Psychological Review
1982, Vol. 89, No. 6, 730–755

0033-295X/82/8906-0730$00.75

Beyond Déjà Vu in the Search for Cross-Situational Consistency

Walter Mischel and Philip K. Peake
Stanford University

Recent efforts to resolve the debate regarding the consistency of social behavior are critically analyzed and reviewed in the light of new data. Even with reliable measures, based on multiple behavior observations aggregated over occasions, mean cross-situational consistency coefficients were of modest magnitude; in contrast, impressive temporal stability was found. Although aggregation of measures over occasions is a useful step in establishing reliability, aggregation of measures over situations bypasses rather than resolves the problem of cross-situational consistency. The Bem–Funder (1978) template-matching approach did not enhance the search for cross-situational consistency either in their original data or in an extended replication presented here. The Bem–Allen (1974) moderator-variable approach also was not found to yield greater cross-situational consistency in the behavior of "some of the people some of the time" either in their original data or in the present study of conscientiousness. Congruent with a cognitive prototype approach, it was proposed and demonstrated that the judgment of trait consistency is strongly related to the temporal stability of highly prototypic behaviors. In contrast, the global impression of consistency may not be strongly related to highly generalized cross-situational consistency, even in prototypic behaviors. Thus, the perception and organization of personality consistencies seems to depend more on the temporal stability of key features than on the observation of cross-situational behavioral consistency, and the former may be easily interpreted as if it were the latter.

The relative specificity versus consistency of social behavior, and the nature and breadth of the dispositions underlying such behavior, probed by Thorndike (1906), Hartshorne and May (1928), Allport (1937, 1966), Fiske (1961), and many others, still remains the focus of controversy in contemporary personality theory. The position one takes on these issues profoundly affects one's view of personality and the strategies worth pursuing in the search for its nature and implications. Few assumptions are simultaneously more self-evident, yet more hotly disputed, than that an individual's behavior is characterized by pervasive cross-situational consistencies. Reviewers of this dispute have repeatedly noted a curious paradox (e.g., Bem & Allen, 1974; Mischel, 1973). On the one hand, compelling intuitive evidence supports the enduring conviction that people are characterized by broad dispositions revealed in extensive cross-situational consistency. On the other hand, the history of research in the area has yielded persistently perplexing results, suggesting much less consistency than our intuitions predict.

Reactions to this consistency paradox have taken two directions. One response has been to challenge the underlying assumptions of traditional trait theories (Mischel, 1968; Peterson, 1968; Vernon, 1964) and search for alternative ways of conceptualizing person variables (e.g., Bandura, 1978; Cantor & Mischel, 1979; Mischel, 1973, 1979) and of studying person–situation interactions through more fine-grained analyses (e.g., Magnusson & Endler, 1977; Moos & Fuhr, 1982; Patterson, 1976; Patterson & Moore, 1979; Raush, 1977). Alternatively, it has been argued that the problems raised reflect

Preparation of this paper and the research by the authors were supported in part by Grant HD MH-0914 from the National Institute of Child Health and Human Development and Grant MH-36953 from the National Institute of Mental Health. We benefited much from the thoughtful comments of more colleagues than we can thank here: We are grateful for their generous help.

Requests for reprints should be sent to Walter Mischel, Jordan Hall, Building 420, Department of Psychology, Stanford University, Stanford, California 94305.

not the inadequacy of traditional conceptualizations of broad traits that yield cross-situationally consistent behaviors but rather the inadequacy of earlier searches for such traits (e.g., Block, 1977; Olweus, 1977; Rushton, Jackson, & Paunonen, 1981). Of the many efforts to pursue this "better methods" approach to the consistency problem, the ones that appear to be most dramatic, and that currently have been greeted as most promising, are the "reliability" solution (Epstein, 1979) and the "template-matching" and "idiographic" solutions proposed by Bem and his colleagues (Bem & Allen, 1974; Bem & Funder, 1978). They are the focus of the present paper both because they deserve serious attention in their own right and because they are excellent exemplars of the methodological solutions proposed for the consistency problem, and their analysis may teach some highly instructive lessons both for research and theory building. In spite of the attention these approaches are attracting currently, the solutions they propose do not resolve the basic issues raised by the consistency paradox.

The present analysis documents this claim, offers data that show the limitations of the approaches of both Epstein and Bem and his associates in the search for behavioral consistency, and argues for a resolution of the consistency paradox based on a theoretical reconceptualization. New data on the temporal stability and cross-situational consistency of conscientiousness in college students are presented and analyzed, guided by a cognitive prototype approach (Cantor & Mischel, 1979; Cantor, Mischel, & Schwartz, 1982a, 1982b). These analyses suggest that the widely shared intuitions of consistency are based on valid observations of behavioral regularities but of a type different from those usually pursued in the effort to resolve the consistency paradox.

On Predicting Most of the People Much of the Time: The Reliability Solution

In Epstein's (1979) view, the consistency debate in psychology, rather than meriting deep and enduring discussion, should never have happened. He sees the consistency issue in personality as ready (indeed, as overdue) for "a solution . . . so obvious that, once pointed out, it reminds one of the fairy tale of *The Emperor's New Clothing*" (p. 1097). The solution is to realize that studies of behavioral consistency rarely sample the behaviors of interest on more than a single occasion. The consistency 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" (p. 1097). In other words, the problems of demonstrating behavioral consistency to support global traits are simply the result of unreliable measurement in past research.

Remembering Reliability

In support of his arguments, Epstein (1979) recently demonstrated that coefficients of temporal stability (e.g., of self-reported emotions and experiences recorded daily and of observer judgments) become much larger when based on averages over many days. Epstein computed split-half reliabilities for samples of behavior varying from 2 days up to about 28 days. He found that as the number of observations included in the composite increased, the split-half reliability also increased. Of course, this phenomenon is a fundamental premise of classical reliability theory (Gulliksen, 1950; Lord & Novick, 1968; Thurstone, 1932), and the increase in reliability through use of aggregated composites is exactly what the Spearman-Brown formula has been used to estimate for years (also see Horowitz, Inouye, & Siegelman, 1979).

The recognition that reliability is important and increases with the number of items aggregated is hardly new to the consistency debate. Even introductory statistics texts routinely intone that we cannot have either validity or utility without reliability, and it is remarkable to suggest that the consistency debate has suffered from mass amnesia for the reliability construct and for the Spearman-Brown prophecy. Far from overlooking reliability, virtually all of the classic, large-scale investigations of cross-situational consistency (e.g., Dudycha, 1936; Hartshorne & May, 1928; Newcomb, 1929) routinely employed behavioral measures aggregated over repeated occasions yet reported findings that triggered the enduring debate (see Peake, Note 1, for review). Moreover, the Office of Strategic Services (OSS) project in World War II, the Michigan Veterans Administration project, the Harvard personologists, and the Peace Corps projects—all large-scale applied assessment projects of the 1940s, 1950s, and 1960s—used aggregated measures, pooled judgments, assessment boards, and multiple-item criteria and nevertheless yielded overall results that raised basic questions about the usefulness and limitations of the traditional personality-assessment enterprise (e.g., Peterson, 1968; Vernon, 1964; Wiggins, 1973). Although those who critically evaluated the state of the field in the 1960s did not overlook the issue of reliability, they nevertheless concluded, as Vernon (1964) did, that "The real trouble (with the trait approach) is that it has not worked well enough, and despite the huge volume of research it has stimulated, it seems to lead to a dead end" (p. 239).

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, then, can Epstein conclude that the consistency debate should never have occurred because it is resolved when reliability is taken into account? Put another way, 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? The answer to this seemingly puzzling shift becomes quite simple once one recognizes that the discrepancy is one of interpretation, not effect. 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.

Distinguishing Temporal Stability and Cross-Situational Consistency

Adequate reliability coefficients can surely be found, as Epstein has demonstrated and as earlier work amply attests. But we have to discriminate clearly between demonstrations of impressive temporal stability on the one hand and cross-situational generality or consistency in behavior on the other. By collecting specific observations over a series of days, and then computing split-half "stability" coefficients, Epstein has accumulated data (Tables 1, 2, 3, of Epstein, 1979) that are relevant to the temporal stability of behavior. But temporal stability has never been a central issue in this debate. As noted by Mischel in 1968, "Considerable stability over time has been demonstrated" (p. 36), and "although behavior patterns may often be stable, they are usually not highly generalized across situations" (p. 282).

Although temporal stability is a fundamentally important phenomenon and merits careful empirical attention, it is not the basic issue of the consistency debate. In our view, the crux of the classic debate is the cross-situational consistency or discriminativeness of social behavior and the utility of inferring traits for the prediction of an individual's actions in particular contexts. The operational distinction between temporal stability and cross-situational consistency is important conceptually because it allows one to postpone (though only momentarily) the problems of psychological similarity by simply looking for the same behaviors over multiple occasions in time. The moment the behavior measures are not identical, the problem of psychological similarity surfaces, and that is the problem that has continually bedeviled the search for cross-situational consistency in a nomothetic trait framework. Although Epstein purports to resolve the consistency debate by using aggregated measures, most of his data are relevant only to an issue that has never been seriously raised.

Epstein presents some data (in Tables 4 and 5 of his 1979 article) that do go beyond the demonstration that aggregation increases reliability coefficients and enhances temporal stability. He presents (in his Table 4) intercorrelations of objective events with each other and with self-rated emotions for a 12-day sample. Because the consistency of self-ratings is also not controversial (Mischel, 1968), only the data intercorrelating aggregated objective events are directly relevant to the issue of behavioral consistency. Table 1, adapted from Epstein's Table 4, shows those coefficients that reached statistical significance. First, note that each of the measures listed in Table 1 is the aggregate of a 12-day sample. As such, the corresponding reliability coefficients (first column) are as substantial as they are noncontroversial.

Evidence for Cross-Situational Consistency?

Now, consider the more interesting question, What is the evidence for cross-situational consistency when highly reliable measures of objectively observed behavior are intercorrelated? Among the 105 relevant correlations computed, the 7 that reached significance (at the $p < .01$ level) were letters written with letters received, calls made with calls received, stomachaches with headaches, heart rate mean with heart rate range, errors with heart rate range, erasures with letters written, and absences with erasures. Significant intercorrelations among such items hardly suggest that the solution to the consistency issue has arrived. Most of the obtained interrelations seem virtually automatic; the more doors I open, the more I tend to close. And we also predict a significant correlation between how often you say "hello" to people and how often they say "hello" to you. Demonstrating some links between bits of behavior (calls made, calls received) that may cohere (often almost by definition and because they are functionally related, virtually demanding each other) does not provide impressive evidence for cross-situational consistencies. Similarly, demonstrating links between bits of behavior that have no apparent conceptual relation (errors with heart rate range) does not provide evidence for the breadth of personality traits. Finally, demonstrating some links between any two bits of behavior that do not even come from the same person seems an odd way to prove consistency within a given person's behavior. It is puzzling how links between the number of letters I write to you, for example, and the number of letters you write to me can speak to my consistency.

