The Stability of Behavior

II. Implications for Psychological Research

SEYMOUR EPSTEIN *University of Massachusetts—Amherst*

ABSTRACT: *Psychological research is rapidly approaching a crisis as the result of extremely inefficient procedures for establishing replicable generalizations. The traditional solution of attempting to obtain a high degree of control in the laboratory is often ineffective because much human behavior is so sensitive to incidental sources of stimulation that adequate control cannot be achieved. An alternative proposal—to investigate higher order interactions—has been no more successful because the number of relevant interactions is often beyond the capacity of experimental investigation. One solution lies in aggregating behavior over situations and/or occasions, thereby canceling out incidental, uncontrollable factors relative to experimental factors. Since such a procedure increases reliability without introducing excessive constraints into the experimental situation, it contributes to the generality as well as the replicability of findings. The value of such aggregation was demonstrated in five studies that examined a variety of data, including the subjective and objective measurement of behavior in the field and in the laboratory. Four kinds of aggregation are discussed, each of which reduces a specific source of error. The degree of aggregation that is required varies inversely with the degree to which the events studied are ego-involving, implicitly or explicitly include an adequate sample of behavioral observation, or have been demonstrated to be robust over incidental sources of variation.*

Some years ago, Koch (1959) concluded that there was a resistance of psychological findings to empirical generalization. Unfortunately, the situation has not improved today. It is becoming increasingly evident that we are rapidly approaching a crisis in research associated with extremely inefficient procedures for establishing reliable generalizations. Not only are experimental findings often difficult to replicate when there are the slightest alterations in conditions, but even attempts at exact replication frequently fail.[1]

The nature of the problem is elucidated, and at least one direction toward a solution suggested, in a series of studies on stability of behavior reported in a previous article (Epstein, 1979c). That article is concerned with the issue of stability in personality, one of the classic debates in psychology. It will be helpful to review briefly the issue and the studies reported in the article before proceeding further.

Mischel, the current leading proponent of the antitrait position, observed that when objectively measured behavior in one situation is correlated with objectively measured behavior in another situation or with scores on a personality inventory, the correlations are almost invariably below .30. This led him to dub such correlations "personality

---

The preparation of this article and the research reported in it were supported by National Science Foundation Grant BNS 78-12336 and National Institute of Mental Health Grant MH-01293. A preliminary version of the article was circulated in May of 1977.

I wish to acknowledge the helpful comments of the following individuals: James Averill, Seymour Berger, Jeffrey M. Delman, Alice Eagly, Donald W. Fiske, Anthony G. Greenwald, Howard Leventhal, Debbie Moskowitz, Jerome Myers, William Schneiderman, and L. Alan Sroufe.

Requests for reprints should be sent to Seymour Epstein, Department of Psychology, University of Massachusetts, Amherst, Massachusetts 01003.

[1] Lest I be misunderstood at the outset, let me foreshadow some of my arguments and note that I do not claim there are no reliable generalizations in psychology. Everyone can probably identify a few to his or her satisfaction. My position is that the overall yield of meaningful information, particularly information that is cumulative, is discouragingly low considering the total amount of research that has been conducted to date. The difficulty with the typical laboratory experiment is not that it cannot yield meaningful generalizations, but that there is no way of establishing within the confines of the experiment that it is likely to have done so. At the same time, there is an unfortunate dearth of replication studies. Moreover, as is demonstrated shortly, the very nature of the paradigm of the single-session experiment is such that very few findings, no matter what their level of statistical significance, are apt to be replicable. Further, in the event that a result is replicable, there is little likelihood that it will be sufficiently general across minor variations in stimulus conditions to identify scientifically useful relationships.

Vol. 35, No. 9, 790–806

0003-066X/80/3509-0790$00.75

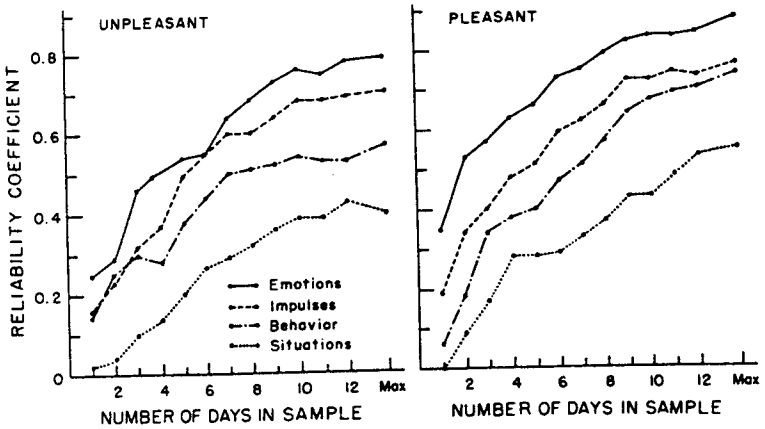

Figure 1. Reliability coefficients in Study 1 (Epstein, 1979c) as a function of the number of days in odd–even samples. The data consisted of daily ratings of the strongest pleasant and unpleasant emotion of the day, of the stimulus variables that gave rise to the emotion (e.g., love and affection), of the response tendencies associated with the emotion (e.g., nurturance), and of whether the response tendency was carried out in behavior; values plotted represent the mean of the correlations for the variables in a category. (From "Traits Are Alive and Well" by S. Epstein. In D. Magnusson & N. S. Endler [Eds.], *Personality at the Crossroads: Current Issues in Interactional Psychology*. Copyright 1977 by Lawrence Erlbaum Associates, Inc. Reprinted by permission.)

coefficients" and to conclude, "With the possible exception of intelligence, highly generalized behavioral consistencies have not been demonstrated, and the concept of personality traits as broad predispositions is thus untenable" (Mischel, 1968, p. 146). He further stated, "I am more and more convinced, however, hopefully by data as well as on theoretical grounds, that the observed inconsistency so regularly found in studies on noncognitive personality dimensions often reflects the state of nature and not merely the noise of measurement" (Mischel, 1969, p. 1014). The finding that personality scales and other self-report procedures are stable, but that objective measures are not, has been interpreted by those in the antitrait camp to indicate that stability of personality lies primarily in the eye of the beholder. Mischel (1969) presented his views on this matter as follows: "How does one reconcile our shared perception of continuity with the equally impressive evidence that on virtually all of our dispositional measures of personality substantial changes occur in the characteristics of the individual longitudinally over time and, even more dramatically, across seemingly similar settings cross-sectionally" (p. 1012). His answer was that the human mind "creates and maintains the perception of continuity even in the face of perpetual observed changes in actual behavior. Often this cognitive construction of continuity, while not arbitrary, is only very tenuously related to the phenomena that are construed" (p. 1012).

An alternative possibility that Mischel and others failed to consider is that the failure of objective events to correlate with each other and with personality scales was often the result of inadequate sampling of the objective events and, therefore, *could* be attributed to the noise of measurement. In most cases the objective data consisted of single behavioral observations, usually obtained in the laboratory. Single items of behavior, no matter how objectively measured, can be expected, like single items in a test, to be low in reliability and to lack generality and therefore to be inadequate to the task of demonstrating stability in behavior and of measuring personality traits. Such reasoning led to the formulation of the following hypothesis and corollary:

Hypothesis: Stability can be demonstrated over a wide range of variables so long as the behavior in question is averaged over a sufficient number of occurrences.

Corollary: Reliable relationships can be demonstrated between ratings by others and self-ratings, including standard personality inventories, on the one hand, and objective behavior, on the other, so long as the objective behavior is sampled over an appropriate level of generality and averaged over a sufficient number of occurrences.

