The Stability of Behavior: I. On Predicting Most of the People Much of the Time

Seymour Epstein
University of Massachusetts—Amherst

One of the classic debates in psychology concerns the stability of personality. With rare exception, studies that have correlated objective behavior on two occasions have obtained coefficients below .30. Not only has the direct measurement of objective behavior failed to provide evidence of stability, but self-report scales in attitude and personality inventories, as well as ratings of behavioral samples by judges (although themselves stable), have produced low correlations with objective behavior. Does this indicate, as some have suggested, that stability of behavior lies primarily in the eye of the beholder? The issue can be resolved by recognizing that most single items of behavior have a high component of error of measurement and a narrow range of generality. In four separate studies it was demonstrated that when measures of behavior are averaged over an increasing number of events, stability coefficients increase to high levels for all kinds of data, including objective behavior, self-ratings, and ratings by others, and that objective behavior can then be reliably related to self-report measures, including standard personality inventories. The observation that it is normally not possible to predict single instances of behavior, but that it is possible to predict behavior averaged over a sample of situations and/or occasions, has important implications not only for the study of personality but for psychological research in general.

A critical issue in personality theory is whether stable behavioral dispositions, or traits, exist. On the basis of everyday observation, it seems evident to most people that they do. Yet the vast bulk of psychological research fails to provide confirmatory evidence. It must be concluded that either the lay view is right and our typical methods of research are lacking, or the research findings are correct and the lay view itself is a phenomenon worthy of study. Not surprisingly, psychologists of both persuasions can be found. The debate on the stability of personality, which is one of the classic debates in psychology, has recently been given new impetus by findings from statistical procedures that have suggested a resolution in the form of what has been called "modern interactionism." It will be demonstrated in this article that modern interactionism does not resolve the issue of stability of personality, no matter what other virtues it may have. A solution proposed in this article lies in an entirely different direction, one that is so obvious that, once pointed out, it reminds one of the fairy tale of The Emperor's New Clothing.
Like the fairy tale, the implications of the solution go far beyond the immediate issue at hand.

This is not the place for a lengthy review of the debate. Excellent reviews and discussions are available in papers by Bowers (1973), Ekhammer (1974), Endler and Magnusson (1976), and Magnusson and Endler (1977). Allport (1937, 1961, 1966) can be referred to for a staunch theoretical defense of the trait position and Cattell and Eysenck (1965, 1970) for sophisticated, empirically based trait theories. The flavor of the attacks on the trait position can be represented by the following quotations spanning the period of the debate.

"Training the mind means the development of thousands of particular, independent capacities, the formation of countless particular habits. The working of any mental function or activity will improve others in so far as they possess elements common to it. The amount of transfer will be determined, among other things, by these common elements."

(Thurstone, 1960, p. 248)

Over and over, a battery of tests designed to measure traits such as persistence, aggressiveness, or honesty yields results so unreliable and unpredictable (when compared with other criteria) that one is led to question the actual existence of the general trait. In numerous cases the instruments are very loosely constructed and are clearly equivocal.

A survey of...a wide variety of frequently used tests suggests that there is a fundamental limitation common to most of them. Trait tests appear to assume that whatever they name has objective reality, many need not be so much as to improve their measures but to improve or change their thought regarding traits."

(Lehman & Witty, 1934, p. 49)

The generality of these (personality) measures, and method and situation was still not high enough to perceptually differentiate the traits of personality. The findings required abandonment of a line of research to which I had devoted seven years of my research."

(Endler, 1964, p. 23)

With the possible exception of intelligence, highly generalized behavioral consistencies have not been demonstrated, and the concept of personality traits as broad predispositions is thus untenable."

(Peterson, 1966, p. 248)

At least earlier, there was nothing silly about the initial assumption of psychologists that everything was joined together until proved otherwise. But since it has now proved otherwise, it seems only fair to give a sporting chance to the counter-assertion that nothing is joined together in this way.

(Endler, 1972, p. 25)

The charge that personality traits do not exist clearly strikes at the very heart of personality theory. One could well argue that if individuals do not have relatively stable behavioral dispositions that differentiate them from other individuals, then the concept of personality itself can be dispensed with."

(Sweder, 1975, in fact, comes close to arguing just that.)

The degree of relevance of the concept "personality" constructed in this "individual differences" framework is not very high."

(Mischel, 1968, p. 30)

It is not our intention to review the position of each here. For present purposes, it will suffice to list the arguments and the kind of evidence cited for and against traits.

"Situationist Position"

According to the situationist position, there is little stability in personality, as behavior is determined almost exclusively by situational variables. The belief is that there is little stability in personality rests on three major sources of evidence. The most important is that when behavior in one situation is correlated with behavior in another situation, the correlations are so low—usually less than .30—that they have been disparagingly referred to by Mischel (1968) as "personality coefficients.""

(Mischel, 1969, further notes, "A correlation of .30 leaves us understanding less than 10% of the relevant variance, and even correlations of that magnitude are not very common and have come to be considered good in research on the consistency of any noncongenital dimension of personality." (p. 101). At the same time there is no dearth of evidence that behavior varies markedly as a function of stimulus or situational variables.

A second source of evidence consists of findings from the appraisal of variances in analysis of variance designs. This procedure was independently introduced into the assessment of stability in personality by Ransch and his colleagues (Rausch, 1963; Rausch, Dittman, & Taylor, 1959; Rausch, Farbman, & Llewellyn, 1960) and by Endler, Hunt, and their colleagues (Endler, 1966; Endler & Hunt, 1968, 1969; Endler, Hunt, & Rozenstein, 1962). They and others (e.g., Argyle & Little, 1972; Ekhammer & Magnusson, 1973; Magnusson, 1971; Moos, 1968, 1969, 1970; Magnusson, Note 1) have demonstrated that the variance attributable to individual differences is usually much smaller than the variance attributable to situations and to the interaction of individuals and situations.

A third source of evidence cited by situationists is that when people rate others, they tend to attribute more stability to individuals across situations than is objectively warranted. (e.g., Bem & Allen, 1974; Jones & Nisbett, 1971; Mischel, 1968; Jones, 1975.) This, of course, can explain how there can be a widespread belief in the stability of personality when, in fact, there is little stability. The study of such bias in person perception has itself become a significant area of research, and has produced a quantity of explanations as to why the phenomenon of falsely perceiving stability in personality occurs. For example, it has been suggested that (a) it is emotionally satisfying to believe that behavior is predictable, particularly when it is someone else’s; (b) it is simpler to classify behavior by people than by situations; (c) people have a need for consistency—i.e., personality theories that assume stability in personality; (d) the behaviors that are taking place; (e) the observer is always present in the situations that he or she observes in real life, thereby presenting the observer with a biased sample; (f) there is a tendency for observers to attribute behavior that elicits the same emotional reactions to them; (g) there is a tendency for judges to generalize from a few attributes that are stable, such as intelligence, to others that are not; and (h) there are more terms for classifying people than for classifying situations, which leads to a bias toward attributing
behavior more often to characteristics of people than to characteristics of situations.

**Trait Position**

The arguments that have been advanced in favor of traits amount primarily to conjectures that if different procedures had been followed in the investigation of stability in personality, higher stability coefficients would have been found. There are also a few studies that can be cited that obtained respectable stability coefficients.

The arguments can be summarized as follows: (a) Many of the studies undertaken to assess stability in personality have been experimental studies, which are better suited to demonstrate change than stability in personality (Bowers, 1973). (b) The unit of analysis is a critical factor that has not been adequately taken into account. What appears to be instability at a phenotypic level of analysis may be stability at a genotypic level (Alker, 1972; Bowers, 1973). To use one of Bower's examples, a woman who is continuously changing her wardrobe may be consistently fashionable. (c) The use of moderator variables would considerably increase reliability coefficients (Alker, 1972). (d) A failure to recognize that some people are more variable than others results in reporting generally low stability coefficients, rather than in noting that at least some individuals are highly stable (Alker, 1972; Ben & Allen, 1974). There is at least one study (Ben & Allen, 1974) that has demonstrated this effect. (e) In everyday life, people determine their own environments, which, in turn, help maintain the stability of their personality. The operation of this effect is not possible in laboratory studies in which individuals are arbitrarily assigned to conditions (Bowers, 1973; Wachtel, 1973). (f) Stability in personality is moderated by an individual's cognitions and, accordingly, will only be found when ideographic procedures are used that take into account the subjective nature of perception (Alker, 1972; Ben & Allen, 1974; Bowers, 1973; Maschke, 1973). (g) Stability may be demonstrable to a greater extent within-subject relations than in between-subject relationships (Alker, 1972).

