Construct Validity Comparisons of Three Methods for Measuring Patient Compliance

K. Michael Cummings, John P. Kirscht, Marshall H. Becker, and Nathan W. Levin

A multitrait-multimethod design was employed to assess the construct validity of three commonly used methods for assessing patient compliance: physiological assessments (e.g., blood chemistries), ratings by health professionals, and patient self-reports. Subjects were patients receiving ambulatory hemodialysis treatments for end-stage renal disease, whose regimen required them to take medications, to follow dietary restrictions, and to limit fluid intake. Study findings indicated that of the three methods examined, the nurse rating approach was the most valid (although it contained only about 50 percent valid variance). Measures derived from physiological assessments contained a substantial proportion of residual error (over 70 percent), and the patient self-report method contained only about 12 percent valid variance (with about 18 percent method-effects variance, and 68 percent residual-error variance). These results make clear the need for additional research directed at developing valid methods for evaluating patient compliance behaviors.

While the problem of compliance with medical recommendations is now widely recognized and studied, relatively little is known about the accuracy of different methods used to measure patient compliance behaviors [1,2]. A conceptual definition of compliance includes both medical recommendation and behavioral performance, with the latter

---

Research for this article was supported by a grant from the Michigan Department of Public Health, Division of Chronic Disease Control, Lansing, Michigan.

Address communications and requests for reprints to K. Michael Cummings, Ph.D., Research Scientist, Roswell Park Memorial Institute, 666 Elm Street, Buffalo, New York 14263. John P. Kirscht, Ph.D. and Marshall H. Becker, Ph.D., M.P.H. are Professors of Health Behavior and Health Education at the School of Public Health, The University of Michigan, Ann Arbor. Nathan W. Levin, M.D. is Chief of Nephrology, Henry Ford Hospital, Detroit, Michigan.

judged in light of the former. At issue in measuring compliance is an assessment of patient behaviors.

Relatively simple to measure are those behaviors in which a formal record is made of an individual's actions. For example, assessments of appointment-keeping or of receipt of an immunization are relatively straightforward (although not all visits and services are accurately recorded). However, most patient behaviors are not subject to direct observation and regular recording in medical records. As a result, investigators have had to develop additional ways to measure compliance behaviors, and these methods vary considerably in their levels of directness and apparent validity.

Many researchers have relied upon patient self-reports to assess compliance. While fairly direct, patient reports of their behaviors are subject to several sources of invalidity—the natural desire to report "good" behaviors, for example, and the inability to recall instances of noncompliance. In general, studies which have compared self-reported compliance with other methods of measuring regimen adherence (e.g., physiological measures, such as blood samples and urine tests to detect trace levels of medication; pill counts; health outcomes) have found that patients tend to overestimate their compliance with recommendations [3,4].

Judgments by health professionals have also been employed as a method for estimating patient compliance. While it seems reasonable that health professionals would be in a good position to evaluate the compliance behaviors of their patients, studies of physicians' ability to predict compliance in patients have revealed only chance levels of accuracy [5–7].

Physiological assessments of cooperation with therapy include such approaches as measuring changes in body weight, seeking traces of drugs or metabolites in blood or urine samples, obtaining reductions in blood pressure, and achieving desired health outcomes (i.e., disease control). As measures of patient compliance, such assessments have the advantage of being relatively unaffected by human judgments; however, they are often costly to obtain—and, if patients know they are to be tested, they may alter their behaviors [8]. Moreover, physiological assessments and health outcomes are often influenced by factors unrelated to patient compliance behavior (e.g., natural recuperative processes, physical characteristics of the individual, time when the measure is taken).

Without some knowledge of the relative accuracy of the various methods used to evaluate patient compliance, one risks making inappropriate judgments in research directed at uncovering the causes of

noncompliance or at determining the therapeutic effectiveness of different medical recommendations. Certainly, not all methods of measurement are equally good; to permit rational choices among them that will lead to subsequent application, it is necessary to assess their relative strengths and weaknesses. In addition, the continued improvement of assessment methods needs to be guided by an evaluation of the comparative success of different approaches in the past.

