JOURNAL OF RESEARCH IN PERSONALITY 27, 49–87 (1993)

Assessing Construct Validity in Personality Research: Applications to Measures of Self-Esteem

RICHARD P. BAGOZZI

University of Michigan

This article addresses the assessment of convergent and discriminant validity in personality research. Four approaches are compared and contrasted for the analysis of classic multitrait-multimethod data, where three or more traits are measured with indicators derived from three or more methods. The approaches are the Campbell and Fiske criteria, the confirmatory factor analysis model, the correlated uniqueness model, and the direct product model. Pros and cons of the approaches are pointed out through a reanalysis of data originally collected by Van Tuinen and Ramanaiah, where global self-esteem, social self-esteem, and orderliness were each measured by true–false inventories, multipoint inventories, and simple self-reports. The Discussion considers guidelines for choosing among the approaches and further addresses procedures for nontraditional multitrait–multimethod data.

Measures of personality traits or states reflect measurement error as well as the theoretical content presumed to underlie the traits or states. In turn, measurement error can be conceived to consist of random and systematic components. Thus, one might represent measure variance as the sum of true or theoretical variance, plus random error and systematic error.

A common source of systematic error in personality research is method error. Method error refers to variance attributable to the measurement procedure(s) rather than to the construct of interest, and examples include halo effects, social desirability distortions, acquiescence tendencies, evaluation apprehension, demand artifacts, and key informant biases associated with peer or expert ratings (e.g., Campbell, 1955; Funder, 1989; Ganster, Hennessey, & Luthans, 1983; Nicholls, Licht, & Pearl, 1982;

Appreciation is expressed to two reviewers for comments and suggestions on an earlier version of this article. Correspondence and reprint requests should be sent to Richard P. Bagozzi, School of Business Administration, University of Michigan, Ann Arbor, Michigan 48109-1234.

0092-6566/93 $5.00

Paulhus, 1989; Rosenthal & Rosnow, 1969; Seidler, 1974; Winkler, Kanouse, & Ware, 1982).

Personality researchers should be concerned with measurement error because such error can have serious confounding influences on the interpretation of empirical research (e.g., Campbell & Fiske, 1959; Fiske, 1982). It is well known that random error frequently attenuates the observed relationships among variables in statistical analyses and therefore may produce errors in inference. Less well known is the possibility that random error can actually inflate parameter estimates under some circumstances in multivariate analyses, depending on the pattern and magnitude of such errors among predictors (e.g., Bollen, 1989). Likewise, method error may suppress or magnify relationships among variables and contribute to Type I or Type II errors if not taken into account (e.g., Bagozzi, Yi, & Phillips, 1991).

Because measurement error (i.e., random error and method variance) provide potential threats to the interpretation of research findings, it is important to validate measures and disentangle the distorting influences of these errors in the course of testing personality theories. This can be achieved by using multiple measures and multiple methods in measurement and hypotheses testing (e.g., Campbell & Fiske, 1959). Using a single measure of each variable in a theory under test does not permit one to take reliability into account in analyses. Similarly, with only a single method one cannot distinguish substantive (i.e., trait) variance from unwanted method variance, because each attempt to measure a concept is contaminated by irrelevant aspects of the method employed.

Construct validity, which is defined broadly as the extent to which an operationalization measures the concept it is supposed to measure (e.g., Cook & Campbell, 1979), is a central issue in personality research (e.g., Ozer, 1989). Given multiple measures obtained with multiple methods, construct validation can be assessed through an inspection of the multitrait–multimethod (MTMM) matrix, the correlation matrix for different concepts (i.e., traits or states) when each of the concepts is measured by different methods (e.g., Campbell & Fiske, 1959). Without assessing construct validity one cannot estimate and correct for the confounding influences of random error and method variance, and the results of theory testing may be ambiguous. That is, a hypothesis might be rejected or accepted because of excessive error in measurement, not necessarily because of the inadequacy or adequacy of theory.

In recent years, there has been an explosion in procedures advocated for the investigation of construct validity, and at least a dozen can be identified.[1] Unfortunately, the assumptions upon which the procedures

[1] The procedures are, in rough chronological order: the classic criteria of Campbell and Fiske (1959), the analysis of variance (e.g., Boruch & Wolins, 1970), first-order confirmatory

are based vary widely, their implementation is confusing to all but the most technically oriented researchers, and the interpretation of findings harbor numerous pitfalls. Little guidance exists on when and how to select among procedures. Individual articles in both the psychometric and applied psychology literatures generally focus on only one procedure and give passing reference to alternatives without fully considering the standards and trade-offs of each. Moreover, little integrative critical commentary exists on the approaches. As a consequence, we lack a coherent approach to construct validity, and the body of knowledge reflected in empirical work is rather piecemeal and inconclusive.

The present article compares and contrasts state of the art methods for investigating construct validity in personality research. The aim is to point out pros and cons of leading procedures and suggest guidelines for conducting construct validation studies. Empirical illustrations are performed through reanalyses of data originally examined by Van Tuinen and Ramanaiah (1979). Van Tuinen and Ramanaiah (1979) used a MTMM matrix to investigate three traits—global self-esteem, social self-esteem, and orderliness—measured by three methods: true–false inventories, multipoint inventories, and simple self-ratings.[2] A sample of 196 undergraduate psychology students was used. Table 1 presents the MTMM matrix.

To facilitate the presentation, each procedure is described, illustrated, and critiqued, in turn, before considering another procedure. The procedures are the Campbell and Fiske method, the first-order confirmatory factor analysis model, the correlated uniqueness model, and the direct product model. Because of serious shortcomings pointed out by others (e.g., Bagozzi, Yi, & Phillips, 1991; Schmitt & Stults, 1986), we will not

---

factor analysis (e.g., Werts & Linn, 1970; Jöreskog, 1974), exploratory factor analysis (e.g., Golding & Seidman, 1974; Jackson, 1975), the generalized proximity function (e.g., Hubert & Baker, 1979), smallest space analysis (e.g., Levin, Montag, & Comrey, 1983), the direct product model (e.g., Browne, 1984), the second-order confirmatory factor analysis model with measures loading indirectly on traits and methods (e.g., Marsh & Hocevar, 1988), the correlated uniqueness model (e.g., Kenny, 1976; Marsh, 1989), the first-order confirmatory factor analysis model with separate factors for traits, methods, and measure specificity (e.g., Kumar & Dillon, 1990), and panel models (e.g., Bagozzi & Heatherton, 1992).

[2] To measure simple self-ratings, the researchers presented each subject with a description of the respective traits, and responses indicating the extent to which one felt he or she possessed the traits were recorded on seven-point rating scales. Global self-esteem was measured with (a) the 25-item true–false short form of Coopersmith's (1967) Self-Esteem Inventory (Crandall, 1973) and (b) the Tennessee Self-Concept Scale, a 90-item inventory using five-point rating scales (Fitts, 1965; Fitts, Adams, Radfor, Richard, Thomas, & Thompson, 1971). Social self-esteem was measured with (a) the 20-item true–false JPI Self-Esteem Scale (Jackson, 1970) and (b) Eagly's (1967) revised version of the Janis–Field Feelings of Inadequacy Scale (Hovland & Janis, 1959), a 20-item inventory. Orderliness was measured by (a) the 20-item true–false PRF Order Scale (Jackson, 1967) and (b) the 20-item CPS Order Scale (Comrey, 1970), which used seven-point response items.

TABLE 1

PEARSON PRODUCT-MOMENT CORRELATIONS AMONG MEASURES OF GLOBAL SELF-ESTEEM, SOCIAL SELF-ESTEEM, AND NEED FOR ORDER AS PROVIDED BY TRUE–FALSE, MULTIPOINT, AND SIMPLE SELF-RATING SCALES (AFTER VAN TUINEN AND RAMANAIAH, 1979)

Measures	A1	A2	A3	B1	B2	B3	C1	C2	C3
A. True–false inventory									
1. Global self-esteem	(.83)								
2. Social self-esteem	.58	(.85)							
3. Need for order	.17	.14	(.74)						
B. Multipoint inventory									
1. Global self-esteem	**.75**	.45	.23	(.93)					
2. Social self-esteem	.72	**.74**	.16	.65	(.91)				
3. Need for order	.09	.06	**.68**	.25	.08	(.85)			
C. Simple self-rating									
1. Global self-esteem	**.58**	.53	.14	**.62**	.68	.09	(.63)		
2. Social self-esteem	.47	**.74**	.10	.40	**.69**	.07	.58	(.74)	
3. Need for order	.22	.18	**.63**	.34	.22	**.56**	.30	.23	(.82)

Note. $N = 196$. Reliability coefficients are in parentheses. Monotrait–heteromethod correlations (i.e., convergent validity coefficients) are shown in boldface. Heterotrait–monomethod correlations are enclosed by solid triangles, and heterotrait–heteromethod correlations are enclosed by dashed triangles.

discuss the use of the analysis of variance, exploratory factor analysis, smallest-space analysis, the generalized proximity function, and the second-order confirmatory factor analysis model with measures loading directly on trait and method factors. We reserve for the Discussion the treatment of the hierarchical confirmatory factor analysis model with measures loading indirectly on traits and methods, the first-order confirmatory factor analysis model with separate factors for traits, methods, and measure specificity, and panel models. The latter three procedures make more assumptions and/or require a greater number of measures for each trait–method combination than the procedures considered in detail herein and therefore are more difficult to implement in practice. Hence our decision to place less emphasis on them.