Culver (1971) makes a related point in noting that although the concept of personality implies that, for example, an organization of variables within an individual, among a large number of studies she reviewed, "not a single published study attempted even minimal inquiry into the organization of personality variables within the individual" (p. 209).

As to studies that have reported evidence for stability in personality, we will not consider investigations relying on self-report inventories, as such studies can simply demonstrate that people's beliefs about their behavior are consistent, which is a far cry from demonstrating that the behavior itself is consistent. Nor will an attempt be made to review all behavioral studies that have reported positive findings. Our purpose will be served adequately by a review of three series of carefully conducted studies. On the basis of the findings in these studies, it is possible to formulate an integrative hypothesis that can account for all the known results to date. This will be followed by a presentation of four studies that were explicitly undertaken to test the hypothesis.

Block (1971, 1977) conducted a series of studies in which the stability of personality variables was examined during different periods between childhood and adulthood. In one set of studies, records were obtained from the archives of the well-known longitudinal studies at Berkeley of children who were in junior high school. The subjects were then intensively interviewed when they were in their mid-thirties. Judges rated each subject for the different periods on the California Q Set (Block, 1973). To ensure that stability coefficients would not be artificially inflated by response sets, memory, or other sources of rater bias, different judges rated the subjects at the different periods. For the period from junior to senior high school, 58% of 114 personality variables produced stability coefficients that were as great as .35. In a study of 114 personality variables, some of the correlations were as high as .70. For the period from senior high school to age 30 and above, 29% of 114 personality variables yielded stability coefficients of .35 or greater, with some as high as .61.

In a study of younger children by Block and Block (1977), observations were made during the children's 4th, 6th, and 8th years of life. Two or three judges, who were the children's nursery school teachers, observed the children for 3 hours a day over a period of 5-9 months and then rated each child on a modification of the California Q Set. Rater bias was controlled by having the children rate the children at different periods. The average stability coefficient for 100 Q items was .48, with several over .60, and a few as high as .70. When the Q items were grouped into broader scales by factor analysis, the mean stability coefficients rose to .56.

In a series of studies on aggression in young boys, Olweus (1974, 1975, 1976) has found that some boys in the 6th grade, and again 3 years later, on a number of variables pertaining to aggression, such as tendency to start fights with peers, ratings for each boy were averaged over 3-10 raters. Memory effects and response bias were evaluated, controlled by statistical procedures, and by examining subgroups in which there was no overlap in raters. The mean stability coefficient for the 3-year period was .66. When correlations were corrected for attenuation due to rater unreliability, the stability coefficients rose to about .80. In a second similar study in which 85 13-year-old boys were the subjects, even higher stability coefficients were obtained. In other studies, it has been observed that peer ratings, teacher ratings, and self-ratings on a specially designed aggression inventory were all highly intercorrelated, suggesting the existence of a broad dispositional trait of aggression.

The classic series of studies by Harshorne and May (1928, 1929) and Harshorne, May, and Shuttleworth (1930) on honesty, often cited as evidence against the existence of stability in personality, was an enormous project that spanned half a dozen years, employed a large team of researchers, and tested a national sample of over 8,000 children. Among the behavioral items that were assessed were cheating in a classroom, cheating during a game, stealing money, lying, and falsifying records of athletic performance. The average intercorrelation of 23 subscales used as part of a total character score was found to be .33, which led Harshorne and May to conclude that honesty in any given situation has low predictive value for honesty in any other single situation. This conclusion is invariably cited as evidence against the existence of stability in personality. What is generally not known is that Harshorne and May combined several tests of honesty into a single score, the reliability coefficient increased to .73, and they concluded, 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 (Harshorne & May, 1928, p. 133).

Further support for the existence of a broad trait of honesty in the Harshorne and May data is provided by a factor analysis by Burton (1963), who observed that a general factor of honesty accounted for nearly 50% of the total variance.

**Interactionist Position**

According to the interactionist position, the question of which is more important, the situation or the person, is a meaningless one, as behavior is always a joint function of the person and the situation. In its applicability to the issue of stable individual differences, the interactionist position can be viewed as a compromise between the trait position and the situational position, for it acknowledges the existence of behavioral stability, but only within situational constraints.

The so-called weak interactionist position was derived from findings on apportioning variances in an analysis of variance design. As already noted, it was observed by a number of psychologists that the interaction of individuals and situations accounted for more variance than either source of variance by itself. It was consequently argued that an interactionist position should supplement both the trait and the situational position (Hewers,
1973; Ekhammer, 1974; Endler, 1966; Endler & Hunt, 1968; Magnusson, Note 1).

As for a more general interactionist viewpoint that goes beyond statistical interaction (see Magnusson & Endler, 1977, for a thorough review of this position), it is assumed that individuals and situations are interdependent. That is, the individual's cognitions and perceptual processes, as much as the objective characteristics of the stimulus, determine the meaning of the stimulus. From this viewpoint, behavior can best be viewed as a transaction between the individual and the stimulus, each influencing the other. Since behavior never occurs in a vacuum but always occurs in a situational context, it is meaningless to talk about characteristics of an individual's behavior without specifying the situation in which the behavior occurs. To understand and predict behavior, it is, accordingly, just as necessary to have a classification system for situations as for individuals and, most important, to know how individuals interpret different kinds of situations.

Evaluation of the Arguments For and Against the Existence of Traits

Evaluation of the Situational Position

The case of the situationalists against traits rests mainly on empirical grounds. They note that the concept of traits is not unreasonable but add that it is also not supported by the facts. The strongest evidence they cite against traits consists of the low stability coefficients obtained when data are derived from direct behavioral measurement. However, it is important to recognize that in almost all cases the correlations were based on single items of behavior. Such correlations actually have little relevance for the existence of traits, as no trait theorist believes that a trait can be inferred from a single instance of behavior. A trait is a generalized tendency for a person to behave in a certain manner over a sufficient sample of events and does not imply that he or she will exhibit trait-relevant behavior in all situations or even on all occasions in the same situation. From another standpoint, single items of behavior have a high component of error of measurement, thereby limiting the possibility of replication, and a high component of situational uniqueness, thereby limiting the possibility of generalization. It is possible with such a procedure for individuals to be perfectly reliable over time, as indicated by a reliability coefficient of .100, and yet have individual differences contribute a relatively small proportion of total variance. For example, consider a situation in which all individuals come out exactly the same in a 50-yard dash and in a 500-yard dash. As the mean difference between events will be many times greater than the differences among racers within events, it is evident that the variance due to individuals will be a small proportion of the variance due to situations, let alone the total variance.

The argument that judges sometimes attribute stability to people that is not there describes an interesting phenomenon, but cannot establish that there is no stability in behavior apart from such bias. In fact, the assumption that under most circumstances there is stability in personality can account for a bias to assume there is stability even when there is not. Further, the same factors that contribute to the perception of personality stability also contribute to its actual stability. If it is assumed that there is a fundamental need for people to establish orderliness and predictability in the world of their experience and that they accomplish this through their habits of perception (see Epstein, 1973; Note 2; Mischel, 1973), then the same cognitions and perceptions could contribute not only to people's ratings of stability but to the stability of their actual behavior as well.

Finally, it should be recognized that the null hypothesis cannot be proven by the failure of many studies to demonstrate stability in personality. The possibility remains that with new understanding and new approaches, the conditions for demonstrating stability in objectively measured behavior will be established.

