers on the results of their research. At that time he was one of the very few established psychologists to speak out on behalf of a "backwoods psychologist" conducting research at the University of North Dakota. Conducting research, it should be noted, that remained successfully unpublished for years. (Two other of my psychological sponsors at that time were Harold Pepinsky--who, fittingly, had developed the very concept of psychological sponsor--and Hank Riecken, who had anticipated so much of the work on the social

psychology of the psychological experiment, and who was responsible for the financial support of the National Science Foundation for the work I was doing in those early days.

My first communication from Don Campbell was in a letter he wrote on December 1, 1958, in which he agreed to contribute to a symposium on the problem of experimenter bias at the forthcoming APA. A long correspondence followed in which he gave invaluable advice on organizing the symposium and later, on publishing a book on the topic. Recently re-reading this correspondence showed me just how good a mentor Don Campbell was, even by mail.

I take pride, in primitive identification with Don, that we both spent time studying at UC Berkeley (he a lot and I a little); that we both taught at Ohio State (he a lot and I a little); and that we both published in unrefereed journals (he a little and I a lot).

It seems consistent with the Campbellian spirit for me to discuss today some matters that are substantive and some matters that are methodological. We begin with the substantive. Those who know me best will be surprised: I am not presenting the results of our most recent studies of covert communication in classrooms, clinics, courtrooms, or laboratories. Instead, I want to propose a compact

"theory" of the mediation of teacher expectation effects. I will describe the theory, and in doing so suggest a research agenda for its investigation. We will have a brief look at the nature of the theory, consider some structural and dynamic features, and the role of (a) various channels of communication, (b) molar versus molecular variables, (c) redundancy versus specificity, (d) channel discrepancy, and (e) interactional synchrony. Finally we consider direct interventions to test the theory and some future directions.

The Affect Effort Theory of the Mediation of Teacher Expectation Effects: A Research Agenda

The affect effort theory states that a change in the level of expectations held by a teacher for the intellectual performance of a student is translated into (a) a change in the affect shown by the teacher toward that student and, relatively independently, (b) a change in the degree of effort exerted by the teacher in the teaching of that student. Specifically, the more favorable the change in the level of expectation held by the teacher for a particular student, the more positive the affect shown toward that student and the greater the effort expended on behalf of that student. The increase in positive affect is presumed to be a reflection of increased liking for the student for any of several plausible reasons (Jussim, 1986). The increase in teaching effort is presumed to be a reflection of an increased belief on the part of the teacher that the student is capable of learning so that the effort is worth it (Rosenthal & Jacobson, 1968; Swann & Snyder, 1980).

Structural Features

The affect/effort theory is consistent with the theoretical writings of most of the workers in this area of research (e.g., Brophy, the Coopers (Harris and Joel), Darley, Deaux, Dusek, Fazio, Good, Jones, Jussim, Miller, Snyder, Swann, Turnbull, Zanna,

and others), most of whom would probably find it congenial. In addition, the conceptual distinction between the affect and effort factors maps nicely onto the affect/cognition distinction recently under fruitful debate by Lazarus (1984) and Zajonc (1984). The neuroanatomic evidence, in particular, gives a strong Bayesian prior probability to the likelihood of the importance and relative independence of the affect and effort factors.

The affect/effort theory is also consistent with (but not directly demonstrated by) the results of a recent set of 31 meta-analyses investigating the older four-factor "theory" of the mediation of interpersonal expectancy effects (Harris & Rosenthal, 1985). Although our meta-analytic work has given strong support to each of the four "factors" of climate, input, feedback, and response opportunity, there are virtually no data available to permit us to conclude that these four "factors" are, in fact, relatively orthogonal. We plan to do principal components analyses of a large set of variables serving to define the four factors. The prediction is that most of the dozens of variables involved will turn out to load substantially either on the affect (roughly climate) or the effort (roughly input) component, after varimax rotation. Our prediction is not that only two "significant" components will emerge, but rather that

our two components of affect and effort will dominate over other emerging components.

Dynamic Features

The emergence of two relatively orthogonal and relatively important (in the sense of the sum of the squared factor loadings) components of affect and effort provides necessary but not sufficient evidence for the theory. It is also necessary to show that the magnitude of teacher expectation effects depends upon a differential increase in positive affect and teaching effort directed toward those students for whom more favorable expectations have been created experimentally, compared to the students of the control group.

The specific predictions from affect/effort theory are that there will be a substantial positive correlation (a) between the favorableness of the expectation induced and the increase in positive affect and teaching effort, and (b) between the increase in positive affect and teaching effort and the increase in subsequent student intellectual performance. Any theory of the mediation of interpersonal expectancy effects must provide evidence for the relationship between (a) expectations and the mediators *and* (b) mediators and the behavior of the expectee or target (Rosenthal, 1981).

