Journal of Personality and Social Psychology
1984, Vol. 47, No. 5, 1074-1090

A Failure to Replicate the Bem and Allen Study of Individual Differences in Cross-Situational Consistency

William F. Chaplin
Auburn University

Lewis R. Goldberg
University of Oregon

Bem and Allen (1974) purportedly found evidence that, by using self-report measures of cross-situational consistency as moderator variables, it was possible to substantially increase the size of correlation coefficients computed among measures of each of two personality traits. The present study was undertaken to (a) replicate the Bem and Allen finding on a larger set of personality traits, (b) determine if the results hold differentially for self-report, other-report, and objective personality measures, and (c) compare different methods for dividing subjects into high and low consistent groups. One hundred twelve subjects (64 men and 48 women) were divided into high and low consistent groups using three different methods. Within each group correlations were computed among and between 6 self-report measures, 10 other-report measures, and when possible a few objective measures, for each of eight traits. In general, none of the methods of consistency classification replicated the Bem and Allen finding. In cases where the high consistent group had a larger average correlation among the measures than their low consistent counterparts, the size of that difference was negligible. Moreover, there were about as many instances in which the low consistent group actually had larger average correlations. Finally, there was essentially no agreement among the various consistency classification methods in dividing subjects into high and low consistent subsamples.

Among the most frequently cited reports in the recent personality literature is that by Daryl Bem and Andrea Allen (1974), "On predicting some of the people some of the time: The search for cross-situational consistencies in behavior." The basic conclusion drawn from this study is that there are large individual differences in cross-situational consistency, specifically that for any one personality trait some individuals will be situationally consistent, whereas others (perhaps the majority) will not be. Bem and Allen invoked two concepts to explain why subjects may not be cross-situationally consistent on a construct defined by the experimenter: (a) lack of agreement between the subjects and the experimenter on the equivalence class of relevant behaviors that serve to define the construct, and (b) lack of agreement among the subjects on the "difficulty level" of each of the behavior–situation units within the equivalence class. (Stated psychometrically, interitem response "inconsistency" can stem from differences between the experimenter and the subjects in the perceived homogeneity of the items in a scale, and/or from disagreements among the subjects as to the popularity or difficulty level of each of those items.)

In an empirical study, Bem and Allen selected two constructs, friendliness and conscientiousness, and asked their subjects (64 Stanford students in an introductory psychology class) to rate both their overall level, and the extent of their cross-situational variability, on each trait. In addition, these subjects completed an inventory called the Cross-Situational Behavior Survey (CSBS), which had been developed by Bem and Allen. The CSBS consisted of 86 statements that descibe both a behavior (e.g., talking) and a situation

This project was supported by Grant MH-32585 from the National Institute of Mental Health and Grant BNS-7826208 from the National Science Foundation. The authors wish to acknowledge the enormous help of Melissa Finch and Antonia Ristorcelli in collecting these data; Molly Stafford, Elizabeth Epler, and Tom Reischl in analyzing them; and Auke Tellegen, Oliver John, Steven Golding, James Lamiell, Philip Peake, and Edward Diener in understanding them.

Most of the analyses were carried out at the University of Illinois during the first author's tenure there as a visiting assistant professor.

Requests for reprints should be sent to William F. Chaplin, Department of Psychology, Auburn University, 4082 Haley Center, Auburn, Alabama 36849.

in which the behavior might occur (e.g., with a stranger). Twenty-four of the items were focused on the construct of friendliness and 23 were focused on conscientiousness. For each item the subject rated, on a 7-point scale, the likelihood that he or she would engage in that behavior in that situation.

As criteria of cross-situational consistency, the investigators examined the correlations among self-reports, mothers' reports, fathers' reports, and reports from a peer, plus a few situational tests of friendliness (frequency and duration of vocalizations, mean ratings by others after a group discussion, and latency in initiating a conversation with an experimental confederate in a waiting room) and of conscientiousness (promptness in returning course evaluation forms, extent of completion of course-related reading, and the neatness of the subject's hair and clothing, and of his or her living quarters).

The results of this study were quite dramatic; indeed, it is rare in psychology for a theoretical conjecture to be verified so strikingly. For the trait of friendliness, the mean correlation among six situations was .57 for the low variability, versus .27 for the high variability subjects. For the trait of conscientiousness, the mean correlation among six situations (the neatness situation here excluded) was .45 for the low, versus .09 for the high variability subjects. Seemingly then Bem and Allen's hypothesis that self-reported cross-situational variability will be associated with actual cross-situational variability was confirmed.

In the years since these results were reported a number of similar studies have appeared in the literature. Some of these used Bem and Allen's paradigm to try to resolve the trait–situation debate (e.g., Campus, 1974; Kenrick & Stringfield, 1980; Mischel & Peake, 1981, 1982; Olweus, 1980; Underwood & Moore, 1981), whereas others explored the utility of self-reported consistency as a moderator variable for improving the validity of personality measures (e.g., Cheek, 1982; Lippa & Mash, 1981; Snyder & Monson, 1975; Tunnell, 1980; Turner, 1978; Vestewig, 1978). However, with only a few exceptions (Amelang & Borkenau, 1982; Rushton, Jackson, & Paunonen, 1981; Tellegen, Kamp, & Watson, 1982) there has been no effort to examine critically the original Bem and Allen investigation. This lack of criticism is particularly surprising because most of these subsequent investigators have reported results that are much less striking than those reported by Bem and Allen (1974).

Moreover, there are several aspects of this intriguing study itself that suggest that it be interpreted cautiously. First, the generality of the Bem and Allen study is limited in two important ways. Bem and Allen investigated only two personality dimensions. The extent to which these findings will generalize to other dimensions of personality is as yet unknown. Also, Bem and Allen conducted their research on a sample of 64 Stanford undergraduates. It is possible that their results, obtained from this small and intellectually homogeneous sample, are unrepresentative of more diverse populations. Thus, it is extremely important that the Bem and Allen hypothesis be tested on a wider range of personality characteristics and a more heterogeneous sample of individuals before their findings become accepted.

Second, the most disquieting aspect of the Bem and Allen study is their failure to confirm the hypothesis with either one of the two alternative measures of self-reported cross-situational consistency for both of the traits under study. Bem and Allen (1974) elected to include in their study two alternative measures of self-reported cross-situational consistency. The first was the response to a simple question: "How much do you vary from one situation to another in how friendly and outgoing (conscientious) you are?" Responses were obtained on a 7-step scale that ranged from "Not at all" to "Extremely." A second measure of self-reported cross-situational consistency was derived from the CSBS; this measure was the ratio of the individual's variance in his or her ratings of the 24 friendliness (or the 23 conscientiousness) behavior–situation items to the total variance of that individual's ratings across all 86 CSBS items. Dubbed an "ipsatized variance index," this measure was recommended by Bem and Allen to subsequent investigators as "the most promising candidate for the moderating variable in any future work" (p. 516, Footnote 6). Yet, although not stressed in the original article, this ipsatized variance index, although

associated with actual cross-situational consistency for the trait of conscientiousness, was not so associated for friendliness. And the simple variability rating scale, which was highly associated with actual cross-situational consistency for the trait of friendliness, was not so associated for conscientiousness. That is, the measure of cross-situational consistency that worked for friendliness failed for conscientiousness, whereas the alternative measure that worked for conscientiousness failed for friendliness. Had Bem and Allen phrased their conclusions in this way, their results might now be viewed as less confirmatory of their hypothesis than is presently believed. What is at issue, however, is not the language of their exquisitely written report but the more fundamental question of the replicability of their findings. In general, if alternative measures of the same construct fail to produce the same pattern of the results, then there is a good deal left to be explored before any findings become accepted as facts.[1]

A final issue raised by the Bem and Allen study concerns the type of data for which their procedures are most effective. Cattell (1957) has suggested a useful categorization of the types of data used to study personality, on the basis of the source from which the information is obtained. One type, called Q-data, is information obtained from the person being assessed. These self-report data can be gathered with questionnaires, personality inventories, or self-rating scales. L-data are obtained when an observer assesses an individual's characteristics on the basis of what the individual does in everyday life. Finally, T-data are obtained from the subject's behavior in a standard situation. Q- and L-data share the use of humans as data integrators. That is, information obtained from self and other reports is generally based on global impressions that the raters presumably form on the basis of a large number of more specific observations. T-data, however, generally require much less integration by the observers, and thus are frequently viewed as more objective.