Evaluation of the Trait Position

The arguments in defense of traits are, for the most part, speculations that if things had been done differently, stability in personality might have been demonstrated. Although the proposals may be of interest in suggesting new directions for personality research, they do not constitute an adequate defense of the position that there is stability in personality. All these proposals can do is indicate that the issue should not be closed until they have been tried. Unfortunately, though some of these proposals have been tested, they have failed to fulfill their promise (Bem, 1972).

In our own research, in which we compared mean within-subject correlations to between-subjects correlations for the same variables, we have generally found the former to yield smaller values.

Finally, one must agree with Bem (1972) that the issue of stability in personality cannot be resolved by dispute but only by data. "And if these separate indices permit Atler to predict behavior across situations better than .30, Mischel will fold up his tent and steal away" (p. 18).

How is one to evaluate the findings in a few studies that have demonstrated stability in behavior in comparison to the vast number of studies that have obtained no positive findings? One possibility is that when enough studies are done, a few are bound to produce significant results by chance alone. An alternative is that the few studies that produced positive findings were better conceived and conducted than the many that failed (see Block, 1977). A careful analysis of the studies that succeeded could uncover the critical conditions for demonstrating stability in behavior. It is noteworthy, in this respect, that all three series of studies that succeeded examined relatively extensive samples of behavior. In the series of studies reported by Block (1971, 1977) and Bem (1971, 1974, 1977a, 1977b), judges rated behavior over a relatively long period. In the Harshorne and May studies (1928, 1929), reliability was found only when a sufficient number of behavioral items were combined into a single index. As previously noted, single items of behavior, like single items in a test, tend to have a high component of error of measurement. Thus, it may be that procedures for reducing error of measurement are critical for demonstrating stability in
personality. Further evidence in support of this conjecture is that in the studies by the Blocks and by Olweus, ratings from several judges were combined, which is one means of reducing error of measurement. In addition, the Blocks found that when they combined a number of single measures into a broader one, stability coefficients markedly increased, a finding also reported by Harshorne and May. Thus, a possibility that must be considered is that the critical factor separating the studies that succeeded in establishing stability from those that failed to do so is the steps taken to reduce error of measurement by obtaining adequate samples of behavior in the former studies.

Evaluation of the Interactionist Position

As already noted, there have been problems with the statistical and methodological procedures employed by the interactionists. Moreover, the claim that the interaction of individuals and situations accounts for more variance than either individuals or situations alone has not been uniformly upheld (see Santer, Smith, & Diener, 1975). In a review of a large number of studies, Sarason and his colleagues found that individuals, situations, and their interaction did not account for very much of the total variance. Rather, the predominant source of variance was error variance. Thus, which of the first three sources of variance is most important is a moot point, as often none accounts for a satisfactory amount of variance.

The more general interactionist position does not involve so much a new position as a reawakening of interest in an old position (see Ekkerham, 1974). Murray (1938), in his classic studies of personality, not only endorsed an interactionist view at a theoretical level, but developed a classification system for response tendencies and for subjectively and objectively defining stimulus situations. Even more to the point, his concept of themes identified a unit of interaction between a response disposition and a stimulus variable. There is little question but that behavior is a joint function of the person and the situation. However, this is irrelevant to the question of whether a reasonable degree of stability in individual behavior can be demonstrated when the behavior is averaged over a sufficient sample of situations. It is noteworthy, in this respect, that interactionists have been no more successful than others in breaking the 30-personality barrier. If they are to succeed in doing so, they will need a sufficient sample of people in particular situations. That is, interactionism does not replace the need to reduce error of measurement by sampling, but simply determines what it is that must be sampled. Without temporal reliability, meaningful generalization cannot be established, and this applies equally to the demonstration of stable effects for individual differences, for situations, and for their interaction.

The main focus of this article is the stability of individual differences. If our assumption about reducing error of measurement through averaging over observations is correct, it should be possible to routinely break the presumed 30-personality barrier by averaging behavior over a sufficient number of events. It should be noted that this was done implicitly in the studies by the Blocks and Olweus, as when a judge makes a single rating after observing a child on many occasions, the single rating can be viewed as an intuitive averaging.

The above analysis of what is necessary to demonstrate stability in behavior, simple as it is, is different from previous explanations that have assumed that the crucial factor is the kind of data obtained. Thus, Block (1977) noted that there is impressive evidence for the stability of personality in ratings by judges of real-life behavior (R data) and in self-reports (S data) and that there exists impressive evidence for the coherence of the two forms of data. He further observed, however, that there is no evidence for stability in personality when the data consist of the direct measurement of objective behavior (T data) and that such data are, at best, tenuously related to the other two kinds of data. Block concluded, "It is now incumbent upon us to consider . . . what strategies are likely to extend the realm of coherence so as to include as well the domain of T-data" (p. 63). Mischel (Note 3) also expressed concern over the lack of relationship between objective data and data derived from ratings by others and from self-report. He stated.

An important test, although surely not the only one—of the utility of concepts about personality dispositions remains their ability to predict the individual's behavior in specific situations. Unless R and S data predict T data appreciably, the links between trait impressions and specific behavior-in-situations remains tenuous (p. 4).

If our reasoning is correct, the following hypothesis and corollary should define the conditions for routinely establishing stability in all kinds of data, including ratings by others, self-ratings, and objective behavior, and for relating ratings by others and self-ratings to objective behavior.

Hypothesis

Stability can be demonstrated over a wide range of variables so long as the behavior in question is averaged over a sufficient number of occurrences. This applies equally to data derived from the direct measurement of objective behavior, from self-reports, and from ratings by others.

Corollary

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

Empirical Testing of the Hypothesis and Corollary

A series of studies has been completed that examined the temporal reliability and in some cases the validity of data derived from self-observation, observation by others, and the direct measurement of objective behavior. In all cases, a similar procedure was followed, that is, behavior was observed on several occasions, and single observations were treated like single items on a test. More specifically, stability coefficients were first determined for
a 1-day sample by correlating each subject's scores on Day 1 with each subject's scores on Day 2. Next, coefficients were determined for a 2-day sample by correlating the mean of a subject's scores on Days 1 and 3 with the mean of the subject's scores on Days 2 and 4, and so on, until the mean of a subject's scores on all odd days was correlated with the mean of the subject's scores on all even days. This made it possible to examine split-half stability coefficients for each variable as a function of the number of observations that were averaged. Such a procedure is analogous to that which compares the reliability coefficients of tests of different lengths.

Study 1: Stability of Self-Recorded Data

Method

Study 1 used data available from a previous investigation of emotions in everyday life. The selection of variables in the study was influenced by a theory of the self-concept that assumes that a person's significant postulates about self and world can be inferred from events that elicit emotions (Epstein, 1973, 1976, in press). For each 14 men and 14 women university students kept records, on specially devised forms, of their most pleasant and unpleasant experience each day. At the end of each week, they met with interviewers who answered their questions and checked for errors. The number of recording days completed by subjects varied from 24 to 34. The best pair of each form consisted of a blank page for describing the incident in narrative form. This was followed by a 90-item adjective checklist for recording emotions and a 66-item checklist for recording response tendencies and for noting whether they had been carried out. Emotions and impulses were scored on a 6-point scale for intensity, with a blank signaling no intensity. Behavior carried out was scored on a 2-point scale for carried out or not carried out. Items were collapsed into broader scales by a factor analysis of the data. These broader scales constituted the data that were analyzed in the present study. As an example of a scale of emotions, we consider the following six categories: adequate, competent, appreciated, respected, pleased with self, and proud. As an example of a scale of response tendencies, stimulus seeking included the following three items: to try something new and adventurous; to take risks; and to seek stimulation; and to seek thrill and excitement. Scores were obtained by averaging the responses to the items in a scale.

As an alternative, it was decided to score the narratives for the stimulus conditions, or situational factors, responsible for evoking the emotions. Despite successive modifications of Murray’s (1938) system for scoring Thematic Apperception Test protocols and intensive training of judges, it was not possible to obtain satisfactory inter-judge reliability coefficients, as information necessary to make decisions was often unavailable. Rather than drop the scoring of situational factors, it was decided to retain it, with the realization that whatever results were obtained on the reliability of situational factors would underestimate the true stability but would nevertheless be of some interest in indicating a lower limit of such stability.

