Psychological Review
1983, Vol. 90, No. 4, 394-402

Some Facets of Consistency: Replies to Epstein, Funder, and Bem

Walter Mischel
Columbia University

Philip K. Peake
Stanford University

Both convergences and continuing disagreements are emerging from current analyses of the cross-situational consistency of behavior. Whether one focuses on the overall aggregate of performance regardless of the situation (Epstein, 1983a), or on cross-situational consistencies among the component behaviors, is a choice that depends on goals and paradigm preferences. Alternative routes in the search for consistency, and the diverse goals that each can serve usefully, are discussed. The data that emerge from the long search for consistency are seen as stable; they have been interpreted quite differently, depending on vantage points, criteria, and purposes, which have changed with the different phases of the "classic debate" and with shifts in the concerns and the issues addressed. Finally, problems in the Bem and Funder (1978) and Bem and Allen (1974) searches for cross-situational consistency were reviewed. These investigations implied resolutions of the consistency problem that were not justified by their studies and were unsupported by the Mischel and Peake (1982) extensions and analyses.

The enduring controversy concerning the nature of personality traits and the consistency of behavior seems to be yielding a curious mix of convergences and disagreements (e.g., Bem, 1983; Epstein, 1983a; Funder, 1983; Mischel & Peake, 1982). For most observers of the controversy, the convergences among the participants sharpen the sense of "déjà vu"; the disagreements generate confusion—and the combination may serve to extinguish all interest in the issues. Unfortunately, this extinction hazard comes just when many researchers are becoming excited about the prospects for genuine progress in the area. Our purpose in this note is to try to avoid the extinction and to encourage the progress. For that reason, we will try to identify the convergences and any substantive differences that emerge from the comments by Epstein (1983a), Funder (1983), and Bem (1983) about our recent analysis of some issues and proposals in the search for cross-situational consistency in social behavior (Mischel & Peake, 1982).

We begin with the areas of agreement. There seems to be general consensus both about the existence of behavioral stabilities and consistencies, on the one hand, and the occurrence of cross-situational discriminativeness in behavior, on the other. The significance, nature, and ways to conceptualize and pursue both these coherences and the discriminative patterns—not the reality of either one—remain at issue and need to be our focus. It is encouraging that current work seems to be focusing more on the clarification of the psychological structure, organization, and construction of behavioral consistencies and of person–situation interactions, and less on debating their existence. Convergence is seen in the move, during the last decades, away from the extreme assumptions that global, highly generalized traits are omnipresent causal entities, that they can be inferred from almost any indicators, and that they determine most of the variance in social behavior most of the time. Simultaneously, it should be equally clear that rejection of such an extreme in no way supports the comparably extreme (and indefensible) opposite assumption that people do not behave in organized, patterned, and potentially predictable ways in particular domains, nor that the "situation" (however construed) is *the* important causal agent for everything of interest.

The Uses of Aggregation

Consensus about these points frees one to face the more challenging task: the analysis in-depth of the stabilities and coherences that do exist, as well as of the discriminations and indiosyncratic patternings that characterize the individual's interactions over time and across situations. Having agreed (we hope) about what is not at issue, we can try to focus on what is. Gordon Allport's persistent question, "What units shall we use?" is still one of the

We thank Albert Bandura, Yuichi Shoda, Brian Wandell, and Jack Wright for their helpful comments on earlier drafts.

Requests for reprints should be sent to Walter Mischel, Department of Psychology, Columbia University, Schermerhorn Hall, 116th Street and Broadway, New York, New York 10027.

field's most compelling challenges and is still unlikely to yield simplistic answers. The discovery of relatively stable psychological structures that can serve as "the units" and that have as much breadth as possible is certainly central for personality psychology. But the routes to find such structures, and the conceptualization of their nature, organization, and operation, can be approached and construed from different (indeed, almost opposite) perspectives. One traditional approach (favored by Epstein, 1979, 1983a, 1983b) seeks to aggregate behavior over occasions and situations to identify and document the stability of individual differences in the resulting aggregate units. The alternative approach seeks to decompose what people "are like" and "do" into more molecular constituent units and to explain the structure and organization of both behaviors and perceptions of people in as fine-grained a fashion as possible (e.g., Mischel, 1973, 1983). Both approaches need to be pursued vigorously within the same field; each is useful, although each addresses different questions and serves different goals.

