The Person-Situation Debate in Historical and Current Perspective

Seymour Epstein
University of Massachusetts at Amherst

Edward J. O'Brien
Marywood College

The person-situation debate remained unresolved for many decades because researchers did not realize that behavior can be situationally specific at the item level and cross-situationally general at the aggregate level. Research that supported the specificity position examined single instances of behavior or did not properly aggregate behavior over items as well as occasions and was therefore unable to provide evidence for the existence of broad, stable response dispositions, or traits. A detailed examination of four classic studies and of a more recent study by Mischel and Peake (1982), all of which examined multiple items on many occasions, revealed that, in all cases, commitment to a procedural paradigm that emphasized relationships among individual items of behavior resulted either in the use of inappropriate procedures or in the misinterpretation of data. The most common problems concerned inappropriate procedures for item selection and retention, the failure to properly take into account reliability, and, most important, the failure to recognize the importance of attending to relationships among aggregates of items. The data from all studies, despite different conclusions by the authors, provided remarkably similar findings, including impressive evidence for stable, broad response dispositions, and for considerably stronger cross-situational relationships than had been reported.

To what extent is behavior situationally specific, and to what extent are there broad generalities in behavior? Since its inception, the field of personality has been divided on this issue. There is a tendency to conclude, when a debate has lasted so long, that the problem is insoluble and that the two sides have likely been emphasizing different aspects of a multifaceted reality, much as in the case of the heredity-environment debate with respect to intelligence. There is some truth in such an assessment, as behavior is unquestionably to some extent general and to some extent specific, and one can choose to study one aspect or the other. However, it would be a mistake to dismiss the debate as not worth further consideration, for there is a specific error in thinking that has characterized the debate and that, unless its persistence is seen in historical perspective, is apt to be continued. A detailed historical review of the person-situation debate is beyond the scope of this article. The interested reader can find additional material in reviews by Bowers (1973), Ekblom (1974), Magnusson and Endler (1977), and O'Brien (1978).

We present a brief sketch of the historical background of the debate, culminating in Mischel's influential book in 1968. Following this, four classic studies are examined in detail. These studies are of particular interest as they examined behavior aggregated over both occasions and situations, but despite having the information necessary for resolving the debate, they exacerbated it instead. We then examine a similar study by Mischel and Peake (1982) that repeated the same error. Next the virtues and limitations of traits are briefly discussed. Finally, we take stock of what issues have and have not been resolved in the person-situation debate and of the role of aggregation in their resolution.

Out of the broader debate on the generality of behavior and the existence of traits, several
more specific issues of concern arose. Included are: (a) a presumed 30 barrier when self-report measures of a trait are correlated with objective measures, (b) surprisingly low correlations between the same behavior observed in different situations, and (c) a presumably paradoxical widespread belief in the existence of traits and in the generality of behavior despite repeated demonstrations that behavior is usually highly situationally specific. We demonstrate that the resolution of all of these issues follows from the recognition that behavior is often highly situationally specific at the individual-item level but general at the aggregate level. This resolution is disputed by Mischel (1983; Mischel & Peake, 1982), who states that recent research on aggregation has added nothing new and therefore has resolved nothing. Mischel and others (e.g., Mischel & Peake, 1982) have often commented about whether the debate has or has not been resolved, withholding specific aspects of the debate the author has in mind. Accordingly, we review the evidence concerning the resolution of each of the more specific issues, as well as of the greater debate on the existence of traits.

Historical Background of the Debate

From its beginning the debate involved a division between those such as Allport (1931, 1937) who viewed behavior as centrally organized and purposeful, and others such as Thorndike (1906) who viewed behavior as mechanistic and composed of discrete habit elements. Thorndike's position is well captured by the following statement on the transfer of training:

Training the mind means the development of thousands of particular, independent capacities, the formation of countless particular habits, for the working of any mental capacity depends upon the concrete data with which it works. Improvement in any one mental function or activity will improve others only in so far as they possess elements in common to it also. The amount of identical elements in different mental functions and the amount of general influence from special training are much less than common opinion supposes. (Thorndike, 1906, p. 248)

Allport (1937) reacted to such statements by noting that the argument for identical elements tends to be circular. If transfer of training occurs, it is assumed that many identical elements were present, and if it does not, it is assumed they were not. Allport further observed that the theory of identical elements cannot account for the observation that situations with many elements in common (a panther inside vs. outside of a cage) often elicit grossly different responses, whereas situations with few elements in common (a panther running toward you in the jungle vs. an out-of-control car careening toward you on a city street) often elicit highly similar responses.

A series of empirical studies in the 1920s and 30s fueled the fires of the person-situation debate. Studies by Hartshorne and May (1928), Newcomb (1929), Allport and Vernon (1933), and Dudycha (1936) were explicitly undertaken to evaluate the degree of generality in behavior. Although their conclusions were mixed, their data were widely interpreted, in balance, to favor the situationist position. Allport was critical of the logical inferences and methodology in the studies that supported the situationist view, but in his own study of expressive behavior with Vernon (Allport & Vernon, 1933) he did not find evidence for a hypothesized single, highly general trait. He was not too distressed by the results from normative studies, however, as he retained the hope that results from an idiographic approach, in which emphasis is on the unique organization of variables within individuals, would be more successful in revealing broad behavioral dispositions.

Not much was added to the debate during the 1940s (Sanford, 1970). However, the next two decades witnessed a flurry of activity that provided further support for the situationist position, culminating in an influential book by Mischel in 1968 that marshaled the various arguments and evidence against a trait position. One major source of evidence consisted of low correlations between objective (non-self-report) measures of the same trait (e.g., Alpazewig, 1954; Broverman, 1964; Burwen & Campbell, 1957; Chown, 1959; Elliott, 1961; Guilford, 1964; Pervin, 1960; Sears, 1963). A second source of evidence consisted of findings that cast doubt on the validity of self-report measures and clinical assessment procedures. The failure to adequately take into account method variance (Campbell & Fiske, 1959), social desirability and response sets (Edwards, 1957), and construct validity (Cronbach & Meehl, 1955) in devising and validating measures of traits was emphasized. In addition, the practice of selecting items for the construction of a theoretical construct was criticized for its potential to introduce bias into the results.
The practical utility of trait measures for prediction was called into question (e.g., Gough, Wenk, & Rozycko, 1965; Lorei, 1964, 1967; McArthur & Stevens, 1955; Mischel, 1965). As if that were not enough, the validity of clinical assessments of broad dispositions was treated with skepticism as a result of findings of low interjudge reliabilities (e.g., Goldberg & Werts, 1966; Howard, 1962; Little & Schneidman, 1959; Marks, 1961), weak relationships between judgments and objective criteria (e.g., Chapman & Chapman, 1967, 1969; Golden, 1964), and a failure to demonstrate the practical utility of clinical assessments (Goldberg, 1959; Horowitz, 1962). A third major influence was the increasing prominence of behavioral assessment and treatment procedures (e.g., Bandura, 1967, 1969; Farber, 1964; Krasner & Ullman, 1965; Paul, 1966; Wolpe, 1958, 1963), which emphasized the situational control of behavior.

The above three themes were effectively merged in Mischel’s (1968) book, Personality and Assessment. Mischel argued for a social learning approach to personality that recognized that the concept of traits, except for broad screening purposes, had proven to be “untenable,” that behavior is highly situationally specific, that self-report measures characteristically produce correlations no greater than .30 with non-self-report measures, that people’s beliefs in traits and in the generality of behavior itself is a paradox that requires explaining, and that behavior can best be explained by concepts referring to operant and respondent conditioning and social learning processes. It is interesting, in this respect, to compare the following statement by Mischel (1968) with the one previously cited from Thorndike (1906):

...it is evident that the behaviors which are often construed as stable personality trait indicators actually are highly specific and depend on the details of the evoking situations and the response mode employed to measure them. (p. 37)

It is important to recognize that Mischel performed an important service in drawing attention to the widespread, inappropriate use of trait theory. For example, it had become a common practice to infer traits from single items of behavior and from single signs in projective tests, and to use trait measures to predict single items of behavior. Unfortunately, Mischel went too far in concluding that the concept of traits had been proven to be “untenable.” Throughout the book, Mischel did not take into account the limiting effect on reliability and generality of single items of behavior and, relatedly, of the value of aggregating observations over occasions and situations in establishing the existence of broad, stable response dispositions. Accordingly, Mischel concluded that, with the exception of intellective abilities, objective data did not support the existence of traits, as it had uniformly been found that interitem correlations were low. He did not consider that interitem correlations in intelligence tests were equally low, and that intelligence tests achieved their high levels of reliability and validity through compounding weak relationships. In any event, Mischel’s book had an important, timely message concerning the widespread misuse of trait concepts and measures, and was, with respect to that issue, deservedly well received.