THE CAMPBELL AND FISKE APPROACH

Rationale

Campbell and Fiske (1959) proposed two aspects of construct validity: convergent and discriminant validity. Convergent validity is the degree to which multiple attempts to measure the same concept are in agreement. The idea is that two or more measures of the same thing should covary highly if they are valid measures of the concept. Discriminant validity is

the degree to which measures of different concepts are distinct. The notion is that if two or more concepts are unique, then valid measures of each should not correlate too highly.

Campbell and Fiske (1959) originally asserted that the most stringent test of construct validity requires that maximally dissimilar methods be employed. Such a practice increases our confidence in the interpretation of evidence for convergent validity in that agreement among putative measures of the same concept are unlikely to be determined by shared method biases. While we agree with Campbell and Fiske's rationale for use of maximally different methods in the assessment of convergent validity, we would argue that such a practice actually makes it easier to achieve discriminant validity. A more stringent test of discriminant validity results when methods are similar. Because similar methods are likely to inflate correlations among measures of different concepts, achievement of discriminant validity under such conditions actually provides strong evidence for the distinctiveness of measures of different concepts. Thus, the most informative tests of construct validity will be obtained when one uses both maximally dissimilar and similar methods in the same investigation. Of course, this recommendation represents an ideal and may be difficult to implement in practice.

To make formal the ideas contained in the definitions of convergent and discriminant validity, Campbell and Fiske (1959) developed four desiderata based on the inspection of the MTMM matrix. First, convergent validity is achieved when the monotrait–heteromethod correlations between the same traits across different methods (i.e., "the validity diagonal" values) are "significantly different from zero and sufficiently large" (Campbell & Fiske, 1959, p. 82). The validity diagonal values represent correlations between measures of the same concept by different methods (see Table 1). Measures of the same concept, no matter how derived, should be highly correlated if they validly measure a common concept. Establishment of convergent validity provides evidence that multiple measures of a concept obtained by multiple methods potentially indicate the same underlying concept. If the validity diagonal values are nonsignificant or too low in magnitude, there is little basis to argue that the measures tap the same concept, and consideration of discriminant validity is not warranted.

However, if convergent validity is attained, this only provides minimal evidence for the construct validity of measures. It is also possible that the measures reflect variance due to other concepts or methods biases and are not unique. Campbell and Fiske (1959) therefore recommended that discriminant validity also be assessed and proposed three criteria to do so. The first stipulates that the validity diagonal values should be higher than their corresponding heterotrait–heteromethod coefficients (see Table

1). That is, each validity diagonal value should be higher than the coefficients lying in the columns and rows of its adjacent heterotrait–heteromethod triangles. In other words, efforts to measure the same concept by different methods should yield higher correlations than efforts to measure different concepts by different methods. For example, in Table 1, $r_{A2,B2} = .74$ (the correlation between the measures of social self-esteem as obtained by the true–false and multipoint inventories) is greater than $r_{A1,B2} = .72$, $r_{A2,B3} = .06$, $r_{A2,B1} = .45$, and $r_{A3,B2} = .16$; similar inequalities should hold for comparisons of each remaining validity diagonal value with the appropriate heterotrait–heteromethod coefficients. If this criterion for discriminant validity fails, the implication is that convergence of measures on any individual concept is dependent on convergence of measures on other concepts and/or confounded with method variance, thereby bringing into question discriminant validity.

The second discriminant validity criterion specifies that the monotrait–heteromethod coefficients should be higher than their corresponding heterotrait–monomethod coefficients. Efforts to measure the same concept by different methods should produce higher correlations than efforts to measure different concepts by the same method. For instance, in Table 1 $r_{B1,C1} = .62$ is greater than $r_{B1,B3} = .25$, $r_{C1,C2} = .58$, and $r_{C1,C3} = .30$ but less than $r_{B1,B2} = .65$, thus revealing one violation of the second discriminant validity criterion; similar comparisons should be made between each remaining validity diagonal value and its corresponding coefficients in the adjacent heterotrait–monomethod triangles. The failure of this criterion points to a confounding of method variance with true variance and suggests problems with discriminant validity.

The final criterion for discriminant validity is that the patterns of correlations should be the same among the heterotrait–monomethod and heterotrait–heteromethod correlations. For example, this criterion is met when we compare the correlations in the heterotrait–monomethod triangle for the true–false inventory in Table 1 to the correlations in the lower heterotrait–heteromethod triangle between the true–false inventory and simple self-rating methods: $r_{A1,A2} > r_{A1,A3} > r_{A2,A3}$ and $r_{A1,C2} > r_{A1,C3} > r_{A2,C3}$. The ordering of correlations within triangles can be compared across all triangles by use of Kendall's Coefficient of Concordance (Siegel, 1956, pp. 229–238). When this criterion holds, correlations among measures of concepts will be independent of methods. But when it fails, method variance will be operative differentially across the correlations.

Illustration of Campbell and Fiske's Criteria

Campbell and Fiske's (1959) criteria were applied to the data in Table 1. Because the validity diagonal values are all large and significantly different from zero ($p < .01$)—i.e., the correlations range from .56 to

.75—it can be concluded that convergent validity is achieved. The first discriminant validity criterion involves 36 comparisons of correlations. Only one of the comparisons shows a violation of the criterion: $r_{B1,C1}$ = .62 is actually less than $r_{B2,C1}$ = .68. As the proportion of failures of the first criterion for discriminant validity is p = 1/36 = .028, which is less than what one would expect at a chance level of .05, and at the same time the magnitude of the violation is relatively small (i.e., Δr = .06), one might conclude that the criterion is met overall. A more stringent standard would be to require that the results of all comparisons of correlations with the validity diagonal values not only be in the proper direction but achieve statistical significance (at say, $p < .05$). For the first discriminant validity criterion, this results in six violations, which is greater than what one would expect by chance.

The second discriminant validity criterion also entails 36 comparisons. Here we find that three violations of directionality can be identified: $r_{A1,C1}$ = .58 is not greater than either $r_{A1,A2}$ = .58 or $r_{C1,C2}$ = .58, and $r_{B1,C1}$ = .62 is in fact less than $r_{B1,B2}$ = .65. Although the differences between the correlations are relatively small (i.e., they range from .00 to .03), the proportion of failures of the criterion is p = 3/36 = .083, which is greater than what one would expect at a chance level of .05. Consequently, achievement of the second discriminant validity criterion is questionable.

The third discriminant validity criterion was examined through a comparison of the rank order of correlations across triangles shown in Table 1. The coefficient of concordance was .222 which results in a nonsignificant χ^2 (2, N = 196) = 4.00. Hence, one cannot reject the hypothesis that the patterns of correlations in the heterotrait–heteromethod and heterotrait–monomethod triangles are the same, and the third discriminant validity criterion is therefore met.

Critique of Campbell and Fiske's Procedure

How well can Campbell and Fiske's criteria be relied upon? One answer to this question can be addressed by evaluating the assumptions underlying Campbell and Fiske's criteria. Four assumptions are noteworthy: namely, the criteria are based on the premises that traits and methods are uncorrelated, methods affect all traits equally, methods are orthogonal, and measures are equally reliable (e.g., Campbell & Fiske, 1959; Schmitt & Stults, 1986). The first assumption may not be unreasonable in practice, as traits and methods are frequently unconfounded. However, one case where traits and methods can be related in personality research is when peers or experts rate subjects and the subjects rate themselves on characteristics for which they and the key informants possess an implicit theory as to the nature or origin of the characteristics.

The other three assumptions behind the Campbell and Fiske procedure

are highly unlikely in most contexts for personality research. Measures of traits will be differentially affected by methods to the extent of heterogeneity in the traits and methods under study. Studies dealing with scale development, particularly with multifaceted constructs (e.g., Carver, 1989), and investigations into construct validation strive for heterogeneity by design. Intercorrelations among methods are difficult to avoid as well. Alternative methods based on self-reports or judgments performed by key informants typically correlate at least at moderate levels. Further, method variance and measure reliability generally vary considerably even when similar instrumentation is used to tap traits. Thus, there is reason to question the assumptions underlying the Campbell and Fiske procedure when applied in typical personality research contexts.

Another problem with Campbell and Fiske's procedure is that no precise standards are provided for ascertaining how well the criteria are met. The rules of thumb offered as to the proportion of violations are rather arbitrary and depend on a qualitative assessment of confirming and disconfirming incidents of differences in observed correlations. By focusing on the number of times selected correlations are greater than others, Campbell and Fiske's procedure neglects the importance of the magnitudes of differences between pairs of correlations.

Reliance on the observed correlations provides a rather imprecise and potentially misleading basis for assessing construct validity. An observed correlation will reflect random error and method biases in addition to the true association among measures of traits. The Campbell and Fiske procedure provides no information as to the separate amounts of variation in measures due to traits, methods, and random error.

CONFIRMATORY FACTOR ANALYSIS MODEL

Rationale

An alternative to the Campbell and Fiske procedure that has seen recent application in personality research (e.g., Bagozzi, 1991) is the confirmatory factor analysis (CFA) model (e.g., Jöreskog, 1974). As applied to MTMM matrix data, the CFA model hypothesizes that the total variation in measures can be written as a linear combination of trait, method, and error effects (e.g., Jöreskog, 1974).