Communication Channels

Affect effort theory predicts that the factor of teaching effort depends most heavily on the verbal channel of communication with some contribution from such nonverbal channels as facial expression, body movement, and tone of voice. The factor of affect, however, is predicted to depend at least as much on the nonverbal channels as on the verbal channel of communication. This prediction is based on the association of cognitive with linguistic functioning and the association of affective with paralinguistic functioning (Buck, 1984; Ekman, 1973; Blanck, Buck, & Rosenthal, 1986).

Overall teaching effort can be defined by the mean ratings made by videotape raters on such variables as amount of material taught, task orientation, teaching effort expended, and active, competent, and professional demeanor. These raters have access to the full videotape, including sound track. Four other groups of randomly assigned raters have access only to (a) the written transcript of what the teachers said; (b) the teachers' faces while teaching; (c) the teachers' bodies while teaching; and (d) the teachers' tones of voice while teaching based on content-filtered speech (Rosenthal, 1987).

Overall positiveness of affect can be defined by the mean ratings made by videotape raters on such variables as warm, friendly, likable, pleasant, caring, and empathic. As in the case of the teaching effort variable, ratings are made by five groups of randomly assigned raters. One of these groups has access to all video and audio information but the remaining groups have access only to (a) the transcript of what the teachers said; (b) the teachers' faces; (c) the teachers' bodies; and (d) the teachers' tones of voice based on content-filtered speech.

Molar Versus Molecular Variables

Affect effort theory predicts that the factor of teaching effort is associated more strongly with more molecular variables involving counting or timing than with more molar, global variables involving overall ratings, while the opposite is true for the factor of positivity of affect. Thus, for example, we predict that teaching effort can be relatively more efficiently assessed by more molecular variables such as time on task, work-related contacts, speech rate, and number of words taught, than by such variables as ratings of teaching effort expended or activity level. In the case of affect, the theory predicts that more molar ratings of e.g., warmth, empathy, or friendliness will better assess affect than will more molecular variables such as smiling, glancing, nodding, leaning, pitch level, or pitch range. This is a counter-

psychometric prediction, since molecular variables tend to be far more reliable than molar variables (Rosenthal, 1966; 1976; 1987). Nevertheless, affect-effort theory predicts that molar variables will correlate more highly with the criterion affect variable[1] than will the molecular variables. We predict this because interpersonally communicated affect implicates the use of many channels of verbal and nonverbal communication and molecular variables tend to be more channel-limited than molar variables. Since the factor of teaching effort depends more heavily on a single channel, the verbal, it will be better indexed by molecular speech-related variables than by more molar variables.

Redundancy Versus Specificity

Affect-effort theory states that effort is characterized by greater simplicity and unity and less potential for conflict and ambivalence than is the case for affect. Therefore, when molar variables are assessed in the verbal, face, body, and tone channels, effort will show greater channel-to-channel redundancy than will affect which will show greater channel specificity. Redundancy is measured either by the eigenvalue of the first unrotated principal component or, more simply, by the average intercorrelation among the four channels of communication.

[1]This variable is defined by the composite variable formed from the principal components analysis but with unit weighting. Rosenthal, 1987, Chapter 5.

Channel Discrepancy

On the basis of a rich clinical tradition (e.g., Bateson, Jackson, Haley, & Weakland, 1956), and of more recent empirical work by Bugental's group (e.g., Bugental, Love, Kaswan, & April, 1971), by De Paulo & Rosenthal (1979), and others, there is reason to suspect that teachers showing greater discrepancies between the channels (e.g., larger differences in positivity expressed between verbal content and body movements or tone of voice) will differ in the magnitude of interpersonal expectancy effects shown. Since channel discrepancies are associated with perceptions of negative affect, teachers showing characteristic discrepancies may show smaller effects of positive expectations that have been induced experimentally.

Although we have been speaking of channel discrepant communication as a trait-like, stable moderating variable, it should be noted that we can also consider it as a state-like, situational, mediating variable. Indeed, we will be examining channel discrepant communications as mediating variables with the prediction that discrepant communications will function as more affectively negative than would be predicted from the mean affective level of the two channels involved.

Interactional Synchrony

Affect/effort theory implies that as a consequence of the increased positivity of affect and of teaching effort that typically follows an increase in favorable expectation there will be an improvement in the rapport or micro-climate of the teacher-student dyad. This increased rapport can be assessed by measures of interactional synchrony and it will predict the magnitude of improvement of students' intellectual performance (Bernieri, Reznick, & Rosenthal, 1988; Bernieri & Rosenthal, in press). Interactional synchrony, then, functions as an additional post affect/effort mediator occurring before increased student performance.

Direct Intervention

An additional strong test of affect/effort theory is possible by attempting to achieve direct experimental control of the mediating factors. We can manipulate experimentally both the affect and the effort factors. Our basic independent variables will be high versus low levels of positive affect and high versus low levels of teaching effort in a 2×2 design. By training teachers to show all four possible combinations of affect and effort we can test directly the effects of both factors on student learning. Although our primary goal would be cross-validation of affect/effort theory, this research would also serve as part of a useful foundation for

future programs of applied research designed to improve student performance by using research results from the literature of interpersonal expectation effects.