Ultimately, Mischel’s attack on the generality of behavior, along with two other developments, provoked a paradigm crisis in personality research. One of the other developments consisted of a series of studies by Endler, Hunt, and their associates (e.g., Endler & Hunt, 1966, 1968, 1969; Endler, Hunt, & Rosenberg, 1962) that demonstrated that the amount of variance accounted for by situations and person–situation interactions was greater than that accounted for by people. This ultimately led to the view that interactionism provides a reasonable resolution of the person–situation debate, as behavior can never be determined by person or situation variables alone, but is always a result of the interaction between them. The other development consisted of research that accounted for the widespread belief in traits in the absence of actual consistency in behavior. It was demonstrated that systematic biases in judgments, the constructs of raters, and the semantic relatedness of trait words were often more responsible for the relationships that were found than the actual consistency of the people who were rated (e.g., Chapman & Chapman, 1969; Jones & Nisbett, 1971; Ross, 1977; Sweder, 1975). Thus, the widespread belief in traits by psychologists and laypersons, alike, could be dismissed as a foible of human thinking.
Although the attack by Mischel focused on the generality of behavior and the existence of traits, others carried the attack further. Some, such as Allen & Potkay (1973), argued that behavior was not only lacking in cross-situational generality, but in temporal stability as well. Shweder (1975), noting the logical implications of both arguments, concluded that it was time to acknowledge that "an individual differences theory of personality" is no longer viable. Thus, personality psychology had reached a crisis in which its very existence as a field of endeavor was being questioned. Some felt it was already dead and all that remained was to bury it. Others held out hope that the patient could be saved with new procedures. Among the new approaches most often recommended were interactionism and idio
graphic procedures. Unfortunately, interactionism proved to be no more successful in breaking the presumed .30 barrier than previous approaches, and limiting personality research to idio
graphic procedures would mean abandoning a major aspect of what had been personality psychology, namely normative studies of group and individual differences. The patient might live, but would lack some major functions.

Suffice it to note that all arguments but one were soon answered satisfactorily. (For a more thorough review of the challenges to trait theory and how they were met, see Epstein, 1979.) The one challenge that could not be answered was why, if there are important generalities in behavior, as trait theorists presume, correlations based on objective measures of personality were almost invariably found to be below .30. So long as substantial correlations could not be obtained with the direct objective measurement of behavior, many psychologists were skeptical about the evidence from self-ratings and ratings by others that supported the generality and stability of behavior. (See Block, 1977, however, for an excellent defense of well-conducted studies that use ratings, and for convincing arguments that such studies provided unequivocal evidence for the temporal stability and generality of behavior.) Attempts to defend the trait position, in light of the low correlations when direct, objective measurement was involved, took the form of speculating that, had studies involved other procedures, such as examining intrasubject rather than intersubject relationships, the .30 barrier would be breached. Unfortunately, no evidence was cited to support such assertions. Although the proposals were often of interest in suggesting new directions in personality research, the debate could not be resolved by disputation or speculation, but only by data. Accordingly, in response to suggestions by Alker (1972) of how the .30 barrier might be breached, Bem (1972) wrote, "And if these separate indices permit Alker to predict behavior across situations better than +.30, Mischel will fold up his tent and steal away" (p. 18).

The paradigm crisis in personality research continued through the latter part of the 1970s, when a resolution was proposed (Epstein, 1979) that followed from the observation that when investigators had examined objective behavior, particularly in controlled laboratory settings, they had often ignored basic psychometric principles that should have alerted them to the limited reliability and generality of single items of behavior and of the importance of aggregating items over both situations and occasions in order to measure traits that are both broad and stable. Such an oversight confirmed Cronbach's (1957) suspicion that two psychologies had developed, one based on test-construction procedures and the other on experimental-behavioral procedures, and that neither showed much interest in, nor sophistication about, the concepts and procedures of the other.

The Proposed Solution

A moment's reflection reveals that a single item of behavior, like a single item in a test, has too narrow a range of generality and too great a degree of error of measurement to produce more than very modest correlations with other presumed measures of a trait. Viewed otherwise, traits cannot be measured by single items of behavior because traits refer to stable, broad dispositions. As dispositions, traits permit actuarial prediction, that is, reasonably accurate prediction of behavior averaged over a sample of situations and occasions, but not the prediction of single behavioral acts. An extrovert can be expected to be extroverted most of the time, but it would be foolish to wager that someone with a high score on extroversion

![Figure 1: Stability](https://example.com/figure1.png)

Figure 1: Stability represents the mean value on Predicting Most

*Psychology, 37*, p.11
will greet you with a slap on the back and play a practical joke on you when you meet him or her at 2:00 p.m. tomorrow. To measure a trait, it is necessary to establish both temporal reliability, or stability, and generality. The former can be accomplished by averaging behavior over sufficient occasions and the latter by averaging behavior over appropriate situations. Parenthetically, note that sampling real-life behavior on different occasions will, unless a deliberate attempt is made to hold the situations constant, also involve sampling behavior in different situations, thereby fulfilling both requirements for the measurement of a trait.

On the basis of such reasoning, Epstein (1977, 1979, 1980b, 1983b, 1984b) conducted a series of studies that investigated a wide variety of data, including intersubject and intrasubject data, self-report measures, ratings by others, and the direct, objective measurement of behavior in real-life and in laboratory situations. In all cases, evidence for high levels of reliability and generality was found for multiple, but not single, items of behavior. As an example, Figure 1 presents the findings from a study in which the following variables were recorded over several occasions: self-rated emotions; physiological reactions, such as heart-rate mean and range; social behavior, such as number of social calls made and letters written; errors and carelessness, such as number of mistakes made in filling out the optical scanning sheets on which responses were recorded; and the occurrence of headaches and

---

*Figure 1. Stability coefficients as a function of the number of days in the odd-even samples. (Values plotted represent the mean of the correlations for the variables in a category. From "The Stability of Behavior: I. On Predicting Most of the People Much of the Time" by S. Epstein, 1979, Journal of Personality and Social Psychology, 37, p. 1117. Copyright 1979 by the American Psychological Association. Reprinted by permission.)*
stomachaches. The results were averaged over different numbers of occasions, and intersubject correlations were computed for odd versus even occasions. Figure 1 shows that the reliability coefficients, based on a single day of observation, tended to be low, but increased to high levels, between .75 and .93, when the data were averaged over 12 days. The correlations based on the data aggregated over only a few occasions also were unstable, as indicated by the erratic form of several of the curves when data were based on 1 or 2 days of observation. Such correlations provide a poor basis for estimating the effects of aggregation by the Spearman-Brown formula. Once data are sufficiently aggregated, however, the curves become highly orderly, and Spearman-Brown predictions are highly accurate. Of particular interest, it was observed that when high stability coefficients were obtained through aggregation, evidence for validity emerged in the form of an increased number of significant and coherent relationships among the variables and their relationships with personality inventories. Many of these correlations were above the presumed .30 barrier. It was thus concluded that a major source of the previously low correlations was the use of unreliable measures of unaggregated behavior. The findings from most of the studies conducted by Epstein were as relevant to the issue of cross-situational consistency in behavior as they were to temporal stability, as the behavior on different occasions was aggregated over the normal range of situations in everyday life. Thus, the data demonstrated that there is enough stability and generality in behavior so that useful statements about broad dispositions could be made without having to specify the eliciting situations. This, of course, is the way a trait is normally defined, and the findings, therefore, demonstrated the utility of such a concept. Of additional interest, aggregation over occasions and situations was found to enhance the reliability of intersubject data, in the form of relationships among variables within individuals, no less than of intersubject data in the form of individual differences on single variables.

To what extent do the findings on aggregation resolve the person-situation debate? The findings indicate that, with appropriate aggregation, (a) the .30 barrier can readily be breached, (b) behavior that is situationally specific and temporally unreliable when based on single observations can often be demonstrated to be highly general and stable when appropriately aggregated, and (c) there are stable, cross-situationally broad response dispositions, or traits, as indicated not only by high split-half reliability coefficients, but by significant correlations between objective measures of behavior and trait scores on personality inventories.

Resistance to the Proposed Solution

The suggestion that the resolution of the person-situation debate lies in the recognition that behavior can be highly situationally specific at the individual-item level and general at the aggregate level received a mixed reception. Abundant confirmatory evidence appeared from a variety of sources (e.g., Cheek, 1982; Eaton, 1983; Moskowitz & Schwarz, 1982; Rushion, Brainerd, & Pressley, 1983). Miselch and his associates conducted research in which they examined multiple items of behavior in multiple situations in multiple individuals and obtained results that replicated Epstein's findings on aggregation (Mischel, 1984; Mischel & Peake, 1982). Mischel accordingly revised his position on traits and correctly stated that, rather than being untenable (Mischel, 1968), traits were weak predictors of behavior in specific situations (Mischel & Peake, 1982). Nevertheless, there was not complete convergence. Mischel continued to maintain that broad cross-situational consistencies remain "elusive," even with reliable measures (Mischel & Peake, 1982, p. 735). In the same article, he insisted that nothing new had been added to the person-situation debate by the recent work.

1 Unlike the four classic studies, aggregation in these studies was not done separately over items and occasions. Records were kept of a sample of situations and responses that occurred in everyday life, and aggregation was simultaneously conducted over situations and occasions. More specifically, behavior on one set of days was averaged and compared with behavior averaged over a different set of days, which invariably included a different sample of situations. Accordingly, only one procedure for aggregation was necessary in order to establish stable, broad response dispositions. For the four classic studies, aggregation over days alone would have established stability, but not generality of behavior, and therefore it could not adequately measure a trait.
on aggregation, as presumably similar procedures had been used and similar analyses conducted in earlier studies, and such studies he noted, had fueled the fires of the debate rather than dampening them. We return to this issue after briefly considering two other sources of resistance.