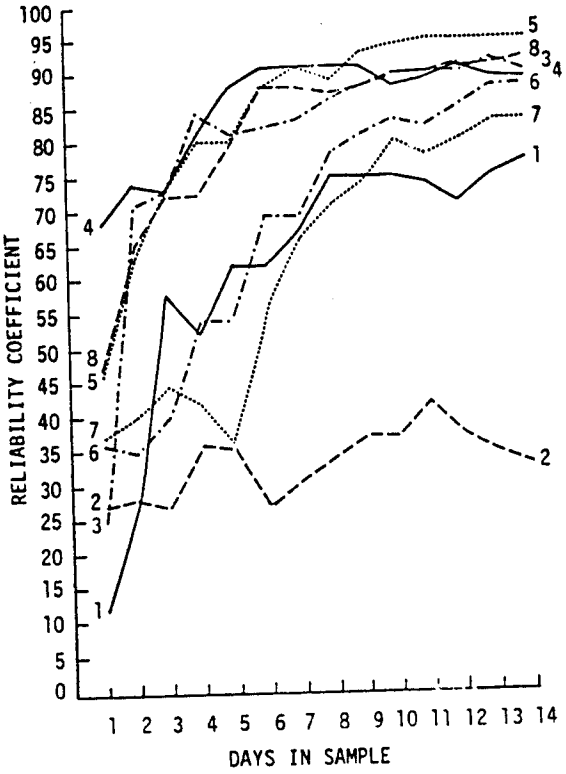

Figure 2. Reliability coefficients for each of the variables in Study 2 (Epstein, 1979c) as a function of the number of days in the odd–even samples. (The data plotted consist of the ratings by observers of a subject's daily behavior on eight variables related to social and impulsive behavior [e.g., "She actively sought out the company of others (a) never, (b) once or twice, (c) three or more times"]. It can be seen in the figure that not until a 3- or 4-day sample was averaged did the reliability coefficients become sufficiently stable to retain their final rank order.)

To test the hypothesis and corollary, I conducted four studies (Epstein, 1979c), each of which sampled behavior on repeated occasions over a period of weeks. In all of the studies, a single behavioral observation was treated as equivalent to a single response on a test, and odd–even stability coefficients were computed for the data averaged over different numbers of observations. More specifically, stability coefficients were first obtained by correlating behavior on Day 1 with behavior for the same variable on Day 2. Next, stability coefficients were obtained by correlating behavior averaged over Days 1 and 3 with behavior averaged over Days 2 and 4 and so on, until each behavioral item averaged over all odd days was correlated with the same behavioral item averaged over all even days. This was done separately for each variable in each study, resulting in a large number of stability coefficients representing a wide variety of different kinds of data. The data in the four studies consisted of self-ratings, ratings by others, objectively measured discrete items of behavior, psychophysiological measures such as heart rate, and responses to standard and specially constructed personality inventories. The results provided unequivocal support for the hypothesis and corollary. For all kinds of data, responses to a single event, with rare exception, produced low stability co-

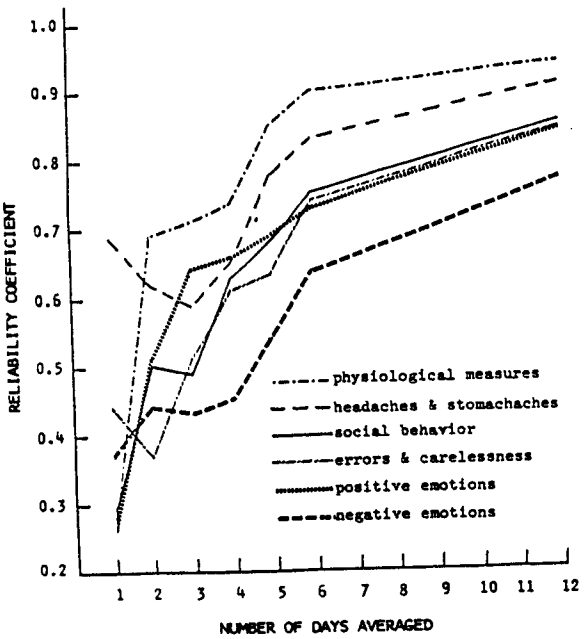

Figure 3. Reliability coefficients for the variables in Study 3 (Epstein, 1979c) as a function of the number of days in the odd–even samples. (The data consisted of daily self-ratings of positive and negative emotions recorded in the classroom; of measures of heart-rate mean and range; of records of discrete objective events of social significance, such as number of social letters written and received; and of objective behavioral events recorded by the instructor, such as number of erasures and omissions made in filling out the answer sheets, number of minutes late to class, and number of times a Number 2 pencil was forgotten. Values plotted represent the mean of the correlations for the variables in a category. The values for the 12-day samples were estimated from the 6-day samples by the Spearman-Brown formula after establishing that predictions from a 3-day sample accurately represented the findings for a 6-day sample. It can be seen in the figure that accurate estimates could not have been made from a 1-day sample, as the data from a 1-day sample were too unstable to reproduce the rank order of the more inclusive data samples.)

efficients, usually below .30. As the data were veraged over an increasing number of events, stability coefficients rose, often to over .80 and sometimes to over .90. The results of the first three studies are summarized in Figures 1 through 3. One can see in Figure 3 that high correlations were not specific to subjective ratings, thereby putting to rest concern about whether stability lies only in the eye of the rater.

Once adequate stability, or temporal reliability, coefficients were obtained, evidence for validity emerged in the form of statistically significant relationships among variables. This was particularly evident in the fourth study, which demonstrated statistically significant and coherent correlations between personality inventories and objective measures aggregated over occasions. There were few statistically significant relationships when the data consisted of single observations, and among these was a lack of coherence. Analysis of the standard deviations of the variables indicated that as the data were increasingly aggregated, the standard deviations decreased, demonstrating that the high standard deviations from the single observations were the result of error of measurement, that is, of transient factors unrelated to the personalities of the individuals.[2]

Thus, the findings from these four studies support the conclusion that the reason previous attempts failed to establish the existence of traits when the criterion consisted of objective behavior measured in the laboratory may simply be that inadequate samples of behavior were obtained in the laboratory. As a result, the laboratory behavior itself was unreliable and therefore incapable of establishing strong relationships. Behavior observed on a single occasion, whether in the laboratory or elsewhere, is apt to be so situationally unique as to be incapable of establishing reliable generalizations that hold over even the most minor variations in the situation. To the extent that this is true, it follows that the laboratory study as normally conducted is often unreplicable. This, of course, should ultimately lead to a crisis in psychological research.[3]

The Crisis in Psychological Research

Greenwald (1976), from his vantage point as editor of the *Journal of Personality and Social Psychology*, recently concluded that there is a "crisis in personality and social psychology, associated with the difficulty often experienced by researchers in attempting to replicate published work" (p. 2). In an earlier article (Greenwald, 1975), he had observed that the same problem very likely exists in all areas of psychology, with the possible exception of psychophysics, neurophysiology, and operant behavior. Lykken (1968), in a discussion of factors that complicate the interpretation of statistical significance, concluded that exact replication of psychological experiments is usually impossible and that other types of replications, which are preferable, face serious problems, most of which are generally ignored. Cronbach, in a presidential address to the American Psychological Association in 1957, predicted a bright future for nomothetic psychological research, so long as sufficient attention was paid to higher order interactions. By 1975 he was less sanguine: "Taking stock today, I think most of us judge theoretical progress to have been disappointing. Many are uneasy with the intellectual style of psychological research. . . . I shall express some pessimism about our predominant norms and strategies and offer tentative thoughts about an alternative style of work" (Cronbach, 1975, p. 116). He offered two alternatives. One was to sample behavior directly in order to make situation-specific predictions for practical purposes.

---

[2] What is treated as error variance varies, of course, with the purpose of an experiment. When interest lies in establishing stable individual differences, situational variability is regarded as error variance, whereas when interest is in the influence of stimulus conditions, individual differences are regarded as error variance.

[3] It should be noted that much of the data presented here and more thoroughly in my first article (Epstein, 1979c) are relevant to cross-situational stability as well as to temporal stability. In my previous article I observed that "as the stability that was demonstrated occurred over the normal range of situational variability in everyday life, it would seem fair to state that a meaningful level of cross-situational stability was demonstrated. Expressed otherwise, it was demonstrated that there is enough cross-situational stability in everyday life so that useful statements about individual behavior can be made without having to specify the eliciting situations. This, of course, is the way a trait is usually defined, and the findings demonstrate the utility of such a concept" (Epstein, 1979c, p. 1122). Despite this conclusion and additional evidence cited from the work of others as well as from my findings in support of the existence of broad traits such as honesty, aggression, and sociability, Mischel (1979) dismissed the relevance of the Epstein (1979c) article for cross-situational stability with the statement that "most of the data provided so far again speak to the issue of temporal stability, not broad cross-situational consistency in social behavior" (p. 742). The unwary reader might be misled into believing that because "most of the data" were devoted to the issue of temporal stability, a considerable amount of evidence was not also presented to document the existence of "broad cross-situational consistency."