Future Directions

We plan to extend the generality of affect/effort theory to other domains: specifically, to the domains of counseling, psychotherapy, medicine, and management. We believe that affect effort theory applies as well to these domains as to the domain of education. The primary conceptual adjustment that must be made is in the nature of the effort factor. For the educational context the effort is teaching effort. For the counseling and psychotherapy contexts, the effort is the effort after understanding. For the medical and management contexts, the effort is problem-solving effort.

Some Methodological Matters

The methodological portion of my talk is designed in part both to comfort the afflicted and to afflict the comfortable. The afflicted are those of us who work in the softer, wilder areas of our field--the areas where the results seem ephemeral and unreplicable, and where the r^2's seem always to be approaching zero as a limit. These softer, wilder areas include those of social, personality, clinical, developmental, educational, organizational, and health psychology. They also include parts of psychobiology and cognitive psychology.

My message to those of us toiling in these muddy vineyards will be that we are doing better that we might have thought. My message to those of us in any areas in which we feel we have pretty well nailed things down will be that we haven't, and that we could be doing a whole lot better.

How Large Must an Effect Be, To Be Important?

There is a bit of good news-bad news abroad in the land. The good news is that more sophisticated editors, referees, and researchers are becoming aware that reporting the results of a significance test is not a sufficiently enlightening procedure to stand alone. More and more we are beginning to see a report of the magnitude of the effect accompanying the p level. The bad news is that we are still not

quite sure what to do with such a report of the magnitude of the effect, for example, a correlation coefficient.

There is one bit of training that all psychologists have undergone. From undergraduate days onward we have all been taught that there is only one proper, decent thing to do whenever we see a correlation coefficient--we must square it. For most of the softer, wilder areas of psychology, squaring the correlation coefficient tends to make it go away--vanish into nothingness as it were. That is one of the sources of malaise in the social and behavioral sciences. It is sad and quite unnecessary, as we shall soon see.

The Physician's Aspirin Study

At a special meeting held on December 18, 1987, it was decided to end prematurely, a randomized double blind experiment on the effects of aspirin on reducing heart attacks (Steering Committee of the Physicians' Health Study Research Group, 1988). The reason for this unusual termination of such an experiment was that it had become so clear that aspirin prevented heart attacks (and deaths from heart attacks) that it would be unethical to continue to give half the physician research subjects a placebo. Now what do you suppose was the magnitude of the experimental effect that was so dramatic as to call for the termination of this

research? Was r^2 .90 so that the corresponding r's would have been .95? No. Well, was r^2 50, .30, or even .20, so that the corresponding r's would have been .71, .55, or .45? No. Actually, what r^2 was, was .0011, with a corresponding r of .034.

Roughly 1 percent of the physicians taking aspirin compared to 2 percent of the physicians taking placebo suffered heart attacks. One way of showing the practical importance of even a small r is by means of a Binomial Effect Size Display (BESD). In such a display, the correlation is shown to be the simple difference in outcome rates between the experimental and the control groups in a standard table which always adds up to column totals of 100 and row totals of 100 (Rosenthal & Rubin, 1982b).

This type of result seen in the physicians' aspirin study is not at all unusual in biomedical research. Some years earlier, on October 29, 1981, the National Heart, Lung, and Blood Institute discontinued its placebo-controlled study of propranolol because results were so favorable to the treatment that it would be unethical to continue withholding the life-saving drug from the control patients. And what was the magnitude of this effect? Once again the effect size r was .04, and the leading digits of the r^2 were .00! As behavioral researchers we are not used to thinking of r's of .04 as reflecting effect sizes of practical importance. But when we think of an r of

.04 as reflecting a 4% decrease in heart attacks, the interpretation given *r* in a Binomial Effect Size Display, the *r* does not appear to be quite so small; especially if we can count ourselves among the 4 per 100 who manage to survive.

These results of biomedical studies are not flukes. For example, the correlation between alchohol abuse and having served in Vietnam is well-known, but the actual correlation is .07 (Centers for Disease Control, 1988). The effects of AZT on survival in treating AIDS are reflected in an *r* of .23 (Barnes, 1986), and the effects of cyclosporine in preventing the rejection of an organ transplant are associated with an *r* of .19 (Canadian Multicentre Transplant Study Group, 1983). The effects of psychotherapy associated with an *r* of .32 are larger than any of these biomedical relationships (Smith & Glass, 1977). Once we begin to think of the correlation coefficient as reflecting the difference in outcome rates between the experimental and control groups we begin to see that we are doing considerably better in our "softer, wilder" sciences than we may have thought we were doing (Rosenthal & Rubin, 1982).

So far, our conversation has been intended to comfort the afflicted. In what follows the intent is a bit more to afflict the comfortable. We consider, first, the topic of replication.

The Meaning of Successful Replication

There is a long tradition in psychology of our urging one another to replicate each other's research. But, although we have been very good at calling for replications we have not been very good at deciding when a replication has been successful. The issue we now address is: When shall a study be deemed successfully replicated?

