Is a Behavioral Measure the Best Estimate of Behavioral Parameters? Perhaps Not.

George S. Howard, Scott E. Maxwell, Richard L. Wiener, Kathy S. Boynton, and William M. Rooney
University of Houston

In many areas of psychological research various measurement procedures are employed in order to obtain estimates of some set of parameter values. A common practice is to validate one measurement device by demonstrating its relationship to some criterion. However, in many cases the measurement of that criterion is less than a perfect estimate of true parameters. Self-report measures are often validated by comparing them with behavioral measures of the dimension of interest. This procedure is only justifiable insofar as the behavioral measure represents an accurate estimate of population parameters. Three studies, dealing with the assessment of assertiveness, students' in-class verbal and nonverbal behaviors, and a number of teacher-student in-class interactions, tested the adequacy of behavioral versus self-report measures as accurate estimates of behavioral parameters. In Studies 2 and 3 self-reports were found to be as good as behavioral measures as estimates of behavioral parameters, while Study 1 found self-reports to be significantly superior.

One of the oldest debates in psychology has dealt with the adequacy of various types of data. This problem often involves a choice between self-report and behavioral data. The controversy dates back to the clash between Titchener's structuralist school of thought and James' functionalist approach to research. Methodologically, the issue was crystallized as the choice between the use of introspective approaches to knowledge versus the implementation of various extraspective techniques. The more behavioral extraspective approaches clearly won the battle for dominance in American psychology. However, since every self-report is, of necessity, somewhat introspective, both traditions are still represented in contemporary empirical approaches to knowing humans.

The status of self-report techniques in modern research is clearly that of a second-class citizen. Critiques of self-report approaches, representing detours on the road to a truly rigorous scientific discipline, are ubiquitous (e.g., Fiske, 1978; Nisbett & Wilson, 1977). Researchers are advised to employ self-reports only if no behavioral index of a construct exists, such as with dogmatism, or if behavioral measures are too difficult or too costly to obtain (cf. Campbell, Dunnette, Lawler, & Weick, 1971). Another related practice, revealing suspicion of subject self-reports, can often be observed when both behavioral and self-report indices of change on a particular dimension are obtained. Researchers often compare these two indices and attribute differences either to problems associated with the self-report measure or to postulation of a concept such as attitudinal lag. This latter interpretation is equally demeaning of self-reports because it contends that "real" (behavioral) change has occurred but that the subject has not yet appre-

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 4, No. 3 Summer 1980 pp. 293–311

ciated that change. In either case, researchers assume that the behavioral measure provides accurate estimates of behavioral parameters and that behavioral/self-report differences can be considered to reflect difficulties associated with the self-report measure.

Why have psychologists become suspicious of introspective (here referring to self-report approaches, not the classical method of introspection) research methods? A line of research has identified a set of contaminants that are referred to collectively as subject response style effects (Millham & Jacobsen, 1978). These contaminants include subject acquiesence, social desirability, memory distortion, selective perception, and others which are assumed to distort and sometimes to invalidate self-report measures. Although this paper recognizes that subject response-style effects are not trivial concerns for proponents of self-report measures, it adopts the position that any demonstration of contamination of self-reports must be weighed against the known contaminants of behavioral measures. That is, the determination of the superiority of a behavioral index relative to a self-report measure should be an empirical consideration—and not determined simply by fiat. In that context, one might consider the somewhat paradoxical-sounding question that the three studies reported herein address, "Is a behavioral measure (relative to a self-report index) the better estimate of behavioral parameters?"

Before considering the adequacy of behavioral measures, some attention should be directed to potential sources of invalidity of behavioral measures. The following is a nonexhaustive listing of some potential sources of invalidity of behavioral estimates:

1. *Method variance* (cf. Campbell & Fiske, 1959)—that portion of variance in the dependent measure that is attributable to the specific method of measurement employed, i.e., different methods may result in systematically different measurements.
2. *Situation variance* (cf. Bem & Allen, 1974)—that portion of variance in the dependent measure that is attributable to the particular situation in which measurements are obtained.
3. *Natural variability within a given method and situation* (cf. Cronbach, Gleser, Nanda, & Rajaratnam, 1972)—that portion of variance in the dependent measure that is attributable to inherent unreliability of a given person in a particular situation, i.e., the same person in the same situation may behave differently on different occasions.
4. *Obtrusive measurement variance* (cf. Webb, Campbell, Schwartz, & Sechrest, 1966)—that portion of variance in the dependent measure that is attributable to the intrusion of the experimenter into the natural environment.
5. *Rater variance* (cf. Campbell & Stanley, 1963)—that portion of variance in the dependent measure that is attributable to the peculiarities and unreliability of raters' judgments.

Therefore, Y_{ijklmn} would be the score of the i^{th} subject obtained from the j^{th} method in the k^{th} situation with obtrusiveness l, and the m^{th} rater.

A mathematical description of these sources of contamination of behavioral measures would be:

$$Y_{ijklmn} = U_i + M_j + S_k + O_l + R_m + e_n \qquad [1]$$

where

U_i is the universe score of the i^{th} subject (cf. Cronbach et al., 1972),
M_j is the effect of the j^{th} method,
S_k is the effect of the k^{th} situation,
O_l is the effect of the l^{th} obtrusiveness,
R_m is the effect of the m^{th} rater, and
e_n reflects natural variability, i.e., variability due to the n^{th} occasion of measurement.

All of the aforementioned sources of contamination are not present in every behavioral measure;

and with appropriate controls, the potential exists for eliminating each of these sources of invalidity. As more of these sources of contamination are controlled, the question posed in this article becomes moot—extensive, unbiased behavioral estimates would be accurate estimates of behavioral parameters by definition. However, inspection of the research literature reveals very few instances where more than one or two of these five contaminants are adequately controlled. The mathematical description is presented to reemphasize that if a researcher is interested in assessing criterion values that generalize across methods, situations, raters, temporal fluctuations, and obtrusiveness effects, then consideration of the impact of these influences upon any measurement is appropriate. However, in some areas of psychological research, there is no attempt to generalize across these dimensions. For example, the behavioral ecologist typically does not wish to generalize across situations, and so what is one researcher's source of contamination can be another researcher's variable of interest.

Much of modern psychology has become overly enamoured with behavioral measures. Researchers often forget that any measure, whether behavioral or self-report, represents an estimate of some construct parameters. Given the existence of uncontrolled sources of contamination present in both behavioral and self-report measures, to arbitrarily select one over the other as a more appropriate estimate of behavioral parameters is inappropriate. The possibility exists that in some instances where discrepancies were found between behavioral measures and self-report measures, the differences might have represented the superiority, rather than the inferiority, of the self-report measures.