The second source of resistance to the conclusion that there are widespread reliable generalities in behavior arose from the belief that if behavior is highly situationally specific, it cannot also be general. Allen & Potkay (1973, 1983), for example, concluded that behavior is far more specific than it is general because they observed, in an investigation of daily emotions conducted over many days, that the variance within subjects was greater than the variance between subjects. They argued, therefore, that any demonstration of stability by aggregating data over many days had to be spurious, as it treated true fluctuations in behavior as error variance. There are two problems with Allen and Potkay’s conclusion. One problem is that they treated all variance in their study that was not between-subjects variance as legitimate intrasubject variance, that is, as a measure of reliable fluctuations. Thus they failed to separate error, or random, intrasubject variance from true state variance. With such a procedure, the poorer the measures they used and the more random the responses elicited, the more they would obtain results supporting their position. The other problem is that they did not recognize that one psychologist’s truth variance is another’s error variance; that is, when the aim is to uncover stable dispositions, momentary states can be regarded as error variance, although for other purposes such states might reasonably be studied in their own right (see Bem, 1977, for similar as well as additional criticisms of the Allen & Potkay, 1973, study). In short, they did not realize that behavior can be either variable or stable, depending on the level of analysis, and a demonstration of one effect at one level, therefore does not rule out the other at a different level.

A third source of resistance, which at first glance appeared to offer more substantive evidence that aggregation produces spurious behavioral stability, was presented by Day, Marshall, Hamilton, and Christy (1983) in a computer-simulation study. These investigators simulated the responses of 200 subjects over 40 trials under a variety of conditions. They concluded that even trivial amounts of actual consistency could, through aggregation, be made to appear substantial. As an example of their procedures, in one study they randomly assigned scores between 1 and 20 to all but 10 of their 200 subjects. To these 10 subjects, they assigned constant numbers across all 40 trials. They were apparently unaware that by assigning random numbers between 1 and 20 to almost all of their subjects, they had fixed these subjects’ true means at precisely 10.5 and thus effectively reduced the variance of this group's true scores to zero. Thus, given sufficient aggregation, all of these subjects’ scores would converge on 10.5, and the source of whatever correlation Day et al. (1983) obtained would have to be the variation contributed by the extremely different scores of the 10 remaining subjects. It is no wonder, then, that the "trivial" number of subjects with constant scores resulted in high stability coefficients after a modest amount of aggregation. Space limitations do not permit a summary of all of their studies. In contradiction to their contention that aggregation produces spurious stability coefficients, their results provided strong support for the value of aggregation, for no amount of aggregation in the absence of an empirical relationship could produce a significant correlation, and any orderliness they built into their data was invariably detected with the aggregated, but not the unaggregated, data. (For a more thorough discussion of the Day et al., 1983, study, see Epstein, in press.)

As noted previously, one source of resistance was from Mischel, who denied that anything new had been demonstrated, since others had examined behavior over repeated occasions and situations before.

As tempting as simple solutions might be, the problems raised by the consistency debate cannot be dismissed as the result of forgetfulness for the basic concepts of measurement error... How can the use of aggregated measures resolve a debate in the 1980s that it was unable to resolve throughout the 1940s, 1950s, and 1960s?... Reliability is doing nothing more (or less) for Epstein than it did for any of the earlier large-scale assessment projects that sought and obtained it (Mischel & Peake, 1982, p. 732).

Mischel and Peake further concluded that the findings uniformly vindicated his position that cross-situational correlations are low, characteristically about .20, and that these low
correlations cannot be attributed to unreliability of measurement.

The results of the Carleton project are not unique. A survey of the early studies of the cross-situational consistency of behavior (Peake, Note 1) indicates that the obtained results are quite consistent with past findings. These early studies, including studies by Hartshorne and May (1928) of honesty, by Newcomb (1929) of introversion-extroversion, by Allport and Vernon (1935) of expressive movements, and by Duda (1936) of punctuality, routinely employed repeated observations of each behavior to increase the reliability of their measures. Each of these investigators reported substantial reliability coefficients, not as a solution to the consistency problem, but as one index of the adequacy (reliability) of their aggregate measures. More importantly, each of these studies obtained cross-situational correlation coefficients of around .20 when using the reliably aggregated measures. Although Epstein proposes aggregation as a potential cure-all for the consistency problem, this cure has been employed routinely for years, and its use only serves to document more clearly the perspicacity of the phenomena of behavioral discriminability (Mischel & Peake, 1982, p. 736).

The existence of discriminability in behavior has never been a source of contention in the person–situation debate. No one contested then or contests now the fact that people behave differently in different situations. The issue has always concerned the existence of broad generalities in behavior. We have always agreed with Mischel that behavior is discriminative, but we have also believed it to be general. The most controversial aspect of Mischel’s position has been his assumption that, given a high degree of discriminability, there cannot be behavioral generality, and as a result, the concept of traits is untenable.

The Four Classic Studies

We examine in detail the four classic studies cited by Mischel for two reasons. They provide an interesting lesson in the sociology of science, for they indicate how a prevalent paradigm, in this case one that emphasized relationships between individual items of behavior, led highly sophisticated researchers to overlook the simplest of truths. In addition, as they examined behavior over both occasions and situations, they contain ideal data for examining the two critical features of traits, namely that they are temporally stable and cross-situationally broad. We demonstrate that when the data from all four studies are properly analyzed, they lead to identical conclusions that confirm the existence of broad, stable response dispositions and that indicate that cross-situational relationships for specific behaviors, reliably measured, are typically much higher than Mischel and others have concluded. This view is supported by Mischel and Peake’s (1982) own data, but is not consistent with their interpretation.

Some Basic Psychometric Principles

Before proceeding further, it will be helpful to outline what is required to properly evaluate the existence of broad response dispositions when the data consist of various items of behavior observed in multiple situations on several occasions. First, it is important to establish that the items selected are all measures of a common construct. The recommended procedure (see Jackson & Paunonen, in press; Loewinger, 1957; Nunnally, 1978) is to identify a population of items on a theoretical basis, select a representative sample of items from the domain of relevant items, and empirically establish that the items are, in fact, measures of a common construct by employing appropriate psychometric procedures, such as factor analysis or item–whole-test correlations. Unless this is done, the inclusion of a few inappropriate items can drastically reduce the average interitem correlation.

The next step is to establish the reliability of each of the items. This can be accomplished by computing split-half, odd–even correlations over the occasions on which the items have been administered, or by using Cronbach’s alpha, which is equivalent to the average correlation among the item scores. With either procedure, the construct validity is not one that is as good as the construct validity of two occasions. However, the behavior is sampled in the studies described above and the reliability is on the basis of reliability formulae, the item of behavior reliability. The observations, very high item reliability, indicate cross-modal relationships between different items on the same scale, and that the same behaviors (whether observed in the same or different situations) and the same behavior relationships can be assessed as well by the number of observations. Investigators vary in their studies in their data or the impossible to compare their results. Henceforth, we use the word "magnitude" to refer to the number of observations.

2 The argument may be repeated that if items are psychometrically selected to measure a common construct, it forces an outcome favoring generality in behavior. The answer to this argument is that all behavior varies along a dimension of generality. Thus, the issue of whether behavior is categorically specific or general is meaningless. A meaningful question is how general a particular kind of behavior is. This question can only be answered by appropriate procedures for measuring the opportunity for demonstrating generality. Unless one uses such a procedure, a failure to demonstrate generality can be attributed to the inadequacy of one’s methods. Expressed otherwise, one cannot demonstrate generality that does not exist, but one can fail to demonstrate generality that does exist. Undoubtedly, some variables are relatively narrow, or cross-situationally specific, others are highly broad, and most fall in between these extremes. As Allport (1937) has noted, which variables do what, both within and across persons, has important implications for personality theory.
PERSON-SITUATION DEBATE

relation among all split-half combinations. With either procedure, occasions are treated as analogous to items in a test. This procedure is not one that would normally be followed in the construction of paper-and-pencil tests, which are usually administered on no more than two occasions to establish test-retest reliability. However, when the same items of behavior are sampled over many occasions, as in the studies reviewed in this article, it is both possible and desirable to establish the reliability of individual items aggregated over occasions in an analogous manner to the usual procedure for establishing internal consistency reliability for responses aggregated over items. The resulting reliability coefficient indicates the degree to which mean performance on an item sampled over several occasions can predict mean performance on the same item sampled over other occasions. As can be demonstrated with the Spearman-Brown formula, the more occasions over which an item of behavior is averaged, the higher is its reliability. Thus, with a sufficient number of observations, it should be possible to obtain very high item reliability. Accordingly, if the investigator wishes to examine, for example, cross-modal relationships (i.e., relationships between different kinds, or modes, of behavior in the same situation), or cross-situational relationships (i.e., relationships between the same behavior in different situations), correlations can be obtained that are not limited by low reliability. It is important to recognize that for the kind of data in question, such limitations in reliability are arbitrarily established by the number of occasions over which an investigator decided to observe and aggregate observations. In the studies that follow, investigators varied widely both within and among studies in the degree to which they aggregated their data over occasions, thereby making it impossible to meaningfully interpret and compare the cross-situational correlations they obtained, unless the correlations are corrected for attenuation, an issue to which we turn next. (Henceforth, unless there is reason to do otherwise, we use the term *cross-situational* in a generic sense to include cross-modal and all other cross-item relationships, as the same principles apply equally well to all. The major distinction we retain is between cross-occasion, or within-item relationships, on the one hand, and cross-situational relationships, generically considered, on the other.)