Epstein also provides coefficients between aggregated objective events and trait-inventory scores (his Table 5). Most of the significant associations he found are between self-reported headaches, self-reported stomachaches, and other self-reported physical complaints and troubles (e.g., muscle tension, autonomic arousal on the Epstein-Fenz Anxiety Scale). It is well known to person-

Table 1
Summary of Epstein's (1979) Statistically Significant Behavior Intercorrelations

Behaviors	Aggregated reliability	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. Heart rate mean	1.00	—														
2. Heart rate range	.88	.74	—													
3. Calls made	.91			—												
4. Calls received	.94			.77	—											
5. Letters written	.79				.42	—										
6. Letters received	.90				.46	.79	—									
7. Headaches	.90							—								
8. Stomachaches	.91							.77	—							
9. Entertainment	.70									—						
10. Errors	.88		.57								—					
11. Erasures	.60					.52	.42					—				
12. Papers missing	NA					.43	.45						—			
13. Absences	NA										.64			—		
14. Lateness	.94														—	
15. Pencils forgotten	.93															—

Note. Adapted from Epstein (1979, Table 4, p. 1114). N = 26 except for pencils forgotten and lateness for which N = 15. NA = not applicable.

Table 2
Effects of Aggregation Over Occasions on Temporal Stability and Cross-Situational Consistency

Measure	Single behaviors	Aggregates
Temporal stability	.29	.65
Cross-situational consistency	.08	.13

Note. The coefficients reported are mean correlation coefficients across all possible comparisons of the designated type.

ality assessors that some people do complain more than others about their well-being (e.g., Byrne, 1964; Mischel, 1981; Mischel, Ebbesen, & Zeiss, 1973). The remaining coefficients are extremely unimpressive. Such objective events as calls made, calls received, letters written, and letters received—the items of behavior that do interrelate—even when aggregated and highly reliable, turn out to correlate significantly with none of the inventories. Epstein's consistency data, far from offering a solution to the problems raised by the consistency debate, demonstrate and illustrate those problems vividly.

The Carleton Behavior Study

An adequate empirical search for behavioral consistency requires data aggregated to achieve reliability. But it also needs to go beyond reliability, beyond temporal stability, and beyond scattered self-report and behavior correlations. It needs to explore cross-situational consistency in behavior with appropriate and reliable behavior measures sampled across a range of presumably similar situations. Furthermore, such a search needs to be informed by some conceptualization—even if rudimentary—of how behavior is organized or should be categorized for particular goals (e.g., Cantor & Mischel, 1979; Mischel, 1973, 1979). 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.

In collaboration with Neil Lutsky, we initiated the Carleton study in an effort to replicate the work of Bem and Allen (1974), extending greatly the behavioral referents and battery of measures employed. In this research, 63 Carleton College volunteers participated in extensive self-assessments relevant to friendliness and conscientiousness. They were assessed by their parents and a close friend and were observed systematically in a large number of situations relevant to the traits of interest. To illustrate the gist of the results as they bear on the issues addressed in the present article, we will focus on the domain of conscientiousness/studiousness (henceforth called simply conscientiousness).[1] The behavioral referents of conscientiousness in this work consist of 19 different measures. For example, the behavioral assessment of conscientiousness included such measures as class attendance, study-session attendance, assignment neatness, assignment punctuality, reserve-reading punctuality for course sessions, room neatness, and personal-appearance neatness. Note that the specific behaviors selected as relevant to each trait were supplied by the subjects themselves as part of the pretesting at Carleton College to obtain referents for the trait constructs as perceived by the subjects; this is in contrast to many studies in which these referents are selected exclusively by the assessors. For each different measure, repeated observations (ranging in number from 2 to 12) were obtained. Thus, for example, we obtained three observations of assignment punctuality and nine observations of appointment punctuality.

This design allowed us to assess both the temporal stability and the cross-situational consistency of behavior using single observations and using measures aggregated over occasions. We were thus able to systematically assess the gains accrued on both temporal stability and cross-situational consistency when we employed the aggregation solution espoused by Epstein. The results of this analysis are summarized in Table 2.

Temporal Stability: Cross-Situational Discriminativeness

First we computed the percentage of significant coefficients among all the possible

[1] Although the focus here is on the conscientiousness data, results that essentially parallel those reported here have emerged from analyses in the friendliness domain and are described in Peake (Note 1).

coefficients of temporal stability. To qualify as a coefficient of temporal stability, the correlation had to consist of two observations of the same type of measure. For example, lecture attendance on Day 1 correlated with lecture attendance on Day 6 is a correlation of temporal stability. This analysis revealed that nearly half of the single-observation temporal-stability coefficients (specifically, 46%) were statistically significant. Note that this is prior to any aggregation to enhance reliability. Here, then, is another clear indication that even at the single-observation level, temporal stability can be demonstrated readily. Given the moderate to high levels of temporal stability among the single observations, it is not surprising either conceptually or empirically that when these single observations are aggregated into composite measures, all of the resulting reliability coefficients are significant (with the mean coefficient being .65).[2]

We performed a similar analysis for all the correlations relevant to cross-situational consistency. At the single-observation level, cross-situational consistency coefficients consist of such correlations as a single observation of appointment punctuality with a single observation of lecture punctuality, or with a single observation of class-note neatness (i.e., with any other single observation except another observation of appointment punctuality). As is now typical for research findings of this type, although the percentage of significant correlations (11%) exceeded chance, the obtained correlations were highly erratic, with a mean coefficient of .08. The critical question then becomes, What gains in cross-situational consistency are evidenced when our more reliable aggregates are intercorrelated?

To address this question, we must examine such correlations as aggregated lecture attendance with aggregated appointment punctuality or aggregated lecture attendance with aggregated appointment attendance. For this purpose, 171 cross-situational consistency coefficients were computed by intercorrelating the 19 different aggregated measures of conscientiousness. Of the 171 coefficients, 20% (35 coefficients) reached significance—a number considerably above chance. Some of these coefficients reached substantial levels. For instance, aggregated class attendance correlates highly with aggregated appointment attendance ($r = .67$, $p < .001$). Furthermore, there are patterns of meaningful coherences among the correlations. For instance, aggregated class attendance correlates significantly with aggregated assignment punctuality ($r = .53$, $p < .001$), with completion of class readings ($r = .58$, $p < .001$), and with amount of time studying ($r = .31$, $p < .05$). These coherences once again testify that behavior is patterned and organized rather than random. But it is just as clear from the results that behavior is also highly discriminative and that broad cross-situational consistencies remain elusive even with reliable measures. For example, whereas aggregated class attendance correlates impressively with the above measures, it does not correlate significantly with aggregated class-note thoroughness ($r = .14$, *ns*), aggregated punctuality to lectures ($r = -.03$, *ns*), or aggregated assignment neatness ($r = -.04$, *ns*). This discriminativeness is further reflected in the fact that for the 19 aggregated measures, the mean cross-situational consistency coefficient was only .13 (see Table 2). Thus, the use of aggregate measures increases the mean consistency coefficient from .08 to .13.[3]

It is possible that our data do not adequately reflect the gains in cross-situational consistency attainable through aggregation because several of our measures evidenced low reliability coefficients even after aggre-

[2] The aggregate temporal reliability coefficients reported here were computed using the split-half technique employed by Epstein (1979) to parallel his procedures. A more efficient procedure involves computing the mean single-observation intercorrelation for the behaviors to be aggregated (Lord & Novick, 1968). Aggregate reliability is then computed using the Spearman-Brown formula, correcting the mean single-observation correlation by the number of observations that are to be included in the composite. With this latter procedure, the mean aggregate temporal stability coefficient is .61 (Peake, Note 1).

[3] Unless the single observations within one aggregate show a stable (albeit low) pattern of association with the observations in the comparison aggregate, substantial increases in validity should not be expected to result from aggregation (Lord & Novick, 1968). That is, gains in cross-situational consistency should not be expected from aggregation when the measures being compared show erratic validity coefficients prior to aggregation. Interestingly, the persistent finding of such erratic relations is precisely what led to the controversy over the breadth and generality of behavioral consistency in the first place.

gation. To apply a more stringent test of the effects of increased reliability for demonstrating cross-situational consistencies, we focused our analysis on those measures whose estimated reliability exceeded an arbitrary selected level of .65. In all, 14 of the 19 original measures met this break-off point, and the mean reliability estimate of these measures is .74. Intercorrelating these 14 measures results in 91 consistency coefficients with a mean level of .14. Thus, restricting our attention to our most reliable measures yields a minimal gain in the mean cross-situational consistency coefficient which increases from .13 to .14.

Although restricting our attention to our most reliable measures does not seem to have a substantial effect on the cross-situational consistency evidenced in the data, one might argue that a mean reliability coefficient of .74 is still too low to evaluate the reliability solution. What would happen if all our mean reliability coefficients were .95? More impressively, what kinds of cross-situational consistency would be evidenced if we were to use perfectly reliable measures? Of course, it is to answer just this type of question that classical test theory tells us to to apply the correction for attenuation. By correcting each of the obtained correlations in our matrix for attenuation due to low reliability, we can estimate the maximal ("true") level of association between each of our measures. In this sense, correcting for attenuation is the logical extreme, or ultimate test, of the "reliability solution" as proposed by Epstein. When we apply the correction as described above, the mean consistency coefficient increases to .20. Thus, if we were to collect a large number of observations for each measure such that the reliability of each of the composites (aggregations) of these observations approached 1.0, the mean level of cross-situational consistency evidenced between these measures still would be unlikely to exceed .20.

In light of the various results summarized to this point, what can one conclude about the promise of the reliability solution for the consistency debate? First of all, it is clear that aggregation of repeated observations in order to obtain adequately reliable measures yields, as expected (and hardly to anyone's surprise), gains in the mean levels of correlations both for measures of temporal stability and for cross-situational consistency in behavior. These gains are most impressive for measures of temporal stability (mean $r = .65$) and document that aggregation over occasions is a useful method for increasing the reliability of a measure. The results also indicate, however, that aggregating observations over occasions does not necessarily lead to high cross-situational consistency (mean $r = .13$ or about .20 if perfect reliability is assumed).[4] Although aggregation over occasions has the desirable effect of enhancing reliability, it does not provide a simple solution to the consistency paradox.

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 the studies by Hartshorne and May (1928) of honesty, by Newcomb (1929) of introversion–extroversion, by Allport and Vernon (1933) of expressive movements, and by Dudycha (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

[4] The summary coefficients of mean temporal stability and mean cross-situational consistency are not strictly comparable because the former always reflect the same response forms in the same situation, whereas the latter often reflect different response forms in different situations. To allow a direct comparison, cross-situational consistency coefficients were computed separately when the response forms in the different situations were the same (mean $r = .28$) and when they were different (mean $r = .12$). Thus, whereas the mean correlation for the same response forms in the same situations over time was .65 (temporal stability), it was .28 for the same response forms in different situations.

pervasiveness of the phenomena of behavioral discriminativeness.

The Carleton data do not imply, however, that there is little coherence among the behaviors studied. Although the overall patterning of correlations is erratic and on the average low level, the results do not suggest that behavior is random and unorganized. As was noted above, some impressive coefficients emerge, and coherent patterns of correlations are apparent among some of the variables. In addition, 78% (133 of 171) of the obtained correlations are positive, and of the 38 negative correlations, only 2 reach statistical significance. Thus, we obtained considerably more positive significant correlations and also obtained considerably fewer negative significant correlations than would be expected by chance. So, whereas the data reflect behavioral discriminativeness, there is also a positive trend, a coherence or gist among the behaviors sampled.