Perhaps this argument can be better appreciated by referring to a measure's strengths rather than its susceptibility to contamination. Suppose, for example, interest is in an individual's level of assertiveness. An unobtrusive behavioral measure of the person's reaction to an unreasonable set of requests made in a telephone call by the experimenter's confederate, such as McFall and Twentyman's (1973) request to borrow a subject's class notes before an exam, has the strength of probably not being contaminated by subject response-style influences. However, suppose the subject, although normally very assertive, has recently passed a test because another student loaned him/her a set of notes. Perhaps his/her recent similar experience will prompt the subject to *want* to "bend over backwards" to try to help the confederate, especially since the subject had just profited by another student's help. The subject would be graded, quite inaccurately, as unassertive. In this example, the behavioral measure might lack the strength of a self-report index, since the behavioral index is specific to only one situation. If the behavioral measure could be taken over a range of situations and tasks, its strength in this regard would have been increased.

The three studies in this paper report data on the accuracy of behavioral versus self-report measures. The first study compared three standard behavioral measures of assertiveness with a standard self-report index in predicting behavioral parameters. Three analyses were performed that compared the self-report with one of the behavioral measures. In all three cases the self-report measure correlated significantly more highly with the criterion measure than did the noncriterion behavioral measure. It should be noted that *all* potential sources of contamination among the three behavioral measures in this study were purposely left uncontrolled. Hence, the lack of similarity among results on the behavioral measures is not at all surprising. The second and third studies provide somewhat more sensitive comparisons between behavioral and self-report estimates by controlling for several sources of invalidity in behavioral measures to provide for a more compelling index of the criterion.

STUDY 1

Spurred by the increased interest in assertiveness as a component of mental health, considerable attention has been devoted to its accurate measurement. To date, there have been at least 10 self-report measures of assertiveness developed, e.g., the Action Situation Inventory (Friedman, 1971); the Conflict Resolution Inventory (McFall & Lillesand, 1971); the Rathus Assertiveness Scale (Rathus, 1973). In addition, several non-self-report indices have been developed. The non-self-report indices fall into three general categories:

1. In vivo unobtrusive behavioral ratings (cf. Bach, Lowery, & Moylan, 1972; Hersen, Eisler, & Miller, 1974; Kazdin, 1974; Weinman, Gelbert, Wallace, & Post, 1972);
2. Ratings by others who regularly observe subjects' in vivo behavior (cf. Arkowitz, Lichtenstein, McGovern, & Hines, 1975; Longin & Rooney, 1975; Martorano, 1973; Serber & Nelson, 1971); and
3. Role-play of standardized tasks (cf. Friedman, 1971; Gulanick, Howard, & Moreland, 1976; Shoemaker & Paulson, 1976).

The typical procedure for studies of assertiveness involves collecting one self-report and one non-self-report measure of subjects' assertiveness to obtain two perspectives on the research question of interest. However, of late several studies have employed multiple measures of assertiveness in the same study. In these instances (cf. Green, Burkhart, & Harrison, 1979) low to moderate intercorrelations among measures seem to be the rule. The present study selected a standard self-report measure and one non-self-report index from each of the three aforementioned behavioral categories. Study 1 treated the non-self-report measures as behavioral measures of assertiveness. The purpose of this investigation was to determine if a self-report measure is superior to a behavioral measure in predicting a subjects' assertiveness scores on independent estimates of behavioral parameters (i.e., the average of the other two behavioral measures).

Method

Subjects

Complete data sets were obtained for 35 (16 male, 19 female) undergraduate students enrolled in introductory psychology classes at a large southwestern university. Subjects received class credit for their participation in the experiment.

Instruments

The College Self-Expression Scale (CSES). The CSES (Galassi, Delo, Galassi, & Barstein, 1974), administered as the self-report measure of assertiveness, is a 50-item scale which utilizes a 5-point Likert-type format on which subjects indicate the frequency that they engage in each behavior. The items sample a variety of assertive situations and the scale has been found to have good reliability and validity (Galassi, Hollandsworth, Radecki, Gay, Howe, & Evans, 1976).

The Interpersonal Behavior Inventory (IBI). The IBI is a modified form of the CSES rewritten to allow for the rating of subjects by others in their social system. The IBI was to be completed by two acquaintances (one male and one female) chosen by the subject. As with the CSES the total assertiveness score was the sum of reported frequency ratings over a variety of assertive situations.

The Behavioral Roleplay Test (BRT). The BRT consists of six interpersonal situations that were presented to each subject via audiotape. Each situation requiring an assertive response was first described and then followed by a line of dialogue (e.g., "May I help you," "Well, I really had a nice time tonight") to which the subject was asked to respond aloud. Subjects' spoken responses to the situations were recorded using a second audiotape recorder. The subjects' re-

sponses were then reordered and recorded on still another tape. Next, trained judges rated these responses with respect to their effectiveness on a 7-point rating scale. Performance scores were obtained by summing the judges' ratings over the six situations. Two undergraduate psychology students served as raters of the audiotapes. First, the raters learned about the properties of an effective response and then practiced rating taped responses from previous research. After demonstrating adequate reliability in practice sessions, the judges rated the responses in this study. The judges listened to the tapes together but rated them independently. Interrater reliabilities for the six situations were .89, .81, .94, .82, .78, and .95.

An in vivo measure of assertiveness (IVA). The IVA, developed by McFall and Twentyman (1973), was also included. A confederate, posing as a classmate, telephoned the subjects and went through a standardized hierarchy of requests concerning an upcoming test (e.g., borrowing the subject's notes 1 week prior to the exam, borrowing the subject's notes until after the exam). The number of the request to which the subject finally gave a refusal response was used as the assertiveness score for that subject. After the last request, the confederate debriefed the subject and explained the real purpose of the call.

Procedure

Subjects were first given a general description of the study along with the requirements for participation. Those students who volunteered to participate completed the CSES, responded to the BRT tapes, and finally received two copies of the IBI along with preaddressed, stamped envelopes in which the questionnaires were to be returned. Subjects were instructed to give one copy of the IBI to a male acquaintance and one to a female acquaintance who knew them well enough to be able to complete the form. Approximately 10 days before the next introductory psychology class exam, subjects were called by a confederate for the IVA assessment.

Data Analysis

Two different data analysis approaches investigated the adequacy of the three behavioral measures relative to the self-report measure. The purpose of both approaches was to assess the extent to which the performance of these behavioral measures was adversely affected by the presence of the previously mentioned sources of contamination.