Because these types of data represent different practical and theoretical approaches to the study of personality, it would be helpful to know if the Bem and Allen results are more likely to appear within any type of data. In their study Bem and Allen relied primarily on L-data, but they included one Q-data variable and two or three T-data measures. The small number of tests selected from each data domain precludes a consideration of this issue on the basis of Bem and Allen's study.

The present investigation was undertaken to (a) replicate the Bem and Allen study and explore the generality of their findings across a larger number of traits, (b) compare the moderator effects of several methods of measuring consistency, and (c) investigate the differential effectiveness of this moderator approach for the three types of data used to investigate personality.

Method

Personality Traits

In their original study, Bem and Allen focused on two personality traits, friendliness and conscientiousness. To extend the generality of their findings, eight traits were selected for this study: (a) friendliness, (b) conscientiousness, (c) assertiveness, (d) honesty, (e) sensitivity, (f) activity level, (g) cultural sophistication, and (h) emotional stability. The first two traits were those investigated by Bem and Allen. Some CSBS items were developed by Bem and Allen for the next three traits, but they were not analyzed in their study. The last three traits cover aspects of human personality that were not considered by Bem and Allen. Overall, these eight traits map fairly comprehensively the domain of human behavior as evidenced by factor analyses of peer-rating data (e.g., Norman, 1963).

Subjects

One source of information about our subjects for each of the eight personality dimensions was extensive descriptions by peers who knew each of them quite well. To facilitate the collection of these descriptions, subjects were recruited from intact living groups. All the major living groups (sororities, fraternities, and dormitories) at the University of Oregon were contacted, told briefly about the nature of our research, and offered a substantial honorarium for participating in this study. Two sororities, five fraternities, and three dormitories (one coeducational and two all-male) volunteered. In the interest of obtaining as diverse a sample as possible, one sorority, one fraternity,

[1] Indeed, Bem and Allen (1974) did not even report the correlation between their two moderator variables, nor did they indicate the extent to which subjects classified as consistent on a trait by one method were similarly classified by the other. The fact that the moderator effect differed so substantially between the two methods suggests that these classifications may have been quite independent.

Table 1
Number of Items and Coefficient Alpha Reliability Estimates for the Eight Scales of the CSBS and the TDA

Scale	CSBS			TDA		
	Number of items	α	rii[a]	Number of items	α	rii[a]
Friendliness	24	.49	.04	56	.80	.07
Conscientiousness	24	.68	.08	48	.86	.11
Honesty	16	.55	.07	22	.62	.07
Sensitivity	17	.67	.11	32	.64	.05
Assertiveness	20	.29	.02	70	.87	.09
Activity level	25	.79	.13	19	.73	.12
Emotional stability	25	.29	.02	41	.83	.11
Cultural sophistication	33	.80	.11	53	.81	.07

Note. CSBS = Cross-Situational Behavior Survey; TDA = Inventory of 415 Trait Descriptive Adjectives.
[a] The rii index is an estimate of the reliability of a single item, based on the Spearman-Brown formula.

and one all-male dormitory were invited to participate. Within each type of living unit, selection was based on the size of the potential sample. The final sample consisted of 64 men (36 from the dormitory and 28 from the fraternity) and 48 women from the sorority, for a total of 112 subjects.

Measures

One of the goals of this study was to investigate the differential effectiveness of the Bem and Allen procedures within each of the three major types of data used to assess personality (Cattell, 1957). Thus, we included a set of 6 self-report measures, 10 other-report measures, and wherever possible a few objective measures for each of the eight traits. In addition, three different measures of self-reported cross-situational consistency were used to divide subjects into high and low consistent groups.

Self-Report Measures

The Cross-Situational Behavior Survey (CSBS). This inventory is a revised form of the original CSBS used by Bem and Allen. The revised inventory consisted of all the items written by Bem and Allen, plus three new sets of items to assess the dimensions of activity level, emotional stability, and cultural sophistication. This new CSBS included 169 items describing specific behaviors in specific situations; subjects rated, on a 7-point scale, the likelihood that they will behave in that manner in that situation. Table 1 shows the number of items and the coefficient alpha reliability estimate for each of the eight CSBS scales. The subject's characteristic status on each of the eight dimensions was assessed by summing the subject's responses across the items on each of the eight scales.

Inventory of 415 Trait Descriptive Adjectives (TDA). This is a set of 415 common personality-descriptive adjectives. The subjects rated themselves on a 6-point scale ranging from extremely inaccurate as a description of themselves to extremely accurate as a descriptor. In addition, the subjects could indicate that the term was neither accurate nor inaccurate as a descriptor for any of four reasons: (a) because she or he is average or neutral on that trait, (b) because his or her behavior depends on the situation, (c) because she or he doesn't know whether the term applies or not, or (d) because the meaning of the term is unclear or ambiguous.

This inventory was originally developed to explore in more detail the "situational" response that is so important in the attribution literature (Goldberg, 1981). The 415 terms were selected from larger sets and include a comprehensive subset of terms whose meanings are familiar to most college students. Scales for each of the eight traits were developed on an independent sample, using a two-step process. First, correlations were computed between each of the 415 terms and the 16 terms that define both poles of the eight traits. Second, an intuitive assignment of each term to one of the eight scales was made by two independent judges. Terms that conformed both empirically and intuitively to the scale definitions were included in that scale. Table 1 shows the number of items and the coefficient alpha reliability estimate for each of the eight TDA scales.

Self-ratings at Time 1. During the first week of the study the subjects rated themselves on each of the eight traits. To obtain these ratings the subjects were asked, "In general, how X are you?", where X was one of the eight traits. The ratings were made on a 7-point scale ranging from "Not at all" through "Moderately" to "Extremely."

Self-ratings at Time 2. At the end of the study the subjects again rated themselves on each of the eight traits in the same manner.

Embedded self-ratings. In addition to providing information about themselves, the subjects were also asked to provide information about a set of their peers who were members of their living group and who were also participating in this study (see below). On one of these peer-rating tasks the subjects were presented with a list of their peers and were asked to rank order the people on the list according to their relative levels on each of the eight traits. Embedded in this list was the subject's own name. Another self-report measure was the subject's

self-perceived position on each of the eight traits, relative to that of his or her peers.

Self-ratings following a group discussion. One phase of this study was a laboratory experiment in which the subjects participated in a group discussion (see below). The primary purpose of this phase was to obtain some objective behavioral measures, but the subjects were also asked to provide self-ratings. Specifically, for each of the eight traits the subjects rated, on a 7-point scale, the degree to which the trait characterized their behavior during the group discussion. The scale was labeled "Not at all" through "Moderately" to "Extremely."

Other-Report Measures

One way of obtaining information about a person's behavior in different situations is to ask many individuals who know the person well to describe that person. Because each rater presumably interacts with a person in different situations, each set of ratings should reflect the ratee's behavior in a different set of situations. Thus, by comparing the ratings provided by various others, some indication of the subject's cross-situational behavioral consistency may be obtained. A total of 10 different other-report measures were used in this study.