A relationship between items of less than perfect reliability can be corrected for attenuation, meaning that an estimate can be made of how high the relationship would be if it were not limited by the unreliability of measurement. The formula for making the correction simply requires the obtained correlation to be divided by the product of the square roots of the reliability coefficients. If reliabilities are respectable, let us say in the vicinity of .80, correcting for attenuation would not make a very great difference. A cross-situational relationship of .30, for example, between variables that each have reliabilities of .80, would increase to .38. On the other hand, if the reliability coefficients are more modest, let us say .50, the correlation will increase from .30 to .60. It follows that an investigator can always obtain evidence of a low cross-situational relationship, given a high true relationship, by employing measures of modest-to-low reliability, which for the kind of study in question can easily be accomplished by simply failing to observe the behavior over sufficient occasions. It is thus important to correct such low cross-situational relationships for attenuation due to the unreliability of measurement, unless this is done, the relationships will be arbitrarily limited by the amount of aggregation that was done over occasions. Reasonably accurate corrections for the kind of data in question can be anticipated so long as the total sample studied is large enough to produce accurate correlations. Normally, this would take at least 100 subjects, but the effect of a smaller sample can be mitigated by averaging over sufficient occasions and by presenting results averaged over many relationships, rather than presenting only single relationships.

Before computing an average interitem relationship, unreliable items should be discarded or the relationships, before averaging, should be corrected for the unreliability of the individual items. As already noted, unless this is done, low interitem correlations would be obtained simply because some items were poorly measured or were not aggregated over enough occasions. The average interitem correlation informs us of how well performance on one item can on the average predict performance on another. Such relationships are
generally low among even good exemplars of a trait, unless they are nearly identical in meaning, not only for reasons of low reliability but because individual items are normally narrow in scope. Although individual items therefore tend not to overlap much with each other, they may well all overlap with a common construct. Thus, despite low correlations with each other, together they can provide a powerful measure of the overall construct. This logic is used to construct intelligence tests of high reliability and predictive validity. It follows that the appropriate test for broad, reliable generalities in behavior is the correlation between aggregates of items, not individual items. A test of how well one aggregate of items can predict another, similar aggregate can be accomplished by computing split-half, odd–even correlations, or Cronbach’s alpha for the total sample of items. A convenient computational procedure is to upgrade the average interitem correlation for the total number of items by using the Spearman-Brown formula.

With this background on what is required to analyze the stability and generality of behavior and to evaluate the existence of traits with data derived from administering many items on multiple occasions, we can now examine the four classic studies cited by Mischel.

The Hartshorne and May Studies of Honesty

The Hartshorne and May studies were one of the most intensive series of investigations of personality ever undertaken. The studies were conducted by a large team of researchers over a 6-year period. The 1928 report alone involved testing a national cross-section of more than 8,000 children with a variety of tests of deception and honesty. The aims of the study were to determine whether a broad trait of honesty existed, what its nature was, and how such knowledge could be used to implement moral training in religious education. It is noteworthy with respect to the issue of how biases and paradigms influence research that the project was conducted under the supervision of Thorndike, who was director of the Division of Psychology of the Institute of Educational Research at Teachers College, Columbia University, under whose auspices the research was conducted. Thorndike was a staunch advocate of specificity in behavior (recall the quote at the beginning of this article) and considered those who believed in generality in behavior to have a misguided, unscientific viewpoint.

On the basis of their findings, Hartshorne and May (1928) concluded that honesty is not a unified trait, but consists, instead, of highly situationally specific habits. At the end of their second volume, they stated, on the basis of their overall evidence, “that honesty or dishonesty is not a unified character trait in children of the ages studied, but a series of specific responses to specific situations” (1929, Vol. 2, p. 243). Not surprisingly, the Hartshorne and May studies were widely cited as evidence that behavior is situationally specific and that the concept of traits had, accordingly, been discredited by the best evidence by far available. Let us carefully examine and evaluate what Hartshorne and May actually did and found.

Hartshorne and May indicated that the initial assumption under which they operated in selecting their items, and which can be assumed to have influenced their decision not to winnow the items by psychometric procedures, was that honesty is “not a generic trait but an accumulation of loosely connected habits” (1928, p. 141). They provided no theoretical rationale for their initial selection of items, but noted, without further elaboration, that some prescreening was done. Thus the initial bias in item selection favored a situationist position, and in the absence of procedures to ensure that the items were measures of a common construct, it can be expected that some might not be. An examination of their table of interitem correlations (1929, Vol. 2, p. 123) does not disconfirm this suspicion. There is a considerable range of relatedness among the items, with several correlations exceeding .40 and others being negative. The reliabilities (1929, Vol. 2, p. 122) are also highly variable, with a range of .24–.87. Among their nine items, two are listed as having “unknown reliability.” These items produced among the lowest correlations in their interitem correlation matrix, which leads one to suspect they have low reliability. The strongest evidence that Hartshorne and May cited in support of their conclusion that honesty is not a general trait was an average interitem correlation of .23. Interestingly, when the 7 items for which reliability

coefficient reliability

...
In general, test for test, we are able to measure deception almost as consistently and with almost as little error as we are able to measure intelligence. Just as one test is an insufficient and unreliable measure in the case of intelligence, so one test of deception is quite incapable of measuring a subject’s tendency to deceive. That is, we cannot predict from what a pupil does on one test what he will do on another. If we use ten tests of classroom deception, however, we can safely predict what a subject will do on the average whenever ten similar situations are presented. (1928, p. 135)

Hartshorne and May added that their findings failed to provide support for a “unified” trait of honesty. They all but defined the possibility of broad traits existing by requiring bimodal frequency distributions, reliability coefficients greater than .90, and a very high level of homogeneity among items, so that the trait could not be subdivided. Given such criteria, not even intelligence would qualify as a trait. Nevertheless, they stated that, in deference to the views of others, they would not claim that honesty is not a trait, but only that it is not a “unified trait,” meaning that it can be subdivided. This nicely, not surprisingly, was lost on their readers, particularly as Hartshorne and May repeatedly emphasized that their results supported a specific position.

There is further evidence for a general trait of honesty in the Hartshorne and May data. They obtained a correlation of .32, uncorrected for attenuation, between teachers’ ratings of overall honesty and a composite of their objective tests of honesty. They noted that this correlation is higher than the correlation of .10 obtained when the teachers’ ratings were restricted to observations of cheating in the classroom. On the basis of these correlations, they concluded that the teachers’ overall ratings had to have been made with respect to some global attribute that extended beyond the classroom in order to have the ratings correlate with their combined objective tests, which included behavior in different settings. Moreover, if their correlation of .32 is adjusted for the unreliability of their measures (.73 for the sum of the objective measures and .50 for the teacher ratings), it is increased to .53, providing impressive evidence for a broad trait of honesty.

Although Hartshorne and May (1928, 1929) stated that honesty consists of a combination of specific habits, when they discussed the broader implications of their study, they consistently referred to honesty as if it were a general trait. Thus, they noted that “deception runs in families in about the same way as intelligence, eye color, or height” (1928, p. 14), that it “goes by gang and classrooms” that “a pupil resembles his friends in his tendency to deceive,” that it occurs less often when an atmosphere of cooperation and good will is established between teachers and friends, and that forced attendance at Sunday School classes in which honesty is taught leads children to behave less honestly (1928, pp. 14-15). To instill honesty, they recommended training in behaviors “characterized by integrity of performance and intelligent grasp of the social significance of honor” (1928, p. 15). Apparently, the usefulness of traits as descriptors of broad dispositions is more than they could resist in their attempts to understand and make practical recommendations for controlling real-life behavior.

A factor analysis of the Hartshorne and May data by Burton (1963) provides additional support for honesty as a broad trait. Burton summed up his findings as follows:

These analyses conform in the main to our hypothesis and lend support to the generality position. In all cases, the first component accounts for at least twice as much variance as the second component. Also, the loadings on this component are all positive and exceed .50, with the exception of the athletic tests for the matrix of uncorrected intercorrelations (p. 487).
Burton noted that Hartshorne and May set an extremely stringent criterion of acceptable evidence for a general trait of honesty, namely a predictive reliability of .90, which nevertheless could have been obtained by eliminating a few tests of particularly low reliability and using a total of 21 tests similar to those retained. Burton also summarized several other studies that provided evidence for a general trait of honesty. He observed that Maller (1934), a co-author of the Hartshorne, May, and Maller (1929) volume, reanalyzed some of the data of Hartshorne, May, and Shuttleworth (1930) and found evidence for a common factor. Brogden (1940), in an examination of the responses of 100 middle-class boys to a variety of tests, including several from the Hartshorne and May (1928) study, also found a clearly defined factor of honesty. Barbu (1951), who conducted a series of studies of honesty in 14-year-old boys, found an average intercorrelation of .46, which led him to conclude that there is a general trait of honesty.

It may be concluded that the overall evidence provides strong support for a general trait of honesty and that the average correlation between individual items corrected for unreliability is substantial.

The Newcomb Study of Extraversion

Newcomb (1929) wished to determine whether a broad trait of extraversion could be identified in the behavior of boys at a summer camp. His subjects consisted of two groups of 27 and 24 problem boys who were observed during successive 30-day camp sessions. To obtain behavioral items, Newcomb began by extracting nine subtraits from a review of the literature on extraversion. Included were categories such as a tendency to seek the limelight and a tendency to be impetuous. Thirty items that could be objectively scored were constructed as exemplars of the subtraits. Daily records were kept by counselors of the behavior of each boy on these 30 items. Four items were eventually discarded because of their infrequent occurrence. In the evening, retrospective ratings were made by other counselors of the extent to which they believed the boys had exhibited each kind of behavior. The ratings served as more global measures than the objective records and were considered to provide a reliability check on the daily records. Examples of specific items were “Did he speak before the group at campfire?” and “Did he get into scrapes with other boys?” No clear rationale was presented for the selection of specific items and for their division into subtraits, neither were psychometric procedures used to establish that the nine categories were subtraits of a broader construct, nor that the items within a subtrait were measures of a common construct.