The first approach involved three pairwise comparisons of correlation coefficients. In all three cases, the average of two of the three behavioral measures constituted a criterion variable, with the remaining behavioral measure and the self-report measure as predictors. The average of the two behavioral measures was obtained by first converting raw scores to z-scores, and then averaging the z-scores, to avoid an artificial weighting that would have otherwise simply reflected the variability of a measure. Thus, this approach examined the relative accuracy of a behavioral measure versus the self-report measure in predicting the average of the other two behavioral measures, representing the best independent estimate of true behavioral parameters of assertiveness available in this study.

Results

Table 1 presents correlations of the behavioral averages with both the self-report measure and the other behavioral measure. For two of the three criteria the self-report measure correlated significantly more highly with the criterion measure than did the noncriterion behavioral measure, as can be seen by the values obtained from the Hotelling-Williams test (see Darlington, 1975). The phone call (IVA) measure was dramatically poorer than the CSES in predicting the average of the other two behavioral measures. The BRT was also considerably poorer than the CSES, while the IBI was only slightly poorer. The relatively slight advantage of the CSES over the IBI is probably due in part to the fact that neither correlated highly with the IVA

Table 1
Correlations of Behavioral and Self-Report Measures

Criterion	Behavioral Prediction	Self-Report Prediction	Hotelling-Williams
BI plus B2	.23	.67	2.75**
BI plus B3	.35	.62	1.86*
B2 plus B3	.30	.48	1.41

**$p < .01$ (1-tailed); *$p < .05$ (1-tailed)
Note. BI = Interpersonal Behavior Inventory; B2 = Behavioral Roleplay Test; B3 = In Vivo Assertiveness.

variable, which forms one-half of the criterion in that comparison. Thus, the best prediction of a behavioral average was in all three cases obtained from the self-report measure rather than from the remaining behavioral measure. These data suggest that these behavioral measures are strongly affected by sources of contamination, since prediction of estimates of behavioral parameters was actually worse when made from the information contained in a single behavioral measure than from the information in a self-report measure.

The second approach utilized confirmatory factor analysis (Jöreskog, 1978) to investigate the correlation matrix for the three behavioral measures and the self-report measure, as shown in Table 2. A model specifying one factor with no restrictions on factor loadings was found to provide a very adequate fit ($\chi^2 = .146$, $df = 2$, $p > .50$). Factor loadings and communalities appear in Table 3. The fact that a single factor model fit the data quite well suggests that these four measures shared a single common underlying dimension, which might be labeled general assertiveness. In addition, the factor loadings in Table 3 reveal that by far the best indicator of the assertiveness construct identified here was the self-report measure (CSES), which correlated .97 with the assertiveness construct. The next best measure was provided by the IBI, with the BRT and the IVA providing substantially lower correlations. Only for the self-report measure was the common variance larger than the unique variance.

Discussion

Both data analysis approaches failed to demonstrate convergent validity for the behavioral measures. The self-report measure, on the other hand, appeared to provide a much better measure of assertiveness. Since the behavioral mea-

Table 2

Measure	Correlations Among the Four Assertiveness Measures		
	BI	B2	B3
BI			
B2	0.31*		
B3	0.15	0.22	
SR	0.64***	0.45**	0.31*

***$p < .001$ (1-tailed); **$p < .01$ (1-tailed); *$p < .05$ (1-tailed)
Note. BI = Interpersonal Behavior Inventory; B2 = Behavioral Roleplay Test; B3 = Phone Call; SR = CSES Self-Report

sures investigated here suffer from all of the previously mentioned sources of contamination, the present data do not suggest that behavioral measures are necessarily inferior to self-report measures. Nevertheless, the data convincingly demonstrate that the self-report measures are not necessarily inferior to behavioral measures, and that, in fact, self-report measures may evidence considerable superiority.

One possible explanation of these results is that the various measures have different reliabilities, which would affect their validity coefficients. All other things being equal, reliability increases monotonically with the number of observations (e.g., items, samples, ratings; cf. Epstein, 1979). If all single observations (whether self-report item or behavioral observation or judged rating) have the same validity, then the most valid measure would be the measure which had the most items (or observations, or ratings, and so forth). Inspection of Table 3 reveals that the measure with the lowest factor loading was the IVA, which was based upon a single observation; the BRT had the next lowest loading and was based upon six situations; while the CSES and IBI, which were based upon 50 items, demonstrated the highest factor loadings. Although the data are consistent with this reliability explanation, there are various alternative explanations possible. For example, any single self-report rating may be more valid than a single behavioral observation. In any event, in this study there was a demonstration of the superiority of a self-report measure, whether that superiority was due to either the scale length or the validity of individual observations. Since single behavioral observations are often more difficult and more costly to obtain than are multiple items from a self-report measure, superior validity coefficients for self-report measures (as in the present study) might be obtained in a very cost-effective manner.

The pattern of intercorrelations among the measures of assertiveness in this study is completely consistent with the patterns found in previous investigations which employed the same or similar measures (e.g., Green, Burkhart, & Harrison, 1979). Therefore, when measuring assertiveness and leaving all potential sources of contamination of behavioral measures uncontrolled, there is reason to believe that self-report estimates are superior to behavioral estimates of behavioral parameters. Leaving all sources of contamination uncontrolled maximized the possibility of demonstrating an instance in which a self-report index would be superior to a behavioral measure. Having made that point, Study 2 considered a situation in which behavioral estimates would possess the maximum possibility of demonstrating superiority over a self-report estimate: where the effects of all sources of contamination of behavioral measures, save one, were controlled. (As noted earlier, in the case in which all sources of contamination are controlled, the behavioral estimates would, by definition, have to equal the behavioral parameters and be a perfect predictor. Unfortunately, a strong case could be made that that state of affairs extremely rarely, if ever, occurs.)

STUDY 2

Initially, it was hoped that assertiveness would also be investigated in Study 2, but there seemed

Table 3
Factor Loadings and Communalities

Variable	Loading	Communality
BI	0.65	0.43
B2	0.46	0.21
B3	0.32	0.10
SR	0.97	0.94

to be no way to gain the necessary control over all but one of the contaminants of behavioral measures. Consequently, Study 2 compared the adequacy of a behavioral observation system versus self-reported estimates of university students' classroom behavior. Although not the primary motivation for moving to this new content domain, the investigation of several domains of behavior in this series of studies gives more support to the implications of the findings for the adequacy of self-report and behavioral estimates of behavioral parameters in general.