The aggregation approach serves to demonstrate stable individual differences in average performance, using data aggregated over occasions and situations within a domain. Appropriately extensive sampling and aggregation is the classic strategy to identify and establish reliable individual differences within a behavior domain. By removing situational variance, these procedures accentuate temporally stable individual differences at a situationless level. Guided by test theory, such aggregation is highly useful; for some goals, including screening decisions and research on individual differences, it is essential (Mischel, 1968). As we noted in our original analysis, these procedures can be used to convert a mean coefficient of .13 among the 19 behavioral measures obtained by Mischel and Peake (1982) into a reliability estimate of .74. That coefficient informs the researcher about the *temporal stability* of subjects' total scores on the whole "test." It should not, however, divert attention from the relations among the components (i.e., the links from behavioral measures to behavioral measures). Such correlations may or may not be interesting, depending on the purpose of the particular research. If the goal is to describe, predict, and analyze the organization and patterning of behavior in a domain (as it often has been in the history of the controversy), relations among indexes of reliably assessed behavior from situation to situation will certainly be of interest. If the goal is to assess the individual's average performance in the domain, regardless of the context or situation, such coefficients are less informative and sometimes irrelevant. Under these conditions, the reliability coefficient favored by Epstein becomes an important index.

Avoiding Preemptive Conceptions and Solutions

The aggregation route, clearly valuable for certain purposes, becomes preemptive when one views a focus on the correlations among the components as misleading and treats psychological situations as mere "items" (Epstein, 1983a) or insists that only the aggregated total is indicative of subjects' "true" position (e.g., Rushton, Brainerd, & Pressley, 1983). A focus on the agregated total is consistent with the traditional trait theory emphasis on "typical" behavior and the search for "true scores." Unfortunately, it also easily leads one to view situational effects as "error," which is appropriate for some goals but not for others, as Mischel and Peake pointed out. Epstein calls the persistent search for consistency from one situation to another the "stability of confusion" (1983a, pp. 179, 182) and sees it as the mistake that has misled the consistency debate since its beginnings. He derogates the search for predictability and consistency in behavior from situation to situation as an aim that is simultaneously unattainable and trivial, of no more psychological interest than any single item on a paper and pencil inventory (even when each situation is reliably assessed over multiple occasions). To reiterate, whether or not that judgment is justified depends on the goal of the assessment and the theory of the assessor, which in turn dictate the units one selects and the nature of the aggregation one seeks or avoids, as we discussed in our previous paper.

The aggregation approach eliminates contexts through aggregation so that "the randomness in any one measure (error variance) is averaged out over several measures, leaving a clearer view of what a person's true behavior is like" (Rushton et al., 1983, p. 23). But why does the systematic removal of the role of situations, a treatment of the contexts of life as "error," necessarily yield a "truer" glimpse of social behavior or of how we should construe the individual? Depending on one's purpose, the within-person variance—the interactional effects of persons with the conditions of their lives—may be as much part of the "true" fabric of human behavior, to be understood and analyzed, as is the abstracted categorization of a person's average performance in relation to a comparison group on the summary score of a more or less arbitrarily created test battery. And that is one basis of our disagreement with Epstein (1979, 1983a) and with a preemptive overgeneralization of the trait approach. To paraphrase George Kelly (1955), the hazard for the trait psychologist is to leave the individual sitting on his or her—or really, the assessor's—continuum.

In our view, some of the "error" variance is random variability, reflecting merely unstable measurement due to momentary instabilities in the spe-

cific occasion. One method to reduce such error is to sample behavior repeatedly, which was attempted in the Carleton study by multiple measurement over occasions (Mischel & Peake, 1982). But some of the variance also lumped traditionally into the "error" category reflects genuine person–situation interactions that are stable over time and are psychologically interesting. For some goals, including the analysis and clarification of behavioral regularities, these interactions *are* the phenomena of interest, the data that demand attention and clarification. They require careful distillation and theory-guided psychological explanation, not gross psychometric aggregation to average out their existence, and that was one of our main points.

Epstein (1979, 1983a, 1983b) has repeatedly proposed aggregation as the basic route for the resolution of the consistency debate. In its original form, Epstein's (1979) proposal focused on aggregation over occasions (repeated measures over time); the data he presented were largely relevant to demonstrations of temporal stability, not to cross-situational consistency as defined in earlier phases of the analyses of the consistency problem (e.g., Bem, 1972; Hartshorne & May, 1928; Mischel, 1968, 1973; Newcomb, 1929). Moreover, Epstein repeatedly treats the issues of temporal stability and cross-situational consistency as if they were the same. Yet, it is cross-situational consistency, not temporal stability, that has been controversial throughout the history of this "classic debate." Mischel and Peake (1982) incorporated the Epstein proposal to aggregate over occasions to reduce error of measurement arising from momentary instability. The results we obtained, although reasonably reliable, did not yield the implied cross-situational consistency resolution. In response, Epstein (1983a) suggests that appropriate aggregation must be not only over occasions but also requires aggregating across situations (i.e., removing situational effects by averaging them out). He redefines the consistency resolution as requiring only the demonstration of reliability in the aggregate units, not consistency from situations to situations (even when the behavioral assessments of the components are reliable). No one has questioned seriously the ability to form reliable behavioral composites from even modestly interrelated component behaviors. The more challenging problem for researchers since the recognition of consistency issues at the start of the century has been to understand and to predict both the distinctive coherences and the discriminations that characterize behavior at different levels of generality and to identify the links between such patterns and the psychological processes that account for them.[1]