Peer ratings and peer rankings. Each of the subjects was rated and ranked by a set of their peers who were also members of the subject's living group. To select this set of peers all subjects were asked to complete a preliminary questionnaire, which included the names of all the people in a given living unit who participated in the study. The subjects were asked to rate everyone on the list on two 5-point scales: (a) how well they knew the person (familiarity), and (b) how well they liked the person (evaluation). The subjects were asked to make both ratings relative to the people on their list.

The selection of the raters was based on the information obtained from the preliminary questionnaire. A computer algorithm was developed to select the subject's raters on the basis of the following conditions: (a) Each subject had to be rated by, and had to rate, either 6 (dormitory), 8 (fraternity), or 9 (sorority) individuals. The number of peer raters selected for each group was determined by the largest number of raters that the least well-known subject could be assigned, given the constraint that the raters needed to know the subject quite well. The assignments were not necessarily reciprocal, although most of the subjects both rated and were rated by some of the same peers. (b) Each person was only allowed to rate people whom she or he indicated she or he knew at least at an average level relative to the members of his or her living group (i.e., ratings of "3" or greater on the preliminary questionnaire). No attempt was made to control the extent to which the rater indicated that she or he liked the person, but given the high degree of relation between familiarity and evaluation, most subjects rated individuals whom they liked at least an average level relative to the other members of the group.

To obtain the peer ratings and peer rankings, each subject received an individualized 16-page questionnaire. On the first eight pages the subjects were asked to provide ratings for each of their selected set of peers on each of the eight traits. Ratings were made on a 7-point scale. On the last eight pages the subjects were asked to rank order their selected set of peers on each of the eight traits. Ties were not permitted. For each subject, the peer ratings, and the rankings, were averaged across his or her set of raters, and this average rating and average ranking was included in the analyses.

Other ratings. The next set of other-reports were obtained from six additional people who knew the subjects in different situations, including the subject's mother, father, a former teacher, a neighbor from his or her hometown, a friend who attended the University of Oregon but who was not a member of the subject's living group, and a friend who did not attend the University of Oregon. The subjects provided the names and addresses of one person for each of the six relations. Each person nominated by the subject received a cover letter informing him or her about the general nature of the study and requesting his or her help; these letters were signed by the subjects. Recipients were told that these ratings would be confidential and that the subjects would not be allowed to see them. A questionnaire accompanied the cover letter, on which the rater was instructed to rate the subject on a 7-point scale, ranging from "Not at all" through "Moderately" to "Extremely," on each of the eight traits. The form of the question was "In general, how X is he/she?", where X was one of the eight traits. The rate of return of these questionnaires was quite high; specifically, 85% of the mothers, 80% of the fathers, 77% of the teachers, 79% of the neighbors, 67% of the university friends, and 71% of the nonuniversity friends returned their questionnaires.

Ratings during laboratory sessions. The last two other-reports were obtained in conjunction with the laboratory group discussion (see below). During the discussion seven observers watched the subjects through one-way mirrors, and at the end of the session they rated each of the participants on the eight traits. These observer ratings were made on 7-point scales ranging from "Not at all" through "Moderately" to "Extremely." The ratings were averaged across the seven observers, and that average was used in our analyses. In addition, at the end of the discussion the participants rated each other on the same set of eight scales. Again, the average of these ratings was used in the analyses.

Objective Measures

The primary source of our objective measures was a laboratory session,[2] in which a subsample (45 men and 25 women, for a total of 70 subjects) of the subjects participated. A laboratory session began when 3 subjects, typically one from each living unit, reported to three different rooms in the psychology building. In each room the subject was greeted by an experimenter, who recorded the time of the subject's arrival (a measure of conscientiousness). The experimenter told the subject that the study was concerned with the acquaintanceship process,

[2] We are indebted to Melissa Finch for directing this part of the study and to Craig Apperson, Holly Brown, Rosemary Hazen, Mike Jenks, Dan Rhodes, Antonia Ristorcelli, Paula Thibedeau, Rick Thorpe, and Kathy Ventres for their work as experimenters, observers, confederates, and troubleshooters.

and that the subject would be participating in a group discussion with 3 or 4 strangers. To start the session, each of the group members, including the subject, was to give a two-min talk about himself or herself to provide a basis for the later discussion.

After being informed about the experiment, the subject completed the Multiple Affect Adjective Checklist (MAACL), developed by Zuckerman and Lubin (1965). The MAACL was designed to assess three aspects of "emotional stability"—anxiety, depression, and hostility. It consists of 132 adjectives, and subjects are asked to check all those that describe how they are feeling at that moment. Each of the MAACL scale scores, when corrected for baseline level (see below), was included as a measure of emotional stability in an anxiety-provoking situation (the subject had just been informed of the two-min talk).

Following the administration of the MAACL, another subject (actually a confederate) appeared, ostensibly to take a test. The experimenter looked for the test, failed to find it, and asked the subject and the confederate to wait in the room while the experimenter obtained the test. The experimenter stayed away for slightly more than two min. The confederate was sitting so that a clock with a second hand was visible. The confederate noted how long (in seconds) it took the subject to initiate a conversation. This was used by Bem and Allen as a behavioral assessment of friendliness. After two min the experimenter returned and told the subject that the group discussion room was now ready. The subject was informed how to reach the room, and sent on his or her way. This procedure took place in all 3 rooms, with 3 different experimenters and 3 different confederates.

On arriving at the group discussion room, the subject was greeted by the discussion leader. The subject was told that because we were interested in the acquaintanceship process, we did not want him or her to talk to any of the other subjects until the experiment started. Instead, the subject was asked to select a magazine from among 15 that were spread out on a table and to read the magazine while waiting for the experiment to begin. The magazine selected by the subject was used as a measure of cultural sophistication. Five of the magazines had been rated as high on this trait (*Atlantic, Harpers, New Republic, Fortune,* and *American Artist*), five had been rated in the middle (*Newsweek, Time, U.S. News & World Report, Psychology Today,* and *Sunset*), and five had been rated as low (*Gossip, Hollywood, True Detective, Reader's Digest,* and *People*). The subject was assigned a score of 3, 2, or 1 if she or he chose a high, medium, or low magazine, respectively.

After all 3 subjects had arrived and had selected a magazine, the group session began. The leader again explained that the study was about the acquaintanceship process, and that each person would begin by talking for two min about himself or herself. The subjects were told that they could talk about anything they liked except for the experiment and their role in it. The subjects were observed from behind one-way mirrors during the entire session, and they were so informed. To break the ice and to set an example, the leader began with a two-min talk about herself. After the leader finished, each subject in turn gave a two-min talk.

Each group included at least 3 people, usually 2 males and 1 female. If a subject failed to appear, that subject's confederate acted as if she or he was an experimental participant. Occasionally 2 confederates were used. A few of the groups included 4 persons, consisting of 3 subjects and another one recruited from a psychology class. If all 3 subjects appeared and no one else was available, the leader filled in two min after the last talk with additional "instructions."

After the leader finished her talk, the subjects gave theirs; if there were confederates, they were always the last to talk. During each talk, two observers counted the number of times the subject blinked his or her eyes. The observers also counted the number of eyeblinks for that subject for the two min immediately following the subject's talk. This eyeblink measure was used to assess the degree to which the subject was nervous while talking (Young, 1973). As a measure of emotional stability, the number of eyeblinks during the talk, minus the number of eyeblinks after the talk, was computed. In addition, the number of times the subject had to be prompted by the leader during the talk was recorded.