The behavior domain selected for Study 2 involved students' activities (both verbal and nonverbal) in selected undergraduate classes. This domain enables the obtaining of measures of a "hard" behavioral criterion: the number of teacher/student interactions, student/student interactions, and study behaviors.

The specific question addressed in Study 2 involved the relative accuracy of student self-reports of their classroom behavior compared with 20-minute behavioral observations in predicting behavioral parameters. That is, when all sources of contamination of behavioral estimates save natural variability over time (or the effects of limited sampling) were controlled, what was the accuracy of behavioral estimates relative to self-reported estimates?

Method

Subjects

Twelve undergraduate students attending a large southwestern university served as observers in this study. Each of the 12 observers began attending one of the three target classes at the beginning of the semester. The observers selected three to five different subjects in the target classes. The target classes were Abnormal Psychology, Human Sexuality, and Educational Psychology. Forty-eight undergraduate psychology students who were enrolled in one of three classes at the University of Houston served as subjects for this experiment: 9 of the 60 students in the first class, 22 of the 280 students in the second class, and 17 of 20 students in the third class. All the students in all three target classes were informed about the observations before the observers collected any data. All students agreed to be observed.

Instruments

The observers used a naturalistic observation system to directly record frequencies of their subjects' behaviors. The observers recorded and coded the nonverbal and verbal behavior streams of their subjects independently. To be included in the record, each nonverbal activity was required to be at least 15 seconds long. Studying behavior, defined as reading a text or taking notes, was the nonverbal behavior of interest in this study. The formal conversation units (student/teacher interactions, student/student interactions) did not require a minimal time limitation. However, the observers tallied new conversation units only after their subjects were quiet for at least 15 seconds. Some periods of prolonged and continuous conversations did occur infrequently and they were represented with only one conversation unit. Nevertheless, Willems, Alexander, Norris-Baker, Stephens & Wiener (1979) have shown that systems like the one above, which are based on variable interval sampling procedures, yield frequency records that are equivalent to continuous time sampling. Wiener, Willems, & Howard (in prep.) have also found that these observation systems are virtually nonreactive.

The observers participated in an extensive training program (3 weeks in length), which consisted of readings, lectures, and practice sessions. Instruction began with a detailed explanation of naturalistic observation and the molarity levels of the activities under investigation. The trainees read about the system's procedures and practiced recording and coding behaviors of other university students. After observers demonstrated proficiency in the use of the observation system (90% interobserver agreement), they began recording the behavior

of their chosen subjects. The observers were unaware of the specific hypotheses of the study, and they were instructed not to communicate with their subjects.

Procedures

Observers recorded activity records each week on each of the 48 subjects for 9 weeks. During each of the 20-minute observation periods an observer recorded the number of studying activities, conversations with other students, and conversations with the professor that the subject performed.

During the last week of classes, the observers presented the subjects with a questionnaire that solicited self-reports of classroom behavior. Each subject reported the number of studying behaviors, student conversations, and professor conversations that he/she engaged in during class. The behaviors were defined on the questionnaire exactly as they were defined for the observers. In addition, each self-report questionnaire provided the subject with the average number of times each activity was performed by students in his/her particular class during the 5th week of the semester (Week 2 of the study). This was done to allow subjects to estimate the frequency of the target behaviors by a hypothetical "average student."

To ascertain if the manner in which the self-report question is asked can influence the accuracy of the data, a second form of each question was asked. For each behavioral category subjects were asked to indicate the number of students in that class who engaged in that in-class behavior more frequently than they. The enrollment in their class was also provided.

Data Analysis

This study tested the usefulness of direct observations of behavior and self-reports of behavior in predicting a behavioral criterion. This was accomplished through a correlational procedure and an absolute deviation method.

Correlational analysis. All data in the correlational analyses were expressed as ratios in order to make the data from the three classes comparable. For the behavioral measure, the denominator for each ratio was the total number of specific behaviors performed during each observation session in that particular subject's class. The numerator was the number of specific behaviors performed by the subject. Thus, for each observation, each subjects' behavior was expressed as a ratio of the total behavior of the sample of the class to which he/she belonged. These ratios were calculated for studying behavior, conversations with students, and conversations with professors. An examination of the distributions of these scores showed them to be skewed toward the lower end of the scale. An arcsine transformation was applied to the data to avoid the effects of the skewness in the correlations that followed.

Next, three criterion variables were established by calculating the averages across the nine transformed observation scales. This was done for each subject. One average was calculated for studying behavior, one for conversations with other students, and one for conversations with the professor.

The self-report data were also expressed in ratios. The first self-report scale, the average number of executed behaviors that the target student reported he/she performed during a class period, was divided by the total average reported in the sample of that student's class. The resulting distribution was also skewed to the lower end of the scale so that the arcsine transformation was also applied to it. These measures were constructed for studying behavior, conversations with other students, and conversations with the professor.

The second self-report scale, the number of other classmates that were reported to engage in more behaviors than the target student, was also expressed in a ratio. Each target student's estimate was divided by the total number of students in the target student's class. The skewness of the resulting distribution required the arcsine

transformation to be applied to the data. These measures were constructed for studying behavior, conversations with other students, and conversations with the professor. Pearson product-moment correlations were calculated between the behavioral criterion measures, the behavioral measures for each of the nine observations, and the two self-report scales, for all three behavioral variables.

Deviation analysis. A new set of behavioral criteria were constructed for the deviation analysis. The raw behavioral frequencies were averaged across all nine observation sessions for each subject. This was done for studying behavior, conversations with other students, and conversations with professors. Thus, each subject's behavior was represented as three average behavioral variables.

The first self-report scale, the average number of executed behaviors that the target student reported he/she performed during a class period, was also used in the deviation analysis. Each subject's scores were divided by 2.5 to make them comparable to the raw score behavioral criterion. The behavioral observations were 20 minutes long and the classes that the students reported on were 50 minutes long. Thus, the behavioral criterion for each of the three dimensions represented the number of behaviors per average 20 minute session.

Next, absolute deviations were calculated between each of the nine observations of each subject and the average behavioral observations of each subject. This was done for all three behavioral variables. The absolute deviations were averaged across subjects resulting in average absolute deviations for all nine observation sessions. In addition, the deviation between the modified self-report scale and the average behavioral observation was also calculated for each subject. The self-report deviation was averaged across subjects.