The other was to employ psychological laws as general guidelines and to weight the contribution of additional sources of influence intuitively, more in the manner of an artist than of a scientist. Unfortunately, Cronbach did not consider how to obtain the general laws given the problems he had observed as a result of the prevalence of higher order interactions.

Two Commonly Proposed Solutions and Why They Often Fail to Work

Achieving high levels of experimental control. The dominant paradigm in psychological research consists of conducting laboratory experiments in which independent variables are manipulated and the effects on dependent variables observed, while all relevant incidental sources of influence are presumably controlled. At face value, it seems that this model should be as successful for the social sciences as it has been for the physical sciences. Yet, there are serious problems in the establishment of adequate control in the social sciences, despite its ready attainment in the physical sciences. Why should this be the case? First, one must consider that adequate control in the social sciences is often impossible because of the extreme degree of situational specificity of much human behavior. No matter what one controls, it is highly likely that variables which were considered unimportant, or were undetected, could influence the outcome. It is instructive, in this respect, to examine the variables that Campbell and Stanley (1966) list as inadequately controlled in different experimental designs. The list is so extensive, and the nature of many of the variables so subtle, that one is forced to the conclusion that no experimental design is adequate to the task of adequately controlling all potential sources of incidental influence.

The problem can be elucidated from the following vantage point. Consider that in many psychological experiments the potency of the experimental manipulation, or the ego-involvement of the subjects, is relatively low and exerts no greater influence than unrecognized factors unique to the experimental situation. Consider further that the achievement of a small error term by controlling the variables that can be controlled magnifies the contribution of uncontrolled situation-specific variables no less than that of experimental variables. As a result, highly significant situation-specific effects that lack generality and, similarly, replicability, are apt to be found. Thus, increasing control often increases the problem control seeks to reduce. I explore this problem in greater detail later. A further problem associated with increased control is that even if the results are not complicated by unrecognized situation-specific variables, the more behavior is constrained by controls, the narrower the range of generalization of the findings (Campbell & Stanley, 1966; Lykken, 1968; Tryon, 1973). As a result, the findings are apt to be of insufficient generality to be scientifically useful.

Investigating higher order interactions. A second solution that has been proposed for increasing the reliable prediction of behavior is to study interactions (Alker, 1972; Bowers, 1973; Cronbach, 1957; Endler & Hunt, 1968; Magnusson & Endler, 1977; Mischel, 1973). The argument is that since behavior is a joint function of the person and the stimulus, both sets of variables must be included in the same experimental design, and predictions must be made for persons within the constraints of relevant properties of stimuli. As already noted, Cronbach (1957) initially believed that reasonable predictability of human behavior could be achieved by the investigation of higher order interactions. After attempting such a solution, he came to the following conclusion: "Once we attend to interactions, we enter a hall of mirrors that extends to infinity. However far we carry our analysis—to third order or fifth order or any other —untested interactions of a still higher order can be envisioned" (Cronbach, 1957, p. 119). He noted that the problem is a highly general one in psychological research and cited examples from cognitive psychology, animal experimentation, human learning, social psychology, and personality research. He noted that if there are seven important variables in experiments on voluntary delay, as has been claimed, "those seven variables can give rise to 120 interactions, a number beyond the reach of a direct experiment" (Cronbach, 1957, p. 120). The prevalence of higher order interactions attests to the extreme situational specificity of behavior and indicates that experiments conducted in a single situation cannot be relied on to produce laws that are general across minor variations in stimulus conditions. It follows that the single-session experiment is often unreplicable because incidental, unidentified variables, such as how subjects were recruited and sampled, the personality and appearance of the experimenter, and the particular time and location at which the study was conducted, may interact with an experimental

manipulation to produce a unique outcome or may otherwise restrict its generality.

An Alternative Solution: Aggregation

Given the widespread failure of increases in experimental control and the study of higher order interactions to produce reliable generalizations, where does one turn next? As demonstrated in the four studies referred to above, although the prediction of specific behavioral acts in single situations is often unattainable, the prediction of behavior averaged over a sample of situations and/or occasions is often attainable. In fact, given some degree of relationship to begin with, the law of sampling distributions informs us that any level of prediction can be achieved as long as there is sufficient aggregation of appropriate data. Strangely enough, this fundamental law of sampling statistics has been widely ignored with respect to the sampling of stimuli and occasions in laboratory studies, although it has been well recognized with respect to the sampling of individuals. Every beginning psychology student knows that findings from a single individual should not be generalized to other people. By what logic does one then assume that findings from a single stimulus situation or from a restricted range of stimuli can yield generalizations that hold for other stimuli and stimulus situations? The logic of being able to predict the mean of sampling distributions better than that of individual scores obviously applies as well to samples of stimuli and stimulus situations as to samples of individuals. As already noted, an effective procedure for reducing the situational specificity of findings, and thereby increasing their replicability and generalizability, is to conduct experiments in which data are averaged over stimuli, situations, and/or occasions. In addition to increasing the generality and replicability of findings, a procedure in which behavior is sampled over stimuli and occasions has the virtue of permitting replicability and generalizability to be assessed and to then be taken into account in interpreting the findings from a particular study and in planning further studies.

The view that behavior aggregated over stimuli and/or occasions, unlike single instances of behavior, can provide an adequate basis for establishing reliable generalizations is not a new one. Brunswik (1947, 1956) observed many years ago that it is no more defensible to generalize from a single stimulus to stimuli in general than it is to generalize from a single subject to subjects in general. Hammond (1954, 1955) and more recently Pervin (1968, 1977) have repeated this message with reference, respectively, to research in clinical psychology and research in personality. Tryon (1973), after performing a series of studies comparing the predictive power of single items and multiple items, concluded, "The findings of these eight studies uniformly reveal that responses of individuals to discrete situations of any sort are virtually unpredictable. To seek laws of underlying factors that would predict and control such responses therefore seems to be a fruitless objective in psychology" (p. 285). Lykken (1968) noted that exact replication of an experimental situation in all of its detail is impossible because of the difficulty in controlling incidental factors. More recently, Hogan, DeSoto, and Solano (1977), Epstein (1977, 1979c), and Green (1978) observed that the principle that single items have lower predictive power than the average of many items applies as well to laboratory behavior as to responses on paper-and-pencil tests. Campbell (1957) and Fishbein and Ajzen (1974) noted that a crisis in social psychology which resulted from low correlations between attitudes and behavior could be attributed to limitations in measurement of the behavior, which was inadequately sampled. More recently, Ajzen and Fishbein (1977) demonstrated in a review of a large number of studies that when adequate samples of behavior were obtained, high correlations between attitudes and behavior were often also obtained.

How Serious Is the Limitation of Being Unable to Predict Single Items of Behavior?

How serious a limitation is imposed on psychology as a science if general laws cannot be used to predict particular instances of ordinary behavior? To many psychologists who are intent on predicting nonexceptional behavior in specific situations, the limitation must appear to be great indeed. Yet, there is no reason to consider the limitation greater than that imposed on physics by the physicist's inability to predict the location of a single molecule in time and space. One might even take heart from the thought that it may be to humankind's advantage that our ability to predict is not greater. Science works not through the establishment of microscopic certainty, but through probabilistic prediction. The physicist is able to generate important laws about the behavior of gases by considering the aggregate behavior of molecules. What is uncertainty at the micro level can be certainty at

the macro level. An example closer to home is provided by the averaging of evoked cortical potentials. The effect of a single stimulus on the electrical activity of the brain cannot be detected because of background noise, which corresponds to error of measurement. However, when responses are averaged over a sufficient number of repeated stimulus presentations, a smooth, orderly curve that represents the brain's response to the stimulus appears. Thus, again we see the principle illustrated that when extraneous variance cannot be eliminated by experimental control, it can often be canceled out through aggregation.