The CFA model is perhaps best introduced by way of a diagram. Figure 1 presents an intuitive description of the CFA model. The three hypothesized traits—global self-esteem, social self-esteem, and orderliness—and the three methods—true-false inventory, multipoint inventory, and simple self-rating—are drawn as circles. These correspond to factors in factor analysis. Note that each trait factor is connected to three boxes with arrows. The boxes represent the actual observed measurements obtained

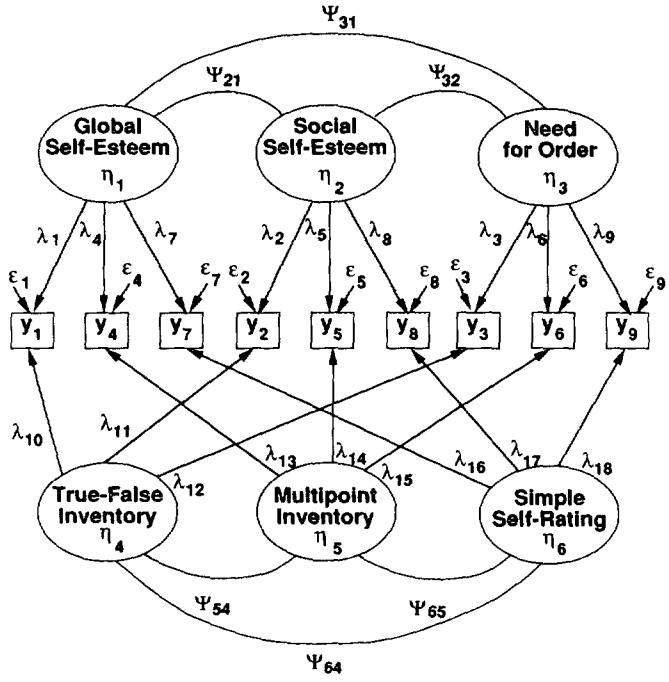

FIG. 1. Illustration of the confirmatory factor analysis model for three traits (global self-esteem, social self-esteem, need for order) and three methods (true-false inventory, multipoint inventory, simple self-rating).

by Van Tuinen and Ramanaiah (1979), of which a total of nine result for the three traits obtained by the three methods. For example, y_1 refers to the measurement of global self-esteem by the true-false inventory, y_4 to the measurement of global self-esteem by the multipoint inventory, and y_7 to the measurement of global self-esteem by the simple self-rating. Each measurement has three arrows terminating into it. The arrows from the trait factors to measures stand for variance in the measures that is due to the underlying trait; the nine λs connected to these arrows are factor loadings relating trait factors to observed measures. The arrows from the methods to measures reflect variance that is due to the procedures used to obtain responses; the nine λs attached to these arrows are factor loadings relating method factors to observed measures. The nine short arrows with ε_i at the origins represent variation in the measures that is due to random error plus measure specificity. Finally, the curved lines connecting pairs of factors indicate correlations between factors and are designated as ψ_{jk}.

If we interpret each measure as an observation whose variation we

desire to explain, we can interpret the CFA model in Fig. 1 as displaying the sources of that variation in three senses: variation due to trait (i.e., the theoretical concept of interest), method (i.e., the measurement procedures), and error (i.e., unexplained random fluctuations). More formally, the general form of the CFA model for the MTMM matrix with r traits and s methods can be expressed through two sets of equations (e.g., Jöreskog, 1974):

$$\mathbf{y} = [\Lambda_T \Lambda_M] \begin{bmatrix} \eta_T \\ \eta_M \end{bmatrix} + \varepsilon \tag{1}$$

$$\Sigma = \Lambda_T \Psi_T \Lambda'_T + \Lambda_M \Psi_M \Lambda'_M + \theta,$$

where $\mathbf{y}$ is a vector of $r \times s$ observed measures for r traits and s methods, $\eta = [\eta_T \eta_M]$ is an $(r + s) \times 1$ vector of trait and method factors, ε is a vector of $r \times s$ residuals for $\mathbf{y}$, Σ is the implied variance–covariance matrix for $\mathbf{y}$, Ψ_T is an $r \times r$ correlation matrix for traits, Ψ_M is an $s \times s$ correlation matrix for methods, Θ is the vector of unique variances for ε, $\Lambda_T = [\Lambda_1, \Lambda_2, \ldots, \Lambda_s]'$, Λ_j is an $r \times r$ diagonal matrix with trait factor loadings for the r traits measured by the jth method, and

$$\Lambda_M = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & 0 \\ 0 & 0 & \cdots & 0 \; \lambda_s \end{bmatrix},$$

where λ_j is an $r \times 1$ vector of factor loadings for the jth method.

Four useful hypotheses to examine with respect to method and trait effects in the CFA model are the following (e.g., Widaman, 1985):

Model 1. The model hypothesizing that only unique variances in measures of personality traits are freely estimated (i.e., the null model). This model hypothesizes that the observed measures correlate zero in the population.

Model 2. The model hypothesizing that variation in measures can be explained completely by traits plus random error (i.e., the trait-only model). This model assumes that method variance is negligible and that the measures reflect only trait and error variance.

Model 3. The model hypothesizing that variation in measures can be explained completely by methods plus random error (i.e., the method-only model). This model assumes that trait variance is negligible and that the measures reflect only method and error variance.

Model 4. The model hypothesizing that variation in measures can be

explained completely by traits, methods, and error (i.e., the trait-method model). This model is the structural equation operationalization of the MTMM matrix as shown in Eqs. (1) and (2) and Fig. 1.

The four models defined above can be tested with statistical programs such as EQS (Bentler, 1989) or LISREL (Jöreskog & Sörbom, 1989). The overall fit of each model can be tested by using the maximum likelihood chi-square statistic provided in the outputs of these programs. Assuming that the models are valid and that the observed measures follow multivariate normal distributions, the statistic is asymptotically distributed as a chi-square variable with its associated degrees of freedom. A nonsignificant chi-square indicates that one can accept (fail to reject) a model. Interpretation of a significant chi-square, which suggests lack of fit, may be ambiguous in the sense of reflecting model misspecification and/or violation of assumptions such as multivariate normality (e.g., Mulaik et al., 1989). In addition, large values of the chi-square test relative to degrees of freedom can result from large sample sizes. It is useful to use chi-square measures to compare models in nested sequences of hypotheses (e.g., Bentler & Bonett, 1980; Mulaik et al., 1989).

The nested hypotheses implied by Models 1–4 above can be tested by comparing chi-square values. A model is nested in another when constraints placed on the latter yield the former. The difference in chi-square values for the models will be distributed chi-square with degrees of freedom equal to the difference between the two models. A test of the significance of trait variance is provided by comparing chi-square tests between Models 1 and 2 and between Models 3 and 4. Similarly, a test of the significance of method variance is provided by comparing Models 1 and 3, as well as Models 2 and 4.

In addition to testing formally for trait and method effects, the CFA models can be used to partition the variance in measures in diagnostically useful ways and to estimate parameters that provide insights into measurement properties and construct validity. The partitioning of variance into trait, method, and error is revealed, respectively, in the squared factor loadings in Λ_T and Λ_M, and in Θ. Further, as we illustrate in the empirical analyses below, useful information is provided in parameter estimates for correlations among traits and among methods, as well as in error variances and factor loadings.

Before we illustrate the CFA model, it is important to point out its advantages over the Campbell and Fiske procedure. Under the CFA model, a variety of measures of fit are provided for an overall model, whereas no omnibus test is possible for the Campbell and Fiske procedure. Moreover, estimates and tests of significance of parameters are derived for the CFA model, and formal tests of trait and method effects are

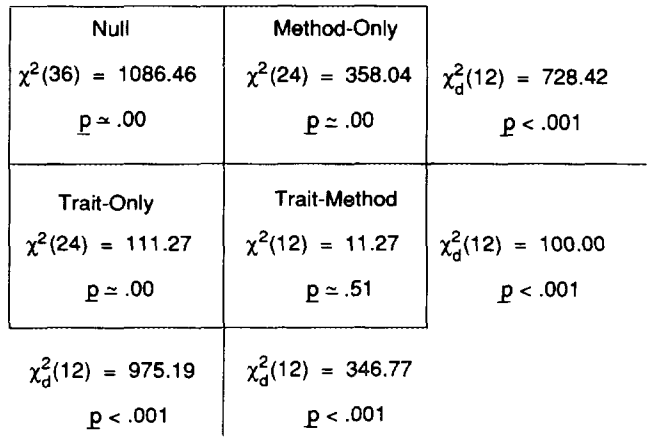

FIG. 2. Summary of nested additive confirmatory factor analysis tests for trait and method effects.

possible. These features are not part of the Campbell and Fiske procedure. Likewise, the CFA model yields a partitioning of variance into trait, method, and error components, but the Campbell and Fiske procedure only suggests qualitative hints as to the presence of trait and method effects. Finally, with regard to the restrictive assumptions noted above for the Campbell and Fiske procedure, it should be noted that methods can correlate freely and affect measures to different degrees under the CFA model, and the reliability of measures can be freely estimated, rather than assuming that they are equal to unknown values. Indeed, by imposing certain restrictions, it is possible to estimate correlations between traits and methods.[3]

Illustration of the CFA Model

We applied the CFA model to the data in Table 1.[4] Figure 2 summarizes the chi-square goodness-of-fit measures for the four models described

[3] For many CFA models, it is possible to permit all traits and methods to intercorrelate freely among themselves and achieve identification in a technical sense. However, in practice such a specification nearly always leads to empirical identification problems (e.g., Marsh, 1989). However, if one has reason to expect that either traits or methods are orthogonal, this can be exploited through appropriate constraints and the correlations between traits and methods can be estimated, if desired.