Although the temporal reliability of the individual items was computed, the results were not presented. However, the average odd–even reliability coefficient for the individual items was reported to be .78, with a range of .46–.93. Interitem correlations were not disattenuated for unreliability, and in the absence of information on the reliabilities of the individual items, it is not possible to make such corrections retroactively. However, it is likely that the average interitem correlation was considerably reduced by the inclusion of items with low reliability.

Rather than compute the entire matrix of intercorrelations for the 26 items for both groups of subjects, which would have been a formidable undertaking at the time, Newcomb elected to compute a sample of 56 correlations between items within the same category, and 56 control correlations between items in different categories. The mean of the 112 interitem correlations for the daily records was .14. The difference between the correlations within and between categories was negligible. The mean interitem correlation for all of the ratings was .45, and the mean correlation for the ratings within categories was only slightly higher than that for between categories. A coefficient of internal consistency reliability was not computed, neither were relationships between aggregates of items examined. Newcomb (1929) attributed the lower interitem correlations for the records than for the ratings to the greater objectivity of the former. He assumed that the ratings were influenced by halo effects and by the raters’ implicit theories of how the different variables were related. The data that were more objectively recorded, he argued, more accurately represented the true state of affairs, which was that the interitem correlations were too low to justify the concept of a broad trait of extraversion. Moreover, he noted that the finding of the same results against a basis of summation seems just and extremely difficult such measures as we have.

The appendix is a strong support of the item structure, construction of .40 and thus looking at it, one can see one analysis, showing a broad average of an intercorrelation of .81. The correlation is between categories and thus the relationship of extraversion.
finding that relationships between items within the same category were no greater than between items in different categories argued against the existence of even of subtraits. On the basis of the overall evidence, Newcomb (1929) summarized his view: "The conclusion that seems justifiable, therefore, is that if introvert and extrovert types do exist, they do not show any distinction that is clearly measurable by such means of recording observed behavior as were used in this study." (p. 113).

The interitem correlations, presented in an appendix, ranged from −.64 to .79. Given the strong negative correlations, it is unlikely that the items are all measures of the same construct. A clue that there is a core of a broad construct, however, is provided by the observation that 29% of the correlations fall between .48 and .79. The .14 interitem correlation is thus very likely a severe underestimate resulting from the inclusion of inappropriate items and items of low reliability. However, even if one accepts the .14 average interitem correlation, it does not rule out the existence of a broad trait of extraversion. With 26 items, an average interitem correlation of .14 produces an internal consistency reliability coefficient of .81. Thus, had Newcomb examined the correlation between aggregates of items, he would have obtained evidence for a broad trait of extraversion.

Fortunately, we were able to reanalyze Newcomb’s study, as the raw data are presented in his Tables 1 and 2. Accordingly, the entire correlation matrix for the merged groups was generated with the aid of a computer. The mean interitem correlation was .13, which is similar to the .14 reported by Newcomb for his smaller sample of correlations. The correlation matrix, subjected to an oblique factor analysis, produced four clearly defined factors, none of which resembled any of Newcomb’s subtraits. Thus, Newcomb’s finding that interitem correlations were no greater within than between categories can be attributed to the limitations imposed by intuitively assigning items to subtraits. The four factors uncovered by factor analysis were labeled, in descending order of percent variance accounted for, Boldness and Assertiveness, Uncooperativeness and Insecurity, Activity Level, and Talkativeness. The alpha internal consistency reliabilities for the four corresponding scales composed of items with the highest factor loadings were .71, .78, .68, and .81, respectively. If the same scales were composed of 20 items similar to the ones included, all of the scales would have internal consistency reliability coefficients greater than .87. All of the scales were positively and significantly correlated with each other except the second scale. Thus, three of the scales can be considered to be components of a broader extraversion scale. The remaining scale, Uncooperativeness and Insecurity, apparently measures an independent dimension that may be related to the selection of problem boys for the study.

In conclusion, the Newcomb study, when appropriately analyzed, provides evidence for a broad trait of extraversion that, like intelligence and honesty, is not unified, but is divisible into subtraits. The initial failure to find evidence for a broad trait and for subtraits can be attributed to relying on intuitive procedures for the selection and categorization of items, to a failure to take into account the unreliability of items, and to an emphasis on relationships among individual items of behavior rather than among aggregates of items.

The Allport and Vernon Study of Expressive Movement

Allport and Vernon (1933) examined 40 simple expressive movements in 25 male subjects in three experimental sessions. Most tests were presented on more than one occasion, thereby allowing temporal reliability to be established. Each test yielded several scores, which were combined into 34 composite measures. Examples of single scores are verbal speed, drawing speed, area of writing, length of stride, and finger-grip pressure. An example of a composite variable is speed of tapping, which consisted of tapping with the right and left fingers in one session and tapping with the right finger in another. Twenty-seven of the measures were objectively scored and the others consisted of more global ratings.

Although a factor analysis of 26 variables is not recommended with only 51 subjects, the problem is mitigated by highly reliable data that were aggregated over a month of daily observations. More important, the factors obtained were coherent and interpretable.
No theoretical basis or systematic procedure was described for selecting items. Selection was based on intuitive judgment and practical considerations regarding administration and scoring. However, the items had to be positively correlated with each other before being combined into composite scores. Thus, procedures for item selection ensured coherence within, but not among, composite scores. An examination of the interitem correlation matrix for composite items reveals that the correlations varied widely, with a range of .67 to .47. Some of the composite items produced predominantly negative correlations with the others and should not have been included as measures of a common construct. The composite item reliabilities ranged from .67 to .96, with a median of .83, indicating a fairly respectable reliability for most items. Allport and Vernon decided not to disattenuate their interitem correlations for unreliability of measurement because, with an $n$ of only 25, reliability coefficients could be assumed to be too inaccurate to provide accurate disattenuation estimates.

It appeared to Allport and Vernon that 24 of their objective variables might reasonably be considered to be exemplars of a broad trait of "vitality." In order to test this hypothesis, they computed the average interitem correlation among the 24 items and obtained a coefficient of internal consistency reliability by upgrading the average interitem correlation with the Spearman-Brown formula. The average interitem correlation was .055, and the internal consistency reliability coefficient was .59. Although the value of .59 provided evidence for a global trait of vitality, they could not accept such a conclusion, given the low average interitem correlation. They were suspicious of the .59 internal consistency reliability coefficient and thought it might be an artifact resulting from their use of the Spearman-Brown formula. Accordingly, they randomly divided their 24 measures into two tests of 12 items each and correlated them, which produced a correlation of .58, confirming the Spearman-Brown calculation. However, they still could not accept the conclusion that their data supported the existence of a global trait when the average interitem correlation was only .055. They stated:

Theoretically, this sum of the 24 variables actually measures a psychomotoric factor with a considerable degree of validity. Moreover, the variables themselves possess independent reliability of but little over +.80, on the average, so that if our general factor were corrected for attenuation, it might well obtain a much higher validity. Nor is this as absurd as it appears at first sight. . . . And yet in a case like this, it seems to the writers more psychologically sound to admit that the variables possess an almost negligible average correlation of +.055, rather than to claim that they have measured a general factor with a validity of near .90. (pp. 45–46)

The assumption that a "validity of near .90" would occur after disattenuation may have helped Allport and Vernon feel better about rejecting a conclusion they did not wish to accept, but it is clearly a gross exaggeration. As their reliability coefficients were relatively high to begin with, correcting for attenuation would most likely have raised their internal consistency reliability coefficient of .58 to no more than .70.

Having rejected the presence of a global trait, Allport and Vernon were content to find evidence for traits of lesser scope. They noted that support for such traits was provided by the reliability and cohesiveness of their composite measures and by coherent clusters of intercorrelated measures, which they identified as an Areal factor (expansiveness), a Centrifugal factor (outward movement tendency), and an Emphasis factor (forcefulness of movement). Allport and Vernon considered that the average interitem correlations for these "factors" of .33, .30, and .25, respectively, and their coefficients of internal consistency reliability of .82, .75, and .82, provided satisfactory evidence that their intuitively constructed factors corresponded to traits.

Several aspects of the Allport and Vernon study warrant comment. First, their reliance on casual, intuitive procedures for selecting items did not rule out the inclusion of inappropriate items, which could have drastically lowered their average interitem correlation. Second, their small sample of 25 subjects tested with the same items on only two occasions resulted in reliability coefficients that themselves were too unreliable to be used for disattenuating their interitem correlations, and this further contributed to low interitem-correlations. Third, the study provides a dramatic illustration of the same source of confusion...
characteristic of the other studies. Namely, Allport and Vernon, who were trait theorists, no less than Harshbrough and May, who were specialists, overemphasized the importance of average interitem correlations and de-emphasized the importance of their internal consistency coefficients based on aggregates of items in evaluating evidence for the broad response disposition that they sought. Given Allport and Vernon's average interitem correlation of .055, a reliable test could be constructed of a broad global trait of vitality. The Spearman-Brown formula reveals that with 100 items similar to those employed, an average interitem correlation of .055 would produce an internal consistency reliability coefficient of .85. Nor is the use of 100 items as prohibitive as it may first seem, as many of the items, such as striding across a room, took only a few seconds to administer. Moreover, the .055 interitem correlation, in all likelihood, is a gross underestimate of what would have been obtained had appropriate item-selection procedures been followed and poor items discarded. If the average interitem correlation were raised to only .10, a .85 coefficient of internal consistence reliability could be obtained with 50 items.