Results

Correlation Analysis

Table 4 shows that the second self-report scale, the number of classmates who performed more behaviors, was the poorest indicator of the behavioral criterion. It also shows that the first self-report scale, the average number of behaviors performed by a student, was a good predictor of the behavioral criterion for both the studying behavior variable ($r = .61$, $p < .001$) and the conversations-with-other-students variable ($r = .62$, $p < .001$). The correlations between the self-report scales and the behavioral criterion was low for the conversations-with-the-professor variable ($r = .17$ for the first self-report scale and $r = .33$ for the second self-report scale). The median correlation between the behavioral criterion and the nine behavioral observations was higher than either self-report scale for this variable (median $r = .67$; mean $r = .59$). Apparently, students' self-reports of their conversations with professors were not reliable indicators of the behavioral criteria. With respect to the studying behavior variable and the conversations-with-other-students variable, the first self-report scale was as effective in predicting the behavioral criterion as were the nine individual observation measures. The median correlation for the nine observational measures and the behavioral criteria was .59 (mean $r = .56$) for studying behavior and .58 (mean $r = .50$) for conversations with other students. The correlation between the first self-report scale and the behavioral criteria was .61 for studying behavior and .62 for conversations with other students.

Deviation Analysis

The lower the deviations in Table 5, the lower the absolute difference between the measure and behavioral criteria, that is, the better the accuracy. For studying behavior, the median deviation of the nine observation sessions was 1.88 and the deviation for the self-report measure was 1.53. For conversations with other students, the median deviation of the nine observation sessions was 1.17 and the deviation for the self-report measure was 1.28. For conversations with professors, the median deviation for the nine observation sessions was .33 and the deviation for the self-report measure was .50. In summary,

Table 4
Correlations of the Behavioral Criterion Measures with the Nine Observation Session Measures and the Two Self-Report Indices

Behavioral Criteria	Self-Report Scale 1	Self-Report Scale 2	Observation Sessions 1	2	3	4	5	6	7	8	9
Study	61	20	63	44	61	73	69	44	53	43	59
Conversations with Students	62	26	78	38	04	55	67	58	54	61	79
Conversations with Professors	17	33	51	56	00	74	74	79	59	75	67

Note. Decimal points omitted.

the self-report measure was the best approximation of the studying behavior criteria, the self-report measure was about as effective as the behavioral measure in approximating the conversations-with-other-students criteria, and the self-report measure was not as effective as the behavioral measure in approximating the conversations-with-professors criteria.

The descriptive nature of the deviation analysis leaves some doubt about its conclusions. To reduce the ambiguity, it was necessary to conduct some tests of statistical significance. The absolute deviations of the nine observations of each subject with that subject's average behavioral variables were calculated and averaged. This was done for each of the three variables. Each subject was assigned a pooled average absolute deviation for the behavioral measures and absolute deviations for the self-report measures. For each of the three variables, comparisons were made between the absolute deviations of the self-report scale and the behavioral measure. The comparisons focused on the ordinal properties of the deviation scores. Table 6 displays the results. Overall, the self-report estimate resulted in a more accurate estimate in 61 instances; the

Table 5
Average Absolute Deviations for the Nine Observation Session Measures and the Two Self-Report Scales

Observation Session	Behavior Criteria: Study	Conversations with Students	Conversations with Professors
1	1.88	1.17	.30
2	1.84	1.23	.29
3	1.81	1.00	.33
4	1.53	.83	.57
5	1.72	.96	.32
6	2.03	1.12	.36
7	2.01	1.26	.36
8	2.57	1.31	.40
9	2.00	1.20	.31
Self-Report Scale 1	1.53	1.28	.50

behavioral estimate was superior in 59 cases; and the estimates were equally accurate in 24 instances. A sign test (which excluded ties) resulted in no significant differences for any of the variables. Comparisons employing t tests were also nonsignificant for two of the three variables, with only the comparison for study behaviors approaching significantce (t (43) = 1.97, $.10 > p > .05$) such that self-report estimates were more accurate than behavioral observations.

Discussion

Study 2 tested the proposition that subjects could accurately self-report the frequency of their in-class verbal and nonverbal behaviors. Results of the deviation score analysis indicated that subjects' self-reports were as accurate as the behavioral estimates that were contaminated by the effects of limited sampling alone. The results of the correlational approach were less straightforward. Although the self-reported estimates yielded higher correlations with behavioral criteria than did the behavior estimates in two instances (conversations with students and study behavior), the magnitude of difference was not large in an absolute sense. However, in the one instance (conversations with professors) in which the frequency of behavior was low, the behavioral samples correlated more highly with the behavioral parameters than did the self-report index by a substantial margin.

Overall, it is concluded that in this study, time limited behavioral samples and subjects' self-reported estimates of their in-class behavior are equally accurate. This finding is rather surprising, given that the effect of all other sources of contamination of behavioral measures were controlled. The equivalence of self-reported estimates is more startling when it is considered that, unlike Study 1, in Study 2 behavioral estimates were *not* independent estimates of behavioral criteria. Each behavioral estimate was also one of nine values that formed the behavioral criterion score. This lack of independence could have only served to spuriously inflate the behavioral estimates of behavioral criteria in the direction of greater accuracy.

The results of the second self-report measure (the number of students who perform the behavior more frequently than the subject) serve as a warning. Overall, the correlations with behavioral criteria for the second self-report measure were not as high as those of the self-reported frequency of behavior ratings. Quite possibly, the manner in which a self-report estimate is worded is critical to its accuracy. This finding recalls the work of McReynolds and Stegman

Table 6
The Average Absolute Deviations Between Observations and Behavioral Criteria, and Between Self-Report Measures and Behavioral Criteria

Variable	Number of Subjects with More Accurate: Self-Report Deviation Scores	Number of Subjects with More Accurate: Behavioral Deviation Scores	Number of Subjects with Equal Self-Report and Behavioral Deviation Scores
Study Behaviors	24	20	4
Conversations with Students	18	26	4
Conversations with Professor	19	13	16
All Variables	61	59	24

(1976), who found that if a person were asked how fearful he/she was of snakes, that rating would not predict the person's behavior very well, but if people were asked instead to predict their own behavior, that rating would be extremely accurate. Since there is substantial evidence that fear and avoidance are deynchronous processes (cf. Rachman & Hodgson, 1974), low fear/avoidance correlations are to be expected and in no way suggest the inadequacy of self-report fear rating (as had been frequently assumed in the past). Thus, the manner in which a self-report questionnaire is phrased can be critical for its accuracy.