[4] In all analyses to follow, the correlation matrix of variables is used as input for didactic purposes. In general, it is preferable to use the covariance matrix as input in order to obtain correct chi-square measures and estimates of standard errors (Cudeck, 1989). Under certain conditions (e.g., when a model is scale invariant and no constraints are imposed on correlations among factors and on error variances), use of a correlation matrix as input may

above and presents the relevant chi-square difference tests for the nested model comparisons. The first thing to note is that the trait–method model fits quite well: $\chi^2(12, N = 196) = 11.27, p \simeq .51$. This is the model hypothesizing that variation in measures can be explained as an additive function of trait, method, and error effects. To formally test for method effects, we compare the null model to the method-only model and the trait-only model to the trait–method model. It can be seen that the addition of method effects results in a significant improvement in fit for both comparisons: $\chi^2_d(12, N = 196) = 728.42, p < .001$ and $\chi^2_d(12, N = 196) = 100.00, p < .001$. To formally test for trait effects, we compare the null model to the trait-only model and the method-only model to the trait–method model. The results show that the addition of trait effects leads to a significant improvement in fit for both comparisons: $\chi^2_d(12, N = 196) = 975.19, p < .001$ and $\chi^2_d(12, N = 196) = 346.77, p < .001$. Thus, we find that significant amounts of both trait and method variance are present.

Before we examine the nature and extent of trait and method effects and the degree of error in measures, it is informative to scrutinize the overall magnitude of information accounted for by the CFA model from a practical standpoint. One way to do this is to compute the noncentralized normed fit index (NCNFI).[5] The NCNFI is defined as

$$\text{NCNFI} = \frac{(\chi^2_0 - df_0) - (\chi^2_j - df_j)}{(\chi^2_0 - df_0)},$$

where χ^2_0 is the chi-square value for the null model, χ^2_j is the chi-square value for a focal model (e.g., the trait–method model), df_0 is the degrees of freedom for the null model, and df_j is the degrees of freedom for the focal model. The NCNFI is a modification of an index originally proposed by Bentler and Bonett (1980). By subtracting the degrees of freedom from their corresponding chi-square values, we correct for any bias that is due to small samples (e.g., Bentler, 1990; McDonald & Marsh, 1990). Computation of the NCNFI for the trait–method model yields a value of 1.00, which is greater than the .90 rule of thumb suggested as a minimum satisfactory level by Bentler and Bonett (1980). Therefore, the trait–method model accounts for a significant proportion of variance from a practical point of view. Had the NCNFI fallen below .90, we would have

yield correct asymptotic standard errors and chi-square values. This holds for some of the models considered in this article.

[5] The NCNFI has been termed the "relative noncentrality index" (RNI) by McDonald and Marsh (1990). The NCNFI will be equal to the normed fit index in large samples. Bentler (1990) and McDonald and Marsh (1990) performed simulations showing the properties of the NCNFI and comparing it with many other indexes.

concluded that a significant proportion of variance remained unexplained, bringing into question the adequacy of the model. It should be acknowledged that the NCNFI is a heuristic, not a statistic. Its sampling distribution is unknown, and the .90 rule of thumb should be regarded as only a rough guideline. A final point to make is that, if one desires to ascertain practical relevance while incorporating a penalty based on the number of parameters estimated (i.e., the complexity of a model), then one may employ the Tucker and Lewis (1973) index or a noncentralized adaptation where appropriate degrees of freedom are subtracted from the chi-square values (Bentler, 1990; McDonald & Marsh, 1990).

A relatively common outcome when performing a CFA of the trait–method model is the presence of negative error variance estimates. Indeed, the findings for the trait–method model summarized in Fig. 2 revealed three negative error variances, albeit all nonsignificant. One solution to the problem of negative error variances is to fix these to zero. We will discuss the pros and cons of such a practice and consider other alternatives below under "Critique of the CFA Model." For now, we illustrate this remedy on the data at hand. When multiple error terms show negative values, it is sometimes sufficient to fix only the highest value to zero and to rerun the trait–method model. This was in fact done (i.e., $\Theta_{\varepsilon 2} = -1.28$, $SE = 1.10$, was constrained to zero). The results show that the overall model fit well with this specification (i.e., $\chi^2(13, N = 196 = 13.99, p \simeq .37)$, and, importantly, no negative error variances occurred and no improper solutions were found for any parameter estimates.

Table 2 presents the parameter estimates for the trait–method model. The factor loading matrix in the top of the table corresponds to the pattern shown in Fig. 1, in which 18 factor loadings are estimated. Note that the factor loadings for traits are relatively high for the multipoint inventory of global self-esteem and the three measures of orderliness and are moderately high for the true–false inventory and self-rating of social self-esteem. These findings suggest that the measures are reasonable indicators of their respective trait factors. The factor loadings for the true–false inventory and self-rating of global self-esteem and the multipoint inventory of social self-esteem are quite low. These results suggest that the measures are poor indicators of their respective factors. Notice also that the factor loadings of methods for all measures of global and social self-esteem are quite high. This indicates that methods biases strongly influence the measures of self-esteem. The factor loadings of methods on the three orderliness measures are quite low, pointing to small methods biases.

As can be seen in the bottom of Table 2, all traits are distinct (i.e., each is correlated with the others at a level significantly less than 1.00). Note that the correlations among traits are corrected for measurement

TABLE 2

SUMMARY OF PARAMETER ESTIMATES FOR THE ADDITIVE CONFIRMATORY FACTOR ANALYSIS MODEL WITH THREE TRAITS AND THREE METHODS

	Traits						Methods					
	Global self-esteem		Social self-esteem		Order		True-false inventory		Multipoint inventory		Self-rating	
Construct	FL	*SE*	FL	*SE*	FL	*SE*	FL	*SE*	FL	*SE*	FL	*SE*
Global self-esteem by												
True–false inventory	.21	.18	.00[a]		.00[a]		.83	.08	.00[a]		.00[a]	
Multipoint inventory	.64	.13	.00[a]		.00[a]		.00[a]		.77	.12	.00[a]	
Self-rating	.05	.17	.00[a]		.00[a]		.00[a]		.00[a]		.82	.07
Social self-esteem by												
True–false inventory	.00[a]		.51	.12	.00[a]		.74	.08	.00[a]		.00[a]	
Multipoint inventory	.00[a]		.17	.13	.00[a]		.00[a]		.90	.09	.00[a]	
Self-rating	.00[a]		.55	.09	.00[a]		.00[a]		.00[a]		.71	.09
Order by												
True–false inventory	.00[a]		.00[a]		.84	.06	.20	.08	.00[a]		.00[a]	
Multipoint inventory	.00[a]		.00[a]		.79	.07	.00[a]		.12	.08	.00[a]	
Self-rating	.00[a]		.00[a]		.69	.07	.00[a]		.00[a]		.33	.08
	Factor intercorrelations											
Traits												
Global self-esteem	1.00[a]											
Social self-esteem	−.32	.23	1.00[a]									
Order	.26	.09	.00	.11	1.00[a]							
Methods												
True–false inventory	.00[a]		.00[a]		.00[a]		1.00[a]					
Multipoint inventory	.00[a]		.00[a]		.00[a]		.98	.04	1.00[a]			
Self-rating	.00[a]		.00[a]		.00[a]		.85	.04	.92	.05	1.00[a]	

Note. FL = factor loading. All parameter estimates differing significantly from zero are underscored.

[a] Fixed at the value reported.

TABLE 3
PARTITIONING OF VARIANCE INTO TRAIT, METHOD, AND ERROR FOR THE ADDITIVE CONFIRMATORY FACTOR ANALYSIS MODEL

Construct	Variance component		
	Trait	Method	Error
Global self-esteem by			
True–false inventory	.04	.69	.27
Multipoint inventory	.41	.60	.00[a]
Self-rating	.02	.68	.32
Social self-esteem by			
True–false inventory	.26	.55	.19
Multipoint inventory	.03	.80	.17
Self-rating	.31	.51	.18
Order by			
True–false inventory	.71	.04	.26
Multipoint inventory	.62	.02	.37
Self-rating	.48	.11	.42

[a] Constrained parameter.

error. The bottom of Table 2 also displays the disattenuated correlations among methods, where it can be seen that all methods correlate at very high levels. Indeed, the correlation between the true–false and multipoint inventories (Ψ_{21} = .98, SE = .04) and the correlation between the multipoint inventory and the self-rating (Ψ_{32} = .92, SE = .05) are not significantly different than 1.00.

Table 3 illustrates the partitioning of variance for the measures into trait, method, and error components. The findings show that only the measures of orderliness achieve reasonably high levels of trait variance, and all measures of global and social self-esteem exhibit very high levels of method bias. Error variance is generally low. The above findings for the particular CFA performed herein should be interpreted in the light of the shortcomings pointed out below. Note also that the disturbance term does not, strictly speaking, represent only random error. Measure specificity will in general be confounded with random error in the CFA and other approaches discussed in this article.

Critique of the CFA Model

A major shortcoming of the application of the CFA model to MTMM matrix data is the all too frequent occurrence of ill-defined solutions. In their examination of 435 MTMM matrices based on actual and simulated data, Marsh and Bailey (1991) report that 77 percent resulted in improper

solutions. Marsh (1989) identifies four types of ill-defined solutions common to CFA investigations of the MTMM matrix: "underidentified or empirically underidentified models . . . , failures in the convergence of the iterative procedure used to estimate parameters, parameter estimates that are outside their permissible range values (e.g., negative variance estimates called Heywood cases), or standard errors of parameter estimates that are excessively large" (p. 339).