Results

Between-subjects reliability coefficients: The stability of individual differences. In Table 1, stability coefficients for pleasant experiences are presented for the mean of all odd days, compared to the mean of all even days, for Day 1 compared to Day 2, and for the last day compared to the next-to-last day. Also included are the means and standard deviations for all variables for the data averaged for all days. Table 2 presents the same data for unpleasant experiences. It can be seen in both tables that for a 1-day sample, most reliability coefficients are above .30, and it is true for the last 2 days as well as for the first 2 days of the study, indicating that a bias toward greater stability did not develop over the course of the study. On the other hand, when the mean of all odd days is correlated with the mean of all even days, most coefficients are .70, and some are above .90. Emotions, which were the central focus of the study and on the basis of which events were selected by the subjects, had the highest stability coefficients, with a mean of .79 for pleasant and .68 for unpleasant experiences.

Note. Correlations of .37 and .48 are significant, respectively, at the .05 and .01 levels, N = 28.

Table 1

<table>
<thead>
<tr>
<th>Variable</th>
<th>M</th>
<th>SD</th>
<th>All odd vs. all even days</th>
<th>Day 1 vs. Day 2</th>
<th>Last vs. next-to-last day</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emotion</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Happy</td>
<td>.94</td>
<td>.52</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
</tr>
<tr>
<td>Kindly</td>
<td>.91</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
</tr>
<tr>
<td>Calm</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
</tr>
<tr>
<td>Adequate</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
</tr>
<tr>
<td>Unified</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
</tr>
<tr>
<td>Energetic</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
</tr>
<tr>
<td>M</td>
<td>62.0</td>
<td>35.0</td>
<td>35.0</td>
<td>35.0</td>
<td>35.0</td>
<td>35.0</td>
</tr>
</tbody>
</table>

Behavior

<table>
<thead>
<tr>
<th>Variable</th>
<th>M</th>
<th>SD</th>
<th>All odd vs. all even days</th>
<th>Day 1 vs. Day 2</th>
<th>Last vs. next-to-last day</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nurturance</td>
<td>.72</td>
<td>.43</td>
<td>.43</td>
<td>.43</td>
<td>.43</td>
<td>.43</td>
</tr>
<tr>
<td>Exuberance</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
</tr>
<tr>
<td>Pleasure</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
</tr>
<tr>
<td>Affiliation</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
</tr>
<tr>
<td>Problem solving</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
</tr>
<tr>
<td>Stimulus seeking</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
</tr>
<tr>
<td>Achievement</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
</tr>
<tr>
<td>M</td>
<td>52.1</td>
<td>27.3</td>
<td>27.3</td>
<td>27.3</td>
<td>27.3</td>
<td>27.3</td>
</tr>
</tbody>
</table>

emotions. It is not evident why positive emotions should produce higher reliability coefficients than negative emotions, but this finding has held up in all further studies. Perhaps subjects are more confused about their negative than their positive feelings.

As expected, the lowest reliability coefficients were obtained for situations. Yet, even here, the mean correlation was well above the .30 level, more than half the correlations were significant at the .01 level, and several were above .70. The finding of stable
Table 2
Between-Subjects Reliability Coefficients, Means, and Standard Deviations for All Unpleasant Experience Variables in Study 1

<table>
<thead>
<tr>
<th>Variable</th>
<th>M</th>
<th>SD</th>
<th>All odd vs. all even days</th>
<th>Day 1 vs. Day 2</th>
<th>Last vs. next-to-last day</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emotion</td>
<td>43.7</td>
<td>27.4</td>
<td>.91</td>
<td>.25</td>
<td>.45</td>
</tr>
<tr>
<td>Blocked</td>
<td>36.0</td>
<td>25.0</td>
<td>.80</td>
<td>.37</td>
<td>.38</td>
</tr>
<tr>
<td>Fragmented</td>
<td>60.9</td>
<td>33.5</td>
<td>.88</td>
<td>.52</td>
<td>.67</td>
</tr>
<tr>
<td>Depressed</td>
<td>24.1</td>
<td>23.5</td>
<td>.77</td>
<td>.26</td>
<td>.34</td>
</tr>
<tr>
<td>Angry</td>
<td>70.7</td>
<td>32.7</td>
<td>.72</td>
<td>.13</td>
<td>.14</td>
</tr>
<tr>
<td>Frightened</td>
<td>44.0</td>
<td>21.8</td>
<td>.72</td>
<td>.09</td>
<td>.10</td>
</tr>
<tr>
<td>Inadequate</td>
<td>25.8</td>
<td>19.8</td>
<td>.72</td>
<td>.12</td>
<td>.13</td>
</tr>
<tr>
<td>Tired</td>
<td>37.0</td>
<td>29.2</td>
<td>.79</td>
<td>.26</td>
<td>.29</td>
</tr>
<tr>
<td>Situation</td>
<td>41.3</td>
<td>26.6</td>
<td>.70</td>
<td>.11</td>
<td>.08</td>
</tr>
<tr>
<td>Loss of love</td>
<td>12.8</td>
<td>12.4</td>
<td>.54</td>
<td>.01</td>
<td>.04</td>
</tr>
<tr>
<td>Noxious stimulation</td>
<td>11.0</td>
<td>11.1</td>
<td>.53</td>
<td>.06</td>
<td>.01</td>
</tr>
<tr>
<td>Frustration</td>
<td>46.2</td>
<td>18.2</td>
<td>.50</td>
<td>.04</td>
<td>.03</td>
</tr>
<tr>
<td>Isolation</td>
<td>4.6</td>
<td>6.0</td>
<td>.44</td>
<td>.08</td>
<td>.06</td>
</tr>
<tr>
<td>Inconsideration</td>
<td>15.9</td>
<td>11.3</td>
<td>.42</td>
<td>.19</td>
<td>.08</td>
</tr>
<tr>
<td>Attack</td>
<td>13.3</td>
<td>12.6</td>
<td>.40</td>
<td>.16</td>
<td>.10</td>
</tr>
<tr>
<td>Identification</td>
<td>11.5</td>
<td>11.5</td>
<td>.29</td>
<td>.01</td>
<td>.10</td>
</tr>
<tr>
<td>Failure</td>
<td>18.4</td>
<td>12.5</td>
<td>.78</td>
<td>.04</td>
<td>.10</td>
</tr>
<tr>
<td>Immorality</td>
<td>6.1</td>
<td>6.9</td>
<td>.06</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>Accidental injury</td>
<td>4.4</td>
<td>6.2</td>
<td>.40</td>
<td>.01</td>
<td>.07</td>
</tr>
<tr>
<td>Impulse</td>
<td>14.4</td>
<td>10.9</td>
<td>.88</td>
<td>.64</td>
<td>.30</td>
</tr>
<tr>
<td>Problem solving</td>
<td>26.2</td>
<td>18.9</td>
<td>.85</td>
<td>.04</td>
<td>.03</td>
</tr>
<tr>
<td>Stimulus reduction</td>
<td>33.5</td>
<td>16.6</td>
<td>.82</td>
<td>.08</td>
<td>.08</td>
</tr>
<tr>
<td>Affliation</td>
<td>31.4</td>
<td>22.4</td>
<td>.82</td>
<td>.08</td>
<td>.08</td>
</tr>
<tr>
<td>Counteraction</td>
<td>39.0</td>
<td>21.0</td>
<td>.75</td>
<td>.06</td>
<td>.04</td>
</tr>
<tr>
<td>Physical escape</td>
<td>29.9</td>
<td>25.0</td>
<td>.68</td>
<td>.06</td>
<td>.12</td>
</tr>
<tr>
<td>Mental escape</td>
<td>49.1</td>
<td>24.0</td>
<td>.68</td>
<td>.06</td>
<td>.29</td>
</tr>
<tr>
<td>Self-punishment</td>
<td>18.3</td>
<td>16.1</td>
<td>.66</td>
<td>.18</td>
<td>.12</td>
</tr>
<tr>
<td>Tension discharge</td>
<td>18.1</td>
<td>17.6</td>
<td>.64</td>
<td>.05</td>
<td>.59</td>
</tr>
<tr>
<td>Achievement</td>
<td>13.5</td>
<td>10.9</td>
<td>.64</td>
<td>.43</td>
<td>.12</td>
</tr>
<tr>
<td>Aggression</td>
<td>33.4</td>
<td>29.2</td>
<td>.51</td>
<td>.24</td>
<td>.08</td>
</tr>
<tr>
<td>Nurturance</td>
<td>19.4</td>
<td>22.8</td>
<td>.51</td>
<td>.13</td>
<td>.18</td>
</tr>
<tr>
<td>Withdrawal</td>
<td>32.7</td>
<td>19.4</td>
<td>.70</td>
<td>.16</td>
<td>.18</td>
</tr>
<tr>
<td>Behavior</td>
<td>28.6</td>
<td>20.4</td>
<td>.87</td>
<td>.04</td>
<td>.45</td>
</tr>
<tr>
<td>Affiliation</td>
<td>6.5</td>
<td>6.7</td>
<td>.87</td>
<td>.33</td>
<td>.67</td>
</tr>
<tr>
<td>Problem solving</td>
<td>8.2</td>
<td>7.5</td>
<td>.82</td>
<td>.13</td>
<td>.17</td>
</tr>
<tr>
<td>Stimulus reduction</td>
<td>8.7</td>
<td>5.7</td>
<td>.79</td>
<td>.06</td>
<td>.28</td>
</tr>
<tr>
<td>Counteraction</td>
<td>8.5</td>
<td>7.5</td>
<td>.73</td>
<td>.16</td>
<td>.57</td>
</tr>
<tr>
<td>Mental escape</td>
<td>12.7</td>
<td>9.2</td>
<td>.65</td>
<td>.08</td>
<td>.03</td>
</tr>
<tr>
<td>Achievement</td>
<td>3.1</td>
<td>3.7</td>
<td>.60</td>
<td>.06</td>
<td>.04</td>
</tr>
<tr>
<td>Nurturance</td>
<td>3.6</td>
<td>5.9</td>
<td>.55</td>
<td>.06</td>
<td>.04</td>
</tr>
<tr>
<td>Withdrawal</td>
<td>5.3</td>
<td>5.3</td>
<td>.43</td>
<td>.56</td>
<td>.00</td>
</tr>
<tr>
<td>Aggression</td>
<td>1.4</td>
<td>2.4</td>
<td>.35</td>
<td>.04</td>
<td>.34</td>
</tr>
<tr>
<td>Self-punishment</td>
<td>1.8</td>
<td>2.3</td>
<td>.27</td>
<td>.81</td>
<td>.04</td>
</tr>
<tr>
<td>Tension discharge</td>
<td>3.1</td>
<td>3.7</td>
<td>.27</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>Physical escape</td>
<td>1.2</td>
<td>1.9</td>
<td>.57</td>
<td>.15</td>
<td>.25</td>
</tr>
</tbody>
</table>