The Dudyca Study of Punctuality

The Dudyca (1936) study is the only study that examined cross-situational generality, strictly defined, as relationships between identical behaviors in different situations. In all cases the dependent variable was minutes late or early to a scheduled event. Unobtrusive records were kept of the arrival times of 307 male and female college students at various events, including (a) 8:00 a.m. classes, (b) meals at the dining commons, (c) individual conferences, (d) extracurricular activities (e.g., college band and group singing), (e) vespers (compulsory Sunday evening religious services), and (f) entertainment events, including watching athletic events. The number of subjects observed for the different events and the number of occasions observed per student varied widely. The number of subjects ranged from 228 for vespers to 38 for extracurricular activities. The number of observations per subject ranged from 34 for 8:00 a.m. classes to 4 for vespers.

No rationale was presented for, nor were psychometric procedures used in, item selection. Apparently they assumed that a trait of punctuality applies equally well to all events that can be timed. Thus, no consideration was given to whether a voluntary activity was involved, as in attending an extracurricular event, or whether the event was compulsory, as in attending Sunday evening church services. The fact that these two events were negatively correlated (-.19) was ignored, and both were assumed to be equally valid measures of the same construct. Another unfortunate decision was to treat early and late arrivals along a single continuum. Thus, arriving at an event 10 min early on one occasion and arriving 10 min late on another occasion, and they were treated as equivalent to coming on time on both occasions.

No information was presented on the reliability of the individual items over occasions, nor was a coefficient of internal consistency reliability computed. In order to determine whether the data provided support for a general trait, Dudyca relied on the average relationship between individual items. The average interitem correlation was .19, with a range of -.44 to .19 among the correlations. Dudyca, who believed in traits, was disappointed by the low interitem correlation, and therefore speculated that a Pearson Product-Moment Correlation might not be the most appropriate statistic for detecting a general component among the measures:

Thus far we have been considering evidence which seems to favor the conclusion that conduct is specific; that is determined by particular situations rather than general traits of character of personality. Harshbrough and May using the same type of evidence decided in favor of the specificity of honesty and deceit, as we have already noted. However... the particular conclusion which they drew may be due entirely to the kind of statistical procedure they used. (p. 28)

It occurred to Dudyca that contingency coefficients might succeed where correlation coefficients had failed. However, when he dichotomized his data and computed contingency coefficients, the results turned out to be the same. Undaunted, he turned next to chi-square (apparently unaware that contingency coefficients and chi-square are closely related) which finally gave him what he had been seeking. Of the 13 relationships, 5 were significant
at the .05 level, 3 were between the .10 and .20 level, and 5 were not significant at the .20 level. Viewing these results, Dudycha concluded that chi-square was a more appropriate statistic for detecting relationships than correlations and contingency coefficients, although he did not indicate why. He interpreted the results from the chi-square analysis as follows:

In other words, in five cases, the agreement is slight and insignificant, whereas in eight the association between the behavior of students in one situation and another is significant or nearly so. We may say that, on the average, the association between the behavior of students in one situation and another is significant more often than not. This being true, we are led to conclude that the evidence points to punctuality as a general trait. (p. 30; italics in original)

An examination of Dudycha's interitem correlation matrix reveals that six relationships were significant at the .05 level and two at between the .05 and .20 level. Thus, the results from the correlation coefficients in terms of the number of significant relationships correspond to the results from the chi-square analysis, so Dudycha could have saved himself the trouble of further computation.

The average interitem correlation of .19 obtained by Dudycha is very likely a considerable underestimate because it includes inappropriate items and items that are very likely of low reliability, as they were aggregated over only a few occasions. In the absence of information on the reliability of the individual items, the .19 average interitem correlation cannot be disattenuated for unreliability. Nevertheless, using the .19 figure, the internal-consistency reliability for Dudycha's 30 items, computed with the Spearman-Brown formula, is .82. Thus despite its shortcomings, Dudycha's study, like the others, provides strong evidence for a general trait, although ironically it is not the evidence that Dudycha himself cited.

**Overall Conclusions From The Four Classic Studies**

All of the researchers made the error of evaluating the existence of broad response dispositions on the basis of the average relationship between single items of behavior, and either did not compute, or de-emphasized, relationships between aggregates of items. Second, the procedures used in item selection were deficient in all of the studies. There seemed to be little awareness that including a few poor items could drastically lower the average interitem correlation. Items were typically selected on the basis of face validity and availability, rather than sampled from a carefully articulated domain on the basis of a theoretical analysis of the nature of a construct. Nor were procedures used to ensure that the items finally retained were measures of a common construct. Typically, some items were included that were either unrelated or negatively related to the other items. Third, the reliability of the individual items was usually not adequately taken into account. With one exception, reliability coefficients were either not computed or, if computed, they were not used to disattenuate the correlations among items, despite the observation that some of the items were much less reliable than others. Fourth, although it was uniformly noted that interitem correlations were low, Mischel and Peake's conclusion that cross-situational correlation coefficients are typically in the vicinity of .20 is unjustified, because the average interitem correlations were underestimates as a result of the inclusion of inappropriate and unreliable items, and many of the relationships were either completely cross-modal, or cross-modal in addition to cross-situational. Nevertheless, an inspection of the interitem correlation matrices revealed that correlations in the vicinity of .40 were common, and that, had inappropriate items been eliminated and reliability properly taken into account by either disattenuating relationships or eliminating unreliable items, average interitem correlations, except for the Allport and Vernon study, would have been in the vicinity of .40. Fifth, and most important, when coefficients of internal consistency were computed, unequivocal evidence was found for cross-situationally broad response dispositions or traits. Expressed otherwise, it was uniformly demonstrated that, although it was not generally possible to predict with much accuracy performance on one item from performance on another item, it was possible in all cases to predict with high accuracy mean performance on a group of items from mean performance on another aggregate of items across occasions, thereby indi-
PERSON-SITUATION DEBATE

indicating the presence of stable, broad response
dispositions, or traits.

The Carleton Behavior Study

The Carleton study was undertaken by Mischel and Peake (1982) to investigate behavioral consistency while adhering to the requirements for a properly conducted investigation, which according to the authors,

requires data aggregated to achieve reliability... needs to explore cross-situational consistency in behavior with appropriate and reliable measures sampled across a range of presumably similar situations... and needs to be informed by some conceptualization... of how behavior is organized.... That is exactly what we have been trying to do over the last 4 years, studying behavioral consistency among college students at Carleton College in Northfield, Minnesota. (p. 734)

One of the major conclusions of the Carleton study is that the presumed "reliability solution" proposed by Epstein fails to provide evidence of impressive levels of cross-situational generality in behavior. "Although aggregation over occasions has the desirable effect of enhancing reliability, it does not provide a simple solution to the consistency paradox" (Mischel & Peake, 1982, p. 736). Here, as elsewhere, Mischel makes the error of considering only one kind of aggregation, aggregation over occasions, and, relatedly, one kind of reliability, namely temporal reliability. He fails to consider aggregation over items and its effect on generality, as indicated by internal consistency reliability. As a trait refers to a stable, broad response disposition, it is necessary to consider both kinds of reliability when assessing the existence of traits. Simply establishing the temporal reliability of individual items is clearly insufficient.

The results of the Carleton study have been widely cited by Mischel (1983, 1984; Mischel & Peake, 1982, 1983) as providing evidence against the widespread existence of broad generalities in behavior. Because the study has been described by Mischel as a model study that examined objective behavioral data, that carefully selected an ecologically representative sample of exemplars of traits, and that appropriately took into account principles of reliability, it might appear to be a definitive study clearly establishing that there is relatively little generality in behavior. The study therefore warrants particularly careful examination. The Carleton study was conducted as a dissertation by Peake (1982) under the supervision of Mischel. As the study has not been adequately described in readily available sources before, it is described in some detail here.4

The subjects were 63 undergraduate male and female students at Carleton College who were monitored over a 10-week period with respect to behaviors deemed relevant to friendliness and conscientiousness. In order to obtain an ecologically representative sample of such behaviors, a group of subjects was asked to list the 20 most significant behaviors they had engaged in during the past 48 hours. They then categorized their items according to 10 personally relevant traits of their own choosing. Friendliness and conscientiousness were selected for study because they were listed frequently and had a relatively large number of exemplars. Immediately a problem with the procedure becomes apparent. By requiring subjects to list 20 different behaviors in a 2-day period, subjects were forced to include trivial, routine events, such as eating meals and attending classes. There is reason to suspect that the investigation of such events is less revealing of significant personal attributes than the investigation of more ego-involving, personally significant behaviors (Allport, 1943; Epstein, 1980b, 1983a). Thus, even if the study found little evidence for behavioral generality and the existence of traits, the results might be specific to routine, relatively trivial behaviors. In any event, one has to begin somewhere, and a well-conducted study of an ecologically representative sample of even routine, non-ego-involving behaviors would be of considerable importance. Unfortunately, the initial intention of obtaining an ecologically representative sample had to be abandoned when it became evident that many of the items listed could not be measured. Moreover, once initial measures were selected, others were added because they could easily be obtained by using the same procedures. Thus the final selection