We have already mentioned the problem of negative error variances and additional comments are in order. For the data used as an illustration in this article, negative but nonsignificant error variances were the only ill-defined solutions to arise. A negative error variance is, of course, impossible theoretically and points to serious problems. Often negative error variances will be nonsignificant, suggesting that no random error exists. However, because one normally expects at least a small amount of residual variance in self-report data, the presence of nonsignificant error variances should in the general case lead one to conclude that overfitting or a misspecified model is the case (e.g., Maxwell, 1977, p. 58; Van Driel, 1978). There is perhaps one exception to this generalization which applies to personality research. When measures of factors are formed as the sum of many well-chosen items, it is possible that this will reduce considerably the residual variance. Indeed, a particular measure so formed may exhibit nonsignificant random error. We would expect, however, that this would be a relatively rare event. Thus, while one might tolerate the occurrence of a single measure showing a nonsignificant error variance in CFA applications, when the measure is formed as the sum of many items, it would seem unwise to accept more than one such occurrence in a CFA application to the MTMM matrix. And when measures of factors consist of a single item or the sum of a small number of items, we would argue that the presence of even a single nonsignificant error variance points to an overfitted or misspecified model. For the data investigated in this paper, proper solutions were found when the error variance corresponding to the multipoint inventory of global self-esteem was fixed to zero. As this measure of self-esteem was formed as the sum of 90 five-point items, it could be argued that random error might be low and that most of the variance in the measure is due to trait and method effects.

Rindskopf (1983) proposed that negative error variances can be avoided in the CFA model by creating a new factor for each error term in the model such that the factor loading corresponding to each new factor is the square root of the error. This will guarantee that the error variance will be non-negative. One problem with this procedure is that it can lead one to accept a misspecified model. For this reason, Jöreskog and Sörbom (1989, p. 215) counsel against imposing constraints to ensure that non-

negative parameter estimates for error variances do not arise. Some evidence can be found showing that the Rindskopf parameterization is equivalent to simply fixing the offending error variance to zero (e.g., Dillon, Kumar, & Mulani, 1987). Note, however, that programs such as EQS (Bentler, 1989) derive optimal parameter estimates while assuming nonnegativity.

Another shortcoming to point out with regard to the CFA model is that the partitioning of variance into trait and method components may not, in general, yield "trait-free" and "method-free" interpretations (Kumar & Dillon, 1992). This is because the individual factor loadings take on different values corresponding to the distinct trait–method pairings. For example, factor loadings concerning a trait can vary across methods, and the corresponding variation cannot be attributed solely to the trait factor. Since each factor loading is specific to the particular trait–method combination, the associated variation is not really "trait-free" or "method-free." If the correlations among traits and the correlations among methods approach zero, the variance due to traits will be reflected in the trait loadings and the variance due to methods will be reflected in the method loadings.

However, as the correlations among traits and among methods increase, trait and method variance will be confounded. For example, a general trait factor may underlie traits such that traits are highly correlated and substantial variance in measures is primarily due to traits, while methods are relatively distinct. In such circumstances, application of the CFA model can misleadingly yield highly correlated methods, accounting for much variation in measures (e.g., Marsh, 1989). However, a good fitting CFA model in this case should not be believed because the apparent method effects are really confounded with trait effects from a general trait factor. That is, correlations among method factors may represent the convergence of the general trait factor across methods, rather than true relationships among methods. Since many applications of the MTMM matrix involve substantially correlated traits and/or methods, the interpretation of results from a CFA model should consider the potential confounding noted above.

A related and final issue to mention with respect to the use of the CFA model is that researchers sometimes jump to applications and interpretations of the trait–method model without considering the possibility that variation in measures could be a function of only traits and random error. That is, although true method effects may be absent, when the trait–method model is fit to data, the results may misleadingly show the presence of method effects. A good fitting model in such cases reflects confounding similar to that noted above. One way to avoid making false inferences in this sense is to carefully examine the trait-only model. The very high

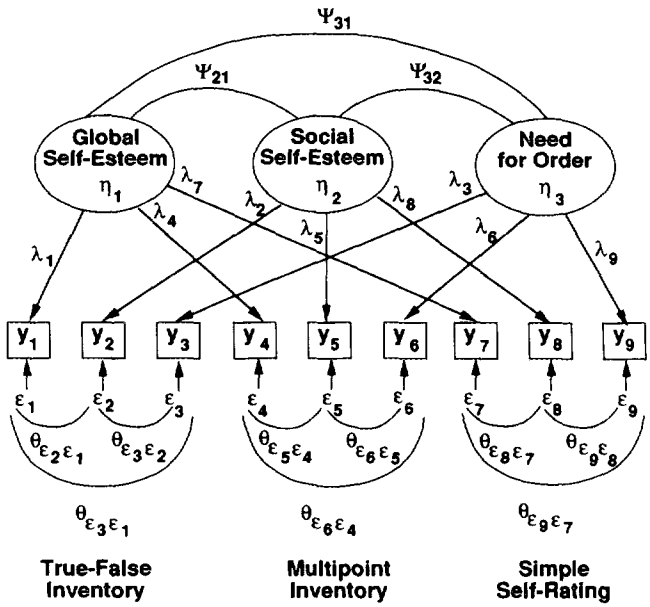

FIG. 3. Illustration of the correlated uniqueness model for three traits (global self-esteem, social self-esteem, need for order) and three methods (true-false inventory, multipoint inventory, simple self-rating).

correlations found among methods for findings summarized in Table 2, coupled with the unexpected nonsignificant correlation between global and social self-esteem, suggest that the inclusion of method factors may constitute overfitting. Indeed, although the fit of the trait-only model is unacceptable based on the chi-square test (i.e., χ^2 (24, N = 196) = 111.27, $p \simeq .00$), the results of the NCNFI = .92 indicate that the model fits well as a practical matter. Overfitting with the CFA model is more common than generally recognized. Bagozzi and Yi (1990, 1991) found that 9 of 11 MTMM matrix investigations of job satisfaction in the applied psychology literature and 2 of 4 investigations of consumer behavior in the marketing literature revealed such outcomes, yet the trait-only model could not be rejected on the basis of the NCNFI.

CORRELATED UNIQUENESS MODEL

Rationale

As a remedy for problems of overfitting and ill-defined solutions, Marsh (1989) proposed a new approach which he termed the correlated uniqueness (CU) model (see also, Kenny, 1976, 1979). Figure 3 presents the CU model as applied to the personality traits studied by Van Tuinen and

Ramanaiah (1979). The interpretation of Fig. 3 is similar to that noted for the CFA model (Fig. 2) except for the meaning of method effects. The effects of methods under the CU model are represented as correlations among error terms. This permits one to capture differential impacts of each method on the multiple measures corresponding to that method.

Three advantages of the CU model over the CFA model are the following. Most importantly, the CU model seldom produces ill-defined solutions. For example, only 2% of the 435 MTMM matrices examined by Marsh and Bailey (1991) exhibited improper solutions. A second advantage of the CU model is that methods are not assumed to be unidimensional as under the CFA model. The confounding of method variance with trait variance is avoided (when this is due to common trait variation across methods and traits are highly correlated). Finally, when four or more traits are measured with at least three methods, one can test the assumption that all correlated uniquenesses associated with one particular method can be explained in terms of a single, unidimensional method factor. This can be done, for instance, by comparing goodness-of-fit indices for the alternative approaches. It turns out that the CFA model with correlations among methods constrained to be zero is a special case of the CU model. For cases where three traits and three methods are used, the models are identical. But when four or more traits are examined, more parameters are associated with each method under the CU model than the CFA model with orthogonal methods.

Illustration of the CU Model

The CU model shown in Fig. 3 was applied to the data in Table 1. The model provided a poor fit to the data based on the chi-square test: X^2 (15, N = 196) = 80.04, $p \simeq .00$. Nevertheless, from a practical standpoint, little variation remains to be explained in the data: NCNFI = .94. Table 4 summarizes the parameter estimates for the CU model. The first thing to note is that all factor loadings of measures on traits are high (range: .72 – .92; mean: .81). This demonstrates that the traits relate strongly to measures and provides support for convergent validity. Error variances are generally low, with the values for the simple self-rating showing the highest values and approaching moderate levels. The correlated uniquenesses for the true–false and multipoint inventories are all nonsignificant. This shows that method effects are negligible for these two methods. All three correlated uniquenesses for the simple self-ratings are significant and thus reveal method effects. However, the magnitudes of the correlations are generally small, suggesting relatively weak method effects. Finally, the correlations among factors show that global and social self-esteem are highly correlated (Ψ_{21} = .81, SE = .03), but the evidence indicates that global self-esteem and orderliness are quite distinct (Ψ_{31} =

TABLE 4a
SUMMARY OF PARAMETER ESTIMATES FOR THE CORRELATED UNIQUENESS MODEL

	Traits					
	Global self-esteem		Social self-esteem		Order	
Construct	FL	*SE*	FL	*SE*	FL	*SE*
True–False inventory						
Global self-esteem	.88	.06	.00[a]		.00[a]	
Social self-esteem	.00[a]		.83	.06	.00[a]	
Order	.00[a]		.00[a]		.89	.06
Multipoint inventory						
Global self-esteem	.83	.06	.00[a]		.00[a]	
Social self-esteem	.00[a]		.92	.06	.00[a]	
Order	.00[a]		.00[a]		.75	.06
Self-rating						
Global self-esteem	.72	.06	.00[a]		.00[a]	
Social self-esteem	.00[a]		.75	.06	.00[a]	
Order	.00[a]		.00[a]		.72	.07
	Factor intercorrelations					
Global self-esteem	1.00[a]					
Social self-esteem	.81	.03	1.00[a]			
Order	.28	.08	.19	.08	1.00[a]	

Note. FL = factor loading. U = unique variance or covariance. All parameter estimates differing significantly from zero are underscored.