After the initial talks were completed, the leader informed the subjects that they would have 10 min to chat among themselves and get to know each other. The leader again informed the subjects that they would be observed during this time, asked them not to talk about the experiment, and then left the room. During the 10-minute discussion, each subject was observed by two coders, who recorded his or her frequency of vocalizations, and the total time that the subject talked. A vocalization was counted every time the subject said something after at least 2 s had elapsed since his or her last vocalization. To assess the duration that the subject talked, a stopwatch was started every time the subject spoke and allowed to run until the subject stopped speaking; the watch was started again when the subject spoke again. Because the watches were not reset between vocalizations, they measured the total time that the subject spoke. Bem and Allen used similar measures of frequency and duration of vocalization as indicators of friendliness; in the present study, these measures were also used to assess assertiveness. In addition, each subject's duration score divided by his or her frequency score was included as another measure of assertiveness. For these measures the scores used in the analyses were the averages of the two observers' recordings. In general the agreement between the two observers was perfect.

After 10 min had elapsed, the leader returned and administered two questionnaires. The first was the MAACL, which provided baseline data against which to evaluate the subject's MAACL scores obtained during the earlier anxiety-provoking situation. The second questionnaire was a self- and peer-rating form. The subjects were asked to rate themselves and each of the other group participants on each of the eight personality dimensions. The ratings were made on 7-point scales ("Not at all" through "Moderately" to "Extremely"). The format was identical to that used to collect other ratings in this project, and to the format used by Bem and Allen. However, for the self-ratings, the subjects were asked to make their ratings strictly on the basis of their behaviors during the group session. While the subjects were rating themselves and the other subjects, the group leader and all of the coders were also rating the subjects on the eight personality traits. In addition, the leader and the coders rated the

subjects on how well they were dressed, by using a 7-point scale ("Not at all well dressed" through "Average" to "Very well dressed"), and on how neatly they were dressed, again with a 7-point scale ("Very sloppy" through "Average" to "Very neat"). The subjects' scores on these measures were the average ratings received from the seven raters. The former was used as a measure of cultural sophistication, and the latter of conscientiousness.

When the subjects had completed their questionnaires, the leader carefully checked their forms, thus permitting only one subject at a time to leave the room. One of the coders, who was unknown to the subject, was positioned down a hall and around a corner from the group discussion room. (The subjects could only walk in one direction down the hall.) As the subject approached, the coder started walking toward the subject and glanced at him or her as they passed. The coder noted the subject's behavior while they were passing in the hall, and made the following rating: 0 = no contact; 1 = eye contact only; 2 = eye contact and smile; and 3 = eye contact, smile, and vocalization. This rating was used as another measure of Friendliness.

In addition to the measures from the laboratory session, three other objective measures were used. One of these was the subject's grade point average, which was obtained with the subject's permission from a copy of his or her transcript. The other two measures were maximum-performance tests, which were administered under carefully controlled testing conditions.

The Test of Implied Meanings. The Test of Implied Meanings (TIM) was developed by Sundberg (1966) to measure the ability to understand a person's feelings and intentions on the basis of the manner in which she or he speaks (e.g., tone of voice, emphasis). The TIM includes a recording of 40 statements, 20 spoken by a male and 20 by a female. Each statement is repeated, with a 12-s pause for recording answers between each statement. The content of the statements includes common complaints or problems heard when interviewing patients. The subject is asked to select from four alternatives the implied meaning of each statement, and subjects are encouraged to guess if they do not know the answer. The key for the TIM is based on the implied meaning, which the speaker was trying to convey; however, for two of the items this meaning did not agree with the modal response of the pilot subjects, and thus only 38 of the statements are actually scored on the test. Scores on this test were used as an objective measure of sensitivity.

Oregon General Knowledge Test. The Oregon General Knowledge Test (OGKT) was developed specifically for this project by Antonia Ristorcelli. Designed to test knowledge in a variety of fields, it was used as an objective measure of cultural sophistication. The test consists of 180 multiple choice questions, selected from a much larger pool. On the basis of pilot studies, items that were too difficult or too easy for college samples, or that did not correlate significantly with their respective scales, were eliminated. Two different aspects of cultural sophistication are measured by the OGKT. The first, called traditional/academic, concerns matters with which an educated person is traditionally supposed to be familiar. To measure this type of cultural sophistication, there are 10 scales: (a) classical literature, (b) contemporary literature, (c) poetry, (d) art, (e) classical music, (f) history, (g) geography, (h) science, (i) current events, and (j) etiquette. The other aspect of cultural sophistication, called contemporary popular, concerns knowledge of contemporary American society. To measure this second type of cultural sophistication there are five scales: (a) contemporary music, (b) movies, (c) television shows, (d) sports, and (e) general popular culture. Scores from the two major subtests of the OGKT, traditional/academic and contemporary/popular were used in this study. Coefficient alpha reliability estimates for the two subtests were .88 and .79, respectively. More detailed information about the OGKT can be found in Ristorcelli, Chaplin, and Goldberg (1982).

Table 2 lists the objective measures which were available for each of the traits.

Measures of Consistency

Three different measures of the subjects' self-reported cross-situational consistency were used in this study. Two of them are identical to the measures used by Bem and Allen, whereas the third focused on the consistency of highly specific everyday behaviors. The measure of consistency recommended by Bem and Allen is the ipsatized variance index (IVI). This measure was derived from the CSBS and is the degree to which the subject's responses to the set of items in a scale varies, relative to the variability of the subject's responses to the total item pool. These variabilities are computed after the items scored in the opposite direction are reversed. Thus, each subject had eight ipsatized variance indices, one for each trait, which were computed from the scales on the CSBS.

The other measure of consistency used by Bem and Allen was a direct self-rating of variability (SV). These ratings were obtained from the subjects when they provided the self-ratings on trait level at Time 1. Subjects were asked to respond to the question, "How much do you vary from situation to situation in how X you are?" where X was one of the eight traits, by rating themselves on a 7-point scale ranging from "Not at all" through "Moderately" to "Extremely."

The third measure of consistency, the Consistency Questionnaire (CQ), is an unpublished instrument which has been under development for the past 10 years. Various samples of subjects have been asked to generate behavior–situation items that seem to them to elicit extremely large individual differences in intraindividual consistency. Other samples of subjects have been presented with lists of these items, each posed as a choice between two or three alternative behaviors (e.g., the use of a bathtub versus a shower for bathing). Subjects are asked to indicate which alternative they typically prefer, and then—most important—to indicate the degree of consistency of their behavior in this situation, using a 4-step rating scale ([1] Always, [2] Usually, [3] More often than not, and [4] No real consistency).

In the present version of the CQ, 122 behavior–situation items are grouped in eight clusters: (a) handedness (3 items), (b) food and eating (40 items), (c) grooming and attire (16 items), (d) sleeping (7 items), (3) news and entertainment (19 items), (f) seating habits (17 items), (g) walking and driving (10 items), and (h) miscellaneous habits (10 items). Individual differences in

Table 2
Summary of the Objective Measures Available for Each of the Eight Traits

Trait	Measures
Friendliness	Latency to start conversation*
	Frequency of vocalizations
	Duration of vocalizations
	Smiling in passing score
Conscientiousness	Average ratings of neatness of dress
	Grade point average
	Promptness to arrive at the laboratory
Honesty	None
Sensitivity	Test of Implied Meanings
Assertiveness	Latency to start conversation*
	Frequency of vocalizations
	Duration of vocalizations
	Duration/frequency
Activity level	None
Emotional stability	Eyeblinks during talk minus eyeblinks after talk*
	No. of prompts by group leader*
	Pre- minus post- MAACL anxiety score*
	Pre- minus post- MAACL depression score*
	Pre- minus post- MAACL hostility score*
Cultural sophistication	Average rating of how well-dressed
	Traditional/academic knowledge score (OGKT)
	Contemporary/popular knowledge score (OGKT)
	Grade point average
	Type of magazine chosen

* Scored in the reverse direction

mean item responses across all 122 items range from less than 2.0 to over 3.5. Coefficient alpha reliability estimates of these average consistency values in various samples are well above .90. Unlike the other two measures of consistency, the CQ yields only one global consistency score, rather than scores for each trait.