Note. Correlations of .37 and .48 are significant, respectively, at the .05 and .01 levels. N = 28.

Figure 1. Between-subjects reliability coefficients in Study 1 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 "Traits are Alive and Well" by S. Epstein. In D. Magnuson & N. S. Endler (Eds.), Personality as a construct: Current issues in international psychology. Copyright 1971 by Erlbaum, Hillsdale, N.J. Reprinted by permission.]

Stability of Behavior: I
Within-subject reliability coefficients: The stability of the organization of variables within an individual. No matter what else personality is, it is widely recognized that it involves an organization of variables within an individual. Yet, despite universal agreement on this point, Carlson (1974) found not a single study that addressed itself to the organization of variables within the individual. The first step in such an inquiry might be to establish whether such organization can be demonstrated to have a reasonable degree of stability. If what has been observed about error of measurement among individuals applies to the organization of variables within individuals, then the stability of such organization should be demonstrable when data are averaged over sufficient observations but not when they consist of single observations.

For each of the 28 subjects, correlations were obtained across variables within a category. A high correlation for a subject indicated that his or her profile was stable over days. Pleasant and unpleasant experiences were treated separately, and the data across variables within a category were averaged for different numbers of days. It can be seen in Figure 3 that when profiles were derived from single observations, the average intrasubject stability coefficient was less than .25 over all categories. As the number of days from which individual profiles were derived was increased the stability coefficients also increased. The average correlation for profiles based on the maximum number of odd days versus the maximum number of even days varied between .60 and .76, depending on the category. Relatively high stability coefficients were even obtained for profiles of situations. Table 3 presents the ranges and means of the correlations of the profiles within individuals for the different categories, with the data based on the maximum number of days in the odd–even samples. It is apparent that there are marked individual differences in the stability of profiles, with some subjects exhibiting almost perfect stability, and other subjects exhibiting a lack of stability. It may be concluded that within-subject reliability coefficients provide evidence for a relatively high degree of stability of the organization of variables within individuals when the data are derived from sufficient observations but provide no such evidence when the data are derived from single observations. It is further concluded that there are marked individual differences in the degree to which individuals exhibit stability in their personality profiles.

**Study 2: Stability of Behavior Observed by Others**

**Method**

The data for this study were obtained from an investigation of 35 pairs of women students at Smith College served as subjects and observers in the study. Each pair was selected for study, and the observers had to have ample opportunity to observe the subject during and outside of class. The observer kept records of the subject's behavior on eight items related to impulsivity and sociability. Ratings were made on a 3-point scale for frequency of occurrence. The following is an example of one of the items: "She actively sought out the company of others: (a) never, (b) once or twice, or (c) three or more times."

**Table 3**

<table>
<thead>
<tr>
<th>Category</th>
<th>Least reliable subject within category</th>
<th>Most reliable subject within category</th>
<th>Mean reliability for all subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unpleasant experience</td>
<td>Subject no.</td>
<td>r</td>
<td>Subject no.</td>
</tr>
<tr>
<td>Emotions</td>
<td>8</td>
<td>28</td>
<td>.24</td>
</tr>
<tr>
<td>Situations</td>
<td>10</td>
<td>10</td>
<td>-.03</td>
</tr>
<tr>
<td>Impulses</td>
<td>12</td>
<td>14</td>
<td>.18</td>
</tr>
<tr>
<td>Behavior</td>
<td>12</td>
<td>2</td>
<td>-.17</td>
</tr>
<tr>
<td>Pleasant experience</td>
<td>Subject no.</td>
<td>r</td>
<td>Subject no.</td>
</tr>
<tr>
<td>Emotions</td>
<td>6</td>
<td>12</td>
<td>.13</td>
</tr>
<tr>
<td>Situations</td>
<td>9</td>
<td>2</td>
<td>-.04</td>
</tr>
<tr>
<td>Impulses</td>
<td>7</td>
<td>2</td>
<td>-.15</td>
</tr>
<tr>
<td>Behavior</td>
<td>7</td>
<td>10</td>
<td>.24</td>
</tr>
</tbody>
</table>
the 1st and 2nd day were compared with correlations between the next-to-last and last day. The latter correlations were slightly lower, indicating that the establishment of a rating set during the study could not have contributed to the increase in correlations that occurred with increasing observations.

A further possibility for contamination of the stability coefficients lies in the inferential processes of the observers. That is, a judge's view on stability could have influenced her ratings independent of the behavior of her subject. To the extent that this occurred, it would most clearly be revealed in the ratings of variables that required the most inference. To examine this possibility, variables were sorted into three groups according to the degree of inference required on the part of the observer. It was found that what differences occurred were in the direction of the items which required the least inference producing the highest stability coefficients. Items 1 and 4, which required no inference, produced reliability coefficients of .90 and .89, respectively, which were among the highest obtained (see Figure 4). As demonstrated later, other findings also indicate that the more objective the data, the higher the stability coefficients rise as a function of averaging over observations.