Aggregation of the Data—Without Aggregation of the Issues

Cross-situational consistency coefficients of the sort we are finding can be construed as evidence either for the relative discriminativeness of behavior or for its coherence, and as evidence either for a stable thread of individual differences or for the need to take account of situations seriously. How one reads the results depends on the particular purposes of the research or assessment task. Recently, however, it has become increasingly common to interpret mean coefficients of the sort obtained at Carleton as ample evidence for the consistency of behavior. By aggregating measures of behavior in particular situations into a single composite, or "multiple-act criterion," substantial internal reliability coefficients are readily obtained (Fishbein & Ajzen, 1975). The problem here is not whether reliability will increase by using various types of data aggregation but how to select the appropriate type and level of data aggregation for particular research problems. Automatic aggregation into overall composites simultaneously risks aggregation of the conceptual issues.

Aggregation of observations over occasions—Epstein's (1979) basic admonition—certainly is a requisite for adequate reliability. No one would contend that a person's attendance at today's psychology lecture at 9 a.m. is an adequate index of that person's tendency to attend psychology lectures. By measuring lecture attendance on repeated occasions, a more reliable index of lecture attendance will result. Of course, aggregation per se need not stop at aggregation over occasions (Epstein, 1980). The investigator who wants to amass high correlation coefficients could forge ahead and aggregate behavior across different response forms and even across situations, arriving at the now popular multiple-act criterion (Fishbein & Ajzen, 1975; Jaccard, 1974; McGowan & Gormly, 1976; Rushton et al., 1981). Here, again, these aggregations will be appropriate and useful for some purposes but not for others.

Consider the case for aggregating across response forms. In many situations a particular dimension of interest may be assessed through several behavioral manifestations or types of responses. In our data at Carleton College, for instance, one could assess cross-situational consistency not just through the intercorrelation of the 19 measures discussed so far but by treating particular settings (e.g., classrooms) as the situations and aggregating the multiple response forms sampled within them. In that case, conscientiousness in classroom situations may be indexed by such measures as lecture attendence, lecture punctuality, note thoroughness, and so forth. The investigator who is more interested in "conscientiousness in classroom situations" than in the variations between measures within this situation could appropriately aggregate across the response forms. The composite measure of conscientiousness in the classroom could then be compared with measures of conscientiousness, aggregated across response forms, in other situations (e.g., at appointments). By aggregating across occasions and response forms in the Carleton data, the mean number of specific observations per situation became 20. The mean cross-situational consistency coefficient for these aggregated measures was .18. Aggregation across response forms within situations provides a useful increment, but hardly a resolution of the paradox.

After measures have been aggregated over occasions, and across response forms within situations, it is possible to aggregate even further, combining measures across situations. One can treat the specific situations simply as "error" and aggregate across them to form a single composite score and, as the Spearman-Brown formula predicts, convert our average .13 cross-situational consistency coefficient into an internal reliability estimate of .74. But aggregating across situations in this way cancels the variance and specificity due to situations, thereby bypassing the problem of cross-situational consistency instead of solving it; that "solution" merely treats situations as errors to be averaged out rather than as psychological units to be taken into account. Although achieving reliable measures by sampling over occasions is of self-evident value, further aggregation, across response forms or across situations, must be dictated by the assessor's goals and by a priori theoretical considerations about psychological similarity (equivalence groupings) and about the level of generality—the units—at which assessment should occur for particular purposes (e.g., Cantor & Mischel, 1979; Mischel, 1977). Although such aggregation is useful for making statements about mean levels of behavior across a range of contexts, cross-situational aggregation also often has the undesirable effect of canceling out some of the most valuable data about a person. It misses the point completely for the psychologist interested in the unique patterning of the individual by treating within-person variance, and indeed the context itself, as if it were "error."

The persistent pursuit of consistency by attempting to treat situations as error is a curious paradox in a field committed to a focus on individuality and the pursuit of personology (e.g., Carlson, 1971). On the one hand, the importance of attention to the within-person patterning of attributes and behavior—the crux of the idiographic approach and the uniqueness of the person—has long been recognized (e.g., Allport, 1937). On the other hand, this patterning is aggregated out and treated as if it undermined the basic phenomena of personality psychology. As a personologist (or clinician) I may be less interested in aggregating a child's total aggressiveness across situations than in noting that she is aggressive with her sister, but not with her brother, or is aggressive only when teased in a particular way, but never when in the presence of her father. In sum, aggregation may be ideal for canceling out many influences and describing mean level differences between individuals, but such lumping is obtained at the cost of much valuable information—often the most interesting information—about the individual in the particular contexts in which he or she lives. For the clinician as well as for the personologist, aggregation is often the route to weak generalizations about people in general but bypasses the uniqueness, specificity—and predictability—of individuality to which a science of personology is ostensibly devoted.

The conceptual problems that must be addressed when aggregating various types of data are essentially problems of psychological similarity. Evidence for higher average coefficients for temporal versus cross-situational consistency suggests that when the situations are as close as possible to identical (i.e., changed only by time), there is impressive average stability. But when situations become even somewhat dissimilar, the patterns become more complex and uneven, average coefficients become much lower, and consistency can no longer be assumed. The need to search for ways to identify similarity then becomes manifest (see Lord, 1982, and Magnusson & Ekehammar, 1973, for interesting examples). Few problems in psychology seem more basic than that of psychological similarity, (e.g., Tversky, 1977), and its resolution ultimately should have much to say to the study of situational equivalences and the categorization of behavior. Theory-guided aggregation requires identifying psychological equivalences and the psychological similarity among situations, not just averaging everything that can be summed.

In our view, the challenging problems in the consistency debate require more than searching for significant correlations and recognizing coherence in the obtained results. We need to understand why the obtained coherences emerge and when and why expected coherences do not. The technologies of psychometrics supply us with ample methods for distilling the coherence among our measures, for accentuating the mean levels of individual differences we have identified, and for focusing on their gist. As our analysis continues, we intend to employ these various technologies in hope of fully illuminating the psychological significance of these coherences. However, we simultaneously intend to pursue the oft-neglected alternative path of attempting to understand the discriminativeness that also clearly exists in our data and that we believe demands an interactional perspective that treats situations as sources of meaningful variance. For instance, in our research at Carleton College, we plan to search for consistency at different levels of abstraction–generality in the data, from the most "subordinate" or molecular to increasingly broad "superordinate" molar levels, guided by a cognitive prototype and hierarchical-levels analysis of the sort proposed by Cantor and Mischel (1979) and Cantor et al. (1982a). We also plan to explore the comparative usefulness of measures specifically designed to tap such cognitive social-learning person variables as the individual's relevant competencies, encodings, expectancies, values, and plans (Mischel, 1973, 1979).

Such analyses should help illuminate processes that underlie both the significant and the nonsignificant coefficients yielded by personality research—the "uneven and erratic patterns" of behavior that characterize person–situation interactions. Aggregation of repeated observations over occasions will aid in these analyses by providing a more accurate picture of the significant and nonsignificant links that characterize the data. The reliability solution, rather than providing a simple answer to the issues raised regarding the cross-situational consistency of behavior, highlights their complexity. Rather than resolving the consistency debate, the reliability solution underlines the need to seek alternative conceptualizations of personality that might lead us to a better assessment, understanding, and appreciation of both the coherence and the discriminativeness of human behavior.

On Predicting More of the People More of the Time: The Template-Matching Solution

Another solution for the consistency problem is offered by Bem and Funder (1978). Their aim was to utilize Block's California Child Q-Set to find a common language for the description of both persons and situations and to develop a template-matching technique that allows one to examine the interface of person and situation characteristics. Because of the great promise it offers, and the substantial attention it already has attracted, Bem and Funder's contribution merits careful analysis and a thorough examination of its findings and implications. In this analysis the focus will be on the content area of delay of gratification, the domain Bem and Funder chose to illustrate their approach, as it speaks to the consistency problem, and thus will deal only with their first study.

The Bem–Funder Study

Bem and Funder note the failure to obtain more impressive cross-situational consistency coefficients in personality research generally and for delay of gratification in particular. They suggest that the inconsistencies obtained are due to the erroneous equation of situations that are superficially similar but functionally different. To identify situations that are functionally dissimilar though apparently similar, they propose examining the rated personality characteristics associated with the behavior in these situations. Their ultimate hope is to increase the predictive power of trait information by matching the individual's personality characteristics with the "personality of the situation," defined by Q-sort ratings that supply "portraits" of both the individual and the setting. They see their work as relevant to the consistency issue by showing that situations that seem alike may actually be dissimilar, as evidenced by discrepant patterns of Q-sort correlates. We therefore should not expect behavior in such situations to be highly intercorrelated. Moreover, they argue that only when situations are characterized by similar Q-sort portraits should we expect—and find—high intercorrelations among the behaviors displayed in them.

Their procedure and results require close consideration. They exposed 29 preschool children to a version of the traditional delay-of-gratification situation. Although Bem and Funder retained the basic features of the delay paradigm (e.g., Mischel & Ebbesen, 1970), there was a possibly important difference: The experimenter remained in the room with the child during the delay period rather than leaving the room as in the traditional paradigm. Correlations were computed between a child's delay time and each of the 100 items in the Q-set (i.e., the parents' trait ratings of this child).

Some of the correlations they obtained were highly consistent with previous findings on the rated characteristics of children who are high versus low in their ability to wait in the standard delay situations (see Table 3, reproduced from Table 1 of Bem & Funder, 1978, p. 490). Bem and Funder note the existence of these expected correlations (e.g., "Has high standards of performance" is related to duration of delay, as earlier research suggests it should be). However, they emphasize that the rest of the portrait

> introduces a more *dissonant note*, a picture of the long-delaying child as *not very intelligent, not verbally fluent, not eager to learn, or not open to new experiences*—a child, moreover, who is *not self-assertive, cheerful, interesting, or creative.* The very strong negative relationship between delay and rated intelligence is particularly inconsistent with theories of ego control. (p. 490; emphasis added)

Table 3
Q-Item Correlates With Delay Scores

Item	*r*
Positively correlated	
Has high standards of performance for self	.48***
Tends to imitate and take over the characteristic manners and behavior of those he or she admires	.39**
Is protective of others	.39**
Is helpful and cooperative	.36*
Shows a recognition of the feelings of others (empathic)	.35*
Is considerate and thoughtful of other children	.34*
Develops genuine and close relationships	.31*
Negatively correlated	
Appears to have high intellectual capacity	−.62***
Is emotionally expressive	−.56***
Is verbally fluent, can express ideas well in language	−.50***
Is curious and exploring, eager to learn, open to new experiences	−.49***
Is self-assertive	−.47**
Is cheerful	−.43**
Is an interesting, arresting child	−.43**
Is creative in perception, thought, work, or play	−.40**
Attempts to transfer blame to others	−.37**
Behaves in a dominating way with others	−.34*
Is restless and fidgety	−.31*
Seeks physical contact with others	−.31*
Is unable to delay gratification	−.31*

Note. From "Prediciting More of the People More of the Time: Assessing the Personality of Situations" by Daryl J. Bem and David C. Funder, *Psychological Review,* 1978, *85,* 485-501. Copyright 1978 by the American Psychological Association. Reprinted by permission.
* $p < .10$ (two-tailed). ** $p < .05$ (two-tailed). *** $p < .01$ (two-tailed).