The measurement literature traditionally distinguishes three types of validity: construct validity, concurrent validity, and predictive validity. Construct validity refers to the relationship of an observed variable to a theoretical construct (or "concept"); concurrent and predictive validities both involve relating an observed measure to an assumed valid criterion variable.

Our research for this article involved examining the validity of three commonly used methods for assessing compliance among hemodialysis patients with taking medications, following dietary restrictions, and limiting fluid intake. At issue were the validity of physiological assessment (blood chemistries and weight gain between dialysis treatments), ratings by health professionals, and self-reports by the patients. Because each method of measuring compliance relies on indirect assessments of patient behaviors, as previously noted, the methods may be subject to measurement error. For example, although most studies of compliance with medical regimens by hemodialysis patients have relied exclusively on physiological outcomes as a basis for evaluation, these measures cannot be assumed to be valid indicators of patient behaviors. We do not believe that appropriate criterion variables exist for validating measures of compliance in dialysis patients. Hence, our study of the validity of compliance measures focused on construct validity.

METHODS

BACKGROUND AND SUBJECTS

The study group consisted of those patients with end-stage renal disease receiving hemodialysis treatments provided in two outpatient dialysis clinics located in southeastern Michigan. Participation in the study was limited to persons over 18 years of age who had been receiving these treatments for a minimum of 3 months. Of the 120 eligible study participants, 4 refused to participate (n = 116, a 97 percent response rate).

Patients ranged in age from 21–76 years (mean = 54.8 years). Fifty-four percent were male, and half were white. The median educational level achieved was high school graduation, and median family income was $10,000–$10,999 per year, with 16 percent receiving less than $5,000 annually. Most of the participants were married (67 percent); 9 percent had never married; and the remaining 24 percent were widowed, separated, or divorced. The average length of time patients had been obtaining dialysis treatments was 29 months; most subjects (87 percent) were receiving dialysis treatments three times per week. Only a quarter of the patients (n = 28) indicated that they expected to have a kidney transplant in the future.

COMPLIANCE MEASURES

Patient compliance behaviors relative to taking phosphate-binding medication, following dietary restrictions (especially with regard to consumption of foods high in potassium), and limiting fluid intake were examined using three measurement approaches.

Physiological Assessments

Dialysis patients need to control the amount of phosphorus in their blood since their kidneys can no longer perform this function adequately. Phosphate-binding medications are used to help control the amount of phosphorus which is absorbed in the intestinal tract. Thus, the serum phosphorus level (SPHL) provides an indirect measure of patient compliance with instructions for taking phosphate-binding medicine. In this study, compliance with taking phosphate-binding medicine was assessed by averaging each patient's SPHLs over a 2–3 month period. For the great majority of patients, SPHLs were obtained routinely once every month.

Since patients with renal failure are unable to excrete potassium adequately, dietary restrictions are imposed to help the patient maintain proper levels of potassium in the blood. Thus, serum potassium level (SPL) is an indirect indicator of compliance with dietary restrictions on food high in potassium (primarily fruits and vegetables). In this study, compliance with dietary restrictions was evaluated by averaging the patients' SPLs over a 1–2 week period. A total of three measures were obtained on each patient. SPLs were assessed routinely each time the patient came for dialysis.

To avoid excessive fluid buildups, dialysis patients are restricted in the amount of fluid permitted between dialysis treatments. The degree of fluid restriction depends on the patient's ability to excrete fluid.

Thus, interdialysis weight gain is an indirect measure of patient compliance with fluid restrictions. In this study, compliance with limiting fluid intake was appraised by averaging the patients' between-dialysis weight gains over a 1–2 week period. Between-dialysis weight gains were calculated by subtracting from each patient's predialysis weight the previous treatment's postdialysis weight. Pre- and postdialysis weights were obtained routinely at each treatment session.