4 See Jackson and Pauwens (in press) for a critical evaluation and reanalysis of the Peake dissertation, which, although based on somewhat different procedures than we have used, arrives at similar conclusions.
of items was determined, as in the other studies, largely by the intuitive impressions of the investigators and by practical considerations. Moreover, as in the other cases, no statistical procedures were used to ensure that the items that were retained were measures of a common construct. The interitem correlation matrix for conscientiousness shows that negative correlations as low as -.36 were included, suggesting that some of the items may be inappropriate. This suspicion is reinforced by examining the content of the items. It is difficult to comprehend, for example, why "neatness of desk" and "amount of time studying" should be considered measures of a common construct of conscientiousness. One might even argue that the more a student uses his desk for studying, the less likely it is to appear neat. (An informal investigation of colleagues' desks revealed that some of the most highly conscientious professors had the shaggiest desks.) Various procedures were used to obtain data. Included were (a) the direct, objective measurement of behavior in a natural environment, such as using a video camera to establish time of arrival at a large lecture class and to record the number of people with whom the student left; (b) archival data, such as records that indicated whether students had signed out books from the library by a prescribed date; (c) "laboratory observations," which consisted of observations of behavior in two interviews scheduled for the study; (d) self-reported behavior, or the responses to seven questions filled out at the end of the day on special forms; and (e) observer ratings, which were unobtrusive ratings by two observers who visited the student's dormitory room and rated items such as neatness of desk, neatness of bed, and number of visitors present. Unfortunately, the amount of aggregation over occasions varied among these, and no criterion was set for reaching a given level of reliability. Moreover, single observations of a component behavior in one cell with other components of some cell at a different time were coded. Most observer ratings were done in pairs, or three, or four, or five.

repeatedly referred to as a behavioral study and the examples cited are of objective measures, much of the data is based on self-ratings. For example, the daily report form included items such as: "How many hours did you spend studying today?" "What percentage of all of the work scheduled for your classes tomorrow did you complete?" "Please list the classes that you attended today and the approximate number of times you talked in each." "Approximately how much time today did you spend socializing with friends?" It is with no little irony that one notes that Mischel and Peake (1982), who, in the very same article in which they dismiss self-report data by others, including frequency counts of days on which headaches were experienced, as subjective, rely on even more ambiguous self-report data, which they cloak under the rubric of "behavioral measurement." In any event, it is evident that Mischel and Peake did not follow ideal procedures for item construction and selection. They retained items that should have been excluded, and their data are not as objective as implied by their summary descriptions. Finally, it should be noted, that although the relationships are referred to as cross-situational, most of them are not cross-situational, but cross-modal, or both. Neatness of desk and neatness of bed, for example, do not refer to the same behavior in different situations, but to different behaviors in the same setting.

Mean interitem correlations for conscientiousness and friendliness were .13 and .08, respectively. Mischel and Peake noted that, even when disattenuated for unreliability, the average interitem correlations for conscientiousness were .14 and .15.

The same principles of aggregation apply, of course, to cross-modal as well as to cross-situational relationships. The concept of a trait implies both cross-modal and cross-situational generality. We nevertheless emphasize a distinction between the two concepts for two reasons. One is that Mischel has emphasized the distinction and has discounted findings by others on cross-modal generality as irrelevant to the issue of cross-situational generality. The other is that, given the occurrence of response-set effects and limited behavioral repertoires in people, as well as evidence indicating greater stability of responses than of situations in the behavior and experience of individuals (Epstein, 1979b), there is reason to believe that cross-situational consistency tends to be greater than cross-modal consistency.
tiousness and friendliness rose to only .20 and .14, respectively. An inspection of the interitem correlation matrices reveals that the relationships vary markedly. For conscientiousness, the range is .73 to -.36, with many correlations above .40. The results are similar for friendliness. Thus, the low average interitem correlations can be partly attributed to the likelihood that conceptually inappropriate items were included. Mischel and Peake (1982), in a footnote, acknowledged that many of their interitem correlations were not strictly cross-situational and that when they confined their relationships to strictly defined cross-situational ones, the average interitem correlation for conscientiousness increased from .13 to .28. Strangely, although they disattenuated the initially low average correlation they obtained for all of the items and noted that even with disattenuation the average interitem correlation remained low, they did not disattenuate their interitem correlations for the strictly cross-situational relationships. When we disattenuated their .28 cross-situational interitem correlation for conscientiousness by using their reported average item reliability coefficient of .65, we obtained an average interitem correlation of .43. The use of an average reliability figure for disattenuating the average interitem correlation provides, at best, a rough approximation to the results that would be obtained if a more accurate procedure had been followed in which each of the relationships is individually corrected. Accordingly, when the data from the Peake (1982) dissertation became available, we corrected all cross-situational interitem correlations for the unreliability of the individual items. An independent classification of the items as cross-situational produced the identical average uncorrected interitem correlation of .28 reported by Mischel and Peake for the conscientiousness domain and a figure of .38 for the friendliness domain. The respective disattenuated correlations were .53 and .64. It may be concluded that the data from the Peake dissertation provide highly impressive evidence of substantial cross-situational relationships when the data are disattenuated for unreliability and confined to a strictly defined interpretation of cross-situational consistency, a conclusion opposite to that arrived at by Mischel and Peake (1982). In common with the authors of the other studies reviewed, Mischel and Peake overemphasized the importance of relationships between individual items and underestimated the importance of relationships between aggregates of items when assessing the presence of broad response dispositions. Even if one accepts their .13 figure, uncorrected for unreliability, an internal consistency reliability coefficient of .74 is obtained for conscientiousness. Although they acknowledged that this figure has implications for assessment, they deemphasized its importance with respect to generality in behavior, noting that aggregating across situations treats differences among individual items as error, “thereby bypassing the problem of cross-situational consistency rather than solving it” (1982, p. 738). They further noted that whether one chooses to aggregate across items or not should depend on one’s goals and that their goal was to investigate discriminability between individual items. Obviously, if one wishes to examine relationships among individual items, one cannot aggregate across the individual items. However, as was previously noted, the issue that provoked the crisis in personality research was not whether there is discriminability in behavior—no one ever doubted that people behave differently in different situations—but whether there is a reasonable degree of generality in behavior, enough for a concept of broad response dispositions, or traits, to be tenable and useful. The problem of establishing whether there is generality in behavior is not bypassed, but resolved, by appropriate aggregation over items.

Although Mischel and Peake (1982) recognized that aggregation over occasions is important for establishing temporal reliability in behavior, they did not recognize that aggre-

---

*6 The question may be raised as to whether disattenuation is appropriate under the circumstances. It should be considered that Peake (1982) aggregated behavior in different situations over different numbers of occasions. The fewer the occasions aggregated, the lower the reliability. Thus, many cross-situational correlations were low simply because of a high degree of error of measurement. Had all items been sampled over sufficient occasions, reliability would have been uniformly high, and the same high cross-situational correlations that were obtained following disattenuation would have been obtained without disattenuation.
gation over items is equally important for establishing generality in behavior. As was noted previously, because the concept of traits as stable, broad response dispositions requires demonstrations of both stability and generality, aggregation over both occasions and items is necessary. As Epstein has repeatedly noted (e.g., 1977, 1979, 1983a, 1983c), the emphasis on average correlations among single items of behavior, whether reliably assessed or not, in evaluating broad response dispositions has been a consistent source of misunderstanding throughout the long history of the person-situation debate. That Mischel and Peake continue in this tradition is well illustrated by their conclusion, based on their observation of low interitem correlations, “that broad cross-situational consistencies remain elusive even with reliable measures” (p. 735). As demonstrated in all of the studies reviewed, including Mischel and Peake (1982), when data are appropriately aggregated over both items and occasions, broad cross-situational consistencies are not only not elusive, but are robust to the point that they were uniformly detectable, even with procedures that were far from optimal.

Conclusions

The Significance of Traits

As broad dispositions, traits cannot be expected to be very good predictors of individual acts. However, this is not sufficient reason to deny their utility. As we have repeatedly noted, there are two reasons why traits are of limited value in predicting single behavioral acts. Single behavioral acts tend to be (a) low in reliability and (b) low in generality. Given the low reliability of single acts, nothing can be expected to predict them well. After all, if they cannot even predict themselves well, how can anything else predict them well? Given reliable behavioral acts, such as can be obtained by aggregating the same act over occasions, the relationship of an item to a trait will be influenced by how well the item represents the trait in question. Peripheral items are less representative of a trait than central, or prototypical, items. Now consider that in the studies reviewed, cross-item relationships between reliable items of behavior were frequently about .40. It follows that the relationship between a trait measure and a reliable prototypical item should be somewhat higher. In other words, it can be inferred from the evidence presented here that traits have received an undeservedly poor reputation as predictors because they were expected to do the impossible. That is, their value as predictors of specific kinds of behavior has been seriously underestimated because of a failure to take into account the limitations in reliability and trait relevance of the single kinds of behavior they were to predict.

The major value of traits lies, however, not in their usefulness in predicting specific behaviors, but in their value as predictors of aggregated behavior, that is, of behavior in the long haul averaged over many situations, occasions, and responses. Relatedly, although traits are often not discernible in individual acts, this does not make them unimportant with respect to individual lives. Just as one item by itself cannot provide a strong measure of a trait, but the accumulation of many can, one unexceptional act by itself is not apt to significantly affect the life of a person or of his or her acquaintances, but the accumulated consequences of many can.

Some would acknowledge that traits have practical utility as predictors, but would insist that they are nevertheless theoretically sterile constructs that can contribute little to an understanding of the personality of individuals. This is true, in a sense, for traits are simply summary behavioral descriptions. Given nothing more than a person’s standing on a list of traits, one learns nothing about the organization of the person’s personality or about the reasons or motives for the person’s behavior. Traits do not explain behavior, but themselves require explaining. They are, nevertheless, an important first step in understanding an individual’s personality, precisely because they identify important regularities in behavior that have to be explained (see Buss & Craik, 1983, for a similar view). Orderliness in behavior, of course, cannot come out of nowhere. Given that traits are neither situation-, nor response-specific, they cannot be attributed to habits or be explained by mechanistic principles. One of their major sources can be assumed to lie in the constructs of individuals. To understand an individual, according to Epstein (1973, 1980a), requires reconstructing the organization of the implicit theories and behavioral patterns.