Bem and Funder then confront the consistency issue by comparing the Q-sort portrait from their version of the delay situation with the portrait that emerged from a study of "gift delay" by Block (1977) in an effort to show why behavioral consistency has not been found in these two situations. In Block's procedure, the child sits in front of a gaily wrapped gift that is not to be opened until he or she has completed a puzzle. The measure of delay is the length of time the child waits before reaching out and taking the gift. Bem and Funder compare their own results with the Q-sort items that were associated with long delay time on Block's measure and note that the two Q-sort portraits

> show very little overlap, and the dull-passive-obedient cluster of items that emerged in our experiment is completely absent from the Block data. What we have here, then, are two situations that appear conceptually equivalent but that are functionally quite different, and it would appear that different subsets of children are delaying in the two settings. Typically, one learns only that behavior across two theoretically similar situations is disappointingly inconsistent. By collecting Q-sort data, however, one can see in exquisite detail the nature of that inconsistency and draw plausible inferences about its source in the nonoverlapping features of the settings. (p. 491)

The Bem–Funder approach promises to help resolve the consistency issue by identifying those situations that are psychologically equivalent rather than merely superficially similar, but in fact, their data provide no evidence whatsoever for cross-situational consistency. The design of their study does not allow them even to make any estimates about potential consistency because evidence for cross-situational consistency can only be obtained if the same subjects are observed in at least two or more situations. Rather, their conclusions are based on a comparison of the

Q correlates of delay behavior in two different delay situations, for two different sets of children, where samples were not even matched in age. Their paper thus raises the prospect of improving evidence for cross-situational consistency in behavior, but it does not take the essential step of doing so. That is why the present authors proceeded to try to fill this void and tested the Bem–Funder promise empirically with an appropriate design for assessing cross-situational consistency.

The Mischel–Peake Study

We attempted to replicate the Bem–Funder study, with some important additions. We exposed children to the Bem–Funder paradigm using identical procedures, the same immediate and delayed rewards, and subjects of the same age drawn from the same population (Stanford University's Bing Nursery School), whom we tested at about the same time of year as those in the Bem–Funder study. These children also participated in the standard delay paradigm (described in Mischel & Ebbesen, 1970), which differs from the Bem–Funder procedure only in that the experimenter is absent during the delay period. This design made it possible for us to assess the consistency of the children's delay behavior across the two situations (experimenter present versus absent) and to compare the Q correlates for the two situations systematically. The children participated in the two situations in random order and within a period of 3 weeks. To enhance reliability, the sample size was nearly twice that in the Bem–Funder study.

Delay behavior in the two situations correlated at a modest level ($r = .22$, *ns*). According to Bem and Funder's reasoning, this low level of association provides an ideal test of their proposition on two counts. First, the two situations appear to be similar (indeed, the two vary only in the experimenter's presence versus absence—an aspect apparently so trivial that Bem and Funder employed it as a substitute for the standard Mischel delay paradigm). Second, behavior in the two situations proves to be not strongly intercorrelated empirically. Applying the Bem–Funder Q-sort approach to the two situations should reveal the "distinctive features" that make these seemingly similar situations psychologically different, showing their "unique personalities" and revealing why behavior in the two situations is not more closely intercorrelated.

Our investigation yielded results that surprised us greatly. In spite of our considerable efforts to conduct an exact replication of the Bem–Funder delay study, the data we obtained did not support their basic findings; it reversed them. The major distinctive Q correlates obtained by Bem and Funder are shown in the first column of Table 4; the Q correlates yielded by our replication are given in the second column. The resulting portrait, far from replicating the Bem–Funder distinctive features, is the opposite of the one they drew. For example, on the identical measure (i.e., with experimenter present), the high-delay child, rather than being intellectually dull and uneager to learn, appears to have high intellectual capacity ($r = .23$, $p < .10$) and is curious and exploring, eager to learn, and open to new experiences ($r = .27$, $p < .05$).[5] Bem and Funder's strong negative correlations between high delay and verbal fluency; self-assertiveness; and creativity in perception, thought, work, or play are all lost ($r = .06$, .09, and .00, respectively). The remaining "distinctive features" also fail to reach statistical significance.

Interestingly, it is only the distinctive features that are lost or reversed in the replication. The Q-sort correlates that are theoretically consistent with the previously established view of the "ego strength" correlates of high delay remain intact, indicating that the delay measure employed had validity even though the Q correlates reported as distinctive by the Bem–Funder methodology turned out to be unstable. Thus, traditional ego-strength correlates for the delay measure (Mischel, 1974) such as "has high standards of performance for self" and "develops gen-

[5] Bem and Funder's criterion for determining the weightings of Q-sort items was statistical significance at the $p < .10$ level (two-tailed). We have adopted this nontraditional level here only for purposes of replication and to compare the obtained Q-item correlates with theirs. For all other correlations and analyses in this article, we use the more traditional $p < .05$ (two-tailed) criterion unless otherwise noted.

Table 4
Comparative Correlates for Significant "Distinctive" Items of the Bem and Funder (1978) Delay Situation

Item	Bem and Funder (1978) r	Mischel and Peake replication r
Appears to have high intellectual capacity	−.62***	.23*
Is verbally fluent, can express ideas well in language	−.50***	.06
Is curious and exploring, eager to learn, open to new experiences	−.49***	.27*
Is self-assertive	−.47**	.09
Is cheerful	−.43**	−.19
Is an interesting, arresting child	−.43**	−.17
Is creative in perception, thought, work, or play	−.40**	.00
N	29	52

* $p < .10$ (two-tailed). ** $p < .05$ (two-tailed). *** $p < .01$ (two-tailed).

uine and close relationships" were of the expected sort ($r = .25, p < .05$ and $r = .23, p < .10$, respectively). In the same vein, the replication yielded the expected correlations between high delay time and such ratings as "is competent and skillful" ($r = .38$, $p < .005$), "is planful, thinks ahead" ($r = .36, p < .01$), "uses and responds to reason" ($r = .27$, $p < .05$), and "is reflective, thinks and deliberates before he or she speaks or acts" ($r = .26, p < .10$).

In sum, we not only found that the Bem–Funder delay situation and the classical delay situation were not reliably differentiated by distinctive Q-sort portraits but also found that the two situations share highly similar and familiar patterns of correlations. The correlations obtained for the "standard delay" (experimenter absent) procedure were generally similar to those for the Bem–Funder (experimenter present) version and were basically coherent with the expected ego-strength portrait (Mischel, 1966, 1974). Namely, these were low-level but statistically significant associations between the standard-delay measure and such ratings as "has high standards of performance for self"; "uses and responds to reason"; "is curious and exploring, eager to learn, open to new experiences"; "can acknowledge unpleasant experiences and admit to own negative feelings."

This total pattern of results has considerably more significance than a failure to replicate the Bem–Funder distinctive portraits. The data more importantly speak to the value of Bem and Funder's approach for resolving the consistency issue in the directions they proposed. According to Bem and Funder, here are two situations revealed by their method to have similar correlates and therefore expected by their analysis to yield cross-situational consistency in behavior. Yet, in fact, the intercorrelation for behavior in these two situations is only .22 in our data. Disappointingly, the Bem–Funder approach, at least in the delay context, provides no perceptible increment in the usual level of cross-situational consistency obtained, regardless of whether one views that level as not very high or not very low.

Although the obtained Q-sort portraits proved neither stable nor distinctive, we must still consider the utility of the template-matching technique for making predictions about the delay behavior of children within our study. So far, we have considered only the Q-sort correlates of the measures, without applying the template-matching procedures proposed by Bem and Funder to the data. Their weighting procedure was intended to allow them to transport "the relevance information . . . from one subject sample to another" (p. 492) and thus to "generate templates that retain their situation-characterizing properties while simultaneously characterizing most closely the particular sample of individuals whose behavior is to be predicted" (p. 492). Given that the distinctive features of the Q-sort portrait for the Bem–Funder measure proved to be so unstable, it is not surprising that application of the original Bem–Funder template weights to our replication turned out to have no predictive value, producing a negligible correlation ($r = .05$, *ns*). More distressing is the fact that within the replication sample itself, efforts to cross validate internally the template-matching technique did not succeed in spite of our attempts to adhere closely to the original procedures.[6]

[6] Although our focus here is only on the application of the Bem–Funder template-matching approach to the

Old Cookbooks in New Templates?

In our view, the significance of the results discussed so far is not that they document the empirical limits of a particular study but rather that they illustrate the boundaries of a more general approach with a formidable history and tradition. To understand the failure of the Bem–Funder strategy for solution of the consistency problem, we must begin by examining more closely the strategy itself. They hoped their strategy would be a new route to uncover the "personality of the situation"—in this case, the delay-of-gratification situation. But though they conceptualized and described their efforts in the language of person–situation interaction and hoped to capture the personality of the delay situation, their method only described the correlates of *performance* in that situation (i.e., the ratings associated with duration of waiting time). To justify describing the Bem–Funder approach as an assessment of the personality of situations would require that one at least also Q sort the *situations* independently, not just the people who perform in them (Hoffman & Bem, Note 2). No set of correlates for performance, no matter how described, can do more than illuminate some of the specific ways particular kinds of people are likely to perform or to seem different from other kinds on particular measures. The search to characterize what kinds of people are likely to do well on particular tasks, situations, or measures is exactly what traditional personality assessment—both clinical and actuarial—has been about for decades (e.g., the classic work of Murray and the Harvard personologists). Such a search for performance correlates to characterize the meaning of a response pattern and to predict individual differences is the essence and staple of the trait approach (Mischel, 1968).