Reasons for the Widespread Acceptance of a Questionable Paradigm

The question may be raised that if the psychological experiment as normally conducted is in fact as poor a procedure for establishing replicable generalizations as I have claimed, how is it possible that this observation has been ignored for so long? There are at least two reasons. One is that personal and social factors support the paradigm. The other is that a failure to distinguish between concurrent and temporal reliability has led researchers to believe they could establish the conditions for assuring replicability without having to replicate.

Personal and social factors. There is no more fundamental requirement in science than that the replicability of findings be established. Yet, in psychology few replication studies are attempted, and of these only a small proportion are published (see N. C. Smith, 1970, for a more extensive discussion of this issue and for other arguments that the experimental method as normally practiced has serious limitations). Some journals state that they do not accept replication studies, and others implicitly follow a similar policy. Replication studies are particularly apt to be rejected when they cast doubt on accepted conclusions. As Greenwald (1975) has observed, findings in psychology often have a faddish quality about them. At one point, studies that support a particular phenomenon are favored, and at another point, studies that refute the phenomenon are favored. Greenwald has also provided evidence of a general bias against studies that support the null hypothesis. Accordingly, it is impossible to interpret correctly levels of statistical significance in the studies that are accepted.

As for personal factors, the psychologist who invests his or her time and energy in replication studies runs the risk of being considered uncreative. If he or she persists, there may be difficulty in getting the work published. The personal consequences in a university setting are then anything but subtle, for a low rate of publication jeopardizes promotion and tenure. Moreover, for the person who makes no attempt to replicate his or her own findings, there is little danger that others will demonstrate that the findings are unreplicable because in the rare event that replication is undertaken by someone else, discrepant results can almost always be dismissed as the result of minor procedural differences.

A further reward associated with the psychological experiment as normally conducted is that it has an aura of scientific respectability about it because it uses the same paradigm as the physical sciences, with an emphasis on manipulation, control, objectivity, precision, and mathematics, at least in the form of statistics. Since experiments in the physical sciences that investigate a single stimulus on a single occasion produce results that are replicable and generalizable, it seems reasonable that the same should be true in psychology. As a result, the single-event experiment that examines a narrow range of stimuli is highly popular in psychological research and enjoys a high priority for publication. Moreover, there are practical advantages to such experiments. They fit into the schedules of researchers and college-student subjects alike, and they can be run at a greater rate than multistimulus, multioccasion experiments.

The failure to distinguish concurrent from temporal reliability. Perhaps the most important reason for the widespread belief in the replicability of psychological experiments in the absence of replication is that experimentalists often fail to distinguish between concurrent and temporal reliability. One can easily be misled into thinking that replicability has been established when a highly statistically significant effect has been obtained. After all, a statistically significant finding at the .01 level indicates that if the experiment were identically reproduced a hundred times, in no more than one case would a difference as large as that obtained be expected by chance. A critical assumption here is that the experiment would have to be reproduced identically in all its minor relevant details, including subtle contextual ones. The .01 reliability estimate would not hold if all factors that could affect outcome were not held constant. As it is normally impossible to achieve complete con-

ol, the reliability figure can only be assumed to apply to concurrent reliability—meaning that something happened in the specific situation which was not likely a chance occurrence—and does not establish that the effect can be attributed to the experimental variables of interest. Expressed otherwise, the reliability estimate refers to a theoretical condition that cannot be met with respect to actual replication, and therefore the reliability figure, no matter what its level of significance, does not signify replicability.

Concurrent and temporal reliability can be compared with concurrent and predictive validity. A test can have high concurrent validity, as revealed in its correlation with an external criterion when both are administered on the same occasion, and low predictive validity, as indicated by its correlation with a criterion administered on a later occasion. Obviously, changes can occur between sessions, and identical situations cannot necessarily be established on different occasions. By the same token, an experiment can have high concurrent reliability, as demonstrated by its ability to establish highly significant effects in a single experimental session, and low temporal reliability, as indicated by its inability to establish the same elationships on different occasions.

That experimentalists have failed to take seriously the concept of temporal reliability, a rudimentary concept in test construction, brings to mind Cronbach's (1957) admonition that two psychologies have developed—one relying on experimental procedures and the other on correlational procedures—and that neither has paid much attention to the concepts of the other.

The Four Faces of Aggregation

A typical exercise in elementary physics is to have students compute the mean and standard deviation of a series of measurements. The point of the exercise is to demonstrate that the mean of a number of measurements is more reliable and accurate than the individual measurements, as incidental factors cancel each other out when the data are aggregated. The same principle is well illustrated in a series of early psychological experiments undertaken for the ostensible purpose of comparing the accuracy of individual and group decision making.[4] Knight (1921) had students estimate the temperature in a classroom. Although individual estimates were highly inaccurate, ranging from 60° to 80°, the mean rating of 72.4° closely approximated the temperature of 72.0° indicated by a thermometer. Gordon (1924) had students rank a series of weights. The ranks were then averaged over different numbers of subjects, and the averaged ranks were correlated with the true rank. A greater amount of averaging resulted in a higher correlation with the true rank. The average correlations for ranks averaged over 1, 5, and 50 subjects were, respectively, .41, .68, and .94. Several years later, Stroop (1932) repeated Gordon's study and obtained virtually identical results. Stroop demonstrated that the results did not indicate the superiority of group judgment, as had been claimed (actually no group decisions had been made), but simply illustrated the principle that reliability, and therefore validity, could be increased by reducing error of measurement through averaging. It made no difference whether the averaging was done over many people with single trials or over many trials with fewer people. In both cases the increase in reliability as a function of the total amount of data aggregated closely corresponded to predictions from the Spearman-Brown formula.

Aggregation accomplishes two purposes: It reduces error of measurement, and it broadens the range of generalization of the findings. Thus, aggregation should be an excellent procedure for establishing replicable generalizations. There are four forms of aggregation: aggregation over subjects, aggregation over stimuli or stimulus situations, aggregation over time, which includes aggregation over trials and over sessions, and aggregation over modes of measurement. As indicated by the Spearman-Brown formula, the same amount of aggregation, no matter what its form, should produce the same increase in reliability. However, the form of aggregation determines whether the increase will be mainly in concurrent or in temporal reliability and establishes the nature of validity and generality of the findings. A further consideration in deciding on a form of aggregation is the availability of subjects and the cost in time and effort of testing many subjects a few times in comparison with testing a few subjects many times.

Aggregation over subjects. Aggregation over subjects requires little comment, as it is the one form of aggregation commonly practiced in psychological research. Because of its familiarity, it can serve as a model for analyzing the effects of other

[4] I wish to express my appreciation to Martha L. Epstein for bringing these studies to my attention.

forms of aggregation. Aggregation over subjects cancels out the effect of the uniqueness of individual subjects. By testing a large sample of subjects and averaging their responses, not only is the stability of the findings increased, but the generality of the findings is both increased and defined with respect to the population from which the sample was drawn. The greater the number of subjects averaged, the more stable the mean, and the more capable the mean therefore is of establishing coherent relationships with other variables, as reliability is a prerequisite for validity. As a result, it is not surprising that a macropsychology which deals with the means of very large numbers may succeed where a micropsychology fails (e.g., Katona, 1979).

To some extent, aggregating over subjects tested on different occasions overlaps with aggregating over occasions and over stimulus situations. This is so because different occasions are represented and different subjects interpret the same stimulus differently, thereby, in effect, increasing the range of effective stimulus representation. However, such aggregation over occasions and over perception of stimulus situations is unsystematic, cannot be statistically evaluated, and is restricted with respect to the effective range of stimulus representation. Aggregating over occasions and over stimulus situations warrants more serious consideration than haphazard treatment as a by-product of aggregating over subjects. Moreover, a serious problem may arise as a result of the imbalance created by systematically aggregating over subjects while failing to aggregate over situations and/or occasions. The larger the number of subjects examined, the smaller the absolute difference between two means required for a statistically significant difference to be detected. As a result, minute incidental effects that are specific to a particular situation or occasion are apt to show up as significant effects and be mistakenly interpreted as more general experimentally induced effects.