Procedures

All of the measures were obtained during a 4-month period. The subjects were met as a group at least once each week when their completed questionnaires were collected and new questionnaires were distributed. In addition, during this time subjects were scheduled to participate in the laboratory session and to take the TIM and the OGKT.

Dividing Subjects into High and Low Consistent Groups[3]

The three measures of consistency were used in conjunction with the self-ratings of trait level and mean scores on the CSBS scales to create three methods of dividing the subjects into high and low consistent groups. With all three methods the consistency classification was done so that the high and low consistent groups had essentially the same mean score on the measures of trait level, as was the case in the original Bem and Allen study.

Specifically, the first method (SR/SV) used the subject's self-ratings of trait level and trait variability, and is identical to one of the methods used by Bem and Allen. The subjects were first classified into seven subgroups on the basis of their response on the trait-level scale. Then, within each subgroup the subjects were divided into high and low consistency groups depending on whether their rating of trait variability was above or below the median for the subjects at the same point on the trait scale.

The second method (CSBS/IVI) is identical to the other method used by Bem and Allen. It used the mean scores on the CSBS scales and the ipsatized variance indices. With this method, the pair of subjects with the highest scores on the CSBS scale were classified as a high consistent and a low consistent subject depending on which one had the highest IVI. Then the pair of subjects with the next highest scores were similarly divided, and so on.

[3] In addition to these methods, other approaches to classifying the subjects into high and low consistent groups are possible. For example, the CSBS level measure could be crossed with self-reported variability, and SR could be crossed with the IVI. Also, measures of level and variability were available from the TDA and these could be crossed with each other and with the SR, CSBS, and CQ measures. Of these additional analyses we examined CSBS level with self-report variability, CSBS level with CQ, and TDA level with TDA variability. Results of these analyses are not reported here because they did not change any of the conclusions based on those we do report; they are available in Chaplin (1981) or from the authors.

Table 3
Correlations Among the Three Consistency Measures

Trait	SV and IVI (N = 112)	CQ and IVI (N = 108)	CQ and SV (N = 108)	M
Friendliness	−.04	.23	−.04	.05
Conscientiousness	.42	.08	.07	.19
Honesty	−.13	−.13	−.07	−.11
Sensitivity	.18	−.01	−.03	.05
Assertiveness	−.09	.00	.13	.01
Activity level	.09	−.02	.03	.03
Emotional stability	.21	−.09	−.03	.03
Cultural sophistication	−.01	−.11	−.13	−.08
M	.08	−.01	−.01	.01

Note. IVI = ipsatized variance index; SV = self-ratings of variability; and CQ = Consistency Questionnaire.

The third method (SR/CQ) used the Consistency Questionnaire scores in conjunction with the self-ratings of trait level. With this method the subjects were again divided into seven subgroups and within each subgroup classified as high and low consistent depending on the position of their CQ score relative to the subgroup median.

Each of the methods was repeated eight times, once for each of the eight traits, separately within the male and female subsamples. This was done to facilitate analyzing the data separately for each sex, because on several of the level and variability measures there were substantial differences between the mean scores of the men and women. (Bem and Allen also divided the subjects into high and low consistent groups separately for each sex.)

Analyses

Correlations were computed among and between the self-report, other-report, and objective measures, separately within the high and low consistent groups, and separately for each trait and each method of consistency classification. The correlations were averaged for each of the six sets of data types (self-reports, other-reports, objective measures, Self × Other, Self × Objective, and Other × Objective), and then averaged again across all six sets for each trait, and across all traits within each data type. By comparing these average correlations, it is possible to check the replicability of Bem and Allen's findings, to assess the relative effectiveness of each of the three methods of consistency classification, and to determine if the Bem and Allen finding is particularly applicable to certain traits or certain types of data. All of the analyses reported here were based on the total sample. Analyses performed separately on the male and female subsamples are available in Chaplin (1981). There would be no major changes in the conclusions of this study if the analyses had been reported separately for each sex, because the few differences that were found can be attributed to the substantial decrease in the sample sizes available for these separate analyses.

Results

Agreement Among the Consistency Classification Methods

Before considering the effectiveness of each of the three methods of consistency classification as moderators of other relations, it is important to consider the degree to which they agree about the classification of subjects into high and low consistency groups. Obviously, if the three methods show a high degree of agreement, then the results obtained with each method will be much the same. Indeed, such a finding would be quite important, because it would suggest that consistency is not a method-specific construct.

In fact, however, the three methods of consistency classification did not exhibit high agreement. Table 3 presents the correlations among the three consistency measures. With the exception of the trait of conscientiousness, for which the correlation between the SR and IVI methods was .42, there was little relation among the three measures.

However, these correlations could be misleading, because measures of trait level were used in conjunction with the consistency measures to classify the subjects into high and low consistent groups. To assess the degree of agreement among the three methods of consistency classification, we examined the conditional probabilities that a person classified as consistent on a given trait by one method was classified as consistent on that trait by another method. These probabilities

are presented in Table 4, separately for the males and females to reflect the actual (within-sex) procedures used for subject classification. Again, the three methods of consistency classification were essentially independent.

Consistency as a Moderator Variable

The previous analyses indicate that the classification of subjects into high and low consistent groups is highly method-dependent. To determine if any one method is superior in terms of its effectiveness at moderating trait relations, we must examine the correlations among and between the self-report, other-report, and objective measures for the high and low consistent subjects for each of the eight traits, separately for each method of consistency classification.

Self-Reports of Variability

Table 5 presents the average correlations based on this method for each of the types of data as well as the average correlation across all the measures and the average correlation within each type of data across all eight traits. All of the original correlation matrices on which Tables 5, 6, and 7 are based are available in Chaplin (1981). The average correlations (across all measures) for each trait provide little support for the Bem and Allen hypothesis. For the traits of honesty, sensitivity, assertiveness, and activity level, the size of the difference between the two average correlations was negligible. Moreover, for the traits of emotional stability, cultural sophistication, and friendliness, there was a reversal of the Bem and Allen finding. For

Table 4
Degree of Agreement Among the Three Methods of Consistency Classification

Trait	SV and IVI	CQ and IVI	CQ and SV	*M*
		Males		
Friendliness	.47	.59	.35	.47
Conscientiousness	.59	.53	.44	.52
Honesty	.44	.51	.51	.49
Sensitivity	.56	.45	.55	.52
Assertiveness	.50	.53	.59	.54
Activity level	.41	.41	.39	.40
Emotional stability	.34	.51	.35	.40
Cultural sophistication	.41	.45	.48	.45
M	.47	.50	.46	.48
		Females		
Friendliness	.50	.57	.49	.52
Conscientiousness	.63	.31	.45	.46
Honesty	.54	.49	.59	.54
Sensitivity	.50	.57	.49	.52
Assertiveness	.38	.49	.41	.43
Activity level	.63	.43	.49	.52
Emotional stability	.46	.57	.49	.51
Cultural sophistication	.54	.61	.43	.53
M	.52	.51	.48	.50

Note. These were conditional probabilities that a person classified as consistent by one of the methods would be classified as consistent by the other. The expected probability if the methods result in independent classification is .50. In cases where the conditional relation $p(A/B) \neq p(B/A)$ the values in the table are the average of the two probabilities. In no case was the discrepancy between the two conditional probabilities greater than .03. IVI = ipsatized variance index, SV = self-ratings of variability, and CQ = Consistency Questionnaire.

the trait of conscientiousness, the two groups tied.