Successful replication is ordinarily taken to mean that a null hypothesis that has been rejected at time 1 is rejected again, and with the same direction of outcome, on the basis of a new study at time 2. We have a failure to replicate when one study was significant and the other was not. Let us examine more closely a specific example of such a "failure to replicate."

Pseudo-Failures to Replicate

The saga of Smith and Jones. Smith has published the results of an experiment in which a certain treatment procedure was predicted to improve performance. She reported results significant at $p<.05$ in the predicted direction. Jones publishes a rebuttal to Smith claiming a failure to replicate. In situations of that sort it turns out often to be the case that, although Smith's results were more significant than Jones's, the studies were in quite good agreement as to their estimated sizes of effect

as defined either by Cohen's d [(Mean$_1$ – Mean$_2$) / σ] or by r, the correlation between group membership and performance score (Cohen, 1977; 1988; Rosenthal, 1984). Thus, studies labeled as "failures to replicate" often turn out to provide strong evidence for the replicability of the claimed effect.

On the odds against replicating significant results. A related error often found in the behavioral and social sciences is the implicit assumption that if an effect is "real," we should therefore expect it to be found significant again upon replication. Nothing could be further from the truth.

Suppose there is in nature a real effect with a true magnitude of $d = .50$ (i.e., [Mean$_1$ – Mean$_2$] / $\sigma = .50$ σ units), or, equivalently $r = .24$ (a difference in success rate of 62% versus 38%). Then suppose an investigator studies this effect with an N of 64 subjects or so, giving the researcher a level of statistical power of .50, a very common level of power for behavioral researchers of the last 30 years (Cohen, 1962; Sedlmeier & Gigerenzer, 1989). Even though a d of .50 or an r of .24 can reflect a very important effect (as we saw earlier in this paper), there is only one chance in four that both the original investigator and a replicator will get results significant at the .05 level. If there were two replications of the original study there would be only

one chance in eight that all three studies would be significant, even though we know the effect in nature is very real and very important.

Contrasting Views of Replication

The traditional, not very useful view of replication has two primary characteristics:

(1) It focuses on significance level as the relevant summary statistic of a study, and

(2) It makes its evaluation of whether replication has been successful in a dichotomous fashion. For example, replications are successful if both or neither $p < .05$ (or .01, etc.), and they are unsuccessful if one $p < .05$ (or .01, etc.) and the other $p > .05$ (or .01, etc.). Psychologists' reliance on a dichotomous decision procedure accompanied by an untenable discontinuity of credibility in results varying in p levels has been well documented (Nelson, Rosenthal, & Rosnow, 1986; Rosenthal & Gaito, 1963, 1964).

The newer, more useful views of replication success have two primary characteristics:

1. A focus on effect size as the more important summary statistic of a study with only a relatively minor interest in the statistical significance level, and

2. An evaluation of whether replication has been successful made in a continuous fashion. For example, two studies are not said to be successful or unsuccessful replicates of each other, but rather the degree of failure to replicate is specified.

Some Metrics of the Success of Replication

Differences between effect sizes. Once we adopt a view of the success of replication as a function of similarity of effect sizes obtained. we can become more precise in our assessments of the success of replication. Replication success could be indexed by the difference between the effect sizes obtained in the original study and in the replication. For example, we could employ the differences in Cohen's *d*'s or the effect size r's obtained. or we could employ Cohen's *q*, which is the difference between *r*'s that have been first transformed to Fisher's *Z*'s. Fisher's *Z* metric is distributed nearly normally and can thus be used in setting confidence intervals and testing hypotheses about *r*'s, whereas *r*'s distribution is skewed and the more so as the population value of *r* moves further from zero. Cohen's *q* is especially useful for testing the significance of difference between two obtained effect size *r*'s (Rosenthal. 1984; Rosenthal & Rubin, 1982a, Snedecor & Cochran, 1980). When there are more than two effect size *r*'s to be evaluated for their variability (i.e.. heterogeneity) we can simply compute the standard deviation (*S*) among the *r*'s or their Fisher *Z*

equivalents. If a test of significance of heterogeneity of these Fisher *Z*'s is desired, a simple χ^2 test of heterogeneity is readily available (Hedges, 1982; Rosenthal & Rubin, 1982a).

Meta-analytic metrics. As the number of replications for a given research question grows, a full assessment of the success of the replicational effort requires the application of meta-analytic procedures. An informative summary of the meta-analysis might be the stem-and-leaf display of the effect sizes found in the meta-analysis (Tukey, 1977). A more compact summary of the effect sizes might be Tukey's (1977) box plot, which gives the highest and lowest obtained effect sizes along with those found at the 25th, 50th, and 75th percentiles. For single index values of the consistency of the effect sizes, one could employ (a) the range of effect sizes found between the 75th (Q_3) and 25th (Q_1) percentile, (b) some standard fraction of that range (e.g., half or three-quarters), (c) *S*, the standard deviation of the effect sizes, or (d) SE, the standard error of the effect sizes.