What Has Happened

The major issue in situationalism has been a failure to see the importance of traits and a simultaneous failure to acknowledge the importance of the individual at the aggregate level. Meaningful categories and their relative stability of behavior are aggregate constructs, which do not exist in the mind of the individual. The examination of different aspects of the concept of intelligence is an aggregate endeavor, dependent upon the combined efforts of many individuals. Were there no aggregate level intelligence, then it is not clear that the individual could exist, for the individual is a single test score. Were there no aggregate variability in the concept of intelligence, then it is not clear that the individual could exist. Were the mind of the individual an aggregate construct, then the individual could not exist, for the individual is a single test score.

Once we recognize that traits are an aspect of the human condition, we must not necessarily conclude that traits are the only source of behavior. For example, the concept of intelligence is a useful one, but it is not the only source of behavior. Similarly, the concept of traits is a useful one, but it is not the only source of behavior. The concept of traits is a useful one, but it is not the only source of behavior. The concept of traits is a useful one, but it is not the only source of behavior. The concept of traits is a useful one, but it is not the only source of behavior. The concept of traits is a useful one, but it is not the only source of behavior. The concept of traits is a useful one, but it is not the only source of behavior. The concept of traits is a useful one, but it is not the only source of behavior. The concept of traits is a useful one, but it is not the only source of behavior.
PERSON-SITUATION DEBATE

What Has and Has Not Been Resolved

The most fundamental issue in the person-situation debate was whether behavior is situationally specific or cross-situationally general, and, accordingly, whether the concept of traits as stable, broad response dispositions is tenable. The answer is that behavior is simultaneously situationally specific and unstable at the individual-item level and general and stable at the aggregate level. Accordingly, it is not meaningful to question whether behavior is categorically specific or general. The meaningful task that remains is to establish the relative stability and generality of different kinds of behavior. Moreover, it should be recognized that the same behavior can often profitably be examined over different levels of generality for different purposes. For example, the Wechsler Intelligence Test contains 10 subtests, which are aggregated into verbal and performance tests, which, in turn, are aggregated into a single test of general intelligence. All three levels of generality provide useful information, and it is meaningless to argue whether intelligence is specific or general, as it is obviously both. Relatively, the tenability of traits as broad, stable dispositions is no longer a debatable issue, for despite unequivocal evidence for specificity, there is also unequivocal evidence for the existence of broad, stable dispositions.

Once it is recognized that behavior can be specific and unstable at the item level and general and stable at the aggregate level, many of the more specific issues that arose during the course of the person-situation debate can be resolved. Thus, the presumed .30 barrier for correlations between self-report and objective measures can be attributed to the widespread practice of examining single items of behavior. It is now apparent that the .30 barrier can routinely be breached when the objective data are appropriately aggregated over situations and occasions.

As for the low magnitude of cross-situational correlations, which Mischel stated were typically in the vicinity of .20, the studies reviewed here demonstrate that the relationships were attenuated by unreliable measures. When reliable measures were used, or relationships disattenuated for unreliability, cross-situational relationships rose, on the average, to about .40, and many were above .60. A moment’s reflection, however, reveals that depending on the similarity and conceptual relatedness of the situations, and the nature of the responses investigated, cross-situational relationships can assume any magnitude. In some recent studies of real-life events (e.g., Diener & Larsen, 1984; Epstein, 1984a, 1984b), a wide range of cross-situational relationships was found, including many much higher than .40. Not surprisingly, it was observed that some behaviors were highly consistent across situations, whereas others were more situationally specific. Mischel and Peake (1982) have attributed considerable importance to their belief that cross-temporal relationships are much greater than cross-situational relationships. They even suggested that the reason people believe in cross-situational consistency is because they generalize from cross-temporal stability. However, a meaningful answer to the question of which kind of relationship is generally greater is impossible, as, in the absence of a metric for equating dimensions of time to dimensions of stimulus similarity, any result can be obtained depending on the particular time intervals, response modes, and homogeneity of the situations investigated.

We next consider the presumed paradox that people believe in widespread cross-situational generality of behavior, despite repeated demonstrations to the contrary. Although this para-

In a recent article, Conley (1984) compared the magnitude of cross-situational and cross-temporal correlations in the same four classic studies reported here and in Mischel and Peake’s (1982) study. He concluded that when the two kinds of relationship are examined in completely parallel form, so that cross-situational relationships are not confounded with cross-temporal relationships, they are of similar magnitude. Peake and Mischel (1984) countered by noting that, in their own data, although internal-consistency reliability coefficients were similar for the two kinds of relationship, this was only because the cross-temporal relationships were aggregated over more items, as the cross-situational relationships had lower average interitem correlations than the cross-temporal relationships. They did not dispute Conley’s findings with respect to the other studies. Our position on this issue, as stated above, is that, depending on the items and time intervals selected, either kind of relationship can be greater than the other.
adox has been widely endorsed (e.g., Bem & Allen, 1974; Mischel, 1973; Nisbett, 1980), it has never been adequately documented. It is based instead on questionably relevant data and on incorrect inference. The inference is that because people believe in traits, they must necessarily believe in strong cross-situational relationships among individual items. As has been demonstrated, the view that traits require strong cross-situational relationships at the item level was a prevalent source of confusion throughout the long history of the person–situation debate. As for the data cited in support of the view that people grossly overestimate the degree of cross-situational generality in behavior, it consists mainly of laboratory studies that investigated relationships that, at best, were only remotely related to cross-situational generality. Examples are the relationship between the height of men and the length of their walking sticks, presented as a series of drawings (Nisbett & Ross, 1980); and the relationship between cloud-seeding and rainfall, presented as a word problem (Ward & Jenkins, 1965). For a more thorough review of this research, see Epstein and Teraspulsky (in press). In the Epstein and Teraspulsky study, people’s judgments of real-life cross-situational relationships were examined for items of behavior for which empirically established correlations were available. The items were extracted from the studies of real-life behavior by Peake (1982) and Newcomb (1929). The findings did not support the view that people indiscriminately assume broad levels of cross-situational generality in behavior. Instead, the subjects functioned as reasonable, intuitive psychometricians, taking into account principles of aggregation over occasions and items.

There is a particularly important issue that should be considered before leaving this subject, and that is whether aggregation bypasses rather than resolves the problem of cross-situational consistency in behavior, as it treats differences among situations as error of measurement rather than investigating them (Mischel, 1983; Mischel & Peake, 1983). Although it is self-evident that aggregating over situations cancels out the differences among the situations aggregated, it should be recognized that aggregation can be conducted within narrow limits no less than within broader limits. Science is not concerned with relationships between isolated fragments of behavior, but with establishing general relationships between variables, no matter how narrowly or broadly defined, which then can be put to use for different purposes, including predicting narrowly defined behaviors.

If one wishes to establish relationships between situationaly specific behaviors, one must sample appropriate exemplars, preferably measured on more than one occasion, so that temporal reliability and generality within the appropriate narrowly defined class of situations can be established and taken into account in evaluating relationships among variables. To return to our example from the Wechsler Intelligence Test, in which three useful levels of generality were noted, performance at the individual-item level was not included among the useful, interpretable levels. Performance on an individual item cannot be interpreted in isolation, because the same item can vary along many dimensions. Moreover, there is no way of assessing the degree of generality in behavior without an index of error of measurement that is derived from the variation among responses to different exemplars of the class of stimuli at issue. One needs an appropriate error term when assessing narrow behaviors no less than when assessing broad behaviors. Thus contrary to Mischel’s view that aggregation over individual items is inappropriate when one wishes to examine narrow, situation-specific behaviors, repeated measurement and appropriate aggregation are absolutely necessary if psychology is to function as a science (which is not to deny that one can aggregate too broadly). It follows that the psychologist who is interested in a specific kind of behavior indicated by performance on a single item, is obligated to determine the critical aspect of the item and then measure the corresponding concept by including more than one exemplar (see Nesselroade & Bartsch, 1977, and Wittman, in press, for further discussion of this issue).

Aggregation, of course, is just a tool that can be used to enhance the temporal reliability and generality of relationships, and it is these latter constructs that are of critical importance, not the aggregation per se. The principles of aggregation as they relate to reproducibility, generalizability, and the enhancement of relationships by compounding redundancies...
The relationship between behavior and personality is a complex one. Behaviorist theories, such as those of Brunswik and his colleagues, focus on the direct relationship between behavior and environmental stimuli. On the other hand, psychodynamic theories, such as those of Freud, emphasize the role of unconscious processes in shaping behavior.

Brunswik's concept of representativeness and his colleagues' work on subjective probability have been influential in this area. The theory suggests that individuals use probabilistic models to predict outcomes based on past experiences.

Recent research has focused on the role of personality in predicting behavior. Personality traits can influence how individuals perceive and respond to stimuli. For example, introverts may be more sensitive to social cues, while extroverts may be more focused on external rewards.

The study of personality and behavior is crucial for fields such as psychology, sociology, and psychiatry. Understanding the relationship between the two can help in the diagnosis and treatment of mental health disorders.

References


Loevinger, J. (1957). Objective tests as instruments of psy-
PERSON-SITUATION DEBATE


Received December 21, 1984
Revision received May 1, 1985