Far from ignoring the value of aggregation, researchers used aggregation to enhance reliability decades ago. But it did not lead them to the conclusion of dispositional breadth that Epstein suggests. Asking how well dishonesty is measured by all the behaviors they tested, Hartshorne and May noted in 1928:

> The average inter-*r* of the entire nine types is .227. Substituting this figure in the Spearman-Brown formula, we get a predicted reliability of .725, the square root of which gives a predicted validity of .851. To repeat, this simply means that, if we had nine more similar tests, . . . the correlation between the measured nine and the unmeasured nine would be .725. Also if we had an infinite number of similar types measured, the correlation between the nine we have and this infinite number (the theoretical true measure of dishonesty) would be .851. (Book 2, p. 125)

Their results anticipated those obtained with much effort by other investigators since them. Their writing showed that they had not forgotten test theory and were sensitive to the complex issues of reliability and sampling and attenuated correlations, (O'Brien, Note 1, cited in Epstein, 1983a, not withstanding). Nevertheless, Hartshorne and May's classic investigation led them to the conclusions "that honesty or dishonesty is not a unified character trait in children of the ages studied" (1928, Book 2, p. 243). They reached this interpretation not because they were confused, or ignorant of test theory and aggregation procedures, as has been implied repeatedly in the recent literature. Instead, their conclusions reflected their recognition of the children's behavioral discriminativeness across tests, as well as of their stability over aggregated measures.

Stable Data and Unstable Interpretations

Although our earlier article addressed several proposals for enhancing cross-situational consistency, we were not exhaustive and did not include the standard psychometric route that guides test construction. As we pointed out:

> The technologies of psychometrics supply us with ample methods for distilling the coherence among our measures, for accentuating the mean levels of individual differences

[1] In his current writing on the consistency problem, Epstein contends once again (1983b, p. 361) that "the debate on the stability of behavior across time and situations could have been resolved a long time ago had investigators been more aware of some basic principles of test construction." But this restatement of ostensible resolution is qualified with a crucial footnote (Footnote 2, p. 361) that reveals the sorts of qualifications and reservations we seek to reiterate. In the footnote, Epstein writes that the "important issues that still require resolution are . . . the extent to which personality is temporally and cross-situationally general for different variables and people . . ." We agree: If this is Epstein's resolution, the convergence among positions has become considerable indeed, and further debate is unnecessary.

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. (Mischel & Peake, 1982, pp. 738–739)

In these analyses, we have used several psychometric methods to enhance the coherences and the predictability in the Carleton data, and each leads to somewhat distinctive answers (e.g., Mischel, 1983; see also Peake & Lutsky, Note 2, and Shoda, Note 3). The general strategy in these analyses is to treat each specific measure of behavior in a situation (e.g., class attendance, assignment neatness) as a test item, albeit an item aggregated over multiple occasions, and to systematically eliminate those items that do not relate empirically to the total at a prescribed level. For example, in one analysis of the data on school conscientiousness, the number of items in our category was reduced from 19 to 12 measures, using a stepwise elimination procedure that adopted a criterion of empirical linkage between the component measures and the total at $p < .01$, one-tailed. By tightening the category with these procedures, the mean level of cross-situational consistency increases to .24, and the increase in mean correlation of specific behavior to total aggregate moves from about .31 to about .43. Further variations are, of course, possible, depending on the particular criterion adopted.[2]

For such goals as test construction, the gains obtained from psychometric procedures of this sort are noteworthy and suggest, as expected, a more coherent patterning among this reduced subset of variables. But caution is required because the increases obtained are practically guaranteed by the procedure adopted; it searches for consistency by systematically ignoring inconsistency and discarding "inconsistent items." The decrease in the breadth of the domain resulting from the excluded behaviors also requires attention. In this regard, recall that the items constituting the sampled domain in the first place were not arbitrarily selected by the researchers in the Carleton study. Rather, they were obtained as the behavioral referents defining the domain (school conscientiousness) as identified by the college students in the community itself (Peake, 1982). Given that a major goal of the research was to explore the links between subjects' perceptions of consistency and the coherences in the observed behavior, this seemed a not unreasonable step. If, instead, one defines a domain as those items in a set of measures that correlate with one another, then certainly one can systematically enhance the correlations that remain.