Study 3. Stability of Directly Measured Objective Behavior

Method

This study, in addition to replicating many aspects of Study 1, examined discrete items of objectively measured behavior. It was conducted in two separate classes as an exercise in research.

Nineteen undergraduate seniors in a class in clinical psychology and 15 first-year graduate students in a class in personality, at the beginning of each class period during the second half of the semester, filled out an adjective checklist on their current emotional state. Following this, they recorded information from daily behavior tally sheets that they had made since the last meeting. In addition, unknown to the class, the instructor kept records of selected items of behavior.

Current emotional state was rated by the students on Sudsman graphic scales anchored at one end with "not at all" and at the other with "very." The scales were identified by clusters of three adjectives, such as happy-arrested-joyous, the clusters having been determined by factor analysis of adjective checklists used in previous, similar studies (e.g., Epstein, 1970).

Results

To evaluate the effects of bias in the self-ratings, it is important to consider that the data can be divided according to level of objectivity in two ways. First, they can be classified as subjective, in the sense that they are descriptions of inner states, and objective, in the sense that they can be observed by an external observer. Second, the latter category can be subdivided according to the degree to which there is the possibility for bias to enter. For example, when a subject records the number of telephone calls or letters he or she has received, the figure may be distorted according to the impression the subject wishes to create. The data can be arranged in three categories of decreasing opportunity for sub-
In Table 4, stability coefficients are presented for a 1-day and a 12-day sample for all variables in the study. Also included are the intercorrelations for a 12-day sample of the objective events with each other and with the negative emotions. The results for positive emotions are not included because they add little to the findings for the negative emotions, producing by large and equivalent opposite results. Given a total sample of 12 days, split-half reliability coefficients could, of course, only be computed for subsamples of 6 days. The reliability coefficients for the 12-day sample are estimates obtained by the Spearman-Brown formula for determining the reliability of a total test from the correlation between its halves. To assess the accuracy of such estimates for the current data, 6-day reliability coefficients were estimated from 3-day samples. In all cases, the estimates were almost identical to the values actually observed for a 6-day sample. The 12-day reliability estimates are presented in Table 4, since the intercorrelations of the different variables with each other are all based on 12-day samples, and is helpful in evaluating the validity coefficients to examine the reliability coefficients for the same number of observations.

All items in Table 4 exhibit a marked increase in reliability from the 1-day to a 12-day sample. For the 12-day sample, 21 of the 23 items have a reliability coefficient of at least .70, and 9 of these are at least .90. Of these 9, only one involves the report of an inner state. The increase in reliability that occurs when data are averaged over events and arranged in order of increasing objectivity results in the following: For the 10 unpleasant inner states, the respective mean coefficients for a 1-day and a 12-day sample are .37 and .71, respectively. For self-recorded behavior that is externally observable, the corresponding figures are .40 and .96. For self-recorded physiological reactions, the corresponding figures are .27 and .94. For examiner-recorded variables, the corresponding figures are .44 and .84. It is noteworthy that all three categories that include externally observable events yield higher reliability coefficients than the category that refers to inner states. That the most objective category, examiner-recorded behavior, attained a somewhat lower level of reliability (.44) than the other two categories of externally observable behavior is attributable to one variable, number of erasures, which had a relatively low frequency of occurrence. The other two variables in this category, lateness and pencils forgotten, obtained reliability coefficients of .94 and .93, respectively. It may be concluded that the increase in reliability as a function of averaging over an increasing number of observations cannot be accounted for by subjective bias associated with self-report, since the same phenomenon was as well demonstrated for objective data. This same conclusion is supported by inspection of Figure 5, which presents the data grouped in a somewhat different fashion. It is apparent in Figure 5 that self-ratings of emotions tend to differ in the direction of producing lower reliability coefficients.

Having established high levels of reliability, let us now examine the data with respect to validity. In Table 4, it is evident that the number of significant correlations of objective events with other objective events and with inner states is well beyond chance. Further, the correlations form coherent patterns. Calls made, calls received, letters written, and letters received are all variables that involve communication with others and are intercorrelated. Erasures, papers missing, and lateness to class, all suggestive of carelessness and lack of organization, are related to the feeling states of tension, powerlessness, and confusion. Heart rate mean, heart rate range, headaches, and stomachaches, suggesting heightened physiological arousal and physiologically reactive, are correlated with unhappiness, confusion, and anger-in but interestingly, not with anger-out. Heart rate
range, but not heart rate mean, despite the
former having a lower reliability coefficient,
is significantly correlated with errors (.57)
and with measures (.32), suggesting a variable of
physiological latency that is associated with
behavioral liability.

It may be concluded that high levels of
stability can be demonstrated for objective
as well as subjective data when the data are
averaged over a sufficient number of events
and that once reliability is established, evi-
dence of validity may not be far behind.

Study 4: Relationship of Personality
Inventories to Behavior

There are four ways of interpreting the
evidence that personality inventories have in
the past produced, at best, low correlations
when evaluated against a criterion of objective
behavior. One is that the two are measuring
nonoverlapping aspects of behavior; another
is that the inventories are inadequate; a third
is that the objective criteria are inadequate;
and a fourth is that both are inadequate.

Let us consider the possibility that the cri-
eria are inadequate, either because the specif-
crism of the form is unreasonable or because they are so narrow and
limited in representativeness as to share little
variability in common with the broader at-
ttributes sampled by personality inventories
(e.g., Davidson & Jaccard, 1975; Egly, 1975;
Fiedheim & Ajzen, 1974; Jaccard, 1974;
McGowan & Gormly, 1976; Weigel & New-
man, 1976). To the extent that either of
these limitations exists, an increase in the
sample of objective behavior comparing the
criteria should produce an increase in the
correlations between personality inventories
and objective behavior. Study 4 was taken
to examine this possibility.

Method

Forty-five undergraduates kept records (on forms
similar to those used in the other studies) of their
feelings and behavior on 14 consecutive days, not
including weekends. Each day, at the same time,
a subject set aside 10 min to rate his or her current
feelings and to obtain three 30-sec samples of pulse
rate. Feelings were rated on bipolar scales with 10
adjectives that were anchored at one end by a
cluster of adjectives, such as happy-cheerful-joyous,
and at the other end by an opposite cluster, such as
unhappy-sad-depressed. A tally was kept of the
following, recorded at the end of each day: number
of social phone calls made and received, number of
social letters written and received, number of social
contacts initiated with groups of three or more people,
number of headaches, and number of stomachaches.

On the morning of each day, hours of sleep during
the previous night were recorded, and a rating was
made of moodiness of sleep.

Before they began recording their daily behavior,
six subjects took a battery of personality inventory
forms including a specific and a general form of a specially
constructed inventory made to resemble as closely as
possible the forms used for daily recording. The
specific form contained specific descriptions of sit-
uation and responses. The general form was similar
in wording to most standard personality inventories
in that it was vague or general in its descriptions
of events and response options. An example of a
item in the general form was, "How often do you make
social phone calls? Consider social calls to include
calls other than business calls." Responses for the
general form were selected from a 5-point scale
where each one was scored at one end with "almost never" and at
the other with "very often." The corresponding item in
the specific form was, "How many social phone calls
do you make, on the average, over a 5-day period?
not including weekends? Consider social calls to
include all calls other than business calls." The
response options were as follows: less than 1, 1-3, 4-
7-9, 10 or more. The general form was always scored
before the specific form. Both forms included all the
items of objective behavior in the daily forms, with the exception of heart
rate mean and heart rate range. Frequency of emotions, in
the form of an adjective checklist containing single
adjectives, was included only in the general form, as it
did not seem reasonable to ask subjects to estimate pec-
ifically how many times they experienced a particular
emotion per day. Subjects responded with estimates of the
frequency they experienced the emotion by check-
ing "never," "rarely," "sometimes," "frequently," or
"hardly always." Scores on the adjective checklists
were obtained by combining individual adjectives into
the same 6-item clusters as in the daily forms.