[a] Fixed at the value reported.

.28, SE = .08), as are social self-esteem and orderliness (Ψ_{32} = .19, SE = .08). These latter findings are more consistent with theory than the CFA findings (compare Tables 2 and 4).

Critique of the CU Model

At least two shortcomings of the CU model should be mentioned. First, the interpretation of correlated uniqueness as method effects is not always clear. Two possible outcomes make the meaning of findings potentially ambiguous: the presence within the same method of (a) significant positive and negative correlations and (b) significant and nonsignificant correlations. The former is incongruous, since it is difficult to conceive of reasons why the same method has opposite effects on measures of different traits when the traits are expected to covary in *either* a positive or negative direction. The latter finding is possible in theory, but in practice is difficult to explain unless one has a priori methodological reasons accounting for

TABLE 4b

SUMMARY OF PARAMETER ESTIMATES FOR THE CORRELATED UNIQUENESSES MODEL

	Unique variances and covariances																	
	True–false inventory						Multipoint inventory						Self-rating					
Construct	U	*SE*	U	*SE*	U	*SE*	U	*SE*	U	*SE*	U	*SE*	U	*SE*	U	*SE*	U	*SE*
True–false inventory																		
Global self-esteem	.23	.05																
Social self-esteem	.01	.03	.31	.04														
Order	.00	.03	.03	.03	.22	.06												
Multipoint inventory																		
Global self-esteem	.00[a]		.00[a]		.00[a]		.31	.05										
Social self-esteem	.00[a]		.00[a]		.00[a]		.05	.03	.16	.04								
Order	.00[a]		.00[a]		.00[a]		.08	.04	−.02	.03	.42	.06						
Self-rating																		
Global self-esteem	.00[a]		.00[a]		.00[a]		.00[a]		.00[a]		.00[a]		.48	.06				
Social self-esteem	.00[a]		.00[a]		.00[a]		.00[a]		.00[a]		.00[a]		.12	.04	.40	.05		
Order	.00[a]		.00[a]		.00[a]		.00[a]		.00[a]		.00[a]		.10	.04	.08	.04	.48	.06

Note. FL = factor loading. U = unique variance or covariance. All parameter estimates differing significantly from zero are underscored.

[a] Fixed at the value reported.

differences in the significance and nonsignificance of correlated uniquenesses for a common method. In sum, whereas a CU model may fit MTMM matrix data well, the presence of one or both of the above outcomes for the correlated uniqueness may be a consequence of capitalization on chance.

A second, broad limitation of the CU model is that it assumes that methods are uncorrelated. This may be reasonable when highly different methods are purposefully chosen in a construct validation study. But for the typical study where different self-reports constitute the methods, methods would be expected to be significantly correlated, perhaps highly so. Even in cases where self-ratings and peer or expert ratings are used, one generally anticipates at least a moderate amount of association between methods because of the common format of items, shared experiences and outlooks, and other factors.

For the data in Table 1, it is interesting to note that the findings for the trait-only model (not shown) and the CU model reveal nearly equivalent parameter estimates for factor loadings, error variances, and correlations among traits. This occurs because apparently no linear method effects exist for the true–false and multipoint inventories and the method effects for the simple self-ratings are relatively small.

DIRECT PRODUCT MODEL

Rationale

Up to this point, we have considered linear models where traits, methods, and error terms have additive effects on measures. It is also possible that methods may interact with traits in a multiplicative way. That is, a multiplicative interaction can occur such that "the higher the basic relationship between two traits, the more that relationship is increased when the same method is shared" (Campbell & O'Connell, 1982, p. 95). Campbell and O'Connell (1967, p. 421) implied that trait–method interactions may be the rule rather than the exception, and some of the conditions governing such interactions will be explored below after a procedure is described for modeling interactions.

Until recently, no unambiguous procedure existed for representing trait–method interactions, and Campbell and O'Connell's ideas remained little more than speculations. The foundation for a formal model representing the multiplicative interaction between traits and methods was developed by Swain (1975). Swain (1975) proposed that

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_m \otimes \boldsymbol{\Sigma}_T, \tag{3}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of the observed measures in a MTMM matrix design, $\boldsymbol{\Sigma}_m$ and $\boldsymbol{\Sigma}_T$ are method and trait covariance matrices, re-

spectively, and $\otimes$ indicates a right direct (Kronecker) product. This model expresses the covariance matrix of measurements as the direct product of a covariance matrix of methods and a covariance matrix of traits. However, the model does not allow for measurement errors or different scales for different measures, oversights that limit the applicability of the model for typical MTMM matrix applications in personality research.

Browne (1984, 1989) extended Swain's (1975) approach to incorporate unique variances and scale factors and proposed the following direct product (DP) model (see also Cudeck, 1988):

$$\boldsymbol{\Sigma} = \mathbf{Z}(\mathbf{P}_m \otimes \mathbf{P}_T + \mathbf{E}^2)\mathbf{Z}, \tag{4}$$

where $\mathbf{Z}$ is a diagonal matrix of scale constants, $\mathbf{P}_m$ and $\mathbf{P}_T$ are method and trait correlation matrixes, respectively, whose elements are particular multiplicative components of common score correlations (i.e., correlations corrected for attenuation), and $\mathbf{E}^2$ is a diagonal matrix of unique variances.

It is possible to give an intuitive description of the DP model as follows. The DP model hypothesizes multiplicative effects of methods and traits such that sharing a method exaggerates the correlations between highly correlated traits relative to traits that are relatively independent. That is, the higher the intertrait correlation, the more the relationship is enhanced when both measures share the same method, whereas the relationship is not affected when intertrait correlations are zero.

Two different processes lead to multiplicative effects. One might be called differential augmentation (e.g., Campbell & O'Connell, 1967, 1982). Here, multiplicative effects are a consequence of an interaction between the "true" level of trait correlation and the magnitude of method bias. A conventional position is that method factors add irrelevant systematic (method-specific, trait-irrelevant) variance to the observed relationships among measures. In other words, sharing a method is expected to increase the correlations between two measures above the true relationship; halo effects and response sets are two common sources of such outcomes. However, not all relationships are exaggerated by sharing a common method; only those relationships that are large enough to be noted are likely to be exaggerated. Campbell and O'Connell (1967, pp. 421–422) provide an example of such effects where ratings (e.g., self-ratings and peer-ratings) are used as methods. Each rater might have an implicit theory and set of expectations about the co-occurrence of certain traits, which lead to rater-specific biases. In such cases, the stronger the "true" associations are between traits, the more likely they are to be noted and exaggerated, thus producing the multiplicative method–effect pattern.

A second process producing multiplicative effects is differential attenuation (e.g., Campbell & O'Connell, 1967, 1982). A conceptual basis for

this view is that using different methods will attenuate the relationships between traits that are better represented when methods are held constant rather than varied. That is, methods are seen as diluting trait relationships rather than as adding irrelevant systematic variance. Not sharing a method attenuates the observed correlations differently, depending on the level of true trait relationships. Suppose, for example, that multiple occasions are used as methods in a MTMM matrix design. This approach was taken by Marsh and Hocevar (1988) in their development of the hierarchical confirmatory factor analysis (HCFA) model. The results of longitudinal studies often show that correlations are lower for longer than for shorter lapses in time, demonstrating an autoregressive process. Accordingly, a high correlation between two traits will be more attenuated over time than will a low correlation (see also, Campbell & O'Connell, 1982, pp. 100–106). In contrast, a correlation of zero can erode no further, and it remains zero when computed across methods (i.e., occasions). Differential attenuation might be expected in key-informant data to the extent that functional relations over time are monitored.

The DP model can be estimated with programs such as EQS or LISREL (e.g., Wothke & Browne, 1990), but certain advantages result when the MUTMUM program is used (Browne, 1990). The MUTMUM program is less cumbersome than EQS or LISREL, provides standard errors for both trait and method correlations (a particular EQS or LISREL run only computes standard errors for trait or method correlations and must be reparameterized and run twice to yield these estimates), and accommodates constraints on both trait and method correlation matrixes.

Campbell and Fiske's (1959) original criteria for convergent and discriminant validity have the following interpretations under the DPM (e.g., Browne, 1984, pp. 9–10). Evidence for convergent validity is achieved when the correlations among methods in $\mathbf{P}_m$ are positive and large. The first criterion for discriminant validity is met when the correlations among traits in $\mathbf{P}_T$ are less than unity. The second criterion for discriminant validity is attained when the method correlations in $\mathbf{P}_m$ are greater than the trait correlations in $\mathbf{P}_T$. The final discriminant validity criterion is satisfied whenever the DP model holds as determined, for example, by the results for goodness-of-fit indices. These interpretations follow from the specification of the DP model, and a demonstration showing this can be found in Bagozzi and Yi (1990, pp. 549–550). More formal tests of most of these conditions as well as other useful hypotheses are possible and will be described below when we consider an example.

Illustration of the DP Model

The MUTMUM program was applied to the data in Table 1, giving the following results: $\chi^2(25, N = 196) = 98.06$, $p = .00$. Table 5 presents

TABLE 5
PARAMETER ESTIMATES FOR THE DIRECT PRODUCT MODEL ANALYSIS

Measures	Communalities	Error	Trait intercorrelations			Method intercorrelations		
			T1	T2	T3	M1	M2	M3
True–false								
Global SE	.86 (.03)[a]	.14	1.00[b]			1.00[b]		
Social SE	.89 (.02)	.11	.78 (.04)	1.00[b]		.93 (.03)	1.00[b]	
Order	.85 (.03)	.15	.33 (.07)	.20 (.07)	1.00[b]	.88 (.04)	.83 (.04)	1.00[b]
Multipoint								
Global SE	.89 (.02)	.11						
Social SE	.91 (.02)	.09						
Order	.88 (.03)	.12						
Self-rating								
Global SE	.85 (.03)	.15						
Social SE	.88 (.03)	.12						
Order	.83 (.04)	.17						

[a] Standard errors in parentheses.
[b] Fixed at the value reported.