The point is that the application of these psychometric procedures can be useful but provides no surprises; they highlight the levels of continuity in the Carleton data that are entirely congruent with earlier efforts of this type to find behavioral consistency (e.g., Allport & Vernon, 1933; Dudycha, 1936; Hartshorne & May, 1928; Newcomb, 1929). By eliminating measures that do not fit empirically, the relevant mean coefficients at Carleton fall even more closely into the by-now typical .20 to .30 range for the specific component behaviors, although each effort has categorized those components somewhat differently (e.g., Mischel, 1968). Exact comparisons are not possible because diverse units that ranged in size from relatively more to less molecular were used by various investigators, but, in general, the levels of the overall results seem stable; it is their interpretation that seems to be unstable. And that instability is understandable, because the evaluation of these data shifts with different goals, perspectives, and reference criteria. Whether one greets the results as evidence for the solution of the consistency problem or for its existence depends on theoretical assumptions, paradigm preferences, comparison standards, and goals. The appropriate conclusions to draw may depend more on vantage point than on further data collection of the same type or on further debate about how to label the size of the resulting numbers. Judgments about whether particular levels of coherence call for cheers or jeers cannot be made sensibly without reference to comparison standards and purposes. Confusion occurs when these shift imperceptibly, as they do over the decades with changing problems and changing aims in the specifics of what is being debated, as discussed in Mischel (1979), and as we consider again in the next section.

Rewriting the Trait Critique

In his analyses, Epstein (e.g., 1979, p. 1098) repeatedly quotes excerpts from critics throughout this century who question the generality and utility of traditional trait conceptions and measures for various reasons. He then overstates their conclusions to imply they argued that there are no traits. Unfortunately, this exaggerated conclusion lingers longer than the subtleties of the more moderated originals that were addressing different issues concerning

---

[2] Currently, we also are continuing to examine the utility of alternative methods of studying distinctive, stable, person–situation interactions in behavioral data. The goal here is to try to go beyond the total aggregates to distinguish any meaningful stable patterns from "error," and a variety of alternative strategies are being attempted. We are looking, for example, at patternings that may exist in subclusters of the variables assessed (e.g., Peake & Lutsky, Note 2). We are especially interested in alternative methods for identifying distinctive patterns that characterize individuals beyond their average levels of behavior (see Wright, 1983; Shoda, Note 3; Wright & Mischel, Note 4).

the nature of the structure of dispositions and the organization of behavior. After reducing that long and complex history to "the charge that personality traits do not exist," Epstein (1979, p. 1098) provides an "Evaluation of the Arguments For and Against the Existence of Traits" (p. 1102), concludes that traits exist, and insists that this conclusion be acknowledged (1983a). But it is not the existence of traits that was or is at issue, a point Mischel and Peake attempted to reiterate and with which we again began this note. How to conceptualize dispositions, how to construe personal qualities and person–context interactions, and the utility of inferring them in various ways for various goals, not the existence of stable differences between individuals, seem to us the concerns that persist (e.g., Mischel, 1973, 1983). These concerns involve paradigm choices and decisions about the problems one wants to pursue and the methods they dictate; they are not resolved by remembering the role of reliability and errors of measurement, as discussed in Mischel and Peake (1982).

Epstein's (1983a, p. 179) most central objection to Mischel and Peake (1982) is that the procedure of "examining the average interitem correlation of a sample of items selected casually by face validity . . . has led Mischel and others to conclude [incorrectly] that the concept of traits as broad cross-situational response dispositions is 'untenable.' " But Epstein's statement goes beyond a critique of Mischel and Peake (1982) to misrepresent the bases of the widespread critical reactions to personality trait assessment and to trait conceptions in the 1950s and 1960s (e.g., Hunt, 1965; Peterson, 1968; Vernon, 1964; Wiggins, 1973). The conjunction of the procedure of examining interitem correlations with the conclusion of the "untenable" concept of traits in Mischel's 1968 review perpetuates a by-now common distortion of history; it muddles the rationale, bases, and conclusions of the trait critique, and, in so doing, trivializes it.