Aggregation over stimuli and/or situations. Aggregating over stimuli and/or stimulus situations cancels out the unique effects associated with particular stimuli and/or situations. It is a well-known principle in test construction that a number of different items which individually have low reliability and validity can be combined into a composite scale which has high reliability and validity. Unfortunately, this principle is often forgotten when items consist of behavior in laboratory or real-life situations rather than responses to paper-and-pencil tests. There is no dearth of evidence that the principle is widely applicable (e.g., Ajzen & Fishbein, 1977; Campbell, 1963; Epstein, 1977, 1979c; Fishbein & Ajzen, 1974; Green, 1978; Tryon, 1973). Despite such evidence, most psychological research consists of the investigation of either single stimuli or a limited range of stimuli that do not adequately represent the domain of stimuli to which the experimenter wishes to generalize (Brunswik, 1956). Nor is the problem restricted to contrived experiments in social psychology and personality. To a lesser extent, it is prevalent in most areas of psychological research, including studies in animal behavior and in sensory and perceptual processes (Brunswik, 1956; Cronbach, 1975; Greenwald, 1975). There is no shortcut for generalizing over stimuli any more than there is one for generalizing over people. In both cases it is necessary to sample adequately the domain over which one wishes to generalize.

Although the limits on generalization imposed by the investigation of a narrow range of stimuli are easily recognized, this is not the case for the limits imposed by the situational context, particularly those aspects of the context that experimenters would not list in their "experimental recipe" (Lykken, 1968). It is widely assumed that such situational contexts are unimportant. Yet there is a growing body of evidence which indicates that this is not so. It is well-known, for example, that even the way rats are handled can affect the outcome of an experiment. Potential sources of error in the situational context of an experiment include the experimenter's personality, sex, status, and vested interest in the outcome (Rosenthal, 1976; Smith, 1970; Eagly & Carli, Note 1). The time of the semester in which a study is carried out can relate to subject selection, to subject motivation, and to whether subjects have had an opportunity to gain knowledge about, and specific attitudes toward, participating in certain experiments. The atmosphere created by an experimental setting, including the use of apparatus, can seriously limit the range of generalization of the findings.

Given the potential influence of contextual effects on experimental outcomes, it follows that no generalization can be accepted until the influence of such incidental sources of variance has been ruled out. The history of past findings and the degree of situational specificity of much human behavior (Mischel, 1968) suggest that relatively few phenomena will be exempt from such influences. Thus,

the need for replication in various settings before a relationship can be accepted as of general theoretical interest can hardly be overstated. In the absence of such replication, results may be completely situation specific, with the critical features of the situation remaining unknown. Aggregation over varying situations within an experiment is one way to reduce the influence of such situationally specific effects.

It should be noted that the arguments advanced here are not the same as the ones advanced by those who advocate ecological validity (e.g., Brunswik, 1947; Gibbs, 1979), which is not to deny the value of the latter for certain purposes. The concern addressed in this article is with aggregation as a method for increasing the replicability and generality of findings by canceling out unrecognized and possibly uncontrollable unique effects. The principle of aggregation can be applied to orthogonal experimental designs as well as to ecologically valid samples of behavior. This is not to deny that for some purposes it is important to maintain distinctions among classes or levels of stimuli and to not cancel out such differences by aggregating over them. In such cases it is nevertheless important to sample adequately stimuli within classes or levels.

Aggregation over trials and/or occasions. Aggregating over trials and/or occasions cancels out the uniqueness of particular trials and/or occasions. Aggregating over trials and occasions thereby increases the reliability of findings and their generality over trials and occasions. To a limited extent, it also increases generality over situations and subjects. This follows from the consideration that no situation or subject is identical on different occasions. Thus, aggregating over occasions can function, to a limited degree, in the same way as aggregating over situations and subjects.

Note that although trials and occasions are included under a single heading because they both involve aggregating behavior over time, they have different implications for increasing reliability. Aggregating over trials cancels out incidental effects associated with specific trials within sessions and thereby primarily contributes to an increase in concurrent reliability, whereas aggregating over occasions cancels out incidental effects associated with specific sessions and thereby contributes to an increase in temporal reliability. Thus, aggregating over occasions is the more important procedure for enhancing replicability. This is not to deny that an increase in concurrent reliability through aggregation over trials can also contribute to an increase in temporal reliability, as the variation that occurs within sessions may, in certain instances, overlap with the variation that occurs among sessions.

It is of interest to contrast the effect on reliability of averaging over stimuli and situations, on the one hand, and of averaging over trials and occasions, on the other. The former increases reliability by achieving redundancy through the use of different items that measure different aspects of the same concept and might therefore be referred to as *conceptual redundancy*. The unique contribution of each individual item is canceled out relative to its contribution to the concept on which all items converge. Thus, reliability is achieved for a concept that is broader than its components. The situation is different for what might be called *item redundancy*, where redundancy is achieved by repeating the same item over occasions. If interest is in a single behavioral item, it can be brought to any desired level of reliability by replicating it a sufficient number of times. The reliability will be restricted to the narrowness and uniqueness of the particular item in question. It is noteworthy that the Spearman-Brown formula, which estimates the increase in reliability that occurs when an increasing number of different items are combined into a scale, can be applied with equal accuracy to estimating the increase in reliability that occurs when single items are replicated over occasions. This was demonstrated when Spearman-Brown estimates were made from the data in the four studies previously reviewed. In almost all instances, estimated values closely approximated values actually obtained, as long as there was a sufficiently stable relationship to begin with.

Although averaging, or blocking, trials within sessions is a common practice, the value of aggregating behavior over sessions has received little attention. This is unfortunate, for as noted above, aggregating over occasions is a powerful technique for increasing temporal reliability, or replicability. The findings in the studies reviewed indicated that the widespread assumption that error variance over occasions is unimportant is incorrect. This raises the question of what the variables are that vary incidentally over occasions. Among these one can assume are variants of the ones previously listed under extraneous influences in the stimulus situation. This follows from the consideration that the same situation will vary *psychologically* over occasions, even though the objective situation remains constant. Consider, for example, the experimenter who conducts the same experiment on dif-

ferent occasions. The experimenter's mood will vary, and he or she will be influenced by intervening experiences. Thus, the "same" experimenter will actually be a somewhat different experimenter. The situational context can also be influenced between sessions by subjects' being exposed to certain course material, by feedback from other subjects, by experiences in other experiments, by failures and successes in examinations, and so on. Not all of these will vary randomly. Some will interact with experimental manipulations, and others will restrict the range of generalization of the findings.

Order and sequence effects over sessions are also important factors that limit generalization. Psychologists tend to believe they have circumvented the complexities introduced by such factors when they conduct experiments on single occasions. For example, they assume their results are general across occasions without considering the possibility that their findings may be applicable only to first-time encounters. Because strong habituation effects are the rule rather than the exception, it can be expected that many findings from single-session experiments are influenced to a significant degree by uncertainty and anxiety associated with first-time encounters. Whatever the sources of incidental variance associated with occasions, they are apparently influential enough so that there is a serious problem in replication when such sources are not reduced through aggregation over occasions.

Although some of the data in the four studies included direct measurement of objective events, the studies were not laboratory investigations. Thus, the possibility remained that the conclusions from these studies might not apply to actual laboratory experiments. To test this possibility, a fifth study was conducted in which a number of simple phenomena, such as reaction time, habituation, conditioning, and motor steadiness, were investigated in the laboratory.[5] Dependent variables included simple behavioral responses and measures of psychophysiological reactivity, such as skin conductance and heart-rate responses. Twenty subjects were each examined on fourteen occasions. At the beginning of the first session, a battery of personality inventories was administered. The design of the study permitted a comparison of the effects of aggregating behavior over stimulus situations on a single occasion with aggregating behavior over occasions for a single stimulus situation. Although the study is not yet fully analyzed, it can be concluded from the analyses which have been completed that aggregating over stimuli and aggregating over occasions are not equivalent. Aggregating over occasions produced higher stability coefficients, while aggregating over stimuli produced stronger correlations with the scales in the personality inventories. The increase in stability coefficients for the laboratory data averaged over occasions was, for most variables, similar to that for the real-life studies previously reported, while a few variables were stable when tested on a single occasion. The stronger relationships with the personality inventories when the data were averaged over stimuli could be attributed to the fact that the personality scales measure broad dispositions and that aggregating over stimuli produces broader measures than aggregating over occasions.