This search can be undertaken through a strategy of construct validity (Cronbach & Meehl, 1955), in which the focus is on the development of a theory about what accounts for the behavior of interest. Or, like Bem and Funder's approach to revealing consistency in delay of gratification, it can be undertaken in an entirely empirical fashion, as in the actuarial and empirical keying strategies pioneered by Paul Meehl (e.g., 1956) and favored for many years in personnel and psychodiagnostic assessment. The "cookbooks" and "atlases" devised two decades ago in this approach, for example, tried to assess the degree of fit between an individual and an ideal personality type (defined by a distinctive Minnesota Multiphasic Personality Inventory [MMPI] test configuration) in order to predict a pattern of criterion performance and characteristics associated with that ideal type in a particular setting (e.g., a Veterans Administration mental hygiene clinic). It is worth remembering in detail Meehl 's (1956) cookbook method for the prediction of personality descriptions and/or other data in particular situations. The logic—and fate—of his approach bears directly on the solutions currently proposed by Bem and his associates:

> In the cookbook method, any given configuration (holists please note—I said "configuration," not "sum"!) of psychometric data is associated with each facet (or configuration) of a personality description, and the closeness of this association is explicitly indicated by a number. This number need not be a correlation coefficient—its form will depend upon what is most appropriate to the circumstances. It may be a correlation, or merely an ordinary probability of attribution, or (as in the empirical study I shall report upon later) an average Q-sort placement. (p. 264)

Twenty-two years later, Bem and Funder (1978) wrote,

> What is being proposed here, then, is that situations be characterized as sets of template–behavior pairs, each template being a personality description of an idealized "type" of person expected to behave in a specified way in that setting. The probability that a particular person will behave in a particular way in a particular situation is then postulated to be a monotonically increasing function of the match or similarity between his or her characteristics and the template associated with the corresponding behavior. (p. 486)

Note again that although their intent is to characterize situations, the operation is actually to seek the correlates of performance in a specific situation. Both Bem and Funder and Meehl assert that the probability that a subject will do a specified act in a particular

delay situation, there also have been attempts to replicate their effort to predict attitude change in the forced-compliance paradigm. These attempts at extended replication so far also seem to have failed (Funder, 1979; Hoffman & Bem, Note 2). A more situation-specific or "contextual" procedure now has been substituted for the global Q-sort descriptions (Hoffman & Bem, Note 2) and seems promising, but it might be wise to await replication of these new efforts before judging them.

situation (or be judged to have particular characteristics) depends on the degree of match or similarity between that subject's scores and a given configuration. The configuration is a "template" for Bem and Funder, a "profile" for Meehl. The behaviors (or outcomes) for Bem and Funder are such things as delay time or attitude change or adjustment to Stanford University. For Meehl, they were Q-sort patterns, other personality descriptions, or any other data (e.g., duration of remission or adjustment to the Minnesota Veterans Administration hospital). The method for determining the degree of match between the subject and the configuration involves a particular weighting procedure for Bem and Funder; it was also achieved quantitatively by Meehl but with the recognition that the particular form (simple correlation, regression, probability attribution) will depend on its appropriateness for the particular purpose at hand. And, like Bem and Funder, Meehl (1956, p. 267) searched for ideal types whose distinctive profiles (recipes) would be related to such criterion data as Q sorts in the cookbooks.

Although the language in the descriptions of the two methods is somewhat different, the Bem–Funder "template–behavior pair" parallels the Meehl "configuration–personality description association." The two approaches seem to overlap a good deal, including a heavy reliance on the same technique—the Q sort. Indeed, Meehl (1956) illustrated his pioneering article on the cookbook method with a study that shows its value for predicting therapists' Q sorts of patients from their match to the cookbook's MMPI curve types (substitute "templates"). For example, for each patient who best fits a given MMPI code, "we simply assign the Q-sort recipe found in the cookbooks as the best available description. How accurate this description is can be estimated (in the sense of construct validity) by Q correlating it with the criterion therapist's description" (p. 268). Empirically, the initial yield was highly encouraging, with a median validity coefficient of .69.

Once the close parallels are recognized, it becomes clear that the Bem–Funder approach shares both the strengths and the weaknesses of its predecessor. The strengths are the bypassing of weak (poor, messy) theory-based predictions for the sake of neat, mechanically assisted empiricism.[7] The weakness is the opposite side of this same coin, that is, the cost of blunt empiricism: Atheoretical approaches yield results that tend to seem promising at first but are notoriously difficult to replicate. Assessors excited by the first actuarial cookbooks and their strong predictions were soon sobered by the frequent failure to replicate the configuration–outcome patterns that at first seemed so useful. Particularly when the sample size is small, and when the associations are purely empirical (with no theoretical hypotheses a priori), correlations with Q-sort items may be extremely unstable and can vanish rapidly, as we saw in the failure to replicate the distinctive Bem–Funder portraits. If personality assessment has led to any firm conclusions, it is that it is generally not worth offering conclusions about actuarially obtained performance correlates unless they are based on careful cross-validation.

Another truism emerging repeatedly from Meehl's once-exciting cookbook approach and from the history of personality assessment more generally is that simple linear combinations of single scale scores often turn out to be more accurate than complex, sophisticated configural models, including the Meehl-Dahlstrom Rules (e.g., Goldberg, 1965). Simple cooking with simple recipes generally works better than more esoteric methods in actuarial assessment. At a minimum, it seems wise to compare the increments that fancier (thus costlier) methods (like template matching) provide when tested against old-fashioned, basic fare (like linear regression to predict a particular behavior in a given situation from any cross-validated items shown to have predicted it before). It may be less exciting, but often it works better (e.g., Tellegen, Kamp, & Watson, 1982). Complex weighting procedures, as in the Bem–Funder template-matching technique, may inadvertently serve to compound error

[7] Our analysis and comments in the present article are restricted to Bem and Funder's (1978) first study, the only one that addressed the consistency issue, not to their efforts to use social-psychological theories to generate predictors about individual differences in the relevant social situations.

(by weighting chance findings) and thus hurt rather than help the enterprise. For example, reviewing efforts to compare a number of regression equations that varied in complexity from linear to higher order, Wiggins (1981) noted,

> As one might expect, there was a tendency for increased predictiveness to be associated with increased complexity of prediction models—in the sample from which predictor weights were derived. When these same equations were applied to an independent sample, however, there was a tendency for decreased predictiveness to be associated with increased complexity of models. In other words, cross-validation appears to wipe out any predictive gains that are apparent in a derivation sample. Perhaps there is a general principle here that should discourage psychologists from overfitting their data with complex equations. (p. 6)

In sum, the Bem–Funder attempt to resolve the consistency issue held out an exciting prospect. We have analyzed it in detail to try to untangle the promise from the outcome. The prospect of a parallel language and methodology for the study of persons and situations remains attractive, but Bem and Funder's efforts toward a solution of the consistency issue in the delay domain proved disappointing both in its method and in its results. When the sample is adequate, one finds the typical, replicable, low-level, theory-consistent Q-sort correlates of delay behavior. No advance is made by use of their methods to show replicable distinctive portraits, to demonstrate improved cross-situational consistency in the domain, or to explain its phenotypic absence. The promise that the Bem–Funder technique would allow demonstrations of impressive cross-situational consistency by identifying measures that have similar rather than distinctive Q-sort correlates still awaits realization. We are pessimistic about the prospect for this effort in its present form not only because of the data we presented but because of the theoretical considerations discussed and the fate of relevant efforts attempted repeatedly in the past.

On Predicting Some of the People Some of the Time: Bem and Allen's (1974) "Idiographic" Solution

In the third approach to the consistency issue to be discussed here, Bem and Allen (1974) argued that the low consistency coefficients so typically found in research reflect the nomothetic fallacy of assuming that all traits are relevant to all people. Instead, they urged adopting an "idiographic stance," studying only the subset of people for whom a given trait is relevant.[8] For this purpose, Bem and Allen separated subjects into high- and low-variability groups, using subjects' self-reports about their variability and behavior. They then tried to document that low-variability subjects are in fact more consistent than high-variability subjects on two traits: friendliness and conscientiousness.

To assess cross-situational consistency, Bem and Allen used measures consisting mostly of global ratings of the subjects made by the subject, a peer, and the subject's parents. In addition, on each trait dimension, several composited behavior measures were employed. Careful inspection of their correlational matrices reveals good support for their predictions on the rating data. For example, for subjects classified as low rather than high variability, raters agreed much more about the subject's level of conscientiousness. But, whereas the technique nicely identifies people for whom the correlations among the global ratings will be substantial (i.e., people about whose trait level raters agree), the results are tenuous when behavior measures are intercorrelated. In fact, on those few measures directly relevant to the issue of cross-situational consistency of behavior, only one correlation was higher for subjects rated a priori as "low variability." Thus, in their analysis of friendliness, the

[8] Bem and Allen presented their work as "idiographic," and it is widely cited as exemplifying the power of idiographic methodologies (e.g., Kenrick & Stringfield, 1980). There is a confusion of terminology here, however. Idiographic usually refers to the unique organization of traits *within* individuals (Allport, 1937), not to the fact that not all characteristics may be relevant to all people. Because Bem and Allen's approach does not speak to within-person trait organization, it seems a misnomer to label it idiographic. Rather, their approach rests on the assumption that a given trait dimension may simply not be relevant for some people, and such irrelevance may be identified by selecting those subjects who rate themselves as highly variable on that dimension. Instead of bearing on the idiographic–nomothetic distinction, this approach, as noted elsewhere, is an instance of the classic moderator-variable strategy (Tellegen, Kamp, & Watson, 1982).

Table 5
Interrater Agreement About Trait Standing for Low- and High-Variability Groups Determined by Self-Report (Below Diagonal) and Ipsatized Variance (Above Diagonal) for Conscientiousness

Measure	Self-report	Mother's report	Father's report	Peer's report
Self-report		.47 (.29)	.51 (.52)	.54 (.50)
Mother's report	.56 (.20)		.67 (.45)	.75 (.35)
Father's report	.74 (.28)	.76 (.29)		.44 (.24)
Peer's report	.66 (.39)	.71 (.29)	.66 (−.12)	

Note. Agreement ratings for high-variability subjects are in parentheses.

correlation between "spontaneous friendliness" and "group discussion friendliness" was .73 for the low-variability group, whereas the same correlation for the high-variability group was .30. However, none of the three low-variability versus high-variability comparisons among behavior measures of conscientiousness fell in the predicted direction. It would seem premature to offer conclusions about the efficacy of the Bem and Allen approach as it speaks to the issue of behavioral consistency on the basis of a single confirming comparison (see Lutsky, Peake, & Wray, Note 3).

In view of the weight given to the Bem and Allen (1974) data in current theorizing about the consistency issue, it seemed important to try to replicate their work as carefully as possible, extending the number and types of behavior measures obtained. For this purpose, their Cross-Situational Behavior Survey (CSBS) was administered to the 63 subjects at Carleton College. Similar ratings of the students were obtained from their parents and from a close friend, using a modified CSBS. Subjects were divided into high- and low-variability groups for both traits on the basis of their self-reported variability and on the basis of the ipsatized variance index—the two techniques proposed by Bem and Allen. For convenience, as well as continuity with the rest of our presentation, we will summarize the findings only for the conscientiousness domain. (Detailed analyses and discussion of these data for both traits are in Peake & Lutsky, Note 4.)

As already noted, the bulk of Bem and Allen's data consisted of interrater agreement across CSBS trait ratings. Table 5 shows that on these measures, their results are nicely replicated regardless of the classification procedure used. Raters agreed more about subjects classified as low variability by either of the Bem–Allen techniques. Using the self-reported variability procedure, the mean intercorrelation for low-variability subjects was .68 compared to .22 for high-variability subjects. The comparable mean coefficients using the ipsatized variance index were .56 and .39 for the low- and high-variability groups, respectively.

These findings replicate Bem and Allen's for those measures that worked best for them, providing support that their techniques for classifying low- versus high-variability people allow one to select those individuals for whom raters will tend to agree when making global personality judgments. But more relevant to the issues of *behavioral* consistency are the cross-situational behavior data from the Carleton project summarized in previous sections on the applications of the reliability solution to the conscientiousness data. Noting the .13 mean consistency coefficient obtained in the Carleton College data after aggregating over occasions, Bem and Allen might reasonably argue that even perfect reliability will be of little value as long as researchers proceed with the nomothetic assumption that all traits belong to all individuals. Rather, now that adequate reliability is established, the search for consistency must adopt the idiographic stance and must select for study only that subset of individuals for whom the particular trait is relevant: Greater consistency should be obtained, but only for those people who are identified as low variability on the trait.