It is surprising that the low consistent group achieved a higher average correlation than the high consistent group for the trait of friendliness, because it was this classification method that worked so well for this trait in the Bem and Allen study. To see if we could replicate their result using only the measures they used, we computed the average of the correlations among the self-reports, mother reports, father reports, average peer ratings, peer ratings in the group discussion, latency to converse, frequency of vocalizations, and duration of vocalizations measures. For the high consistent group this average correlation was .14, whereas for the low consistent group it was .22. However, Bem and Allen combined the group discussion, peer rating, frequency, and duration measures into one index in their study. Our use of these measures separately has probably made them less reliable than the Bem and Allen variable, thus attenuating the correlations involving these measures. When we eliminated those measures before averaging the correlations, we obtained an average correlation of .14 for the high consistent group and one of .31 for the low consistent group correlation. Clearly, the result Bem and Allen achieved for the trait of friendliness in their study is not very stable. The bottom rows of Table 5 indicate whether this method of consistency classification is effective with any particular type of data across all of the eight traits. The high consistent group did have a higher average correlation computed across traits than the low consistency group for the self-report measures and between the self- and other-report measures. A reversal was found for correlations among the objective measures, between the self-report and objective measures, and between other-report and objective measures. For correlations among the other-report measures the two groups tied. None of the differences between the average size of the corre-

Table 5
Average Correlations for the Subjects Classified as High and Low Consistent on the Basis of Their Self-Ratings

Trait	Self-reports	Other-reports	Objective measures	Self × Other	Self × Objective	Other × Objective	*M* (across measures)
Friendliness							
High consistent	.40	.17	.03	.16	.06	.09	.21
Low consistent	.32	.36	−.03	.21	.00	.08	.19
Conscientiousness							
High consistent	.52	.19	−.06	.23	.14	.16	.22
Low consistent	.36	.32	.07	.20	.07	.14	.22
Honesty							
High consistent	.29	.23	—	.08	—	—	.16
Low consistent	.31	.08	—	.12	—	—	.13
Sensitivity							
High consistent	.38	.18	—	.09	.09	.01	.15
Low consistent	.20	.19	—	.11	.09	−.02	.14
Assertiveness							
High consistent	.48	.22	−.11	.19	.05	.07	.17
Low consistent	.46	.15	.02	.17	.14	.06	.16
Activity level							
High consistent	.30	.32	—	.27	—	—	.29
Low consistent	.35	.17	—	.11	—	—	.16
Emotional stability							
High consistent	.19	.18	.13	.16	−.01	−.02	.10
Low consistent	.30	.16	.11	.10	.11	.05	.12
Cultural sophistication							
High consistent	.38	.17	.12	.15	.13	.09	.15
Low consistent	.29	.22	.16	.11	.13	.20	.17
M (across traits)							
High consistent	.37	.21	.06	.17	.07	.07	.18
Low consistent	.32	.21	.09	.14	.09	.10	.16

lations for the high and low consistent groups were larger than .05.

Overall, the results obtained with the self-report method of consistency classification offer little support for the Bem and Allen hypothesis. Only for the trait of activity level was there any indication that individuals classified as high consistent achieve higher correlations among measures of this trait than individuals classified as low consistent. Had we fortuitously selected just this trait for investigation, then we would have appeared to confirm the Bem and Allen results. However, when this result is viewed within the context of the entire set of analyses, chance is the most likely explanation for that finding.

Ipsatized Variance Index

The failure to confirm Bem and Allen's results with the self-report classification method may not be of great significance, because Bem and Allen themselves suggested that so simple a method might not be very effective. Instead, they recommended the ipsatized variance index as the most promising method for future investigations. With this method they achieved confirmatory results for the trait of conscientiousness. Table 6 presents the average correlation when the ipsatized variance index was used to classify subjects into high and low consistent groups.

Unfortunately, these results are quite similar to those based on the self-report method. The average correlations computed across all of the measures for each trait are listed in the last column in Table 6. For only two traits, conscientiousness and assertiveness, did the high consistent group have a larger correlation than the low consistent group. For the traits of honesty, sensitivity, activity level, emotional stability, and cultural sophistication the effect was reversed, whereas for the trait of friendliness the high and low consistent groups tied.

In general, the size of the difference between

Table 6
Average Correlations for the Subjects Classified as High and Low Consistent on the Basis of Their CSBS Ipsatized Variance Index

Trait	Self-reports	Other-reports	Objective measures	Self × Other	Self × Objective	Other × Objective	*M* (Across measures)
Friendliness							
High consistent	.25	.34	.03	.17	.06	.08	.18
Low consistent	.45	.22	.01	.20	.06	.20	.18
Conscientiousness							
High consistent	.58	.28	−.05	.30	.08	.19	.27
Low consistent	.29	.24	.05	.11	.10	.13	.16
Honesty							
High consistent	.30	.10	—	.06	—	—	.10
Low consistent	.28	.19	—	.12	—	—	.17
Sensitivity							
High consistent	.25	.23	—	.09	.03	−.05	.14
Low consistent	.35	.14	—	.12	.13	.05	.15
Assertiveness							
High consistent	.42	.23	−.06	.16	.15	.09	.17
Low consistent	.53	.14	−.01	.22	−.04	.04	.15
Activity level							
High consistent	.32	.22	—	.21	—	—	.23
Low consistent	.34	.29	—	.19	—	—	.25
Emotional stability							
High consistent	.28	.15	.10	.15	−.03	.01	.10
Low consistent	.26	.21	.12	.13	.15	.02	.13
Cultural sophistication							
High consistent	.32	.17	.05	.08	.09	.10	.12
Low consistent	.35	.21	.21	.16	.18	.19	.20
M							
High consistent	.34	.22	.04	.15	.06	.08	.16
Low consistent	.36	.21	.11	.16	.06	.09	.17

the average correlations for the two groups was negligible. However, the largest of these differences, (.27 versus .16) was for the trait of conscientiousness, and it was in the predicted direction. Thus, in this instance one of Bem and Allen's findings was (weakly) replicated. These results also occurred when we examined only those variables that Bem and Allen used in their original study, self-reports, mother reports, father reports, peer reports, and neatness of dress. The average of the correlations among these measures in the present sample was .38 for the high consistent and .27 for the low consistent groups. Omitting the neatness measure, as did Bem and Allen, the correlations were .44 and .36 for the high and low consistent groups, respectively. However, none of these effects were as strong as those reported by Bem and Allen, and when viewed within the context of the entire set of analyses, these findings present even less cause for optimism.

The bottom rows of Table 6 show that the ipsatized variance index was not effective with any particular type of data. Only in the case of the other-report measures were the correlations for the high consistent group larger than those for the low consistent group. Thus, except perhaps for the trait of conscientiousness, the ipsatized variance index method does not appear to be an effective way of increasing the correlations among trait measures.

Consistency Questionnaire

Because neither of the trait-specific consistency measures used by Bem and Allen appear to be generally effective as moderators, let us turn to the method that uses a general measure of consistency for the classification of subjects into high and low consistent groups. The correlations based on this method, which combined the self-report measure of trait level with the Consistency Questionnaire scores in making the classifications, are presented in Table 7. Whereas the high consistent group had slightly higher average correlations across the measures of the eight traits than the low consistent group for all traits except cultural sophistication and sensitivity, the size of these differences was, in all cases, negligible. This same consistent but weak tendency for the high consistent groups to have higher average correlations was found for the average correlations across the eight traits for each combination of data, except that between the other-report and objective measures, for which the high and low consistent groups tied. Overall, however, the results obtained by using this method of consistency classification provided no more support for the Bem and Allen hypothesis than did either of the two methods that were used in the original study.