In addition to the above two specially designed
inventories, the following more standard personality
inventories were administered: the Guilford-Zimmer-
man Temperament Survey (1949), the Eysenck Per-
nosality Inventory (1957), the Eysenck-and-Gray Manifest Anxiety scales
and the Eysenck Personality scales (Eysenck & Eysenck, 1965), and
the Epstein-O'Brien foulability Scale (Epstein, 1976). The
latter three were specially designed for previous
research and have been demonstrated to have some
degree of construct validity (cf., Epstein, 1962, 1976;

Results

The findings on reliability replicate the
results of the previous studies that demon-
strated a marked rise in stability coefficients
as a result of averaging over an increasing
number of observations. In Table 5, the
stability coefficients for ratings of inner states
for a 1-day sample range from .22 for ex-
ternal versus internal direction of attention
to .59 for feeling attractive versus unattractive.
The mean reliability coefficient for the 15
inner states is .45, which is somewhat higher
than in the previous study. The increase in
reliability in the present study for a 1-day
sample can be attributed to the use of bipolar
scales that bring six, rather than three, ad-
jectives to bear on each dimension and that
by contrasting the opposites, elucidate the
construct. For the entire 14-day sample, as

Figure 5. Reliability coefficients as a function of the number of days in the odd-even sample in
"Study 1. (Values plotted represent the means of the correlations for the variables in a category. The
values for the 12-day samples were estimated from the 6-day samples by the Spearman-Brown formula.)"
estimated from the split-half 7-day coefficients by the Spearman-Brown formula, the reliability coefficients for inner states range from .71 to .90 external versus internal direction of attention for feeling attractive versus unattractive, with a mean of .86 for the 15 inner states.

A similar increase in reliability as a function of the increase in number of occasions sampled occurs for the more objective data. For a 1-day sample, the reliability coefficients range from .59 for hours of sleep to .70 for heart rate and mean with an average correlation of .41 for the 12 objective variables. The estimated reliability coefficients for a 14-day sample range from .74 for letters received and letters written to .97 for social contacts initiated, with an average correlation of .88. It is noteworthy that for the objective events, the two kinds of data that were most objective in the sense that they involved nothing more than a count taken under supervised conditions in the classroom, that is, heart rate mean and heart rate range, exhibited the same increase as the other measures, attaining two of the highest reliability coefficients, .94 and .93, respectively. Moreover, although the former had a relatively high stability coefficient for a 1-day sample, the latter did not.

Let us now turn to a consideration of validity by examining the relationships of the personality inventories to the daily recorded events, beginning with the inventories designed to record events on an item-by-item basis. It can be seen in Table 5 that the general form of the inventory has a mean validity coefficient of .51 for objective events, as compared to a validity coefficient of .61 for the specific form. There was apparently a moderate gain in validity as the wording in the self-report inventory was made more precisely to match the criterion. For the specific form, 4 of the 10 validity coefficients are greater than .70, indicating a substantial degree of validity for 40% of the items. These items, with their validity coefficients, are as follows: calls made (.73), calls received (.72), social contacts initiated (.73), and hours of study (.79). Although hours of study requires a degree of inference, as it is not clear exactly what constitutes study, number of letters received and written simply requires a tally of unambiguous events. Thus, the results are not easily explained away by the assumptions that daily tallies of events and self-report estimates over an extended period can be viewed as common method variance, in that they both require reports by subjects. Moreover, if all that is involved is common method variance, the correlations of self-report estimates with daily recordings should be as great for a 1-day sample of the latter as for a 14-day sample, which inspection of the data indicates is not the case, and could have been predicted from the consideration that reliability provides a limiting condition for validity. It is noteworthy, in this respect, that in 9 of the 10 variables of objective events under consideration, the validity coefficients for the specific form of the inventory against the criterion of objective events over a 14-day sample are higher than the matching reliability coefficients for a 1-day sample of objective events.4

Validity can further be examined by noting the relationships of the standard personality inventories with the 14-day samples of the daily records. It should not be expected, of course, that the relationships will be very high, given the different ranges of generalization encompassed by the recording of specific kinds of behavior and the broad personality dimensions measured by the personality inventories. In Table 5 it can be seen that there are far more significant correlations than would be expected by chance and that the relationships are, for the most part, coherent and relatively high for what is usually obtained for validity coefficients against objective criteria. There is a large number of relationships that break the .30 barrier, and no correlation is in an opposite direction from expectancy. All three of the anxiety scales in the Epstein-Fenzi inventory (Fenz & Epstein, 1965) correlate significantly

---

4 For a 1-day sample, the mean validity coefficient for self-ratings of emotions was .49 as contrasted with .47 for a 14-day sample. For the general form of the inventory of objective events, the mean validity coefficients for 1-day and 14-day samples were, respectively, .43 and .45. For the specific form of the same inventory, the corresponding figures were .40 and .41.
in the expected direction with daily ratings of threat and tension. Although it is true that the anxiety scales also correlate with daily ratings of other negative feelings, such as sadness, hostility, and depression, the positive correlations are much smaller and are not generally reliable. It is true that there is a general tendency for negative feelings to cluster together (see factor analysis of similar data reported in Epstein, 1970). One or more of the anxiety scales correlates significantly with daily ratings of sleep, stomachaches, and headaches. Moreover, the scales of muscle tension correlate more highly with headaches, a symptom of muscle tension, than with stomachaches, which is a symptom of autonomic arousal, whereas the reverse is the case for the scale of autonomic arousal. The scales of overcontrolled hostility and undercontrolled hostility are both directly correlated with daily ratings of threat and tension, which is reasonable considering that hostility is often a reaction to threat. Yet, only the scale of overcontrolled hostility is significantly correlated with daily ratings of inhibition, inward direction of attention, feelings of helplessness, and lack of reactivity, all of which can be seen as signs of restraint.

In view of the observation in the previous study that heart rate range was associated with behavioral stability, it is noteworthy that the heart rate range in this study was not significantly correlated with either controlled hostility but not with overcontrolled hostility, whereas heart rate mean is associated with neither. It is of further interest that heart rate range, but not heart rate mean, is significantly associated with the scale of disturbance over hostile feelings (40) and with Eysenck and Eysenck's (1968) scale of neuroticism (37), which suggests that, despite its lower reliability than heart rate mean, heart rate range is a more interesting personality variable.

In conformity with the theory of the self-concept (cf. Epstein, 1973, 1976) that guided the selection of the items for the study, the Epstein-Obrien Self-Esteem Scale (Epstein, 1976) correlates with daily feelings of worthiness (47), integration (31), optimism (55), and alertness (40). Among the Guilford-Zimmerman scales (1949), the scale of sociability is significantly negatively (47), and the scale of sociability positively (40) associated with number of social contacts initiated. The scale of emotional stability in positively associated with daily ratings of integration (44), kindness (41), optimism (44), and soundness of sleep (30) and is negatively associated with frequency of headache (20). The Eysenck (Eysenck & Eysenck, 1968) scale of extraversion is most strongly associated with daily ratings of spontaneity (45), with outgoing feelings (47), and with number of social contacts (42), all of which are recognized elements of extraversion according to Eysenck.9 The reader may judge for himself or herself by further examination of Table 5 whether the results provide evidence of coherent relationships between self-report inventories and records of daily behavior and feelings.

In summary, the correlations of the daily records with a specific form of an inventory that closely matched the items in the daily form were in the vicinity of .60, the correlations with a more general form of the same inventory were in the vicinity of .50, and the correlations with relevant scales on more standard personality inventories were in the vicinity of .40. It should be noted that all of these correlations are highly significant, that all are above the presumed .10 personality barrier, that none was corrected for attenuation due to unreliability that most would be about .10 higher if they were thus corrected. It may be concluded that evidence for a respectable degree of validity in self-report inventories can be demonstrated when the criterion consists of an adequate sample of behavior. This is not to imply that scores on personality inventories permit a high degree of accuracy in predicting behavior. The problem of accuracy in predicting behavior will be discussed later.