To explore this possibility, separate correlation matrices were generated for the 19 measures of conscientiousness for both the high- and low-variability subgroups identified with the Bem and Allen classification procedures. The summary coefficients in Table 6 suggest that the Bem–Allen classification procedures and approach provide no appreciable gain over the traditional yield when one turns from data based on interrater agreements about the subject's conscientiousness to more direct measures of cross-situational consistency in the referent behaviors. Thus, the mean cross-situational consistency coefficients for the low- versus high-variability subgroups are .11 versus .14 using the self-reported variability index and .15 versus .10 using the ipsatized index. Although Bem and Allen subtitled their article "The Search for Cross-Situational Consistencies in Behavior," it is precisely in this search that their technique fails to meet its promise.

What do these data imply? First, note that the current results parallel Bem and Allen's quite closely. In both studies, interrater agreement about the subjects' conscientiousness was substantial for those students classified as low variability, but this agreement was not reflected in substantially higher cross-situational consistency in their observed behavior. We are left then with a replicated paradox: Subjects classified as low variability (consistent) tend to show high levels of interrater agreement when rated on personality indexes by relevant others (intuitively implying some consistency), yet they do not show appreciably higher levels of cross-situational consistency in behavior—the very data on which their variability judgments were presumably based. We believe that this "replicable paradox" is a key component of the puzzle that needs to be solved to help untangle the consistency problem.

Table 6
Overall Mean Correlations by Variability Classification for Correlations Reflecting Interrater Agreements and Cross-Situational Consistency

	Type of correlations	
Type of Classification	Interrater agreement	Cross-situational consistency
Nomothetic (all subjects)	.52	.13
Low variability (high variability)		
Self-reported	.68 (.22)	.11 (.14)
Ipsative	.56 (.39)	.15 (.10)

Note. Correlations for high-variability groups are in parentheses.

Toward a Resolution of the Consistency Paradox

Reviewing the hoary history of the consistency paradox, Bem ánd Allen (1974) "appreciate the sense of déjâ vu that must currently be affecting psychology's elder statesmen now that the 'consistency problem' has suddenly been rediscovered" (p. 507). Déjà vu may be experienced not only at the rediscovery of the consistency paradox but also at the contemporary proposals for its solution. In one sense, this feeling of familiarity has a comfortable side: It is reassurring to see continuity in the fundamental questions and struggles for progress in the field of personality. Each of the "better methods" proposals has a long history but also a major lesson to teach, one that is too often forgotten. But if the study of personality is to be a cumulative enterprise in which knowledge and insight are not merely recycled, then these lessons must be distinguished from premature conclusions about their theoretical implications. In our view, the consistency paradox may now be at the brink of a resolution, if the relevant lessons of our field are properly read and integrated with a theoretical reconceptualization for which the outlines are already available.

The State of the Paradox

Consider again the proposed solutions reviewed here. First, aggregation over occasions, as advocated by Epstein, is a necessary step that will surely increase the reliability of measures, as the Spearman-Brown Formula has long predicted. Further aggregation, across response forms and especially across situations, will serve to highlight stable mean levels of behavior by eliminating the variability due to contexts. We do not doubt the occurrence of stable means, but we are equally impressed by the occurrence of substantial variance around those means. Sampling behavior extensively in a domain often allows useful predictions of individuals' aggregated mean levels of behavior in that domain. The fact that relevant past behavior is often the best predictor of future behavior is not in doubt (Mischel, 1968). Just as the occurrence of stable mean levels of behavior in a domain does not deny within-person variability, so should the appreciation of such variability not be mistakenly read to preempt the existence of stable personal qualities.

But as numerous analyses (including ours) have shown, such increases in reliability of behavior measures within specific situations, rather than assuring impressive increases in the associations between them, suggest again that the discriminativeness of behavior is a valid phenomenon rather than a reflection of poor methodology. It is old wisdom that the prediction of single acts is as difficult and generally as unlikely as it is challenging. (Yet it would hardly be of trivial interest to be able to predict single acts like suicide, coronary death, and homicide.) Usually we must settle for trying to predict an average of repeated observations, to infer a tendency within a given situation. With that goal, one samples the behavior of interest multiply in the situation, as we did in the reported Carleton project analyses. Such sampling of repeated observations within a given setting should not be confused, however, with aggregation across situations in the search for cross-situational consistency. To aggregate cross situationally is to circumvent the issue rather than to demonstrate its resolution. The risk is that such aggregation may tempt one to substitute self-evident psychometrics (which magnify even trivial consistencies by eliminating situational variance) for more complex (and less obvious) psychological analyses of the nature of perceived similarities and appropriate equivalence groupings in the organization of behavior and the construction of personality.

We share Bem and Funder's (1978) desire for a language that can be used to describe both persons and situations commensurately. But we doubt the viability of their approach to this goal as they applied it to their search for consistency in the domain of delay of gratification. Searching to identify the personality of situations and unraveling the structure of person–situation interactions is surely a fundamental problem shared by our field. The long and rich history in the Meehl cookbook tradition shows that understanding situations will require more than identifying the correlates of performance within them. In our view, an adequate resolution of the many issues raised in the pursuit of person–situation interaction will require a theoretical reconceptualization of both personality and situation constructs themselves, not just more clever methods for applying everyday trait terms to people's behavior in particular contexts. We believe that such a reconceptualization will unify the analysis of person characteristics with the analysis of cognitive-learning processes and requires that the person and the situation be analyzed in light of the same psychological principles and not merely described with the same trait terms (e.g., Mischel, 1973). It also requires a deeper analysis of the nature of person and situation categories (Cantor et al., 1982a, 1982b; Cantor & Mischel, 1979). In the absence of an appropriate theoretical framework, the search for consistencies can become an ultimately uninteresting hunt for statistically significant coefficients that neglects their psychological significance and their links to psychologically interesting processes.

Finally, our attempts to identify a subset of individuals for whom conscientiousness is relevant, and who will thus show appreciably more consistency across contexts, guided by Bem and Allen (1974), led to a problematic conclusion. Like Bem and Allen, we found clear support that raters agree well with each other about people who see themselves as generally consistent with regard to the particular dimension. Conversely, raters agree much less about the attributes of people who view themselves as highly variable on the relevant dimension. Less obvious, but more challenging theoretically, is the finding that people's global perceptions of their own overall consistency or variability on a dimension do not appear to be closely related to the observed cross-situational consistency of their behavior. Although interjudge agreement was greater for people who see themselves overall as consistent in conscientiousness, cross-situational consistency in their behavior was not significantly greater than it was for those who see themselves as variable or for the entire group as a whole. This pattern occurred in the Bem–Allen data as well as in the Carleton data.

From Paradox to Paradox?

Our pursuit of cross-situational consistency in behavior through the use of the better methods proposals has brought us full circle. We began by noting the paradox that exists between our intuitions of consistency in behavior on the one hand and research that documents specificity on the other. We reviewed and examined the utility of three of the most popular methodological refinements that have been proposed as possible solutions to the consistency paradox. The end result of our conceptual and empirical endeavors is another paradox that closely parallels the one we set out to resolve. The results of the replication of the Bem and Allen work suggests that raters agree substantially more about persons who identify themselves as cross-situationally consistent. However, these individuals do not show substantially greater cross-situational consistency in behavior than people identified as more variable. Here, again, shared intuitions about persons do not agree with the data.

We might account for these results in a variety of alternative ways. On the one hand, one might suggest that the replicable paradox results from methodological problems commonly associated with the use of behavioral data and dismiss the behavioral results, resting the case for personality structure on the impressive findings among the rating data (e.g., Block, 1977). Alternatively, one might argue that the behavioral data accurately reflect the complex structure of behavior and that the substantial interrater agreements reflect shared theories about persons and other heuristics that bias our judgments about the coherences that actually exist in the behavior of others (e.g., Chapman & Chapman, 1969; Mischel, 1968, 1979; Nisbett & Ross, 1980; Ross, 1977; Schneider, 1973; Shweder & D'Andrade, 1980). Both of these interpretations have some merit. Nevertheless, granting both the methodological problems of behavioral data and the existence of cognitive economics, our perceptions of others are still unlikely to be entirely illusory. They may derive rather directly from the behavior of the individual but not from those aspects that we expect or to which the consistency debate has pointed so far.

Components for a Reconceptualization

Our approach to understanding the consistency paradox is guided simultaneously by a cognitive social-learning conceptualization of behavior organization (e.g., Mischel, 1973) and a cognitive prototype view of person categorization (e.g., Cantor & Mischel, 1979). First, consider the nature of the regularities revealed from the study of behavioral consistency. We read these data as repeatedly showing temporal stability more impressively than cross-situational consistency. Greater temporal than cross-situational consistency seems sensible from the perspective of cognitive social-learning theory. Because the contingencies in a given situation often remain unchanged over time, stability over time is expected and predicted in much social behavior (e.g., Bandura, 1969; Mischel, 1968). Moreover, from this perspective, temporal stability would be expected to the degree that such qualities as the person's competencies, encodings, expectancies, values, and plans endure (Mischel, 1973). The pursuit of durable values and goals with stable skills and expectations for long periods of time surely involves coherent and meaningful patternings among the individual's efforts and enterprises. The degree of cross-situational consistency, however, might be high, low, or intermediate, depending on many considrations, including the structure of the perceived cross-situational contingencies and the subjective equivalences among the diverse situations sampled. Distinctive contingencies may be expected to occur even in slightly different situations, producing high discriminativeness cross situationally. If so, if cross-context behavioral discriminativeness is the rule rather than the exception—the phenomenon rather than the error—then the search for consistency across situations will continue to yield slim results.

But then how can we understand the other side of the consistency paradox, the intuitive conviction of consistency? Our attempts to employ Bem and Allen's idiographic solution to the Carleton College data showed that people who see themselves as consistent on a dimension are indeed rated with greater interjudge agreement by others even though (and this is the key point) their behavior does not necessarily show appreciably greater overall cross-situational consistency. What are the bases—the ingredients—of the seemingly pervasive and shared perception of consistency in a personality disposition if the perception is not related to the level of cross-situational consistency in the reliably observed referent behaviors? To try to answer this question, we turn to the cognitive prototype approach (e.g., Cantor & Mischel, 1979). Guided by cognitive theories of the categorization of everyday objects (e.g., Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976), this prototype approach to the categorization problem appreciates the reality of individual differences but seeks to reconceptualize the nature of the consistencies they reflect in an interactional framework (Cantor et al., 1982a, 1982b; Cantor & Mischel, 1979). The prototype approach recognizes the especially fuzzy nature of natural categories and, along the lines first traced by Wittgenstein (1953), searches not for any single set of features shared by all members of a category but rather for a family-resemblance structure, a pattern of overlapping similarities. The recognition of fuzzy sets also suggests that categorization decision will be probabilistic, with many ambiguous borderline cases that produce overlapping, fuzzy boundaries between the categories, and that members of a category will vary in degree of membership (prototypicality).