The conclusion that traits are untenable did not come from the interitem procedure Epstein claims nor from anything found by Mischel and Peake (1982) or conceivably discoverable in any single empirical effort of the sort they attempted. Instead, the critique of global cross-situational response dispositions as "untenable" (Mischel, 1968, p. 146) came after an extensive review of diverse clinical and assessment data that spoke to the utility and limits of the trait-sign approach for making predictions of specific outcomes for individuals in diverse new situations, using a wide range of common assessment practices for decades of research and real-life application. Explicitly, this critique did not reject the tenability of trait tests and omnibus measures for screening, group decisions, and personality research on individual differences, nor of using omnibus measures for predicting omnibus outcomes. In more than 100 pages, the critique was moderated by many crucial qualifiers and took account of diverse data, as suggested by this summary paragraph that preceded the quote to which Epstein referred:

> In sum, the data reviewed on the utility of psychometrically measured traits, as well as psychodynamic inferences about states and traits, show that responses have not served very usefully as indirect signs of internal predispositions. Actuarial methods of data combination are generally better than clinical–theoretical inferences. Base rates, direct self-reports, self-predictions, and especially indices of relevant past behavior typically provide the best as well as the cheapest predictions. Moreover, these predictions hold their own against, and usually exceed, those generated either clinically or statistically from complex inferences about underlying traits and states. In general, the predictive efficiency of simple, straightforward self-ratings and measures of directly relevant past performance has not been exceeded by more psychometrically sophisticated personality tests, by combining tests into batteries, by assigning differential weights to them, or by employing more complex statistical analyses involving multiple-regression equations. (Mischel, 1968, p. 145)

As this excerpt suggests, the historical critique of the misuses of trait assessment was rooted in much more than a focus on interitem correlations among items casually selected by face validity and on arbitrary laboratory measures. It encompassed the practice of sophisticated assessors using their test batteries to try to predict diverse criteria in real-life contexts with a variety of combinatorial strategies—a practice they themselves found increasingly unjustifiable, excessively hazardous, and subject to judgment errors (e.g., Hunt, 1965; Peterson, 1968; Vernon, 1964; Wiggins, 1973). The issues here were and remain too complex and important—theoretically and in their applications and misapplications—to accept Epstein's version of this history without an attempted correction.

It is also important to recall that the critiques of trait assessment did not question the value of past behavior, or aggregated behavior in a domain, for the prediction of future behavior in similar situations and on closely matched criteria. Indeed, one of the conclusions from Mischel's (1968) review was that "past behavior tends to be the best predictor of similar future behaviors" (p. 135). Likewise, a close convergence between the behaviors to be sampled and those to be predicted on the criterion measures (as advocated, e.g., by Ajzen & Fishbein, 1977, and Epstein, 1983b) was one of the main recommendations emerging from the critique of global trait-sign assessment in the 1960s (e.g., Mischel, 1968, pp. 135–142, 278, 290–294). The assessment critiques of the 1960s were not aimed against careful behavior sampling to predict future similar behavior. They were aimed instead at such common practices as using responses to unstructured, "depth" interviews to predict the individual's future success as a Peace Corps volunteer teaching in Nigeria (Mi-

schel, 1965). At that time, predictions of diverse specific outcomes were readily attempted from indirect signs of pervasive, global dispositions integrated into a broad personality portrait by the assessor and then extrapolated to forecast specific outcomes in specific situations. In personality assessment, one rarely sampled behavior directly from the domain to be predicted. The prediction of future "school conscientiousness" from an obtained sample of school conscientiousness (Mischel & Peake, 1982), probably would have seemed too obvious to require confirmation.

Although some assessors were less bold, it was not unusual to assume two decades ago that one could go from very limited, narrow observations to very broad generalizations—that one could predict from a few indicators or behavioral signs the person's future behavior in new situations far removed psychologically and situationally from the original observed behavior sample. In its extreme forms, almost any indicator seemed to offer a promising royal road into the core personality. And, not uncommonly, one took responses to sentence stems, or to inkblots, and to such specific signs as the use of the white spaces or the color of the inkblots in forming percepts about them to predict such remote outcomes as psychiatric prognosis or rehospitalization. In this atmosphere of reliance on limited data that rarely overlapped directly or closely with the criterion contents, assessors in the 1950s and 1960s often tried to predict specific outcomes in an individual's life and to make significant decisions from personality indicators only loosely linked to the criterion behaviors. Behavior served not to provide relevant samples for predicting related future behavior but as signs of broad dispositions with omnipresent ubiquitous manifestations.