Health-Professional Ratings

One nurse from each clinic was instructed to estimate the degree to which each patient was following instructions for the phosphate-binding medicine, the diet, and the limit on fluid intake. The direction and degree of patient compliance was rated on a seven-point scale, where responses ranged from "poor" to "excellent" compliance. Due to a rotating work schedule for nurses in both clinics, the nurse raters were personally familiar with each patient they rated. It should be pointed out that nurses in the clinic also did have access to the physiological test results of patients.

Patient Reports

As part of an interview designed to assess health beliefs and knowledge about their illness, patients were asked to rate the degree to which they were complying with instructions about their phosphate-binding medicine, their diet, and their between-dialysis fluid intake limit. Patients rated the direction and degree of compliance to each of the three components of the medical regimen on the same seven-point scale as that used by the nurse raters. Interviews were conducted by trained interviewers (not employees of the clinic) while the patients were receiving their dialysis treatments.

Means and standard deviations for each of the compliance measures are shown in Table 1.

ANALYSES

Since theoretical constructs are unmeasured variables, a study of construct validity requires an estimate of relationships between observed measures and hypothetical (unobserved) variables. Such an investigation depends both on a network of relationships within a set of observed measures and on a series of theoretical assumptions about the relationship of specified hypothetical constructs to one another and to the observed measures.

Table 1: Means, Standard Deviations, and Number of Respondents for Study Measures of Compliance

Compliance Measures	*Number of Respondents*	*Mean*	*Standard Deviations*
Serum phosphorus level (mg/dL)	98	6.05	1.68
Serum potassium level (mEq/L)	106	4.96	0.72
Weight gain between dialysis treatments (kilograms)	108	2.59	1.13
Nurse report of patient compliance with taking phosphate-binding medicine	98	4.18	1.55
Nurse report of patient compliance with diet regimen	111	4.07	1.24
Nurse report of patient compliance with fluid restriction	108	3.87	1.45
Patient report of compliance with taking phosphate-binding medicine	98	6.02	1.21
Patient report of compliance with diet regimen	111	4.88	1.47
Patient report of compliance with fluid restriction	100	5.00	1.41

Campbell and Fiske [9] argue that evidence for construct validity exists when convergence occurs among independent measures of the same trait and discrimination is noted among measures of different traits. The authors demonstrate that it is possible to examine convergence and divergence within a matrix of intercorrelations among three or more theoretically unrelated traits measured by three or more independent methods. Evidence for convergence exists when the correlations among measures of the same trait are positive and significantly different from zero. Evidence of discrimination is threefold:

1. The observed correlations among measures of the same trait using different methods should be greater than the correlations among measures of different traits using the same method.
2. Different traits measured by the same method should have a lower intercorrelation than different measures of the same trait.
3. The pattern of trait interrelationship should be the same within and among different measures.