The cognitive prototype approach applied to the consistency paradox suggests that consistency judgments with respect to a category are made not by seeking the average of all the observable features of a category but by noting the reliable occurrence of some features that are central to the category, or more prototypic. That is, we suggest that consistency judgments rely heavily on the observation of central (prototypic) features so that the impression of consistency will derive not from average levels of consistency across all the possible features of the category but rather from the observation that some central features are reliably (stably) present. From this perspective, extensive cross-situational consistency may not be a basic ingredient for either the organization of personal consistency in a domain or for its perception.

The Construction of Consistency

In accord with the cognitive prototype view, we propose that the shared global impression of trait consistency (in self and in others) arises not primarily from the observation of cross-situational consistency in relevant behaviors. Rather, we propose that to assess variability (versus consistency) with regard to a category of behavior people scan the temporal stability of a limited number of behaviors that are most relevant (central or prototypic) to that category for them. Thus, the impression of consistency is based substantially on the observation of temporal stability in those behaviors that are highly relevant (central) to the prototype but is independent of the temporal stability of behaviors that are not highly relevant to the prototype. Conversely, the perception of variability arises from the observation of temporal instability in highly relevant features.

To explore this hypothesis in the Carleton College data on conscientiousness, we predicted that people who judge themselves overall as consistent across situations (low variability on Bem and Allen's, 1974, global self-report measure) will show greater temporal stability but not greater cross-situational consistency than those who view themselves as less consistent.[9] In addition, because we believe that the judgment of consistency is independent of the temporal stability of behaviors that are not highly prototypical, we predicted this difference in temporal stability to be more pronounced on the more prototypic features of conscientiousness than on the less prototypic features.

[9] The measure of global self-perceived consistency was the subject's answer to the question (on a 0–6 scale), "How much do you vary from situation to situation in how conscientious you are about daily matters and responsibilities?"

To test these hypotheses, temporal-stability coefficients were obtained separately on each of the behavioral measures employed at Carleton for subjects who rated themselves as high (versus low) in variability. Subjects who perceived themselves as highly consistent rather than as more variable across situations had somewhat higher temporal stability across all the behavior measures. The mean temporal-stability coefficients were .68 and .55 for those high versus low in self-perceived consistency, respectively, $t(32) = 1.82$, $p < .10$ (two-tailed). There were no appreciable differences in the behavioral cross-situational consistency of those who saw themselves as high ($r = .11$) or low ($r = .14$) in consistency.

Most interesting, and central to our hypothesis, is the linkage between the global self-perception of consistency and the temporal stability of more prototypic behaviors. Ratings of prototypicality were available for 17 of the 19 Carleton behavior measures and allowed us to divide these measures into the "more" and the "less" prototypical (at the median of the total ratings for all items).[10] Table 7 presents the links between the global self-perception of consistency and the behavioral data, divided into more prototypic versus less prototypic behaviors. The pattern of results was exactly as expected by the hypothesis. First, consider the more prototypic behaviors. Students who saw themselves as highly consistent in conscientiousness were significantly more temporally stable on these prototypic behaviors than those who viewed themselves as more variable (low variability, mean $r = .71$; high variability, mean $r = .47$), $t(15) = 2.97$, $p < .025$ (two-tailed).[11] This mean difference is reflected pervasively in the component behaviors. On seven of the nine comparisons between prototypic behaviors computed, subjects who perceived themselves as highly variable showed significantly less temporal stability (with no appreciable difference on the other two comparisons). In contrast to the clear and consistent differences in temporal stability of prototypic behaviors, there was no difference between the self-perceived low- and high-variability groups in the mean temporal stability for the less prototypic behaviors ($r = .64$ and .65, respectively). The two groups differed significantly on only one of the eight less prototypic behaviors (with low-variable subjects showing greater stability), and there were no differences approaching significance on any of the other comparisons. Finally, as expected, there was no relation between self-perceived consistency and behavioral cross-situational consistency, regardless of the prototypicality of the behaviors.

Table 7
Links Between Self-Perceived Consistency and Behavior

	Self-perceived consistency	
Behavioral data	Low variability	High variability
Cross-situational consistency		
More prototypical	.15	.13
Less prototypical	.09	.14
Temporal stability		
More prototypical	.71	.47
Less prototypical	.65	.64

Note. The coefficients reported are mean correlation coefficients across all possible comparisons of the designated type.

[10] The prototypicality of the behavioral measures was assessed by having subjects rate each of the 47 conscientiousness items on the original Cross-Situational Behavior Survey (CSBS) questionnaire for that item's relevance to "conscientiousness" using a 0–6 scale where 0 signifies "not at all relevant" and 6 signifies "extremely relevant." Mean relevance (prototypicality) ratings were computed for each of the 47 items. From this larger group of ratings it was possible to abstract the prototypicality ratings of 17 of the 19 behavioral measures employed in the study. These 17 measures were then rank ordered according to their degree of prototypicality and divided into two groups (more prototypical and less prototypical), depending on whether they fell above or below the median level of prototypicality derived from the complete list of 47 items.

[11] Prototypicality ratings for all 19 behaviors subsequently were obtained from a second questionnaire given to an independent sample of 40 Carleton students. They rated on a 7-point scale how well each behavioral referent exemplifies the category. Analyses of these ratings that parallel the analyses just described provide even stronger support for the hypothesis. For the more prototypic behaviors, the mean temporal stability coefficients of the self-perceived low-variability versus high-variability groups were .68 and .31, respectively, $t(17) = 3.33$, $p < .01$. Parallel analyses for the friendliness domain yielded similar results in support of the hypothesis and are described in Peake (Note 1).

These results support the view that the impression of consistency in behavior may be rooted in temporally stable prototypic behaviors rather than in pervasive overall cross-situational consistencies. The findings suggest that individuals judge their degree of consistency from the temporal stability of the relevant, more prototypic behaviors. Interestingly, the two groups do not differ in temporal stability on the less prototypic behaviors, suggesting that those behaviors do not enter into the judgment of one's variability. It seems then that the locus of the perception of variability may be in the temporal stability of highly prototypic behaviors, regardless of cross-situational consistency. A tendency to overgeneralize from the observation of temporal stability in prototypic features to an impression of overall consistency would certainly be congruent with other tendencies to go well beyond observations in social inferences and attributions (e.g., Mischel, 1979; Nisbett, 1980; Nisbett & Ross, 1980; Ross, 1977; Tversky & Kahneman, 1974).

The consistency debate has been aptly characterized as reflecting a continuous conflict between the findings of research and the convictions of our intuitions (Bem & Allen, 1974, p. 508). After reviewing the issues and data on the debate, Bem and Allen concluded that "in terms of the underlying logic and fidelity to reality, we believe that our intuitions are right; the research, wrong" (p. 510). We hope that our present analysis helps to identify some of the roots of the conflict and the routes toward its resolution. We believe that both the intuitions and the research have validity, but they are based on different data. The intuitions of cross-situational consistency are grounded in data, but these data, we suggest, are not highly generalized cross-situational consistencies in behavior: Rather, we propose the intuitions about a person's consistency arise from the observation of temporal stability in prototypical behaviors. The error is to confuse the temporal stability of key behaviors or central features with pervasive cross-situational consistency and then to overestimate the latter, a common tendency hardly confined to the layperson. Our compelling intuitions are based on consistencies in behavior, but perhaps not on the consistencies that the debate pursued for so many years.

The consistency paradox has been a puzzling and persistent barrier in the search for personality structure. But a theory of personality structure does not require everyone to be characterized by high levels of pervasive cross-situational consistency in behavior. It does require a structure for behavior: The Carleton data suggest to us that such structure may be rooted in the occurrence of temporally stable but cross-situationally discriminative features that are prototypic for the particular behavior category as perceived by the particular person. A close analysis of the patterning and organization of such features within individuals should be most interesting, and we plan such an analysis. We expect that the most consistent and prototypic exemplars of a category like conscientiousness will be those individuals who stably exhibt a number—but not necessarily many—of its prototypic features, as they themselves define that prototypicality. We expect that the particular constellations of features will be idiographically patterned so that no individuals necessarily share the identical configuration, although considerable between-person overlap occurs. Although the exact pattern that defines conscientiousness may not be identical for any two persons, each individual who is characterized as consistently conscientious will display some of its features with temporal stability, albeit with a distinctive cross-situational constellation.[12] If so, the route may be open not only for seeing the uniqueness of each personality (which personologists have long appreciated) but also for ultimately understanding its common structure.

Conclusions

It is tempting to tire of the consistency debate, to trivialize it by focusing on its ob-

[12] If these hypotheses prove valid, they also would help one understand why the issue of cross-situational consistency is a more serious problem for the nomothetic trait psychologist than for the layperson. To the degree that each individual maintains reasonable temporal stability in his or her distinctive pattern of prototypic behaviors, the impression of pervasive consistency may be preserved. Individuals thus may be perceived as consistent on their particular set of stable behaviors even if these behaviors do not map very well onto the total set of referents selected by the trait researcher to define his or her dimension for all people.

vious qualities. Surely each person's behavior shows some consistency, some discriminativeness, some continuity, some change. But this tempting stance has nontrivial consequences, diverting attention from the serious questions that have been raised thoughout the debate by evidence of unexpected discriminativeness in behavior as a function of context and the psychological situation. Unfortunately, the debate has been difficult to resolve not only because the phenomena are exceedingly complex but because the data available are often readily misread and the interpretations far exceed their source. The present analysis has attempted to clarify some of the basic issues and to assess some of the proposals for progress while also offering an alternative direction. The data we have presented, and our analysis of the relevant research by others, suggest the following conclusions:

1. An extensive assessment of conscientiousness in college students showed that even with reliable measures, based on multiple observations of behavior aggregated over occasions, or aggregated further over response modes within situations, mean cross-situational consistency coefficients were of modest magnitude (on average not exceeding about r = .20). In contrast, impressive temporal stability was found (average r = .65) in the same data. Although aggregation of measures over occasions is a useful step in establishing reliability, aggregation of measures over situations may bypass rather than resolve the problem of cross-situational consistency. The demonstration of impressive temporal stability, but of only modest cross-situational consistency, was seen as congruent with extensive other research and with predictions from cognitive social-learning theory.

2. The Bem–Funder (1978) template-matching approach did not enhance the search for cross-situational consistency either in their original data or in our extended replication. These results are not surprising given the history of earlier approaches that shared similar logic and method (e.g., cookbooks for the MMPI). The prospect of a commensurate language for describing persons and situations remains attractive but cannot be achieved by merely describing performance correlates in particular situations. A potential alternative is a prototype analysis in which both persons and situations are characterized by the same theory-informed methods (e.g., Cantor et al., 1982a, 1982b; Cantor & Mischel, 1979).

3. The Bem–Allen (1974) moderator-variable approach successfully identifies people who see themselves as consistent (low rather than high variability) on a dimension: Other people, in turn, are more likely to agree with each other and with the low-variability subjects about their status on the dimension. Global self-report judgments of one's consistency, however, do not seem to be strongly linked to cross-situational consistency in the referent behaviors even when reliably measured. Evidence for greater cross-situational consistency in behavior in "some of the people some of the time" was not found either in the Bem and Allen data or in our study of conscientiousness.