Discussion

We had anticipated that the results of this study would allow us to clarify several issues raised by the Bem and Allen findings. For example, we assumed that the manner in which one obtains information about personality (i.e., self-report, other-report, and objective observation) would have an impact on the effectiveness of self-reported consistency as a moderator variable (cf. Block, 1977; Golding, 1978; Olweus, 1980). Likewise, we hoped that the construct of consistency would be clarified as a result of the comparison of the three methods of consistency classification (cf. Lay, 1977), and we hypothesized that self-reported consistency might be differentially effective as a moderator variable for different traits (cf. Kenrick & Stringfield, 1980). Unfortunately, our results are disappointingly easy to summarize: For each of the three methods of consistency classification and for each of the eight traits, we were able to replicate the Bem and Allen finding about as often as we failed to replicate it. Moreover, in those cases where the high consistent subjects achieved higher average correlations than the low consistent subjects, the size of that difference was negligible.

However, in light of these overall negative results it is quite surprising that we obtained an attenuated version of the same finding as did Bem and Allen by using the ipsatized variance method for the trait of conscientiousness. As shown in the summary table (Table 8), not only did the high consistent group have a larger average correlation (.27) than the low consistent group (.16), but the size of the difference between these two correlations was the second largest obtained in all of the comparisons. Within the context of

Table 7
Average Correlations for the Subjects Classified as High and Low Consistent on the Basis of Their Scores on the Consistency Questionnaire

Trait	Self-reports	Other-reports	Objective measures	Self × Other	Self × Objective	Other × Objective	*M* (across measures)
Friendliness							
High consistent	.31	.31	.15	.20	.03	.09	.19
Low consistent	.43	.23	−.15	.17	.09	.08	.17
Conscientiousness							
High consistent	.44	.29	−.03	.24	.06	.14	.23
Low consistent	.45	.21	.06	.20	.17	.20	.22
Honesty							
High consistent	.33	.19	—	.13	—	—	.18
Low consistent	.24	.11	—	.07	—	—	.11
Sensitivity							
High consistent	.23	.20	—	.08	.23	.07	.14
Low consistent	.39	.16	—	.15	−.08	−.05	.15
Assertiveness							
High consistent	.47	.22	.04	.16	.10	.09	.17
Low consistent	.49	.15	−.18	.21	.10	.05	.16
Activity level							
High consistent	.36	.27	—	.22	—	—	.26
Low consistent	.29	.22	—	.16	—	—	.20
Emotional stability							
High consistent	.38	.21	.15	.13	.10	.02	.14
Low consistent	.16	.11	.02	.13	.13	.01	.08
Cultural sophistication							
High consistent	.38	.16	.16	.11	.14	.12	.15
Low consistent	.23	.21	.12	.11	.15	.18	.17
M (across traits)							
High consistent	.36	.23	.12	.16	.10	.09	.18
Low consistent	.34	.18	.00	.15	.09	.09	.16

the rest of our results, chance would be the easiest explanation for this finding. Nonetheless, since this result replicates the one reported by Bem and Allen and subsequently replicated by Mischel and Peake (1981, 1982), other explanations are possible.

One hypothesis is that whereas Bem and Allen made some effort to ensure that their high and low consistency subsamples did not differ in mean trait level (as assessed by self-reports), they never investigated whether these two groups may not have differed substantially in the variances of one or more of those variables used to demonstrate the moderator effect (e.g., parent reports, peer ratings). Indeed, Bem and Allen never even reported the variances of any of their criterion variables, sharing a long tradition of reportorial negligence with many previous investigators of moderator effects (e.g., Cowden, 1969; Ghiselli, 1956, 1960, 1963; Hoyt & Norman, 1954; Stagner, 1933). Yet, if the two samples generally differ in their variances, any "moderator" effect might simply be an artifact caused by comparing a group containing many high and low values with one containing predominantly middle values. That is, if the potential moderator variable happens to select subgroups that differ in their variance on at least one of the variables being correlated, one would expect a higher correlation in the sample with the larger variance(s) than in the one with the smaller variance(s). In the present study, because all moderator effects were so small, one would not expect—nor did we find—large differences in variances between our high and low consistency subsamples.[4] We would expect, however, that in those studies where moderator effects were substantial (e.g., Bem & Allen, 1974; Mischel & Peake, 1982) that differences in variance may have played a crucial role.

[4] Tables reporting the means and variances for all of our measures of each trait are available from the authors.

Table 8
Summary Table: Differences Between the Average Correlations in the High Consistency and the Low Consistency Groups

Trait	Consistency classification method			
	SR/SV	CSBS/IVI	SR/CQ	*M*
Friendliness	.02	.00	.02	.01
Consientiousness	.00	.11	.01	.04
Honesty	.03	−.07	.07	.01
Sensitivity	.01	−.01	−.01	.00
Assertiveness	.01	.02	.01	.01
Activity level	.13	−.02	.06	.06
Emotional stability	−.02	−.03	.06	.00
Cultural sophistication	−.02	−.08	−.02	−.04
M	.02	−.01	.02	.01

Note. The values in this table are differences between two average correlations; differences with positive signs are in the direction predicted by Bem and Allen (1974). SR = global self-ratings; SV = self-reported variability, CSBS = Cross-Situational Behavior Survey; IVI = ipsatized variance index; CQ = Consistency Questionnaire.

One of the more discouraging findings from the present study was that the three methods of consistency classification showed such low convergence (see Tables 3 and 4). With the exception of the trait of conscientiousness, for which the self-report and IVI measures correlated .42, the correlations were negligible. Likewise, the Consistency Questionnaire score did not correlate with any of the self-report or ipsatized variance index measures.

Our general failure to replicate the Bem and Allen findings precluded an assessment of differential effects among the three types of personality data in their susceptibility to moderation by self-reported consistency. However, as usual (Block, 1977; Epstein, 1979; Goldberg, Norman, & Schwartz, 1980) the correlations among and between the self- and other-report measures for both the high and low consistent subjects were higher than the analogous correlations involving the objective measures. This finding is certainly due, in part, to the low reliability of such objective measures (Epstein, 1977, 1979). Moreover, there is an inherent ambiguity in the meaning of such measures when they are used to assess global personality dimensions (Golding, 1977, 1978). Of the three types of data we investigated, these measures suffer most strongly from the "equivalence class" problem discussed by Bem and Allen as one reason for the apparent low level of personality consistency.

One question raised by the Bem and Allen study is whether consistency should best be viewed as a trait-specific, as opposed to a general, characteristic. Traditionally, searches for groups of individuals who achieve higher correlations than usual among measures of a particular trait have used global (across trait) measures of consistency (e.g., Campus, 1974; Fiske, 1957; Goldberg, 1978). These approaches have met with only limited success, and Bem and Allen's conceptualization of consistency as a trait-specific characteristic has been regarded as potentially more useful. However, the evaluation of these two views is fraught with methodological difficulties. Specifically, self-reports of behavioral consistency are not appropriate for drawing conclusions about the tendency for individuals to be consistent across traits, because assumptions about this tendency are built into each particular consistency classification method. The ipsatized variance index, because it is the ratio of a subject's variability on a particular trait scale to his or her total consistency on all the scales, leads inherently to independence of classification across traits. For example, in the present study the across-trait correlations based on the IVI measure ranged from −.33 to .35, and 68% of them were negative in sign. On the other hand, the across-trait correlations based on the self-ratings of consistency averaged .35 and ranged from .05 to .61, a finding easily attributable to individual differences in extremeness response bias on

the rating scales (Hamilton, 1968). Finally, for the Consistency Questionnaire, classification agreement across traits is mandated, because this questionnaire yields only one global consistency score.

In summary, the findings from our study certainly suggest that the excitement that followed the publication of the Bem and Allen study now needs to be tempered. Not only did we fail to find any generalization of their findings to a larger and more representative sample of traits, but we even failed to replicate their findings for one of the two traits they investigated (friendliness). Moreover, in the one case where we weakly replicated their results, we can not rule out the possibility that this was due to the confounding of trait level and trait variability. What, then, are the implications of these findings for the general conclusions that have been drawn from the original Bem and Allen report?