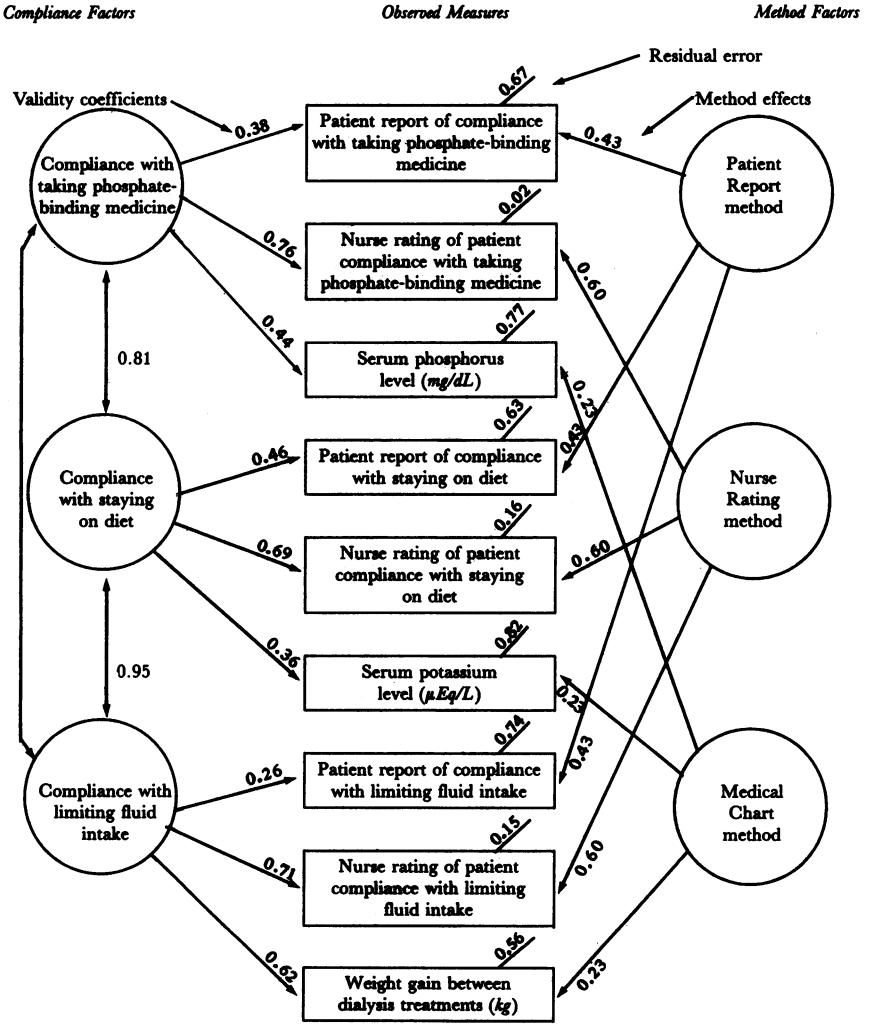
Although the Campbell and Fiske approach represented an important advance, it did not provide precise estimates of construct

validity at the time of its publication; and, in the case of large matrixes, it continues to be quite difficult to interpret. A number of procedures for quantifying the multitrait-multimethod approach have been suggested in the literature [10–12] for assessing construct validity, but truly precise use of this technique seems to be possible only when it is wedded to methods of structural analysis. Many writers [13–16] have discussed applications of structural analysis to multitrait-multimethod problems. Jöreskog [17] developed a powerful maximum-likelihood procedure which partitions the variance of a measure into three parts: (1) valid variance (reflecting what the measure is intended to measure), (2) correlated error variance (reflecting influences other than those that the measure was designed to tap, and which also affect other measures), and (3) residual variance (reflecting that portion of variance which is not otherwise accounted for).

The general form of the structural equation model used in this investigation appears in Figure 1. The variances of the observed measures are represented by rectangles. The circles on the left are linked directly to all of the measures designed to tap identical constructs; thus, the circles represent what the observed measures have in common (which, in this case, is compliance with a particular aspect of the medical regimen). The values for parameters linking measures to the circles on the left are interpreted as "validity coefficients"—the correlation (Pearson's product-moment, r) between the true condition and the observed measures of the condition. The square of the validity coefficient gives the proportion of observed variance that is true variance. The linkages among the circles on the left incorporate the expectation that compliance in one area will be related to compliance with other areas of the regimen.

A similar set of assumptions is involved in the specification of linkages with the circles on the right. These linkages are intended to represent the correlated error component of a measure. A "method effects coefficient" is the correlation (Pearson's product-moment, r) between the method factor and the observed measure. The square of the method effects coefficient gives the proportion of observed variance that is correlated error variance. The correlated error component of a measure is that portion of the variance which reflects influences other than those that the measure was primarily designed to tap and which itself influences other measures as well. Several types of influences may result in correlated errors. The effects of individuals' biases and differences in interpretation are "errors," and if they appear in more than one measure—especially likely to happen when the measures are based on the same method—they result in a spurious degree of correlation among the measures. Different measurement methods activate differ-

Figure 1: Structural Equation Model and Parameter Estimates for Three Types of Compliance Behaviors



Chi-square with 21 degrees of freedom is 24.15
Probability level = 0.286

ent biases, which then act to produce different patterns of correlation among the measures. Examining these patterns makes it possible to obtain estimates of correlated errors associated with a particular method.