Furthermore, data derived from the first session often produced different correlations than data from later sessions, which is consistent with the assumption that responses to new situations are often unique and may not be generalizable to other situations. It was further observed that not all stimulus situations could be combined into a single score. Like different scales within a test, different stimulus situations evoked responses that produced different significant relationships with other variables. Responses to some stimuli could be combined with a gain in reliability and validity, while responses to other stimuli canceled each other out. Such findings provide additional support for the view that principles of test construction should be taken as seriously in the analysis of laboratory data as in the analysis of test data (Cronbach, 1957; Tryon, 1973).

Aggregation over measures. One does not normally select a sample of measurement procedures from a broader population. Obviously, if there were one good measure, it would suffice. In psychological research, unfortunately, the validity of a particular measurement procedure is often questionable, and it is frequently unclear which of several measures is best. Often, each of several measurement procedures can be assumed to measure to some

[5] This study was conducted in collaboration with Robert Alexander, who trained and supervised two undergraduate assistants, John Gmeiner and Michael Kaplan. Alexander supervised the scoring of the records and wrote the computer programs for analyzing the vast amount of data generated by a great many variables measured over fourteen sessions.

degree the concept it purports to measure and to measure to a considerable degree something specific to itself that can be regarded as error of measurement, or method variance (Campbell & Fiske, 1959). To the extent that several such measures are combined, method variance will be canceled out relative to true variance. Thus, where appropriate, aggregating over measurement procedures can be a useful technique for reducing error of measurement.

The need to investigate aggregation empirically. Given the difficulties inherent in the psychological experiment, there is much to be gained by investigating the relative effectiveness of different forms of aggregation under different circumstances. Such investigations could be incorporated to advantage in a single research project. For example, by sampling a range of stimuli and testing subjects on several occasions, it is possible to establish concurrent and temporal reliability and to compare conclusions arrived at from responses to single stimuli or single occasions with conclusions arrived at from data aggregated over stimuli and occasions. An important question is to what extent aggregating data over stimuli within sessions can obviate the necessity for aggregating data over occasions. Different results will undoubtedly be found for different areas of psychological research.

Investigations That Do Not Require Aggregation Over Situations and/or Occasions

There are three kinds of events for which replicable results can be expected without aggregating over stimulus situations and/or occasions. One consists of highly robust phenomena that are insensitive to the influence of social and other incidental factors. Examples include reflex responses and overlearned habitual reactions. A second consists of experimental variables that are so potent or ego-involving for the individuals experiencing them that they eclipse the influence of incidental variables. A third consists of self-ratings and ratings by others that, although themselves made on a single occasion, are based on impressions gathered over an adequate sample of observations in the past.

Intrinsically robust phenomena. As already noted, sources of error associated with incidental factors include knowledge about the experiment and social variables such as transient and stable characteristics of the experimenter. It follows that experimental effects which are insensitive to such influences should produce results that are stable over occasions and situations and therefore replicable even if based on only a single observation. In general, the more a form of behavior is biologically determined, the more it should be insensitive to influence from the above variables. One can therefore expect that research in biopsychology and, to a lesser extent, in psychophysics often qualifies. To the extent that experiments are complex or contrived, they are particularly apt to be influenced by incidental factors. As complexity increases the chain of inference, it provides added opportunity for incidental factors to exert an affect. For example, although habituation by itself may be a highly robust phenomenon, differential rates of habituation for different groups may not be. It is important to recognize that many effects which have been assumed to be invulnerable to the influence of incidental factors in the past have turned out not to be so (Greenwald, 1975). As a result, it is necessary, no matter what the nature of the research, to establish its replicability rather than to assume it. Investigators will undoubtedly find that some phenomena presumed to be robust will turn out to be so when observed on a single occasion, while many others will not.

Potent, ego-involving events. The value of working with potent variables was brought home to me in a series of studies on anxiety in sport parachuting conducted in collaboration with Walter Fenz (e.g., Epstein, 1967; Epstein & Fenz, 1965; Fenz & Epstein, 1967). The results were so consistent among individuals that statistical evaluation was often superfluous. Moreover, the findings in one study, which were often unanticipated and of considerable theoretical interest, invariably held up in all other studies. When I attempted to bring the same variables into the laboratory by substituting threat of an electric shock for fear of a parachute jump and a countdown to the shock for the waiting period before the jump, despite far greater control over incidental variables, the results were far less coherent and not always replicable.

A basic difficulty with many laboratory studies is that they attempt to follow what is good scientific procedure in the physical sciences by examining minute effects with small error terms achieved through the careful control of incidental sources of variance. Unfortunately, as it is impossible to control all sources of incidental variance in psychological studies, what frequently happens is that the small error term obtained magnifies the contribution of unrecognized, situationally unique sources of influence as well as of experimental effects.

Thus, the results that reach significance either lack generality or are produced by incidental factors confounded with the experimental manipulation. In either case, the results are unreplicable. Another way of viewing this problem is to consider that a test of statistical significance consists of the ratio of experimental to error variance. The same increase in statistical significance can therefore be attained by an increase in the numerator (experimental variance) as by a decrease in the denominator (error variance). However, though the final level of significance may be the same, the consequences with respect to replication are not. Results that are significant only with small error terms are apt to be less replicable and generalizable than results that are significant despite large error terms, because the former cannot be expected to hold up under other conditions in which error variance is large or in which there are slight changes in the situational context.

For ethical reasons, it is of course not always possible to examine highly potent, ego-involving events in the laboratory. One solution, in some circumstances, is to find situations like sport parachuting that can serve as natural laboratories for investigating particular phenomena (see Epstein, 1979a, and Gibbs, 1979, for examples of a variety of such studies). It may be possible in such situations to obtain quasi manipulation of an independent variable by arranging testing in relation to a naturally varying condition. In the studies of parachuting, for example, anxiety was varied by testing at different times before a jump. Another approach is to have subjects report significant life events, such as occasions when they were most threatened or most in love. This need not be done in the manner of a subjective interview. Rating scales to test specific hypotheses of theoretical interest can be devised. While self-report techniques have obvious limitations, they also have unique advantages, including accessibility to material that is not otherwise available. The limitations are probably no greater than those for other techniques, each of which has its own advantages and limitations (Epstein, 1979a). As was noted earlier, one of the reasons self-report techniques fell into disrepute was their failure to correlate with laboratory data. Because laboratory data are often situationally specific or temporarily unreliable, the validity of self-report techniques warrants reappraisal.

The selection of variables that are potent and ego-involving for a subject obviously requires a theoretical orientation that takes into account the nature of the subject as well as that of the stimulus. Theory is also necessary to determine the functional equivalence of different stimuli and responses for individuals under varying circumstances, as stability and coherence can be present at a genotypical level when absent at a phenotypical level (Alker, 1972; Block, 1977; Bowers, 1973). This can be illustrated by the findings of Sroufe and his colleagues (Arend, Gove, & Sroufe, 1979; Matas, Arend, & Sroufe, 1978; Waters, Wippman, & Sroufe, 1979). When they aggregated several measures into broadband assessment techniques for measuring developmentally "salient" issues, that is, issues particularly meaningful for children of a particular age, they found strong evidence for genotypical stability across relatively long time spans. For example, resistance to contact with the mother upon reunion at 12–18 months was found to predict ineffectiveness in problem solving at age 2 and low ego resiliency and curiosity at age 5. As observed by Sroufe (Note 2), "Another answer to those who attack personality constructs is not 'How well do you measure?' but 'What did you measure?' " (For an extended discussion of the importance of theory in establishing stability and coherence in behavioral patterns, see Block, 1977.)