If self-report measures are as good as behavioral indices that are contaminated by natural variability (the effects of limited sampling) alone, what might be the impact of decontrolling the other potential sources of contamination of behavioral estimates? Study 3 had three major purposes: (1) to replicate the findings of the impact of natural variability on behavioral measures in a different domain than Study 2; (2) to ascertain an estimate of the impact of situation variance on the accuracy of behavioral measures; and (3) to replicate Bem and Allen's (1974) findings on the cross-situational variability of behavior.

STUDY 3

The logistical problems associated with observing students in two different classes in order to obtain an estimate of cross-situational variability were prohibitive. Consequently, Study 3 considered teachers' ability to self-report the number of teacher-student interactions in two different undergraduate courses. A second reason for investigating the accuracy of teacher self-reported in-class behavior is that Hook and Rosenshine (1979) recently reviewed studies wherein teacher's self-reports were compared with in-class observations of their behavior. Typically, very little agreement between the two estimates was found. Hook and Rosenshine (1979) concluded, "one is not advised to accept teacher reports of specific behaviors as particularly accurate. No slur is intended . . ." However, this conclusion is based upon the assumption that a behavioral measure is a good (or perfect) estimate of behavioral parameters. Study 3 tested the adequacy of that assumption when the number of teacher/student interactions was the dimension of interest.

Method

Subjects

Twenty members of the instructional staff of a large southwestern university participated in this study. Each instructor allowed an observer to attend regular class meetings of two separate undergraduate courses that the instructor taught during the spring semester of 1979. The instructors also agreed to complete a questionnaire at the end of the study. All academic ranks were equally represented in the sample of instructors.

Raters

The university's introductory psychology course offers students the opportunity to fulfill a course requirement by participating as assistants in departmental research. For credit in their introductory psychology class, 20 introductory psychology students served as raters in this study.

Instruments

Self-report questionnaire. During the final week of the semester, each instructor was contacted and asked to complete a questionnaire that requested several self-report estimates of his/her classroom behavior. The instrument consisted of eight separate self-report indices of behavior. Some of these ratings concerned the subjects' own behavioral variability. They were (1) a subject's ratings of his/her variability in the number of verbal interactions across all courses he/she teaches and (2) the subjects' ratings of

their natural variability (daily fluctuations in behavior) for each of the two courses in which he/she was observed. All three of these ratings were obtained on a 7-point scale, which ranged from 1 (not at all variable) to 7 (extremely variable). The remaining self-report items asked the instructors to describe the frequency of their in-class verbal behavior. These indices asked for a single overall numerical estimate of the average number of verbal interactions for a typical class in a typical course and separate single numerical estimates of the average number of verbal interactions per class with students in each course in which he/she was observed. Finally, the questionnaire requested subjects to compare the frequency of their verbal interaction for each course to other university courses. They rated themselves on a 7-point scale, which ranged from "much less than average" to "much more than average," with respect to other classes.

Behavioral measure. For each of the instructor's two classes that the rater attended, separate counts of verbal interactions were recorded. A single count consisted of each instance when instructors responded to a student's question or comment, as well as when instructors replied to comments and questions by students. At the end of each class, the counts were totalled to yield a daily total for each class.

Procedures

Approximately 35 raters were recruited from the university's introductory psychology course. They were contacted and given a description of their duties in data collection. According to their availability, each potential rater selected an instructor who was teaching two courses that semester that they could attend. The rater was permitted to observe an instructor who was teaching a course in which he/she was enrolled but was required to observe that same professor in another course in which he/she was not registered. The rater was constrained to select an instructor who taught two different courses that semester, rather than two sections of the same course. From the original 35 volunteers, 22 students were able to meet the conditions required of them as raters.

Instructors were contacted by one of the experimeters and given a general description of the study and its requirements. Of the 22 instructors who were contacted, 20 agreed to serve as participants. Upon obtaining the instructor's permission, each rater received training in observation techniques from one of the experimenters. Since the nature of the observational procedure was relatively simple, extensive training was unnecessary. Training lasted until each rater demonstrated proficiency (assessed informally) on practice samples of behavior.

All of the courses were taught at the undergraduate level. A large group of courses met twice a week for 90 minutes, while a similar number of classes met three times a week for 60-minute sessions, and a few courses met only twice a week for 60 minutes. Since class length varied, raters kept not only behavioral counts of the number of student-instructor verbal interactions, but recorded class length as well. Each Friday the experimenter collected the behavioral records for that week.

Raters were informed that the data was being collected in a study of classroom verbal behavior but were blind to the specific hypotheses of the study. In addition, they were instructed not to inform other class members of their purpose. In all classes, raters unobtrusively recorded the number of interactions in their own notebooks along with class notes. In some small classes, the instructor requested that he/she be allowed to introduce the rater as an auditor or a visitor. In a few instances this was allowed; otherwise, the rater's presence and purpose was not acknowledged.

At the end of the semester, the professor was contacted individually by one of the experimenters who administered the self-report questionnaire. Any questions about the items were clarified at that time. The instructor was then debriefed as to the purpose and rationale of the study.

Data Analysis and Results

Two different methods of data analysis were again employed to compare direct observations of behavior and self-reports of behavior in predicting a behavioral criterion. The first method, the absolute deviation approach, investigated absolute accuracy of prediction, while the second method, a correlational approach, essentially measured rank order accuracy of prediction and, unlike the absolute deviation method, was insensitive to any consistent tendency toward either underestimation or overestimation. In all cases, the criterion was an average of the behavioral measures, either across time or across both time and situations (classes).

Deviation Analysis

Preliminary to any data analysis, all frequency counts of student-teacher verbal interactions were converted into hourly rates. Since both 50-minute (MWF) classes and 75-minute (TTH) classes were included in the study, this conversion produced rates with a common metric across all classes.

The first step in the analysis was to obtain a situation-specific behavioral criterion score for each teacher. This score was computed for each individual as that person's mean hourly rate of interaction across time (i.e., observational periods) for a particular class. Because approximately one-half of all classes met three times a week while the other half met twice a week, and also because of occasional teacher absences, the number of times a teacher was observed in a particular class differed somewhat from teacher to teacher and from class to class. On the average, each subject was observed 11 times per class, with the range of number of observations for the two classes combined being 17 to 35.

To obtain a measure of absolute accuracy of a single behavioral measure in predicting the behavioral criterion, absolute deviations were calculated for each observation. These values were then summed for each individual and divided by the number of observations for that individual to obtain a measure of the average deviation of a single behavioral measure from the situation-specific behavioral mean. In a similar manner, the absolute value of the deviation of the class-specific self-report from the class-specific behavioral mean was computed for each person. The single behavioral measure was slightly more accurate on the average than the self-report measure, the mean absolute deviation score being 7.3 for the behavioral measure and 9.2 for the self-report. However, the superiority of the behavioral measure was indeed slight, because the difference in means of 1.9 was small compared to the standard deviation of 8.1. In addition, the difference was not statistically significant, either by a dependent t test ($t(38) = 1.62$, *ns*) or a sign test ($z = .163, p = .873$).