TABLE 6
TESTS OF HYPOTHESES FOR THE DIRECT PRODUCT MODEL

Model	Chi-square goodness-of-fit	Chi-square difference test
Baseline	$\chi^2(25) = 98.06$ $p \simeq .00$	
All traits equivalent $\rho_{t21} = \rho_{t32} = \rho_{t31} = 1.00$	$\chi^2(30) = 578.17$ $p \simeq .00$	$\chi^2_d(5) = 480.11$ $p < .001$
Traits 1 and 2 equivalent $\rho_{t21} = 1.00$	$\chi^2(28) = 185.81$ $p \simeq .00$	$\chi^2_d(3) = 87.75$ $p < .001$
Traits 2 and 3 equivalent $\rho_{t32} = 1.00$	$\chi^2(28) = 495.85$ $p \simeq .00$	$\chi^2_d(3) = 397.79$ $p < .001$
Traits 1 and 3 equivalent $\rho_{t31} = 1.00$	$\chi^2(28) = 426.50$ $p \simeq .00$	$\chi^2_d(3) = 328.44$ $p < .001$
All methods equivalent $\rho_{m21} = \rho_{m32} = \rho_{m31} = 1.00$	$\chi^2(30) = 137.14$ $p \simeq .00$	$\chi^2_d(5) = 39.08$ $p < .001$
Methods 1 and 2 equivalent $\rho_{m21} = 1.00$	$\chi^2(28) = 105.04$ $p \simeq .00$	$\chi^2_d(3) = 6.98$ $p > .05$
Methods 2 and 3 equivalent $\rho_{m32} = 1.00$	$\chi^2(28) = 130.37$ $p \simeq .00$	$\chi^2_d(3) = 32.31$ $p < .001$
Methods 1 and 3 equivalent $\rho_{m31} = 1.00$	$\chi^2(28) = 125.30$ $p \simeq .00$	$\chi^2_d(3) = 27.24$ $p < .001$

the parameter estimates. Note first that the communalities are all high and significant, and error variances are relatively low. The criterion for convergent validity is satisfied in that all intermethod correlations are large and significant. The first criterion for discriminant validity is satisfied in that each correlation among pairs of traits is less than 1.00 by an amount greater than twice its standard error. The second criterion for discriminant validity is met since every intermethod correlation is greater than every intertrait correlation. The final criterion for discriminant validity holds if we rely on a measure of practical relevance: NCNFI = .93.

The above assessment of construct validity was based on a visual inspection of correlations. It is desirable to more formally examine specific hypotheses concerning construct validity, reliability, trait effects, and method effects (e.g., Bagozzi & Yi, 1992a). Table 6 shows the results of various hypotheses of interest. The first set of comparisons in the table

focuses on the equivalence of traits and tests whether each intertrait correlation is lower than 1.00 in an absolute sense. This provides a formal test of the first discriminant validity criterion. The null hypothesis maintains that all intertrait correlations are 1.00 (i.e., $\rho_{t21} = \rho_{t32} = \rho_{t31} = 1.00$). A comparison of this null model to the baseline DP model shows that one must reject this hypothesis (i.e., $\chi^2_d(5, N = 480.11\ p < .001)$. This omnibus test indicates that one or more correlations among traits is significantly less than unity. Table 6 shows, further, that the correlations are less than 1.00 between traits 1 and 2 (i.e., global self-esteem and social self-esteem: $\chi^2_d(3, N = 196) = 87.75, p < .001$), traits 2 and 3 (i.e., social self-esteem and orderliness: $\chi^2_d(3, N = 196) = 379.79, p < .001$), and traits 1 and 3 (i.e., global self-esteem and orderliness: $\chi^2_d(3, N = 196) = 328.44, p < .001$).

The second set of comparisons in Table 6 scrutinizes the associations among methods and tests whether each intermethod correlation is lower than 1.00 in an absolute sense. These tests of the equivalency of methods are not related to construct validity, per se, but are useful for discovering any redundancy in methods. The null hypothesis maintains that all intermethod correlations are 1.00 (i.e., $\rho_{m21} = \rho_{m32} = \rho_{m31} = 1.00$). A comparison of this null model to the baseline DP model shows that one must reject this hypothesis (i.e., $\chi^2_d(5) = 39.08, p < .001$). Thus, at least one of the intermethod correlations is less than 1.00. Inspection of the findings in the bottom of Table 6 shows that methods 2 and 3 (the multipoint inventory and self-ratings) are distinct ($\chi^2_d(3, N = 196) = 32.31, p < .001$) and methods 1 and 3 (the true–false inventory and self-ratings) are distinct ($\chi^2_d(3, N = 196) = 27.24, p < .001$). However, we cannot reject the hypothesis that methods 1 and 2 (the true–false and multipoint inventories) are equivalent ($\chi^2_d(3, N = 196) = 6.98, p > .05$).

A number of other hypotheses might also be examined, depending on the needs of the researcher. For instance, a researcher might wish to discover which traits among a set under scrutiny are orthogonal in order to choose promising candidates for a future test where traits will enter as independent variables in a regression analysis. This can be investigated by comparing a model with intertrait correlations constrained to be zero to the baseline DP model. Formal comparisons could be made as well between intertrait and intermethod correlations to see whether the latter are greater than the former, as is required by the second discriminant validity criterion. This can be examined with tests imposing inequality constraints but was not done herein because the significance levels derived do not strictly apply. A proper test could be developed, if desired, based on asymptotic distributions with the analysis of moment structures. Further, tests of the orthogonality of methods might be of interest in some circumstances and can be pursued with a similar strategy to that outlined

above for tests of the orthogonality of traits. Finally, it is possible to test whether the communality of each trait remains constant across methods. This can be accomplished by comparing the baseline DP model to a model fixing the diagonal matrix of errors corresponding to methods to unity.

Critique of the DP Model

One drawback with the DP model is that convergent validity is assessed by a rather global standard (e.g., Bagozzi & Yi, 1990, p. 556). The requirement that method correlations be substantial is a composite indicator of sorts for convergence of multiple measures of each trait. The criterion for convergent validity does not supply information about the degree of convergent validity or point out which measure(s) is satisfactory or not. In this sense, the DP model is less informative than the CFA model.

A related shortcoming of the DP model is that it is not possible to arrive at an estimate of variation in a method due to traits, as is possible with the CFA model. Trait and method variance are confounded in the DP model.

A final point to note is that, on occasion, the DP model and either the CFA model or the CU model can fit the same data set. Bagozzi and Yi (1991) found, for example, that two of four data sets in their study were explained satisfactorily by both the DP and CFA models. However, because improper solutions arose for the CFA models, there is reason to reject these models and accept the DP model. On the other hand, the CU model and the DP model both fit the two data sets in question (not shown in Bagozzi & Yi, 1991). One of these data sets can be accounted for by the trait-only model (Bagozzi & Yi, 1991, p. 438), so it appears that only one data set actually can be explained by both the CU and DP models. Because the trait-only model is more parsimonious than the DP model, we might accept the former and reject the latter for the data set in question.

It thus appears that the DP model and the CFA and CU models can fit the same data, although the likelihood of this happening in practice is unknown. Unfortunately, little is known as well about the conditions under which both models will fit the same data. One decision rule that can be applied until we learn more about the relationship between the two models is to rely on differences in parsimony between the two models. From the point of view of the number of parameters to estimate, the DP model has fewer parameters than the CU (or CFA) model. But it could be argued that linear effects are conceptually more parsimonious than multiplicative effects. A choice between the two, when both fit the same data, will depend on one's interpretation of parsimony.

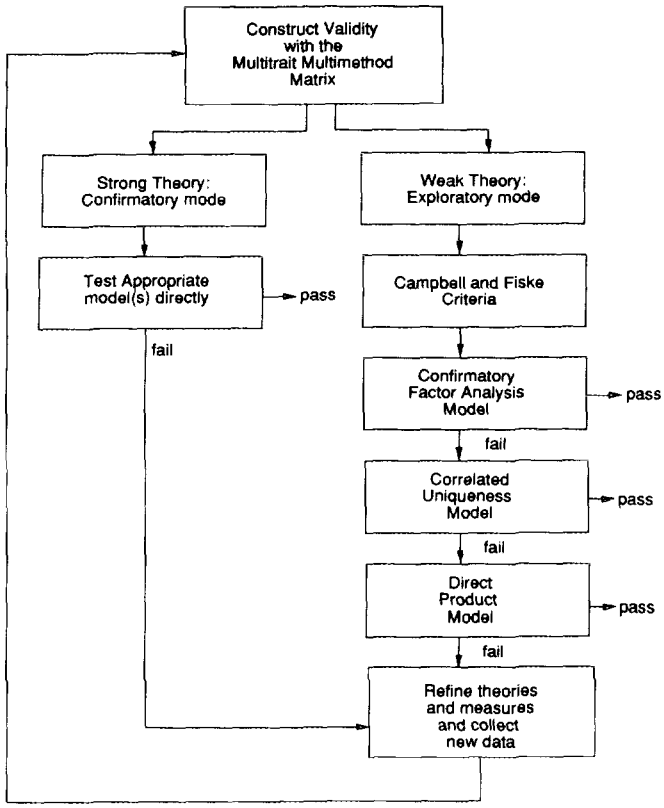

FIG. 4. Guidelines for the Analysis of MTMM Matrix Data

DISCUSSION

The assessment of convergent and discriminant validity is a complex endeavor with many options and many pitfalls. We considered the rationales, assumptions, and pros and cons of four leading approaches to the analysis of MTMM matrix data. Figure 4 presents guidelines for using the procedures in studies of construct validity.