As a slightly more complex index of the stability, replicability, or clarity of the average effect size found in the set of replicates, one could employ the mean effect size divided either by its standard error ($S/\sqrt{k}$ where k is the total number of

replicates), or simply by S. The latter index of mean effect size divided by its standard deviation (S) is the reciprocal of the coefficient of variation or a kind of coefficient of robustness.

What Should Be Reported?

Effect sizes and significance tests. If we are to take seriously our newer view of the meaning of the success of replications, what should be reported by authors of papers seen to be replications of earlier studies? Clearly, reporting the results of tests of significance will not be sufficient. The effect size of the replication and of the original study must be reported. It is not crucial which particular effect size is employed, but the same effect size should be reported for the replication and the original study. Complete discussions of various effect sizes and when they are useful are available from Cohen (1977, 1988) and elsewhere (e.g., Rosenthal, 1984). If the original study and its replication are reported in different effect size units these can usually be translated to one another (Cohen, 1977, 1988; Rosenthal, 1984; Rosenthal & Rosnow, 1984; Rosenthal & Rubin, in press).

Power. Especially if the results of either the original study or its replication were not significant, the statistical power at which the test of significance was made (assuming, for example, a population effect size equivalent to the effect size actually

obtained' should be reported (Cohen, 1988). In addition to reporting the statistical power for each study separately, it would be valuable to report the overall probability that both studies would have yielded significant results given, for example, the effect size estimated from the results of the original and the replication study combined.

The equally likely effect size. A marvelous suggestion has been made by Donald Rubin that would go a long way toward helping us get over our problem with the relative risks of type II versus type I errors. Don has suggested that whenever we conclude that there is "no effect" we report both the effect size and that confidence interval around the effect size that ranges from the effect size of zero to the equally likely effect size greater than the one we obtained. For example, suppose a replicator, Jones, did not reject the null but obtained an effect size of $d = .50$. If Jones had been required to report that his d of.50 was just as close to a d of 1.00 as it was to a d of zero, Jones would have been less likely to draw his wrong conclusion that he had failed to replicate Smith's work who had found a very similar effect size.

Meta-Analytic Procedures: Some Benefits

Any discussion of replication and of the evaluation of the success of a particular replication cannot avoid a more formal consideration of meta-analytic procedures.

In the years 1980, 1981, and 1982 alone, well over 300 papers were published on the topic of meta-analysis (Lamb and Whitla, 1983). Does this represent a giant stride forward in the development of the behavioral and social sciences or does it signal a lemming-like flight to disaster? Judging from reactions to past meta-analytic enterprises, there are at least some who take the more pessimistic view. Some three dozen scholars were invited to respond to a meta-analysis of studies of interpersonal expectancy effects conducted by Don Rubin and myself (Rosenthal & Rubin, 1978). Although much of the commentary dealt with the substantive topic of interpersonal expectancy effects, a good deal of it dealt with methodological aspects of meta-analytic procedures and products. Some of the criticisms offered were accurately anticipated by Glass (1978) who had earlier received commentary on his meta-analytic work (Glass, 1976) and that of his colleagues (Smith & Glass, 1977; Glass, McGaw, & Smith, 1981). These criticisms have been detailed and addressed elsewhere (Rosenthal, 1989). Today, therefore, I want to use the time that remains to note a number of special benefits of meta-analysis. Some of these benefits are well known, but some are not--indeed, some are most arcane.

Most Obvious Benefits

Completeness. Meta-analytic consideration of a research domain is more complete and exhaustive though this does *not* mean that all studies found are weighted equally. Indeed, every study should be weighted from zero to any desired number. These weights, of course, must be defensible. (It will not do to weight all my results +1.00 and all my enemies' results 0.00).

Explicitness. The quantitative nature of the process of obtaining effect sizes, standard normal deviates, and weights, forces explicitness on the analyst. Vague terms like "no relationship," "some relationship," a "strong relationship," "very significant," are replaced by numerical values.

Power. Empirical work has shown that meta-analytic procedures increase power and decrease type 2 errors (Cooper & Rosenthal, 1980).

Less Obvious Benefits

Moderator variables. These are more easily spotted and evaluated in a context of a quantitative research summary. This aids theory development and increases empirical richness.

Cumulation problems. Meta-analytic procedures address, in part, the chronic complaint that social sciences cumulate so poorly compared to the physical sciences.

It should be noted that recent historical and sociological investigations have suggested that the physical sciences may not be all that much better off than we are when it comes to successful replication (Collins, 1985; Hedges, 1987; Pool 1988). For example, Collins (1985) has described the failures to replicate the construction of TEA-lasers despite the availability of detailed instructions for replication. Apparently TEA-lasers could be replicated dependably only when the replication instructions were accompanied by a scientist who had actually built a laser.