A second comparison was employed to compare the magnitude of error in a single behavioral measure and a self-report when cross-situational variability was considered in addition to simple variability over time. For this comparison, the overall behavioral mean (the average for both classes over time) was used in a deviation analysis similar to that conducted for the class-specific behavioral mean. Absolute deviations of single behavioral observations were again computed, but this time from the new criterion, the cross-situational mean. Once again, these deviations were averaged for each subject by dividing by the total number of observations per subject, to provide an index of the average deviation of a single behavioral observation from a cross-situational mean. In a similar manner, the absolute value of the difference between the self-report and the cross-situational mean was computed for each person. As before, the single behavioral measure was very slightly more accurate, on the average, than the self-report. The mean absolute deviation score was 8.3 for the behavioral measure and 8.6 for the self-report. With a standard deviation of 8.0, however, the superiority of the behavioral measure was miniscule. Indeed, the superiority was not statistically significant, either by a dependent t test (t (38) = .28, *ns*) or a

sign test ($z = .24$, $p = .59$). Thus, even when cross-situational variability was considered, there was no evidence to support a difference between the deviation of a single behavioral measure from the criterion and the deviation of a self-report from the criterion.

Correlational Analysis

A second set of analyses was performed contrasting single behavioral measures with self-reports using a correlational approach when natural variability alone or natural variability as well as cross-situational variability was considered. Such an approach, unlike the deviation analysis, did not assume that the two independent methods of measurement, self-report and direct observation, employed the same metric.

In order to assess the accuracy of a single behavioral measure in predicting the situation-specific behavioral criterion, Pearson product-moment correlation coefficients were computed between each single behavioral measure and the situation-specific behavioral average within lower and upper division courses separately. Because of the missing data problem some correlations were based on a very small sample. To avoid the undue influence of such correlations, only those correlations based on at least eight subjects were included in any of the subsequent analyses. The average of all such correlations between a single behavioral measure and the situation-specific behavioral mean was .81.

In order to compare the self-report with a single behavioral measure, Pearson product-moment correlation coefficients between the class-specific self-report and the class-specific behavioral criterion were computed for lower and upper division courses separately. The average correlation between class-specific self-report and class-specific behavioral mean was .80.

A Hotelling-Williams test (Darlington, 1975) was performed to compare the accuracy of the single behavioral measure versus the self-report in predicting the class-specific behavioral criterion. In particular, the average correlation between the single behavioral measure and the behavioral mean was compared with the average correlation between the self-report and the behavioral mean. The test statistic ($z = .1475$, $p = .88$) provided no evidence of a difference in predictive accuracy.

A second correlational analysis was employed to examine relative accuracy when cross-situational variability was considered in addition to simple variability over time. As in the first analysis, Pearson product-moment correlation coefficients between each single behavioral observation and the behavioral criterion were calculated; but now the behavioral criterion was an average across situations as well as time and both upper and lower division courses were included in the same analysis. The average correlation between single behavioral measures and the overall behavioral mean was .77.

In order to compare a single behavioral measure and the self-report, the Pearson correlation between the overall (cross-situational) behavioral mean and the overall self-report was computed, yielding a correlation of .81. The results of a Hotelling-Williams test comparing the correlation of .81 with 0.77 showed the difference to be nonsignificant ($z = .3957$, $p = .79$). Thus, for the cross-situational analysis, as for the situation-specific analysis, results provided lack of support for a difference between a single behavioral measure and a self-report in predicting a behavioral criterion.

Discussion

The behavior observed in Study 3 was the number of teacher-student interactions, which was the criterion on which self-report estimates fared most poorly in Study 2. However, in Study 3, teachers, rather than students, rated the frequency of interactions. The results of the correlation and deviation analyses were so similar that they will be discussed together.

When considering the impact of natural variability alone (or the effect of limited sampling

with a particular situation), behavioral and self-report estimates were about equal. The deviation score analysis revealed a slight advantage (albeit nonsignificant) for behavioral estimates. This was due to teachers' consistent tendency to underestimate the number of their interactions. Of course, this tendency did not influence the correlations of self-reports with behavioral parameters. Despite clear instructions to be as accurate as possible, instructors frequently made statements such as, "I know this estimate is low, but I'll be safe," while completing the self-report questionnaire.

When the effect of contamination due to situation was added, even the effects of the conscious tendency to underestimate were washed out and self-report estimates were fully as accurate as behavioral estimates of behavioral parameters for both deviation and correlation analyses. Similar to Study 2, behavioral estimates in Study 3 were not independent of behavioral parameters, since behavioral parameters consisted of the average of all behavioral estimates. Again, the lack of independence could only have served to spuriously inflate the accuracy of behavioral estimates of behavioral parameters.

CONCLUSIONS

To avoid a misunderstanding of the motivation for or the results of this series of studies, it should be emphasized that the findings in no way impute the value of behavioral measures. All of the sources of contamination of behavioral indices dealt with in these investigations have been known to psychometricians for quite some time. No new sources of contamination, nor any unexpectedly severe levels of invalidity, were observed. Rather, these demonstrations re-emphasize the fact that the construct validity (and generalizability; cf. Cronbach et al., 1972) of any measure is an empirical issue. The form or method of the measure (e.g., self-report, behavioral, physiological), in and of itself, in no way speaks to its construct validity. Sensitive researchers have always been cognizant of the limitations of behavioral measures. Here a case was made for some underappreciated qualities which self-report techniques often possess.

But why is it important to make a strong case for self-report techniques? There are several motivations; the first deals purely and simply with an economy of research effort. As mentioned earlier, when a researcher wishes to approximate certain behavioral parameters, in theory, by controlling for the various sources of contamination, a behavioral measure(s) can be obtained that *must* be superior to self-reported estimates of those parameters. However, the present studies suggest that obtaining those superior estimates in many cases might be cost-prohibitive. If the present studies are representative of most psychological studies, the more costly behavioral indices sometimes gathered are either no better (Studies 2 and 3) or even inferior (Study 1) to the substantially more efficient self-report indices.