Single ratings following multiple or extended observations. When individuals rate themselves on personality inventories, and when observers rate subjects whom they have observed for some time, although the ratings consist of single responses, they represent an intuitive averaging of many observations. As a result, such ratings have the potential for producing highly replicable and valid results. This is not to deny that there may be serious problems of validity in particular circumstances, such as biased recall or purposeful misrepresentation. It is to note that such ratings have a very important advantage in that they need not be limited by insufficient opportunity for observation.

Advantages of a Combined Ideographic–Nomothetic Approach

Not only has psychological research been dichotomized with respect to correlational and experimental procedures (Cronbach, 1957), but it has also been dichotomized with respect to *ideographic* (individual) and *nomothetic* (universal) procedures. These two terms were introduced by Allport (1940) to draw attention to the degree to which he believed

research was overbalanced in the direction of normative studies relative to in-depth investigations of individuals.

Ideographic research and nomothetic research both have advantages and disadvantages. The in-depth and broad investigation of an individual over a sample of situations and occasions permits reliable generalizations to be made over situations and occasions and, most important, permits information to be obtained on psychological processes within the individual. Its obvious weakness is that it does not permit generalizations to be made over individuals. Whatever processes are elucidated within an individual may, of course, be specific to that particular individual. The sampling of people, but not situations or occasions, on the other hand, allows generalizations to be made over individuals but not over situations or occasions. Moreover, it provides no information on psychological processes, as the relationships found among individuals may not hold within individuals.[6]

Fortunately, there is no need to choose between ideographic and nomothetic procedures, as it is possible to integrate the two by examining a sample of individuals with a variety of measures over a sample of situations and occasions. An excellent example of such an approach is presented in Murray's (1938) classic work, *Explorations in Personality*. (See discussion of this work in Epstein, 1979b. For other views advocating a combined ideographic-nomothetic approach, see Bem & Allen, 1974; Lazarus, 1978; and Lazarus & Launier, 1978.) In-depth studies that examine a small group of individuals not only are important to personality research but have much to offer all branches of psychology. First, they are an improvement over typical nomothetic studies that examine individuals in single situations on single occasions, in that by averaging over many situations and/or occasions, reliability and generality of findings are increased. Second, by obtaining replications over stimuli and over trials and/or occasions, concurrent and temporal reliability can be assessed and taken into account in evaluating the results. Not only can this provide important information in its own right (such as by indicating that the reason the null hypothesis was not rejected in a particular study had nothing to do with the theory that was being tested, but was a necessary consequence of the low reliability of the procedures), but it can also provide information on how many situations or occasions should be sampled in future work to obtain adequate reliability. Third, it permits correlations that have been obtained within individuals over variables or over time to be used as data in a between-subjects design, thereby making it possible to compare groups with respect to relationships in addition to comparing them on means. Fourth, it permits between-subjects relationships to be compared with averaged within-subjects relationships for the same variables, thereby providing a direct test of the extent to which nomothetic and ideographic procedures produce similar results.

Too often the highly questionable assumption is made that correlations derived from nomothetic studies of groups of individuals are applicable to processes within individuals. In the five nomothetic–ideographic studies referred to in this article, different, and sometimes opposite, relations were sometimes found between the two procedures. For example, depression and anger were found to be directly related when between-subjects correlations were computed, but were often inversely related when within-subjects correlations were computed.

Because the use of experimental designs in which a sample of individuals is investigated with several measures over a sample of situations and/or occasions demands an increased amount of time and effort, it is necessary to devise efficient procedures for gathering and reducing data. In the four studies summarized at the beginning of this article, most of the data were obtained by having subjects observe their own or other people's behavior. In the first and second studies, which entailed extensive self-observation and the daily filling out of lengthy forms, the project was conducted as a set of exercises in a class on personality research. The other studies, which were less intensive, were conducted as a class project, and a few minutes of time at each class meeting were devoted to obtaining data. In two of the studies, subjects recorded their responses on Opscan sheets, which were automatically transferred to computer cards. A rewarding by-product of the procedure was that the students, functioning in effect as co-investigators,

---

[6] It should be noted that operant procedures are ideographic and are therefore exempt to a large extent from criticisms that were made of the nomothetic laboratory experiment as normally conducted. It remains to be seen how well the generalizations from operant procedures are replicable over time and individuals. It is noteworthy, however, that the intensive study of a few cases, as in the work of Freud, Pavlov, and Skinner, has produced some of the most significant advances in psychology.

found the projects to be interesting learning experiences.

Of course, not all research can be conducted by having subjects observe their own or others' behavior. The fifth study, in which subjects were repeatedly seen in the laboratory, took a great deal of time and effort. However, the results indicated that eleven sessions were superfluous for achieving adequate levels of temporal reliability for between-subjects comparisons. Three or four sessions would have sufficed. If interest is in establishing within-subjects relationships, then responses to a relatively large number of stimuli and/or sessions must of course be obtained.

The Place of the Traditional Single-Session Experiment

Given its evident limitations, what place is there for the single-session experiment that examines a single event which neither is highly potent nor, on the basis of previous evidence, can be assumed to produce stable findings? No findings from such a single study, no matter what their level of statistical significance, should be given much credence. Single studies, however, can be of value by stimulating other studies and by contributing thereby to a population of studies that can ultimately be interpreted in the aggregate.[7] A population of experiments that examines single events has an advantage over a single study that incorporates into its design replications over time and situations, as desirable as the latter experiments are, because the population of experiments allows generalizability and replicability to be established over such important incidental sources of variance as experimenters (Hammond, 1954, 1955), laboratories, and samples of subjects. Further, there are phenomena that cannot be studied with experimental designs that use repeated measures within or over sessions. In studies requiring deception or surprise, only a single exposure to a stimulus may be feasible. Thus, studies of single events on single occasions have their place. However, if they are to be able ultimately to make a contribution in the aggregate, it will be necessary for journals to encourage replication studies, to not be influenced by outcomes in such studies, and to make decisions solely on the basis of the adequacy of the procedures. Until this is done, levels of statistical significance will be misleading, and many widely accepted findings will turn out to be myths, as has too often been the case in the past (Epstein & Burstein, 1966; Greenwald, 1975, 1976; Rosenthal, 1969, 1976).

Finally, it should be noted that aggregation, whether carried out within a single study or over a sample of studies, is not a panacea. Certain systematic effects, such as the sex and the attitudes of experimenters (Rosenthal, 1976; Eagly & Carli, Note 1), may bias an entire group of studies. Moreover, aggregating over such variables can obscure important differences. What initially appears to be noise may turn out to mask important variables. Examining large samples of studies provides one way of detecting the existence of such variables (e.g., Eagly & Carli, Note 1). Given that differentiation, when possible, is preferable to aggregation, does this mean that aggregation is of secondary importance as an experimental technique? The answer is that one must distinguish between potentially meaningful and incidental sources of experimental influence. As I have demonstrated, the value of aggregation lies in removing the influence of incidental variables of no apparent theoretical interest that either cannot be controlled or, if controlled, would yield generalizations too narrow in scope to be scientifically useful. Unfortunately, given the extreme degree of situational specificity in much human behavior, the influence of such incidental variables is far greater than psychologists anticipated it would be, and thus the need for aggregation is considerable.

[7] An interesting recent development, *metapsychology*, is concerned with systematic procedures for aggregating findings over samples of studies (Cohen, 1977; Cooper, 1979; Hall, 1978; Rosenthal, 1978; M. L. Smith & Glass, 1977; Eagly & Carli, Note 1).

REFERENCE NOTES

1. Eagly, A. H., & Carli, L. *Sex of researchers and sex-typed communications as determinants of sex differences in influenceability: A meta-analysis of social influence studies.* Manuscript submitted for publication, 1980.
2. Sroufe, L. A. Personal communication, April 15, 1979.

REFERENCES

Ajzen, I., & Fishbein, M. Attitude–behavior relations: A theoretical analysis and review of empirical research. *Psychological Bulletin*, 1977, *84*, 888–918.

Alker, H. A. Is personality situationally specific or intrapsychically consistent? *Journal of Personality*, 1972, *40*, 1–16.