In addition to the specification of linkages, a constraint was imposed on the model with regard to the linkages among the method factors (circles on the right) and the observed measures (rectangles). The magnitude of the linkages from any one method factor was constrained to become equal for all measures using the same method. This constraint on the model reflects the belief that error due to the effects of a method should be equal for all measures which use the same measurement method. The linkages among the circles on the right were fixed at zero, reflecting our assumption that the method factors are independent. Models of this form have sometimes been called "confirmatory" or "restricted" factor analysis models [16,17].

The assumption of independence among the measurement methods is violated to some extent by the fact that the nurses had access to the physiological measurement information in the patients' medical records. However, the modest correlations (ranging from 0.20 to 0.45) between the nurse rating measures and the physiological measures suggest that the nurses did not base their ratings of patient compliance solely on the results of the physiological tests.

The vectors extending from the right side of each of the nine rectangles represent the residual error component of the measures (i.e., that portion of the variance that can be attributed neither to true effects nor to correlated errors). Included here are the *random* effects of all of the occurrences that may influence a particular score on a measure: e.g., the respondent might have heard or interpreted the question incorrectly; a technician might have read the blood value incorrectly; some error might have been introduced as the answer was being prepared for computer processing. By definition, the residual errors in the model are assumed to be independent.

Maximum-likelihood estimates of all parameters in the structural model were obtained using the LISREL IV computer program. Input data for the structural equation analysis was a matrix of correlations (Pearson's product-moment,r) among the observed measures (Table 2).

To evaluate the efficacy of the model, a Chi-square goodness-of-fit statistic was computed. The Chi-square is a direction function of the discrepancy between the sample correlation matrix and that reproduced by the parameter estimates of the model. The null hypothesis initially proposes that a sample correlation matrix is obtained from a population having the proposed causal structure. If the obtained Chi-

Table 2: Pearson Correlations Among the Study's Measures of Compliance

	Phosphate-Binding Medicine			*Diet*			*Fluid Limit*		
	Self-Report	*Nurse Rating*	*Serum Phosphorus*	*Self-Report*	*Nurse Rating*	*Serum Potassium*	*Self-Report*	*Nurse Rating*	*Weight Gain*
Phosphate-binding medicine									
Self-Report	1.00								
Nurse rating	0.28	1.00							
Serum phosphorus	0.36	0.36	1.00						
Diet									
Self-report	0.30	0.30	0.08	1.00					
Nurse rating	0.22	0.81	0.24	0.36	1.00				
Serum potassium	0.10	0.21	0.10	0.19	0.20	1.00			
Fluid limit									
Self-report	0.29	0.13	0.19	0.27	0.26	0.04	1.00		
Nurse rating	0.19	0.73	0.25	0.34	0.82	0.18	0.21	1.00	
Weight gain	0.07	0.29	0.17	0.24	0.40	0.35	0.04	0.45	1.00

square corresponds to a probability level greater than 0.05, it is concluded that the null hypothesis cannot be rejected [16].

RESULTS

Figure 1 shows the estimated validity, method effect, and residual error coefficients for the nine compliance measures. The validity coefficients ranged from 0.26 to 0.76. The nurse rating method of measuring patient compliance produced validities of about 0.7. Data obtained using the physiological assessments had lower average validity (0.5), while the patient-report method yielded validities of about 0.4.