Finally, psychology has developed a mistrust of people as faithful reporters. Although some areas, such as cognitive psychology, accept self-report data with great ease (cf. Ericsson & Simon, 1978; Newell & Simon, 1972), other areas seem to be bent upon moving still further away from a reasoned consideration of self-reported evidence (e.g. Hook & Rosenshine, 1979; Nisbett & Wilson, 1977). Further, to the extent that all self-reported data are declared suspect, this precludes consideration of some interesting theoretical possibilities, such as Harré and Secord's (1972) ethogenic approach to social behavior. The present set of studies demonstrates that some of the evidence traditionally cited to demonstrate the lack of accuracy of self-reports must be reconsidered.

References

Arkowitz, H., Lichtenstein, E., McGovern, K., & Hines, P. The behavioral assessment of social competence in males. *Behavior Therapy,* 1975, *6,* 3-13.

Bach, R. C. F., Lowery, D., & Moylan, J. J. Training state hospital patients to be appropriately asser-

tive. *Proceedings of the 80th annual convention of the American Psychological Association,* 1972, *7,* 383-384.

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.

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation of the multimethod matrix. *Psychological Bulletin,* 1959, *56,* 81-105.

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching.* Chicago: Rand McNally, 1963.

Campbell, J. P., Dunnette, M. D., Lawler, E. E., III, & Weick, K. E. *Managerial behavior, performance, and effectiveness.* New York: McGraw-Hill, 1970.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: John Wiley, 1972.

Darlington, R. B. *Radicals and squares: Statistical methods for the behavioral sciences.* Ithaca, NY: Logan Hill Press, 1975.

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

Ericsson, K. A., & Simon, H. A. Retrospective verbal reports as data. (Working Paper No. 388, Department of Psychology, Carnegie-Mellon University, Pittsburgh, PA, 1978.

Fiske, D. W. *Strategies for personality research.* San Francisco: Jossey-Bass, 1978.

Friedman, P. H. The effects of modeling and role-planning on assertive behavior. In R. D. Rubin, J. Fernsterheim, A. A. Lazarus, & C. M. Franks, *Advance in behavior therapy.* New York: Academic Press, 1971.

Galassi, J. P., Delo, J., Galassi, M. D., & Barstein, S. The college self-expression scale: A measure of assertiveness. *Behavior Therapy,* 1974, *5,* 165-171.

Galassi, J. P., Hollandsworth, J. G., Radecki, J. C., Gay, M. L., Howe, M. R. & Evans, C. L. Behavioral performance in the validation of an assertiveness scale. *Behavior Therapy,* 1976, *7,* 447-452.

Green, S. B., Burkhart, B. R., & Harrison, W. H. Personality correlates of self-report, role-playing, and in vivo measures of assertiveness. *Journal of Consulting and Clinical Psychology,* 1979, *47,* 16-24.

Gulanick, N. A., Howard, G. S., & Moreland, J. Evaluation of a group program for highly feminine women aimed at increasing androgyny. *Journal of Sex Roles,* 1979, *5,* 811-827.

Harré, R., & Secord, P. F. *The explanation of social behavior.* Oxford, England: Blackwell. 1972.

Hersen, M., Eisler, R. M., & Miller, P. M. An experimental analysis of generalization in assertive training. *Behavior Research and Therapy,* 1974, *12,* 295-310.

Hook, C. M., & Rosenshine, B. V. Accuracy of teacher reports of their classroom behavior. *Review of Educational Research,* 1979, *49,* 1-12.

Jöreskog, J. G. Structural analysis of covariance and correlation matrices. *Psychometrika,* 1978, *43,* 443-477.

Kazdin, A. E. Effects of covert modeling and model reinforcement on assertive behavior. *Journal of Abnormal Psychology,* 1974, *83,* 240-252.

Longin, H. E., & Rooney, W. M. Teaching denial assertion to chronic hospitalized patients. *Journal of Behavior Therapy and Experimental Psychiatry,* 1975, *6,* 219-222.

Martorano, R. Effects of assertive and nonassertive training on alcohol consumption, mood, and socialization of the chronic alcoholic. *Proceedings of the 81st annual convention of the American Psychological Association,* 1973, *8,* 393-395.

McFall, R., & Lillesand, D. Behavior rehearsal with modeling and coaching in assertion training. *Journal of Abnormal Psychology,* 1971, *77,* 313-323.

McFall, R., & Twentyman, C. T. Four experiments on the relative contributions of rehearsal, modeling and coaching to assertion training. *Journal of Abnormal Psychology,* 1973, *81,* 299-318.

McReynolds, W. T., & Stegman, R. Sayer versus sign. *Behavior Therapy,* 1976, *7,* 704-705.

Millham, J., & Jacobson, L. I. The need for approval. In H. London & J. Exner (Eds.), *Dimensions of personality.* New York: John Wiley, 1978.

Newell, A., & Simon, H. A. *Human problem solving.* Englewood Cliffs, NJ: Prentice-Hall, 1972.

Nisbett, R. E., & Wilson, T. DeC. Telling more than we can know: Verbal reports on mental processes. *Psychological Review,* 1977, *84,* 231-259.

Rachman, S., & Hodgsen, R. Syncrony and desyncrony in fear and avoidance. *Behavior Research and Therapy,* 1974, *12,* 311-318.

Rathus, S. A 30-item scale for assessing assertive behavior. *Behavior Therapy,* 1973, *4,* 398-406.

Serber, M., & Nelson, P. The ineffectiveness of systematic desensitization and assertive training with hospitalized schizophrenics. *Journal of Behavior Therapy and Experimental Psychiatry,* 1971, *2,* 107-109.

Shoemaker, M. E., & Paulson, T. E. Group assertion training for mothers: A family intervention strategy. In E. J. Marsh, L. C. Handy, & L. A. Hamerlynck (Eds.), *Behavior modification approaches to parenting.* New York: Brunner Mayel, 1976.

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. *Unobtrusive measures: Nonreactive reserach in the social sciences.* Chicago: Rand McNally, 1966.

Wiener, R., Willems, E. P., & Howard, G. S. *On the reactivity of behavioral pressures*, in preparation.

Weinman, B., Gelbert, P., Wallace, M., & Post, M. Inducing assertive behavior in chronic schizophrenics: A comparison of socioenvironmental, desensitization, and relaxation therapics. *Journal of Consulting* and *Clinical Psychology,* 1972, *39,* 246–252.

Willems, E. P., Alexander, J. L., Norris-Baker, C., Stephens, M. A., & Wiener, R. L. Behavioral frequency versus behavioral time:Frequency is enough. *Journal of Applied Behavioral Analysis,* under review.

Author's Address

Send requests for reprints or further information to George S. Howard, Department of Psychology, University of Houston, Houston, TX 77004.