Least Obvious Benefits

Decrease in overemphasis on single studies. One not so obvious benefit that will accrue to us is the gradual decrease in the overemphasis on the results of a single study. There are good sociological grounds for our monomaniacal preoccupation with the results of a single study. Those grounds have to do with the reward system of science where recognition, promotion, reputation, and the like depend on the results of the single study, also known as the smallest unit of academic currency. The study is "good," "valuable," and above all, "publishable" when $p \leq .05$. Our disciplines would be further ahead if we adopted a more cumulative view of science in which the impact of a study were evaluated less on the basis of p levels, and more on the basis of its own effect size and on the revised effect size and combined probability

that resulted from the addition of the new study to any earlier studies investigating the same or a similar relationship. This, of course, amounts to a call for a more meta-analytic view of "doing science."

B. F .Skinner has been eloquent in his comments on the overvaluation of the single study: "In my own thinking, I try to avoid the kind of fraudulent significance which comes with grandiose terms or profound 'principles.' But some psychologists seem to need to feel that every experiment they do demands a sweeping reorganization of psychology as a whole. It's not worth publishing unless it has some such significance. But research has its own values, and you don't need to cook up spurious reasons why it's important." (Skinner, 1983, p. 39).

"The new intimacy." This new intimacy is between the reviewer and the data. We cannot do a meta-analysis by reading abstracts and discussion sections. We are forced to look at the numbers and, very often, compute the correct ones ourselves. Meta-analysis requires us to cumulate *data*, not *conclusions*. "Reading" a paper is quite a different matter when we need to compute an effect size and a fairly precise significance level--often from a results section that never heard of effect sizes, precise significance levels (or the APA publication manual)!

The demise of the dichotomous significance testing decision. Far more than is good for us, social and behavioral scientists operate under a dichotomous null hypothesis decision procedure in which the evidence is interpreted as anti-null if $p <$ .05 and pro-null if $p >$ .05. If our dissertation p is $<$ .05 it means joy, a Ph.D., and a tenure-track position at a major university. If our p is $>$ .05 it means ruin, despair, and our advisor's suddenly thinking of a new control condition that should be run. That attitude really must go. God loves the .06 nearly as much as the .05. Indeed, I have it on good authority that she views the strength of evidence for or against the null as a fairly continuous function of the magnitude of p. As a matter of fact, two .06 results are much stronger evidence against the null than one .05 result; and 10 p's of .10 are stronger evidence against the null than 5 p's of .05.

The overthrow of the omnibus test. It is common to find specific questions addressed by F tests with $df > 1$ in the numerator or by χ^2 tests with $df > 1$. For example, suppose the specific question is whether increased incentive level improves the productivity of work groups. We employ four levels of incentive so that our omnibus F test would have 3 df in the numerator or our omnibus χ^2 would be on at least 3 df. Common as these tests are, they reflect poorly on our teaching of data analytic procedures. The diffuse hypothesis tested by these omnibus tests usually

tells us nothing of importance about our research question. The rule of thumb is unambiguous: Whenever we have tested a fixed effect with $df > 1$ for χ^2 or for the numerator of F, we have tested a question in which we are almost surely not interested.

The situation is even worse when there are several dependent variables as well as multiple *df* for the independent variable. The paradigm case here is canonical correlation and special cases are MANOVA, MANCOVA, Multiple discriminant function, multiple path analysis, and complex multiple partial correlation. While all of these procedures have useful exploratory data analytic applications they are commonly used to test null hypotheses which are scientifically almost always of doubtful value. The effect size estimates they yield (e.g., the canonical correlation) are also almost always of doubtful value.

This is not the place to go into detail, but one approach to the problem of analyzing canonical data structures is to reduce the set of dependent variables to some smaller number of composite variables using the principal-components-followed-by-unit-weighting approach. Each composite can then be analyzed serially.

Meta-analytic questions are basically contrast questions. F tests with $df > 1$ in the numerator or χ^2's with $df > 1$ are useless in meta-analytic work. That leads to an additional scientific benefit:

The increased recognition of contrast analysis. Meta-analytic questions require precise formulation of questions and contrasts are procedures for obtaining answers to such questions, often in an analysis of variance or table analysis context. Although most textbooks of statistics describe the logic and the machinery of contrast analyses, one still sees contrasts employed all too rarely. That is a real pity given the precision of thought and theory they encourage and (especially relevant to these times of publication pressure) given the boost in power conferred with the resulting increase in .05 asterisks (Rosenthal & Rosnow, 1985).

A probable increase in the accurate understanding of interaction effects. Probably the universally most misinterpreted empirical results in psychology are the results of interaction effects. A recent survey of 191 research articles involving interactions found only two articles that showed the authors interpreting interactions in an unequivocally correct manner (i.e., by examining the residuals that define the interaction) (Rosnow & Rosenthal, 1989). The rest of the articles simply compared means of conditions with other means, a procedure that does not

investigate interaction effects but rather the sum of main effects and interaction effects.

Most standard textbooks of statistics for psychologists provide accurate mathematical definitions of interaction effects but then interpret not the residuals that define those interactions but the means of cells that are the sums of all main effects and all interactions.