On the basis of these results, it can be inferred that single-item measures using the nurse-report method to assess patient compliance with different components of the medical regimen contain approximately 50 percent valid variance. At the low end, the patient-report method results in about 12 percent valid variance; and falling between, the physiological-assessment method yields about 23 percent valid variance.

As indicators of specific constructs, self-reports by patients of compliance in limiting fluids and physiological assessments of patients' serum potassium levels appear to be especially poor measures of the constructs they are intended to tap.

Sharp differences appear in the method-effect coefficients associated with the different methods of measurement. The percentage of total variance attributable to method effects was about 35 percent for the nurse-rating method, about 18 percent for the patient-report approach, and about 5 percent using the physiological assessments.

Inspection of the residual error values reveals a substantial amount of unaccounted-for variance for each of the patient-report measures (60 percent) and the physiological assessment measures (72 percent). However, the observed measures, which include the nurse-rating method, contain only a small proportion of total variance as uncorrelated error variance (11 percent).

One additional set of parameter estimates appears in Figure 1. Linkages among the hypothesized constructs indicate a strong positive association among all three types of compliance. This result suggests that if a patient is found to comply with one aspect of the medical regimen, he/she is likely to comply with other components of the regimen as well.

The parameter estimates obtained for the model shown in Figure 1 fit the data well. The estimated relationships among the measures

(there are 45 such relationships) show a mean deviation from the observed correlations of 0.037 (in no instance was the discrepancy more than 0.19). The Chi-square test comparing the sample correlation matrix with the one reproduced by the parameter estimates of the model yields a nonsignificant Chi-square ($x^2 = 24.14$, $df = 21$, $p = 0.28$).

DISCUSSION

It appears that the nurse rating approach is the most nearly valid of the three methods of measuring patient compliance studied (even though, as mentioned elsewhere, it contains about 50 percent valid variance). The physiological-assessment method does not perform as well with respect to validity, although it is an adequate indicator of between-dialysis weight gain and, thus, of compliance with fluid limiting instructions. The patient-report technique, with only about 12 percent valid variance, seems to be a rather doubtful method of measuring compliance.

The notably high residual-error variances produced by the physiological-assessment measures may represent a significant empirical finding in their own right. Apparently, much of the variation in the physiological-assessment measure (over 70 percent) can be attributed to factors unrelated to patient compliance. For example, the serum potassium level is not very sensitive to changes in dietary behavior. Moreover, both serum potassium and serum phosphorus levels can be influenced by the presence of a coexisting catabolic process or by the degree of adequacy of the dialysis treatment. The high residual-error variance associated with the between-dialysis weight gain measure is less easily understood and is probably the result of several factors: (1) failure to adjust for varying lengths of time between dialysis treatments (they varied from 2 to 3 days); (2) failure to take account of the fact that some patients had urine output; (3) lack of standardized measurement procedures for obtaining pre-postdialysis weight gains; and (4) errors in recording pre-postdialysis weights.

The relatively high method effects in measures obtained from nurses' ratings reflect a tendency on the part of the nurses to classify patients as either "compliant" or "noncompliant" without discriminating among compliance behaviors relative to different aspects of the medical regimen. Correlations between nurse ratings of patient compliance and selected patient characteristics revealed that the nurses consistently rated patients who had more formal education, who were

married, and who were white, as compliant with all three aspects of the medical regimen. Thus, it appears that an important component of the nurse rating which contributes to its consistency across different types of compliance is a set of stereotypes reflecting assumed sociodemographic patient characteristics.

The strong positive correlations obtained among the three types of compliance indicate substantial overlap among the different constructs. These findings suggest that compliance with one aspect of the regimen represents compliance with other components of the regimen as well. This pattern of response is by no means typical of adherence to medical advice; it simply may be more characteristic of behaviors that are similar in form.

The likelihood of generalizing the results of this research is limited both by the sample of patients represented and the measurement methods used. The sample of dialysis patients studied does not represent all dialysis patients, nor are dialysis patients similar to most other types of patients. Further, only three methods of measurement were compared. Since, in overview, external validity can only be deduced, replication is always essential. The same methods and materials reported here should be employed with other types of patients (copies of all study materials may be obtained from the authors), and the research extended to incorporate new methods.