These assumptions and practices were described, referenced, and criticized in detail (e.g., by Hunt, 1965; Mischel, 1968; Peterson, 1968; Vernon, 1964; Wiggins, 1973). The gist of this critique was to question the utility of inferring global, highly generalized dispositions from behavioral signs as predictors of the person's specific life outcomes and future behaviors, and as the basis for major treatment decisions. These critics also urged a search for alternative conceptual routes and called for more limited predictions from data more closely matched to the criterion and from analyses of the criterion situation itself. And they urged renewed attention to what people actually do, to their actions and behaviors in contexts relevant to their lives and to the problems for which they sought help (e.g., Bandura, 1969). These messages were not motivated by a desire to return to an atomistic behaviorism, nor to construe situations as the prime or exclusive causes of behavior. They were expressions of concern that the psychology of tests and test taking and trait labeling that then dominated personality psychology was losing sight of people's discriminative coping behavior and of their interactions in everyday life, that it was in danger of substituting the study of tests and the simplistic categorization of groups for the analysis of persons and their cognitions and actions. They were criticisms guided by the view that consistencies surely exist but that they must themselves be explained rather than invoked as the ultimate explanatory constructs.

The phrase about the untenability of traits to which Epstein's above-cited quote alludes comes from this context. But the critique that it is untenable to make and to use sweeping dispositional inferences about individuals in the context-free way that they were then commonly used, need not be misread continuously as a denial of the existence of stable individual differences. Such a misreading perpetuates unnecessary debates without addressing the substantive points still at issue. Likewise, criticisms of the limits and the easy misuses of trait explanations (e.g., when moral behavior is construed as caused by the moral disposition that was inferred from observing the moral behavior in the first place) need not be mistaken as a rejection of all personality variables that are not directly observable. The case for more carefully specified, contextually anchored units in personality study was originally misheard as a homicidal assault on personality itself (e.g., Craik, 1969). But it is time to overcome such fears and to recognize that our concepts of personality study can be broadened to incorporate meaningful situational variability and stable person–situation interactions without undoing the viability of personality psychology itself.

Responses by Funder (1983) and Bem (1983)

The joint authors of the template-matching solution to the consistency problem in the analysis of delay behavior have responded separately to our critique of their paper, so each must be considered separately although the points overlap. Funder (1983) quarrels with the specifics of what we did, claimed, and concluded. Any sufficiently interested readers can assess his version against ours for themselves by comparing our statements, procedures, and conclusions, step by step, to those he attributes to us, without any aid in decoding. For most readers, a brief summary of the essentials should suffice.

Funder (1983)—as well as Bem (1983)—fault us for focusing our critique exclusively on their first study, the investigation of delay behavior, and for referring to the other two parts of their article only briefly. But, as we said, and as they said, this was the only study in their paper that was relevant to the consistency problem and, specifically, that was relevant to demonstrating how their approach "can help pinpoint the sources of behavioral inconsistency across seemingly similar situations" (Bem &

Funder, 1978, p. 488). Hence, their delay study was the only part of their work that was relevant to our analysis (i.e., to the search for cross-situational consistency). That is why it is the only part on which we focused and why we explicitly (Mischel & Peake, 1982, p. 744) restricted our comments and analysis to it. As we said, our criticisms were not addressed at other uses of template matching but were focused on the claim that the Bem and Funder template-matching approach had enhanced the search for cross-situational consistency. We found no support for their claim either in their delay study or in our extended replication attempt. Our basic point (pp. 740–741) was that although their approach promises to help resolve the consistency issue by identifying those situations that are psychologically equivalent rather than merely superficially similar, their data provide no evidence whatsoever for cross-situational consistency. Indeed, the design of their study did not allow them to even address issues of cross-situational consistency because they obtained only *one* response (delay time in one situation) from each subject. Their paper raised the promise of a new route for finding cross-situational consistency in behavior, but it did not take the essential step of showing it. That is why we proceeded to try to fill this gap and tested the Bem and Funder claims with an appropriate design for assessing cross-situational consistency. And when we undertook that empirical effort, which did provide the necessary data to attempt a test of their proposition, the results failed to support it.

Funder also refers to our effort to study the perception of consistency and its links to the structure of behavior (Mischel & Peake, 1982, pp. 747–752). Here we are trying to use designs that incorporate relatively molecular behavior observations over time and across referent situations with more global measures of perceptions of consistency and variability so that behavioral bases for the perception of consistency and for its organization can be analyzed. In a first effort toward that goal, we proposed—congruent with a cognitive prototype approach (e.g., Cantor & Mischel, 1979; Cantor, Mischel, & Schwartz, 1982)—that the judgment of trait consistency is strongly related to the temporal stability of highly prototypic behaviors. In contrast, we proposed that the global impression of consistency may not be strongly related to highly generalized cross-situational consistency, even in prototypic behaviors. The data so far supported these hypotheses and suggested to us that the perception of personality consistencies may depend more on the temporal stability of key features than on the observation of overall cross-situational behavioral consistency, and the former may be easily interpreted as if it were the latter.