In addition, users of SPSS, SAS, BMDP, and virtually all other data-analytic software are poorly served in the matter of interactions since virtually no programs provide convenient tabular output giving the residuals defining interaction. The only exception to that of which I am aware is a little-known package called Data-Text developed by Arthur Couch and David Armor for which William Cochran and Donald Rubin provided the statistical consultation.

Since many meta-analytic questions are by nature questions of interaction (for example, that opposite sex dyads will conduct standard transactions more slowly than will same sex dyads), we can be hopeful that increased use of meta-analytic procedures will bring with it increased sophistication about the meaning of interaction.

Meta-analytic procedures are applicable beyond meta-analyses. Many of the techniques of contrast analyses among effect sizes, for example, can be used within a single study (Rosenthal & Rosnow, 1985). Computing a single effect size from correlated dependent variables, or comparing treatment effects on two or more dependent variables serve as illustrations (Rosenthal & Rubin, 1986).

The decrease in the splendid detachment of the full professor. Meta-analytic work requires careful reading of research and moderate data analytic skills. We cannot send an undergraduate research assistant to the library with a stack of 5×8 cards to bring us back "the results." With narrative reviews that seems often to have been done. With meta-analysis the reviewer must get involved with the actual data and that is all to the good.

Conclusion

I hope that the methodological section of this paper has provided some comfort to the afflicted in showing that many of the findings of our discipline are neither as small nor as unimportant from a practical point of view as we may have feared. Perhaps I hope, too, that there may have been some affliction of the comfortable in showing that in our views of replication and of the cumulation of the wisdom of our field there is much yet remaining to be done.

References

Barnes, D. M. (1986). Promising results halt trial of anti-AIDS drug. *Science, 234*. 15-16.

Bateson, G., Jackson, D. D., Haley, J., & Weakland, J. (1956). Toward a theory of schizophrenia. *Behavioral Science, 1*. 251-264.

Bernieri, F. J., Reznick, J. S., & Rosenthal, R. (1988). Synchrony, pseudosynchrony, and dissynchrony. *Journal of Personality and Social Psychology, 54*, 243-253.

Bernieri, F. J., & Rosenthal, R. (In press). Interpersonal coordination: Behavior matching and interactional synchrony. In R. S. Feldman & B. Rimé (Eds.), *Fundamentals of Nonverbal Behavior*. NY: Cambridge University Press.

Blanck, P. D., Buck, R., & Rosenthal, R. (Eds.)(1986). *Nonverbal Communication in the Clinical Context*. University Park, PA: Pennsylvania State University Press.

Buck, R. (1984). *The Communication of Emotion*. NY: Guilford Press.

Bugental, D. E., Love, L. R., & Gianetto, R. M. (1971). Perfidious feminine faces. *Journal of Personality and Social Psychology, 17*, 314-318.

Canadian Multicentre Transplant Study Group. (1983). A randomized clinical trial of cyclosporine in cadaveric renal transplantation. *New England Journal of Medicine, 309*, 809-815.

Centers for Disease Control Vietnam Experience Study. (1988). Health status of Vietnam veterans: 1. Psychosocial characteristics. *Journal of the American Medical Association, 259*, 2701-2707.

Cohen, J. (1962). the statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology, 65*, 145-153.

Cohen, J. (1977). *Statistical Power Analysis for the Behavioral Sciences* (rev. ed.), New York: Academic Press.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed., Hillsdale, NJ: Lawrence Erlbaum Associates.

Collins, H. M. (1985). *Changing Order: Replication and Induction in Scientific Practice.* Beverly Hills, CA: Sage.

Cooper, H. M., & Rosenthal, R. (1980). Statistical versus traditional procedures for summarizing research findings. *Psychological Bulletin, 87*, 442-449.

DePaulo, B. M., & Rosenthal, R. (1979a). Ambivalence, discrepancy, and deception in nonverbal communication. In R. Rosenthal (Ed.), *Skill in Nonverbal Communication*. Cambridge, MA: Oelgeschlager, Gunn, & Hain, 204-248.

DePaulo, B. M., & Rosenthal, R. (1979b). Telling Lies. *Journal of Personality and Social Psychology*, *37*, 1713-1722.

Ekman, P. (Ed.) (1973). *Darwin and Facial Expression*. New York: Academic Press.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*, 3-8.

Glass, G. V. (1978). In defense of generalization. *The Behavioral and Brain Sciences*, *3*, 394-395.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-Analysis in Social Research*. Beverly Hills, CA: Sage.

Harris, M. J., & Rosenthal, R. (1985). The mediation of interpersonal expectancy effects: 31 meta-analyses. *Psychological Bulletin*, *97*, 363-386.

Hedges, L. V. (1982). Estimation of effect size from a series of independent experiments. *Psychological Bulletin*, *92*, 490-499.

Hedges, L. V. (1987). How hard is hard science, how soft is soft science? *American Psychologist, 42*, 443-455.

Jussim, L. (1986). Self-fulfilling prophecies: A theoretical and integrative review. *Psychological Review, 93*, 429-445.

Lamb, W. K., & Whitla, D. K. (1983). *Meta-Analysis and the Integration of Research Findings: A Trend Analysis and Bibliography Prior to 1983.* Unpublished manuscript, Harvard University, Cambridge.