Allport, G. W. The psychologist's frame of reference. *Psychological Bulletin*, 1940, *37*, 1–28.

Arend, R., Gove, F. L., & Sroufe, L. A. Continuity of individual adaptation from infancy to kindergarten: A predictive study of ego-resiliency and curiosity in preschoolers. *Child Development*, 1979, *50*, 950–959.

:m, D. J., & Allen, A. On predicting some of the people some of the time: The search for cross-situational consistencies in behavior. *Psychological Review*, 1974, *81*, 506–520.

Block, J. Recognizing the coherence of personality. In D. Magnusson & N. S. Endler (Eds.), *Personality at the crossroads: Current issues in interactional psychology*. Hillsdale, N.J.: Erlbaum, 1977.

Bowers, K. S. Situationism in psychology: An analysis and a critique. *Psychological Review*, 1973, *80*, 307–336.

Brunswik, E., *Systematic and representative design of psychological experiments*. Berkeley: University of California Press, 1947.

Brunswik, E. *Perception and the representative design of psychological experiments*. Berkeley: University of California Press, 1956.

Campbell, D. T. Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, 1957, *54*, 297–312.

Campbell, D. T. Social attitudes and other acquired behavioral dispositions. In S. Koch (Ed.), *Investigations of man as socius*. New York: McGraw-Hill, 1963.

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 1959, *56*, 81–105.

Campbell, D. T., & Stanley, J. C. *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally, 1966.

Cohen, J. *Statistical power analysis for the behavioral sciences*. New York: Academic Press, 1977.

Cooper, H. M. Statistically combining independent studies: Meta-analysis of sex differences in conformity research. *Journal of Personality and Social Psychology*, 1979, *37*, 131–146.

Cronbach, L. J. The two disciplines of scientific psychology. *American Psychologist*, 1957, *12*, 671–684.

Cronbach, L. J. Beyond the two disciplines of scientific psychology. *American Psychologist*, 1975, *30*, 116–127.

Endler, N. S., & Hunt, J. McV. S-R inventories of hostility and comparisons of the proportions of variance from persons, responses, and situations for hostility and anxiousness. *Journal of Personality and Social Psychology*, 1968, *9*, 309–315.

Epstein, S. Toward a unified theory of anxiety. In B. A. Maher (Ed.), *Progress in experimental personality research* (Vol. 4). New York: Academic Press, 1967.

Epstein, S. Traits are alive and well. In D. Magnusson & N. S. Endler (Eds.), *Personality at the crossroads: Current issues in interactional psychology*. Hillsdale, N.J.: Erlbaum, 1977.

Epstein, S. The ecological study of emotions in humans. In P. Pliner, K. R. Blankenstein, & I. M. Spigel (Eds.), *Advances in the study of communication and affect: Vol. 5. Perception of emotions in self and others*. New York: Plenum Press, 1979. (a)

Epstein, S. Explorations in personality today and tomorrow: A tribute to Henry A. Murray. *American Psychologist*, 1979, *34*, 649–653. (b)

Epstein, S. The stability of behavior: I. On predicting most of the people much of the time. *Journal of Personality and Social Psychology*, 1979, *37*, 1097–1126. (c)

Epstein, S., & Burstein, K. R. A replication of Hovland's study of generalization to frequencies of tone. *Journal of Experimental Psychology*, 1966, *72*, 782–784.

Epstein, S., & Fenz, W. D. Steepness of approach and avoidance gradients in humans as a function of experience. *Journal of Experimental Psychology*, 1965, *70*, 1–12.

Fenz, W. D., & Epstein, S. Gradients of physiological arousal of experienced and novice parachutists as a function of an approaching jump. *Psychosomatic Medicine*, 1967, *29*, 33–51.

Fishbein, M., & Ajzen, I. Attitudes toward objects as predictors of single and multiple behavioral criteria. *Psychological Review*, 1974, *81*, 59–74.

Gibbs, J. C. The meaning of ecologically oriented inquiry in contemporary psychology. *American Psychologist*, 1979, *34*, 649–653.

Gordon, K. Group judgments in the field of lifted weights. *Journal of Experimental Psychology*, 1924, *7*, 398–400.

Green, B. F. In defense of measurement. *American Psychologist*, 1978, *33*, 664–670.

Greenwald, A. G. Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, 1975, *82*, 1–20.

Greenwald, A. G. An editorial. *Journal of Personality and Social Psychology*, 1976, *33*, 1–7.

Hall, A. Gender effects in decoding nonverbal cues. *Psychological Bulletin*, 1978, *85*, 845–857.

Hammond, K. R. Representative vs. systematic design in clinical psychology. *Psychological Bulletin*, 1954, *51*, 150–159.

Hammond, K. R. Probabilistic functioning and the clinical method. *Psychological Review*, 1955, *62*, 255–262.

Hogan, R., DeSoto, C. B., & Solano, C. Traits, tests, and personality research. *American Psychologist*, 1977, *32*, 255–264.

Katona, G. Toward a macropsychology. *American Psychologist*, 1979, *34*, 118–126.

Knight, H. C. *A comparison of the reliability of group and individual judgments*. Unpublished master's thesis, Columbia University, 1921.

Koch, S. Epilogue. In S. Koch (Ed.), *Psychology: A study of a science* (Vol. 3). New York: McGraw-Hill, 1959.

Lazarus, R. S. A strategy for research on psychological and social factors in hypertension. *Journal of Human Stress*, 1978, *4*, 35–40.

Lazarus, R. S., & Launier, R. Stress-related transactions between person and environment. In L. A. Pervin & M. Lewis (Eds.), *Perspectives in interactional psychology*. New York: Plenum Press, 1978.

Lykken, D. T. Statistical significance in psychological research. *Psychological Bulletin*, 1968, *70*, 151–159.

Magnusson, D., & Endler, S. Interactional psychology: Present status and future prospects. In D. Magnusson & N. S. Endler (Eds.), *Personality at the crossroads: Current issues in interactional psychology*. Hillsdale, N.J.: Erlbaum, 1977.

Matas, L., Arend, R. A., & Sroufe, L. A. Continuity of adaptation in the second year: The relationship between quality of attachment and later competence. *Child Development*, 1978, *49*, 547–556.

Mischel, W. *Personality and assessment*. New York: Wiley, 1968.

Mischel, W. Continuity and change in personality. *American Psychologist*, 1969, *24*, 1012–1018.

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

Mischel, W. On the interface of cognition and personality: Beyond the person–situation debate. *American Psychologist*, 1979, *34*, 740–754.

Murray, H. A. *Explorations in personality*. New York: Oxford University Press, 1938.

Pervin, L. A. Performance and satisfaction as a function of individual-environment fit. *Psychological Bulletin*, 1968, *69*, 56–68.

Pervin, L. A. The representative design of person-situation research. In D. Magnusson & N. S. Endler (Eds.),

Personality at the crossroads: Current issues in interactional psychology. Hillsdale, N.J.: Erlbaum, 1977.

Rosenthal, R. Interpersonal expectations: Effects of the experimenter's hypothesis. In R. Rosenthal & R. L. Rosnow (Eds.), *Artifact in behavioral research*. New York: Academic Press, 1969.

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

Rosenthal, R. Combining results of independent studies. *Psychological Bulletin*, 1978, *85*, 185–193.

Smith, M. L., & Glass, G. V. Meta-analysis of psychotherapy outcome studies. *American Psychologist*, 1977, *32*, 752–777.

Smith, N. C., Jr. Replication studies: A neglected aspect of psychological research. *American Psychologist*, 1970, *25*, 970–975.

Stroop, J. R. Is the judgment of the group better than that of the average member of the group? *Journal of Experimental Psychology*, 1932, *15*, 550–562.

Tryon, R. C. Basic unpredictability of individual responses to discrete stimulus presentations. *Multivariate Behavioral Research*, 1973, *8*, 275–295.

Waters, E., Wippman, J., & Sroufe, L. A. Attachment, positive affect, and competence in the peer group: Two studies in construct validation. *Child Development*, 1979, *50*, 821–829.