Addressing this part of our work, Funder (1983) provides a summary that fits neither our reasoning nor our results. For example, he charges that we presented only the difference between two correlations within a single behavior domain as the basis for our theorizing about the links between self-perceived consistency and the temporal stability of behavior. He then chides us for not heeding our own advice that it seems premature to offer conclusions about an approach on the basis of a single confirming comparison. In fact, the relevant hypotheses of Mischel and Peake (1982) were not based on the difference between two coefficients in one behavior domain (conscientiousness). As stated on p. 751:

> 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 differencs 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.

Moreover, as stated on the same page, parallel analyses were done and similar results were obtained in a replication analysis of a second behavior domain, friendliness (Peake, 1982). Nevertheless, Funder (1983) insists repeatedly (e.g., p. 287) that our work was based "on a single contrast between two correlations, unreplicated" Funder's version of our research does not correspond with what we reported.

The Missing Theory–Data Connection

Bem's (1983) note argues that

> Mischel and Peake misrepresent the link between the empirical study they examined and the methodological or conceptual strategy we advocated. In particular, they distort the raison d'être of the template-matching technique, relegating its central contribution—literally—to a footnote. And second, they misapprehend the conceptual status of Bem and Allen's empirical study, revealing their misapprehension—literally—in a footnote. Despite, this, however, some of their concluding suggestions are remarkably like those of Bem and Allen. (p. 390)

Let us consider Bem's complaints closely. First, we are accused of distorting the raison d'etre of the template-matching technique and of relegating its central contribution to a footnote. How? Because we confined our analysis to the ways Bem and Funder applied their template-matching method to the search for cross-situational consistency, and thus

dealt only with it. That is true. As already stated in reply to the same point by Funder, our interest in the Bem and Funder work was only in testing the claims that the template-matching approach will allow the prediction of cross-situational consistency in behavior, the prediction of more of the people more of the time. We said this (Mischel & Peake, 1982, p. 739) when introducing the Bem and Funder work:

> 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.

And the first footnote to which Bem refers reiterated the limits of our focus:

> 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. (Mischel & Peake, 1982, p. 744)

In sum, the explanation of why in our paper on consistency our analysis of the Bem and Funder contribution to the consistency problem was limited exclusively to their study on the topic of consistency was because that was the topic of our paper.

Bem (1983), like Funder (1983), skirts our reminder that template matching is a new, promising-sounding label for Meehl's (1956) old and familiar but uncited (by Bem and Funder) cookbook procedures as they were applied decades ago to matching profile types with Q-sort patterns and other outcome data. As we showed (Mischel & Peake, 1982, pp. 743–745), it is based on the same logic and uses the same kinds of data for the same goals selected decades later by Bem and Funder (e.g., Q-sorts, typologies of kinds of people who may be expected to behave in predictable ways in particular kinds of situations). When this strategy was presented by Bem and Funder with the implication that it offered the solution to the consistency problem, it demanded new attention. Now, both Bem and Funder deal with our findings that their illustrative demonstration fails to help with—or even to address—the consistency problem by dismissing their own study on this topic.

Bem's (1983) second point is that Mischel and Peake did not apprehend the conceptual status of Bem and Allen's empirical study. Bem clarifies our misapprehension as follows. After quoting our comment that "because Bem and Allen's approach does not speak to within-person trait organization, it seems a misnomer to label it idiographic" (Mischel & Peake, 1982, p. 745), Bem goes on to write:

> But it is not we who are confused. Our approach *does* speak to within-person trait organization; it *is* idiographic. Ths *study* is not. It—like the ambitious replication it spawned—remains an object lesson for the investigator who fails to heed our advice to go forth and be idiographic. (1983, p. 392)

We are impressed with Bem's candor. His words make it clear that the theory section in Bem and Allen (1974) and their empirical study were unconnected. Reviewing the state of previous consistency research, 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." (1974, p. 510) Now, reviewing his own work, Bem seems to conclude that his intuitions were right; his research, irrelevant.

Finally, consider Bem's (1983) comment that despite our distortion and misapprehension "some of their concluding suggestions are remarkably like those of Bem and Allen. (p. 390). Here he refers to his points "that personality traits are sets of behaviors that each individual has idiographically organized into his or her equivalence classes" (p. 392) and that "individuals traditionally appear inconsistent to personality researchers because an individual's behaviors are not necessarily organized into the nomothetic equivalence class that the investigator necessarily imposes in the very act of selecting which behaviors to measure" (p. 392). But this was not Mischel and Peake's "final prescription for a person-centered approach to personality" (p. 392) nor the hypothesis that we proposed and tested. As discussed on pages 750–752 of our earlier article, we proposed (and found) specific links between the temporal stability of prototypic behavior in a domain and the perception of cross-situational consistency within that domain. That hypothesis should not be confused with Bem's summary.