Lazarus, R. S. (1984). On the primacy of cognition. *American Psychologist, 39*, 124-129.

Nelson, N., Rosenthal, R., & Rosnow, R. L. (1986). Interpretation of significance levels and effect sizes by psychological researchers. *American Psychologist, 41*, 1299-1301.

Pool, R. (1988). Similar experiments, dissimilar results. *Science, 242*, 192-193.

Rosenthal, R. (1966). *Experimenter Effects in Behavioral Research.* New York: Appleton-Century-Crofts.

Rosenthal, R. (1976). *Experimenter Effects in Behavioral Research.* (Enlarged ed.) New York: Irvington.

Rosenthal, R. (1981). Pavlov's mice, Pfungst's horse, and Pygmalion's PONS: Some models for the study of interpersonal expectancy effects. In T. A. Sebeok & R. Rosenthal (Eds.), *The Clever Hans Phenomenon: Communication with Horses, Whales, Apes, and People. Annals of the New York Academy of Sciences*, No. 364, 182-198.

Rosenthal, R. (1984). *Meta-Analytic Procedures for Social Research*. Beverly Hills, CA: Sage.

Rosenthal, R. (1987). *Judgment Studies: Design, Analysis, and Meta-Analysis*. New York: Cambridge University Press.

Rosenthal, R. (1989, April). *Research: How Are We Doing?* Distinguished Lecture presented at the meeting of the Eastern Psychological Association, Boston, MA.

Rosenthal, R., & Gaito, J. (1963). The interpretation of levels of significance by psychological researchers. *Journal of Psychology, 55*, 33-38.

Rosenthal, R., & Gaito, J. (1964). Further evidence for the cliff effect in the interpretation of levels of significance. *Psychological Reports, 15*, 570.

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

Rosenthal, R., & Rosnow, R. L. (1984). *Essentials of Behavioral Research: Methods and Data Analysis*. New York: McGraw-Hill.

Rosenthal, R., & Rosnow, R. L. (1985). *Contrast Analysis: Focused Comparisons in the Analysis of Variance*. New York: Cambridge University Press.

Rosenthal, R., & Rubin, D. B. (1978). Interpersonal expectancy effects: The first 345 studies. *The Behavioral and Brain Sciences, 3*, 377-386.

Rosenthal, R., & Rubin, D. B. (1982a). Comparing effect sizes of independent studies. *Psychological Bulletin, 92*, 500-504.

Rosenthal, R., & Rubin, D. B. (1982b). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology, 74*, 166-169.

Rosenthal, R., & Rubin, D. B. (1986). Meta-analytic procedures for combining studies with multiple effect sizes. *Psychological Bulletin, 99*, 400-406.

Rosenthal, R., & Rubin, D. B. (in press). Effect size estimation for one-sample multiple-choice-type data: Design, analysis, and meta-analysis. *Psychological Bulletin*.

Rosnow, R. L., & Rosenthal, R. (1989). Definition and interpretation of interaction effects. *Psychological Bulletin, 105*, 143-146.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105*, 309-316.

Skinner, B. F. (1983, August). On the value of research. *APA Monitor*, p. 39.

Smith, M. L., & Glass, G. V (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 32*, 752-760.

Snedecor, G. W., & Cochran, W. G. (1980). *Statistical Methods* (7th ed.). Ames: Iowa State University Press.

Steering Committee of the Physicians Health Study Research Group. (1988). Preliminary report: Findings from the aspirin component of the ongoing physicians' health study. *The New England Journal of Medicine, 318*, 262-264.

Swann, W. B., Jr., & Snyder, M. (1980). On translating beliefs into action: Theories of ability and their application in an instructional setting. *Journal of Personality and Social Psychology, 38*, 879-888.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

Zajonc, R. B. (1984). On the primacy of affect. *American Psychologist, 39*, 117-123.

Author Notes

This paper was presented as the Donald T. Campbell Address at the Annual Meeting of the American Psychological Association, New Orleans, August 14, 1989. Preparation of this paper was supported in part by the National Science Foundation and in part by a Fellowship at the Center for Advanced Study in the Behavioral Sciences. I am grateful for financial support provided by the John D. and Catherine T. MacArthur Foundation, and for improvements suggested by Deanna Knickerbocker.

Appendix

I. The Problem

Oh, F is large and p is small
That's why we are walking tall.

What it means we need not mull
Just so we reject the null.

Or Chi-Square large and p near nil
Results like that, they fill the bill.

What if meaning requires a poll?
Never mind, we're on a roll!

The message we have learned too well?
Significance! That rings the bell!

II. The Implications

The moral of our little tale?
That we mortals may be frail
When we feel a p near zero
Makes us out to be a hero.

But tell us then is it too late?
Can we perhaps avoid our fate?
Replace that wish to null-reject
Report the size of the effect.

That may not insure our glory
But at least it tells a story
That is just the kind of yield
Needed to advance our field.