As a study of construct validity, the research for this article treats the set of measures as potentially equivalent—no "true" criterion is available in this situation. Physiological measures do not manifest themselves as the most coherent, consistent measures of adherence. By the standards of validity, these measures could not be taken as "true" or as criteria against which to judge the adequacy of other forms of measurement. Practically speaking, it appears that a multiple assessment of adherence, using a variety of measurement processes, is the best way to assure accuracy in gauging levels of patient adherence to medical instructions.

REFERENCES

1. Gordis, L. Conceptual and methodologic problems in measuring patient compliance. In R. B. Haynes, D. W. Taylor, and D. L. Sackett (eds.). *Compliance in Health Care*, Baltimore: Johns Hopkins University Press, 1979, pp. 23-45.
2. Dunbar, J. Issues in assessment. In S. J. Cohen (ed.). *New Directions in Patient Compliance*, Lexington, Massachusetts: Lexington Books, 1979, pp. 41-57.

3. Feinstein, A., H. F. Wood, J. A. Epstein, et al. A controlled study of three methods of prophylaxis against streptococcal infection in a population of rheumatic children, II. Results of the first three years of the study, including methods for evaluating the maintenance of oral prophylaxis. *New England Journal of Medicine* 260:697, 1959.
4. Gordis, L., M. Markowitz, and A. M. Lilienfeld. Inaccuracy of using interviews to estimate patient reliability in taking medications at home. *Medical Care* 7:49, 1969.
5. Berkowitz, N., M. Malone, M. Klein, and A. Eaton. Patient follow-through in the Outpatient Department. *Nursing Research* 12:16, 1963.
6. Davis, M. Variations in patients' compliance with doctors' orders: Analysis of congruence between survey responses and results of empirical investigations. *Journal of Medical Education* 41:1037, 1966.
7. Haynes, R. B. A Critical Review of the "Determinants" of Patient Compliance with Therapeutic Regimens. In D. L. Sackett and R. B. Haynes (eds.). *Compliance with Therapeutic Regimens*, Baltimore: Johns Hopkins University Press, 1976, pp. 26–39.
8. Evans, R., W. B. Hansen, and M. B. Mittlemark. Increasing the validity of self-reports of behavior in a smoking in children investigation. *Journal of Applied Psychology* 62:521, 1978.
9. Campbell, D. T., and D. W. Fiske. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin* 29:259, 1959.
10. Conger, A. J. Evaluation of multitrait-multimethod factor analysis. *Psychological Bulletin* 75:416, 1971.
11. Jackson, D. N. Multimethod factor analysis in evaluation of convergent and discriminant validity. *Psychological Bulletin* 72:30, 1969.
12. Jackson, D. N. Comments on Conger's "Evaluation of multitrait-multimethod factor analysis." *Psychological Bulletin* 75:421, 1971.
13. Boruch, R. F., and L. Wolins. A procedure for estimation of trait, method and error variance attributable to a measure. *Educational and Psychological Measurement* 30:547, 1970.
14. Althauser, R. P., and T. A. Heberlein. Validity and the Multitrait-Multimethod Matrix. In E. F. Borgatta and G. W. Bornstedt (eds.). *Sociological Methodology*, San Francisco: Jossey-Bass, 1970, pp. 151–69.
15. Alwin, D. F. Approaches to the Interpretation of Relationships in the Multitrait-Multimethod Matrix. In H. L. Costner (ed.). *Sociological Methodology 1973–74*, San Francisco: Jossey-Bass, 1974, pp. 79–105.
16. Burt, R. S. Confirmatory factor-analysis structures and the theory construction process. *Sociological Methods and Research* 2:131,1974.
17. Jöreskog, K. G. A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika* 34:183, 1969.