Beyond Déjà Vu, We Also Hope

All participants in the present set of exchanges seem to share the view that it is unproductive to continue to recycle the debate concerning consistency. We hope that the déjà vus in the debate are ending, to be replaced by a less argumentative, more focused set of questions that dictate increasingly constructive research routes. There may be a useful analogy to the controversy that raged for years in the neighboring area of clinical psychology. That debate raised the question "Does psychotherapy work?" and asked it over and over again. After decades of turmoil about whether the answer was yes, no, or maybe, this omnibus question began to be reconstrued into more specific subquestions that allow subtler issues to be examined. Thus, for example, What kinds of therapy when practiced with what kinds of techniques are likely to help what kinds of problems for what kinds of people under what kinds of conditions? And, perhaps most important, What are the psychological processes

through which therapeutic changes operate? We hope that the area of personality is ready for an analogous restructuring of perhaps its central problems into more finely honed, process-oriented questions. Such a restructuring should stimulate research into the many issues concerning the organization and construction of consistencies, as has been urged by all the participants in the last cycle of discussions. Rather than more debate, we need to continue to pursue the patterns of both coherence and stable discriminativeness found in behavior at different levels of generality and to link those patterns to the perceptions of consistency and to their psychological antecedents and consequences.

Reference Notes

1. O'Brien, E. J. *The stability of personality traits.* Unpublished manuscript (presented in partial fulfillment of the requirements for the Preliminary Comprehensive Examination), University of Massachusetts, January 1978.
2. Peake, P. K., & Lutsky, N. S. *Exploring the generality of "predicting some of the people some of the time": The Carleton Student Behavior Study.* Manuscript in preparation, Stanford University, 1983.
3. Shoda, Y. *Comparing additive and interactive models of assessment in the prediction of behavioral consistencies.* Unpublished research memorandum, Stanford University, April 1983.
4. Wright, J. C., & Mischel, W. *Predicting cross-situational consistency: The role of person competencies and situation demands.* Manuscript in preparation, Stanford University, 1983.

References

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

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

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

Bem, J. Constructing cross-situational consistencies in behavior: Some thoughts on Alcer's critique of Mischel. *Journal of Personality,* 1972, *40,* 17–26.

Bem, D. J. Further déjà vu in the search for cross-situational consistency: A response to Mischel and Peake. *Psychological Review,* 1983, *90,* 390–393.

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.

Cantor, N., & Mischel, W. Prototypicality and personality: Effects on free recall and personality impressions. *Journal of Research in Personality,* 1979, *13,* 187–205.

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

Craik, K. Personality unvanquished. (Review of *Personality and Assessment* by W. Mischel). *Contemporary Psychology,* 1969, *14,* 147–148.

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 confusion: A reply to Mischel and Peake. *Psychological Review,* 1983, *90,* 179–184. (a)

Epstein, S. Aggregation and beyond: Some basic issues on the prediction of behavior. *Journal of Personality,* 1983, *51,* 360–392. (b)

Funder, D. C. Three issues in predicting more of the people: A reply to Mischel and Peake. *Psychological Review,* 1983, *90,* 283–289.

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

Hunt, J. McV. Traditional personality theory in the light of recent evidence. *American Scientist,* 1965, *53,* 80–96.

Kelly, G. A. *The psychology of personal constructs* (Vol. 1 & 2). New York: Norton, 1955.

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

Mischel, W. Predicting the success of Peace Corps volunteers in Nigeria. *Journal of Personality and Social Psychology,* 1965, *1,* 510–517.

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

Mischel, W. Alternatives in the pursuit of the predictability and consistency of persons: Stable data that yield unstable interpretations. *Journal of Personality,* 1983, *51,* 578–604.

Mischel, W., & Peake, P. K. Beyond déjà vu in the search for cross-situational consistency. *Psychological Review,* 1982, *89,* 730–755.

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

Peake, P. K. Searching for consistency: The Carleton Student Behavior Study (Doctoral dissertation, Stanford University, 1982). *Dissertation Abstracts,* 1982, *43,* Pt. B, Section 8, p. 2746. (University Microfilms No. AAD 83-01259)

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

Rushton, J. P., Brainerd, C. J., & Pressley, M. Behavioral development and construct validity: The principle of aggregation. *Psychological Bulletin, 94,* 18–38.

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.

Wright, J. C. *Explorations of the structure and perception of behavioral consistency.* Unpublished doctoral dissertation, Stanford University, 1983.

Received July 9, 1983
Revision received July 11, 1983 ■