One major conclusion was that self-reported consistency is an effective moderator variable for enhancing the correlations among measures of personality traits. This assumption induced a plethora of investigators to incorporate self-report measures of consistency into their studies (e.g., LaPointe & Harrel, 1978; Lippa & Mash, 1981; Tunnell, 1980; Turner, 1978; Turner & Gilliam, 1979; Vestewig, 1978; Zanna, Olson, & Fazio, 1980), because the discovery of a general moderator variable would represent a major advance in the field of personality assessment (Fiske, 1957; Ghiselli, 1956). The results of our study clearly undermine this conclusion, and thus we must add our present results to the long list of failures in the search for moderators (e.g., Brown & Scott, 1966, 1967; Goldberg, 1972; Kellogg, 1968; Stricker, 1966; Wallach & Leggett, 1972).

References

Amelang, M., & Borkenau, P. (1982). *In search of persons with traits: Intra-individual variability, moderator scales, and differential predictability.* Unpublished manuscript, Department of Psychology, University of Heidelberg.

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

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

Brown, F. G., & Scott, D. A. (1966). The unpredictability of predictability. *Journal of Educational Measurement, 3,* 297–301.

Brown, F. G., & Scott, D. A. (1967). Differential predictability in college admissions testing. *Journal of Educational Measurement, 4,* 163–166.

Campus, N. (1974). Transituational consistency as a dimension of personality. *Journal of Social Psychology, 29,* 593–600.

Cattell, R. B. (1957). *Personality and motivation structure and measurement.* Hudson, NY: World Books.

Chaplin, W. F. (1981). *A closer look at the use of measures of consistency as moderators of criterion relationships: An attempted replication and extension of the Bem and Allen study.* Doctoral dissertation, University of Oregon.

Cheek, J. M. (1982). Aggregation, moderator variables, and the validity of personality tests: A peer rating study. *Journal of Personality and Social Psychology, 43,* 1254–1269.

Cowden, J. E. (1969). Prediction enhancement through the use of moderator variables. *Journal of Consulting and Clinical Psychology, 33,* 621–624.

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

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

Fiske, D. W. (1957). The constraints on intra-individual variability in test responses. *Educational and Psychological Measurement, 17,* 317–337.

Ghiselli, E. E. (1956). Differentiation of individuals in terms of their predictability. *Journal of Applied Psychology, 40,* 374–377.

Ghiselli, E. E. (1960). The prediction of predictability. *Educational and Psychological Measurement, 20,* 3–8.

Ghiselli, E. E. (1963). Moderating effects and differential reliability and validity. *Journal of Applied Psychology, 47,* 81–86.

Goldberg, L. R. (1972). Student personality characteristics and optimal college learning conditions: An extensive search for trait-by-treatment interaction effects. *Instructional Science, 1,* 153–210.

Goldberg, L. R. (1978). The reliability of reliability: The generality and correlates of intra-individual consistency in responses to structured personality inventories. *Applied Psychological Measurement, 2,* 269–291.

Goldberg, L. R. (1981). Unconfounding situational attributions from uncertain, neutral, and ambiguous ones: A psychometric analysis of descriptions of oneself and various types of others. *Journal of Personality & Social Psychology, 41,* 517–552.

Goldberg, L. R., Norman, W. T., & Schwartz, E. (1980). The comparative validity of questionnaire data (16PF scales) and objective test data (O–A Battery) in predicting five peer-rating criteria. *Applied Psychological Measurement, 4,* 183–194.

Golding, S. L. (1977). Method variance, inadequate constructs, or things that go bump in the night?

Journal of Multivariate Behavioral Research, 12, 89–98.

Golding, S. L. (1978). Toward a more adequate theory of personality: Psychological organizing principles. In H. London (Ed.), *Personality: A new look at methatheories.* New York: Wiley.

Hamilton, D. L. (1968). Personality attributions associated with extreme response style. *Psychological Bulletin, 69,* 192–203.

Hoyt, D. P., & Norman, W. T. (1954). Adjustment and academic predictability. *Journal of Counseling Psychology, 1,* 96–99.

Kellogg, R. L. (1968). The Strong Vocational Interest Blank as a differential predictor of engineering grades. *Educational and Psychology Measurement, 28,* 1213–1217.

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

LaPointe, K. A., & Harrel, T. H. (1978). Thoughts and feelings: Correlational relationships and cross-situational consistency. *Cognitive Therapy and Research, 2,* 311–322.

Lay, C. (1977). Some notes on the concept of cross-situational consistency. In D. Magnusson and N. S. Endler (Eds.), *Personality at the crossroads: Current issues in interactional psychology.* Hillsdale, NJ: Erlbaum.

Lippa, R., & Mash, M. (1981). The effects of self-monitoring and self-reported consistency on the consistency of personality judgments made by strangers and intimates. *Journal of Research in Personality, 15,* 172–181.

Mischel, W., & Peake, P. K. (1981). In search of consistency: Measure for measure. In M. P. Zanna, E. T. Higgins, & C. P. Herman (Eds.), *Consistency in social behavior: The Ontario Symposium of Personality and Social Psychology* (Vol. 2, pp 187–207). Hillsdale, NJ: Lawrence Erlbaum.

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

Norman, W. T., (1963). Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *Journal of Abnormal and Social Psychology, 66,* 564–583.

Olweus, D. (1980). The consistency issue in personality psychology revisited—with special reference to aggression. *British Journal of Social and Clinical Psychology, 19,* 377–390.

Ristorcelli, A., Chaplin, W. F., & Goldberg, L. R. (1982). *The development and validation of a test of cultural sophistication: The Oregon General Knowledge Test.* Unpublished manuscript, Department of Psychology, University of Oregon.

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

Snyder, M., & Monson, T. C. (1975). Persons, situations, and the control of social behavior. *Journal of Personality and Social Psychology, 32,* 637–644.

Stagner, R. (1933). The relation of personality to academic aptitude and achievement. *Journal of Educational Research, 26,* 648–660.

Stricker, L. J. (1966). Compulsivity as a moderator variable: A replication and extension. *Journal of Applied Psychology, 50,* 331–335.

Sundberg, N. D. (1966). A method for studying sensitivity to implied meanings. *Gawein, 15,* 1–8.

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

Tunnell, G. (1980). Intraindividual consistency in personality assessment: The effect of self-monitoring. *Journal of Personality, 48,* 220–231.

Turner, R. G. (1978). Consistency, self-consciousness, and the predictive validity of typical and maximal personality measures. *Journal of Research in Personality, 12,* 117–132.

Turner, R. G., & Gilliam, B. J. (1979). Identifying the situationally variable subject: Correspondence among different self-report formats. *Applied Psychological Measurement, 3,* 361–369.

Underwood, B., & Moore, B. S. (1981). Sources of behavioral consistency. *Journal of Personality and Social Psychology, 40,* 780–785.

Vestewig, R. (1978). Cross-response mode consistency in risk taking as a function of self-reported strategy and self-perceived consistency. *Journal of Research in Personality, 12,* 152–163.

Wallach, M. A., & Leggett, M. I. (1972). Testing the hypothesis that a person will be consistent: Stylistic consistency vs. situational specificity in size of children's drawing. *Journal of Personality, 40,* 309–330.

Young, P. T. (1973). *Emotion in man and animals* (2nd ed.). New York: Krieger.

Zanna, M. P., Olson, J. M., & Fazio, R. H. (1980). Attitude–behavior consistency: An individual difference perspective. *Journal of Personality and Social Psychology, 38,* 432–440.

Zuckerman, M., & Lubin, B. (1965). Normative data for the Multiple Affect Adjective Checklist. *Psychological Reports, 16,* 438.

Received March 3, 1983
Revision received July 19, 1983 ■