Discussion

Stability of Behavior and Error of Measurement

The classic debate on stability in personality can be resolved by noting that the problem lay in a failure to take into account error of measurement as it relates to temporal reliability. Single items of behavior, no matter how carefully measured, like single items in a test, normally have too high a component of error of measurement to permit demonstration of high degrees of stability. Once this is recognized, the solution to two related problems becomes apparent. First, the contradictory findings between a few studies that have reported stability in personality using data derived from ratings and a much larger number of studies that failed to find evidence of stability using data derived from the direct measurement of behavior can be accounted for by the observation that the former studies examined adequate samples of behavior, whereas the latter examined single events. Second, the failure to relate self-ratings, ratings by others, and personality inventories to the same behavioral criteria can be accounted for by the unreliability of the behavioral criteria, which almost always consisted of single items of behavior, usually measured in a laboratory setting.

After a review of the literature, the following hypothesis and corollary were formulated: (a) Stability can be demonstrated over a wide range of variables, so long as the behavior in question is averaged over a sufficient number of occurrences. This applies equally to data derived from the direct measurement of objective behavior, from self-reports, and from ratings by others. (b) Significant relationships can be demonstrated between ratings by others and self-ratings, including standard personality inventories on the one hand and objective behavior on the other, so long as the objective behavior is sampled over an appropriate level of generality and averaged over a sufficient number of occurrences. Four studies conducted to test the hypothesis and corollary provided unequivocal support for their validity. The studies demonstrated that when single events were examined, there was little evidence for stability, but that when averaging was done over a sufficient sample of events, there is strong evidence for stability, as well as convergence among the different kinds of data. The studies also demonstrated that once high levels of reliability are established, evidence of construct validity is apt to emerge in relationships among the different variables, including ones that do not share common method variance. Thus, error of measurement appears to be the crucial consideration in demonstrating stability in personality and in relating self-ratings and ratings by others to objective data.

It is difficult to believe that with rare exception (e.g., Tryon, 1973), the concept of error of measurement was overlooked throughout the long history of the debate on stability of personality. Perhaps there is a lesson to be learned from this. Can it be that overevaluation of the experimental method, as normally practiced, blinded researchers to the inherent limitations of studying behavior in the laboratory? Given the ease in which laboratory experimental procedures have been held, who would have thought that they often fail to meet one of the most fundamental scientific tests of all, temporal reliability (reliability)?

9 Other evidence consistent with this thesis is provided by Finkbeiner and D'Amato (1974), by Weiss and Silverman (1971), by Magnusson and Heffer (1969), by McGowan and Gormley (1956), by Stouffer (1954), by Tryon (1973), and by Wiegand and McNew (1976).
Who would have dared to think that the emperor was wearing no clothing? In both cases, the solution is not only emotionally unacceptable but seems too simple to be true. Yet having arrived at the solution, one finds that the future is more hopeful, for decisions can neither warm an emperor nor advance the cause of science. (See also Brunswik, 1947, 1956; Hammond, 1954, 1955; Magnusson & Heffer, 1969; Pervin, 1977; and Tryon, 1973, for other expressions of concern about inadequately sampled situations.) The broader implications for psychological research of the low temporal reliability of single observations will be discussed in a second article devoted to this topic (Epstein, Note 2).

Issue of Cross-Situational Stability

It has been argued that the real issue with respect to the existence of traits is the demonstration of cross-situational stability. The demonstration of stability per se may, of course, establish nothing more than the existence of narrow habits, which never has been at issue. It is thus important to consider here the findings reported in this article relative to the issue of cross-situational stability.

It will be recalled that a relatively high level of stability in behavior was demonstrated over a wide variety of behaviors, some narrowly conceived, such as making telephone calls, and others broadly conceived, such as feelings of kindliness and acts of nurturance, all of which could be elicited by a variety of situations. Inspection of the individual cases indicated that the same responses were, in fact, elicited by a variety of situations. As the stability that was demonstrated occurred over the normal range of situational variability in everyday life, it would seem that there is a meaningful level of cross-situational stability demonstrated. Pressed otherwise, it was demonstrated that there is enough cross-situational stability to be observed and that individual behavior can be made without having to specify the eliciting situations. Of course, the way a trait is usually defined and the findings demonstrate the utility of such a concept.

Further support for the existence of broad response dispositions was provided by intercorrelations among the variables. Cohen's clusters of correlations occurred not only among variables sharing the same metaveance, such as ratings of inner feeling, but also among variables assessed by differing methods. Thus, a broadly defined variable, a personality inventory, such as Eysenck's extraversion scale (Eysenck & Eysenck, 1964), produced highly reliable correlations with daily records of feelings of inhibition and desire for seclusion, as well as with a test of social contacts initiated, all of which can be related to the construct extraversion as defined by Eysenck. Number of erasures and papers missing, which were recorded by the examiner without the subject's awareness, not only correlated reliably with each other but correlated reliably with self-reports of feelings of tension, powerlessness, confusion.

The conclusion that there are relatively broad, stable response dispositions, or traits, does not conflict with the behaviorist's position that situations often exert a strong influence on behavior. People obviously do not manifest response dispositions independent of the situation. That is why it is usually necessary to cancel out situational effects, including the basis of such considerations alone, it would not have been surprising that the interactions between single items of behavior and different occasions are usually below .30. To predict individual behavior with reasonableness, correlations in the vicinity of .80 or .90 are required. In the four studies reported in this article, correlations of such magnitude were obtained when the average sample of 14 days was correlated with the average of another sample of 14 days, usually for smaller samples of behavior. It indicates that one can predict average behavior accurately from a similar sample of miming or acting behavior. Moreover, the prediction is not based on some people are more prone than others when there is a reasonable opportunity to do so, and this may include swimming in pools, in lakes, and in oceans.

Further, one cannot test such a cross-situational proclivity to swim by observing a person once in the vicinity of a swimming pool and once in the vicinity of a lake, as there may be many reasons for that person to forgo swimming on a particular occasion. Behavior is obviously determined by more than response dispositions. Given an adequate sample of occasions, however, response dispositions will out.

Issue of Predictability of Behavior

How stable is behavior? It has been demonstrated that there is sufficient stability in behavior so that over a sufficient sample of events, thresholds of consistency become apparent. As noted above, in any one instance behavior is determined largely by the situation. This, of course, allows one to predict situational effects averaged over subjects but not necessarily an individual's rank order, as represented by a correlation coefficient, from one occasion to the next.
obtained for personality inventories. If both measures had sufficiently high stability coefficients, even at least the unreliability of the criterion would not be a limiting factor for validity, as it customarily has been. Finally, it should be noted that not everyone is equally predictable. This was demonstrated in the present article by the finding that within-subject correlations varied over a range that suggested almost no stability in a few individuals and extremely high stability in others, with most individuals demonstrating a moderately high degree of stability. Bern and Allen (1974) observed that a sample of subjects who described themselves as highly consistent produced moderately high stability coefficients, whereas a sample of subjects who reported they were not consistent produced low coefficients. The authors concluded that stability in personality can be demonstrated only for specially selected samples of subjects, which is aptly reflected in their title, "On Predicting Some of the People Some of the Time." According to our findings, it is not necessary to select particular classes of people to demonstrate stability unless one has failed to obtain a sufficient sample of behavior to begin with, in which case one needs individuals with unusually high stability to compensate for a high degree of error of measurement. As demonstrated in the present article, given an adequate sample of behavior to begin with, it should be possible to "predict most of the people much of the time."

Reference Notes


References


Block, J. Advancing the psychology of personal environment: A paradigm shift or improving the quality of current research. In D. Magnusson & N. E. Endler (Eds.), Personality at the crossroads: Current issues in international psychology. Hillsdale, N.J.: Erlbaum.


Cronbach, L. J. Beyond the two disciplines of scientific psychology. American Psychologist, 1959, 14, 701-712.


Jaccard, J. J. Predicting social behavior from personality traits. Journal of Research in Personality, 1974, 7, 357-367.


Received June 23, 1977