4. Congruent with a cognitive prototype approach, we propose that the judgment of trait consistency is strongly related to the temporal stability of highly prototypic (but not of less prototypic) behaviors. In contrast, the global impression of consistency may not be strongly related to overall or average cross-situational consistency, even in prototypic behaviors. Thus, the perception and organization of personality consistencies, we suggest, will depend more on the temporal stability of key features than on the observation of cross-situational behavioral consistency, and the former may be easily interpreted as if it were the latter. Results supporting these expectations were presented.

Before firm generalizations can be reached from the results of the Carleton study just summarized, we will need thorough replications and extensions to assure the robustness and breadth of the conclusions. Even with these strong reservations in mind, the pattern already obtained seems promising, pointing ultimately, we hope, to a structural solution for an enduring dilemma in our field. The consistency paradox may be paradoxical only because we have been looking for consistency in the wrong place. If our shared perceptions of consistent personality attributes are indeed rooted in the observation of temporally stable behavioral features that are prototypic for the particular attribute, the paradox may well be on the way to resolution. Instead of seeking high levels of cross-situational consistency—instead of looking for broad averages—we may need, instead, to identify unique bundles or sets of temporally stable prototypic behaviors—key features—that characterize the person even over long periods of time but not necessarily across many or all possibly relevant situations.

Reference Notes

1. Peake, P. K. *Searching for consistency: The Carleton student behavior study.* Unpublished doctoral dissertation, Stanford University, 1982.
2. Hoffman, C., & Bem, D. J. *Contextual template matching: A progress report on predicting all of the people all of the time.* Unpublished manuscript, University of Alberta, Alberta, Canada, 1981.
3. Lutsky, N., Peake, P. K., & Wray, L. *Inconsistencies in the search for cross-situational consistencies in behavior: A critique of the Bem and Allen study.* Paper presented at the meeting of the Midwestern Psychological Association, Chicago, 1978.
4. Peake, P. K., & Lutsky, N. S. *On predicting "sums" of the person, "sums" of the time.* Manuscript in preparation, Stanford University, 1982.

References

Allport, G. W. *Personality: A psychological interpretation.* New York: Holt, Rinehart & Winston, 1937.

Allport, G. W. Traits revisited. *American Psychologist,* 1966, *21,* 1–10.

Allport, G. W., & Vernon, P. E. *Studies in expressive movement.* New York: Macmillan, 1933.

Bandura, A. *Principles of behavior modification.* New York: Holt, Rinehart & Winston, 1969.

Bandura, A. Reflections on self-efficacy. In S. Rachman (Ed.), *Advances in behaviour research and therapy* (Vol. 1). Oxford, England: Pergamon Press, 1978.

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

Bem, D. J., & Funder, D. C. Predicting more of the people more of the time: Assessing the personality of situations. *Psychological Review,* 1978, *85,* 485–501.

Block, J. Advancing the psychology of personality: Paradigmatic shift or improving the quality of research. In D. Magnusson & N. S. Endler (Eds.), *Personality at the crossroads: Current issues in interactional psychology.* Hillsdale, N.J.: Erlbaum, 1977.

Byrne, D. Repression-sensitization as a dimension of personality. In B. A. Maher (Ed.), *Progress in experimental personality research* (Vol. 1). New York: Academic Press, 1964.

Cantor, N., & Mischel, W. Prototypes in person perception. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 12). New York: Academic Press, 1979.

Cantor, N., Mischel, W., & Schwartz, J. A prototype analysis of psychological situations. *Cognitive Psychology,* 1982, *14,* 45–77. (a)

Cantor, N., Mischel, W., & Schwartz, J. Social knowledge: Structure, content, use and abuse. In A. Hastorf & A. Isen (Eds.), *Cognitive social psychology.* New York: Elsevier North-Holland, 1982. (b)

Carlson, R. Where is the person in personality research? *Psychological Bulletin,* 1971, *75,* 203–219.

Chapman, L. J., & Chapman, J. P. Illusory correlations as an obstacle to the use of valid psycho-diagnostic signs. *Journal of Abnormal Psychology,* 1969, *74,* 271–280.

Cronbach, L. J., & Meehl, P. E. Construct validity in psychological tests. *Psychological Bulletin,* 1955, *52,* 281–302.

Dudycha, G. J. An objective study of punctuality in relation to personality and achievement. *Archives of Psychology,* 1936, *29,* 1–53.

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

Epstein, S. The stability of behavior: II. Implications for psychological research. *American Psychologist,* 1980, *35,* 790–806.

Fishbein, M., & Ajzen, I. *Belief, attitude, intention and behavior: An introduction to theory and research.* Reading, Mass.: Addison-Wesley, 1975.

Fiske, D. W. The inherent variability of behavior. In D. W. Fiske & S. R. Maddi (Eds.), *Functions of varied experience.* Homewood, Ill.: Dorsey Press, 1961.

Funder, D. C. *The person-situation interaction in attitude change.* Unpublished doctoral dissertation, Standford University, 1979.

Goldberg, L. R. Diagnosticians vs. diagnostic signs: The diagnosis of psychosis vs. neurosis from the MMPI. *Psychological Monographs,* 1965, *79* (9, Whole No. 602).

Gulliksen, H. *Theory of mental tests.* New York: Wiley, 1950.

Hartshorne, H., & May, M. A. *Studies in deceit.* New York: Macmillan, 1928.

Horowitz, L. M., Inouye, D., & Siegelman, E. Y. On averaging judges' ratings to increase their correlation with an external criterion. *Journal of Consulting and Clinical Psychology,* 1979, *47,* 453–458.

Jaccard, J. J. Predicting social behavior from personality traits. *Journal of Research in Personality,* 1974, *7,* 358–367.

Kenrick, D. T., & Stringfield, D. O. Personality traits and the eye of the beholder: Crossing some traditional philosophical boundaries in the search for consistency in all of the people. *Psychological Review,* 1980, *87,* 88–104.

Lord, C. G. Predicting behavioral consistency from an individual's perception of situational similarities. *Journal of Personality and Social Psychology,* 1982, *42,* 1076–1088.

Lord, F. M., & Novick, M. R. *Statistical theories of mental test scores.* Reading, Mass.: Addison-Wesley, 1968.

Magnusson, D., & Ekehammar, B. An analysis of situational dimensions: A replication. *Multivariate Behavioral Research,* 1973, *8,* 331–339.

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

McGowan, J., & Gormly, J. Validation of personality traits: A multicriteria approach. *Journal of Personality and Social Psychology*, 1976, *34*, 791–795.

Meehl, P. E. Wanted—A good cookbook. *American Psychologist*, 1956, *11*, 263–272.

Mischel, W. Theory and research on the antecedents of self-imposed delay of reward. In B. A. Maher (Ed.), *Progress in experimental personality research* (Vol. 3). New York: Academic Press, 1966.

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

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

Mischel, W. Processes in delay of gratification. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 7). New York: Academic Press, 1974.

Mischel, W. On the future of personality measurement. *American Psychologist*, 1977, *32*, 246–254.

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

Mischel, W. *Introduction to Personality*, (3rd ed.). New York: Holt, Rinehart & Winston, 1981.

Mischel, W., & Ebbesen, E. B. Attention in delay of gratification. *Journal of Personality and Social Psychology*, 1970, *16*, 329–337.

Mischel, W., Ebbesen, E. B., & Zeiss, A. R. Selective attention to the self: Situational and dispositional determinants. *Journal of Personality and Social Psychology*, 1973, *27*, 129–142.

Moos, R. H., & Fuhr, R. The clinical use of social-ecological concepts: The case of an adolescent girl. *American Journal of Orthopsychiatry*, 1982, *52*, 111–122.

Newcomb, T. M. *Consistency of certain extrovert–introvert behavior patterns in 51 problem boys*. New York: Columbia University, Teachers College, Bureau of Publications, 1929.

Nisbett, R. E. The trait construct in lay and professional psychology. In L. Festinger, (Ed.), *Retrospections on social psychology*. New York: Oxford University Press, 1980.

Nisbett, R. E., & Ross, L. D. *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, N.J.: Prentice-Hall, 1980.

Olweus, D. A critical analysis of the "modern" interactionist position. In D. Magnusson & N. S. Endler (Eds.), *Personality at the crossroads: Current issues in interactional psychology*. Hillsdale, N.J.: Erlbaum, 1977.

Patterson, G. R. The aggressive child: Victim and architect of a coercive system. In L. A. Hamerlynck, L. C. Handy, & E. J. Mash (Eds.), *Behavior modification and families: 1. Theory and research*. New York: Brunner/Mazel, 1976.

Patterson, G. R., & Moore, D. R. Interactive patterns as units. In M. Lamb, S. Suomi, & G. Stephenson (Eds.), *Methodological problems in the study of social interaction*. Madison: University of Wisconsin Press, 1979.

Peterson, D. R. *The clinical study of social behavior*. New York: Appleton-Century-Crofts, 1968.

Raush, H. L. Paradox levels, and junctures in person-situation systems. In D. Magnusson & N. S. Endler (Eds.), *Personality at the crossroads: Current issues in interactional psychology*. Hillsdale, N.J.: Erlbaum, 1977.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. Basic objects in natural categories. *Cognitive Psychology*, 1976, *8*, 382–439.

Ross, L. The intuitive psychologist and his shortcomings: Distortions in the attribution process. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 10). New York: Academic Press, 1977.

Rushton, J. P., Jackson, D. N., & Paunonen, S. V. Personality: Nomothetic or idiographic? A response to Kenrick and Stringfield. *Psychological Review*, 1981, *88*, 582–589.

Schneider, D. J. Implicit personality theory: A review. *Psychological Bulletin*, 1973, *79*, 294–309.

Shweder, R. A., & D'Andrade, R. G. The systematic distortion hypothesis. *New Directions for Methodology of Social and Behavioral Science*, 1980, *4*, 37–58.

Tellegen, A., Kamp, J., & Watson, D. Recognizing individual differences in predictive structures. *Psychological Review*, 1982, *89*, 95–105.

Thorndike, E. L. *Principles of teaching*. New York: Seiler, 1906.

Thurstone, L. L. *The reliability and validity of tests*. Ann Arbor, Mich.: Edwards Brothers, 1932.

Tversky, A. Features of similarity. *Psychological Review*, 1977, *84*, 327–352.

Tversky, A., & Kahneman, D. Judgment under uncertainty: Heuristics and biases. *Science*, 1974, *185*, 1124–1131.

Vernon, P. E. *Personality assessment: A critical survey*. New York: Wiley, 1964.

Wiggins, J. S. *Personality and prediction: Principles of personality assessment*. Reading, Mass.: Addison-Wesley, 1973.

Wiggins, J. S. Clinical and statistical prediction: Where are we and where do we go from here? *Clinical Psychology Review*, 1981, *1*, 3–18.

Wittgenstein, L. *Philosophical investigations*. New York: Macmillan, 1953.

Received November 4, 1981
Revision received June 16, 1982 ■