It is useful to think of the analysis of MTMM data from the point of view of either of one of two goals, based upon whether one has strong or weak criteria for making hypotheses. When one has strong reasons for expecting a particular kind of structure underlying the data or desires information of a specific nature, either the confirmatory factor analysis model, correlated uniqueness model, or direct product model may be tried first. For example, if one has reason to believe that traits and methods interact (e.g., this might be expected when self and expert ratings are

gathered of states or traits and respondents have an implicit personality theory which affects their judgments), then the direct product model should be examined first. If, on the other hand, one believes that traits and methods have additive effects (a likely outcome in many contexts), then either the confirmatory factor analysis model or correlated uniqueness model can be investigated. The former would be preferred when the researcher desires to partition variance into trait, method, and error components. The latter is advantageous when (a) improper solutions result in a confirmatory factor analysis or (b) trait and method variance are suspected to be confounded (e.g., when traits are highly correlated and a general method factor reflects trait variance).

When one lacks a strong rationale for hypothesizing an underlying structure, the exploratory sequence outlined in the right-hand side of Fig. 4 might be appropriate. It is helpful often in these cases to begin with the classic Campbell and Fiske (1959) procedure. Satisfaction of the four criteria suggested by Campbell and Fiske (1959) is incomplete, however, for making definitive conclusions, and the approach rests on unrealistic assumptions. Nevertheless, positive results from the classic analysis provide tentative information that a linear model might capture the relationships in the MTMM matrix. A failure to satisfy the Campbell and Fiske (1959) criteria could stem from many reasons, some of which include excessive random error and unreliable measures, method effects which are highly correlated and/or nonproportional across measures of traits, and unknown relations between traits and methods. Marsh (1988), building on general observations made by Campbell and Fiske (1959, p. 84), suggested that an inspection of the MTMM matrix can point to the likely presence or not of method effects. Specifically, Marsh (1988) proposed that the mean of the correlations in the heterotrait–monomethod triangles be compared to the mean of the correlations in the heterotrait–heteromethod triangles. The larger the difference, the more likely that method effects and/or shared method effects occur. One might perform this comparison separately by methods as well, to discover which methods contribute more than others. In sum, when little a priori information exists to forecast the nature of trait, method, and error effects, it is sometimes helpful to begin with an application of the Campbell and Fiske (1959) criteria.

Whether the Campbell and Fiske (1959) criteria are met or not, we recommend that the confirmatory factor analysis model be applied next in an exploratory investigation. Because the confirmatory factor analysis model overcomes so many limitations of the Campbell and Fiske (1959) approach and at the same time yields a partitioning of variance into trait, method, and error components, it is a potentially informative window into construct validity. We recommend that the trait-only model be run

first in this regard. If it fits satisfactorily, then the addition of method factors should be done warily, as it will lead to overfitting and improper solutions with high likelihood. A poorly fitting trait-only model might be caused by a failure to model method effects, which the results from an analysis of a trait–method analysis should confirm.

As Marsh and Bailey (1991) show, the confirmatory factor analysis model is often unsuccessful in the sense that either iterations fail to converge and no proper solution is possible or else the solution that one finds results in improper parameter estimates such as negative or nonsignificant error variances. Of course, proper solutions can result but the model may deviate significantly from the data. For all these instances, the confirmatory factor analysis model must be rejected. When this happens, the correlated uniqueness model should be explored. Indeed, this model is highly robust and is likely to fit most data sets, assuming the assumptions upon which it is based are met.

In those cases where the correlated uniqueness model fails to account for the pattern of relations in a MTMM matrix, it might be a consequence of interactions between traits and methods. Here the researcher can apply the direct product model. It should be noted that none of the models may fit a particular data set if complex patterns underlie the relationships such as additive effects among some traits and methods and multiplicative effects among others. When this happens and if enough traits and methods exist, it may prove fruitful to explore different models for different subsets of measures.

All the models considered up to this point are applicable to data summarized in the classic MTMM matrix. Each trait is measured by a single indicator from each of multiple methods. A drawback common to the approaches is the property that random error is confounded with specific error in the disturbances. The consequence of this is especially important when the reliabilities of different scales vary, because "such differences will distort inferred relations among the scales, the factor loadings on the latent method and trait factors, relations among the latent factors, and summary statistics that are based on these parameter estimates" (Marsh & Hocevar, 1988, p. 108).

Three approaches that represent both random error and measure specificity, and thus circumvent the confounding mentioned above, are the hierarchical confirmatory factor analysis (HCFA) model (Marsh & Hocevar, 1988), the first-order, multiple-informant, multiple-indicator (FOMIMI) model (Kumar & Dillon, 1990), and certain panel models (e.g., Bagozzi & Heatherton, 1991; Bagozzi & Yi, 1992b). Unlike the procedures presented in this paper, the HCFA, FOMIMI, and panel models have data requirements going beyond the classic MTMM matrix. The HCFA model uses first-order factors to represent latent trait–method combina-

tions, where two or more measures load on each trait–method factor. Thus, the correlation matrix needed for a HCFA model is at least twice as large as the traditional MTMM matrix which uses one measure for each trait–method unit. Under the HCFA model, trait and method factors are modeled as second-order latent variables. A further limitation of the HCFA model that potentially limits its usefulness in practice is the following: constant proportions for each measure k are assumed for (a) trait variance to method variance, (b) trait variance to measure specificity, and (c) method variance to measure specificity (Bagozzi, Yi, & Phillips, 1991, Appendix B). The FOMIMI model uses first-order latent variables to represent trait and method effects, as with the confirmatory factor analysis model. But to capture measure specificity, additional first-order factors are introduced. To achieve an identifiable model, at least two and preferably three measures are required for each trait–method combination (Bagozzi, Yi, & Phillips, 1991). This means that the correlation matrix required for a FOMIMI analysis is at least twice the size of a traditional MTMM matrix. Because the ratio of factors to measures in a FOMIMI model is often larger than what one would like to have in a factor analysis, the potential for overfitting and either failures to converge or improper solutions is great. This property and the large data requirements make the FOMIMI model less useful in practice. Panel models can be used to model random error and measure specificity, while testing for convergent and discriminant validity for measures of two or more traits. The drawback with this approach is that the models become unwieldy when multiple methods are used, because as with the HCFA and FOMIMI models, multiple measures of each trait–method combination are required. Coupled with the need to obtain responses from the same respondents over time, this feature of the approach limits its applicability in practice. The HCFA, FOMIMI, and panel models are important procedures for the investigation of construct validity, but because of their complexity and data requirements, nothing more will be said about them herein. Table 7 summarizes the advantages and disadvantages of these approaches to construct validation as well as the ones scrutinized in more detail in this article.

In sum, many procedures can be used for analyzing MTMM matrix data. No single approach dominates the others. No universal procedure can be recommended. The choice of one or more models will depend on the purposes of the researcher. Nevertheless, it is important to recognize the pros and cons of the different approaches. We presented guidelines for conducting the investigation of construct validity with MTMM matrix data. A researcher must be aware of the assumptions of the different procedures and their implications for the information derived from their application. The choice of a procedure should be guided by the nature

TABLE 7
SUMMARY OF PROS AND CONS WITH REGARD TO CONTEMPORARY PROCEDURES FOR ASSESSING CONSTRUCT VALIDITY

Procedure	Advantages	Disadvantages
Campbell and Fiske (1959)	Intuitive	No precise standards for ascertaining convergent and discriminant validity.
	Easy to apply	Cannot determine degree of trait, method, and error variance.
		Assumes that traits and methods are uncorrelated, methods influence all traits equally, methods are uncorrelated, measures are equally reliable.
Confirmatory factor analysis (e.g., Widaman 1985)	Methods can correlate freely and affect measures to different degrees.	Disturbances reflect both specific and error variances.
	Measures of fit provided for an overall model.	Partitioning of variance may not yield "trait-free" and "method free" interpretations.
	Estimates of tests of significance provided for parameters.	Ill-defined solutions frequently result (e.g., negative error variances).
	Variance can be partitioned into trait, method, error components.	Requires at least three traits and three methods, four traits and two methods, or two traits and four methods.
	Under certain conditions, can estimate correlations between traits and methods.	
Correlated uniquenesses model (e.g., Marsh 1989)	Likelihood of ill-defined solutions low.	Confounds random error with measure speciality.
	Avoids confounding of method variance with trait variance under certain conditions.	Interpretation of correlated uniquenesses may be difficult.
	Possible to test assumption that all correlated uniquenesses associated with one method can be accounted for by a single factor (when at least four traits and three methods exist).	Assumes methods are uncorrelated.
		Requires at least three traits and three methods.
Direct Product model (e.g., Browne 1984)	Provides direct translation of Campbell and Fiske criteria.	Confounds random error with measure specificity.
	Represents interaction of traits and methods.	Trait and method variance confounded.
	Can work for models as small as two traits and two methods.	Degree of convergent validity difficult to interpret.

Second-order confirmatory factor analysis model (e.g., Anderson 1987)	Random error and measure specificity estimated separately.	Assumes ratios of trait variance to measure specificity are identical for any particular measure, regardless of the method. Requires at least twice, and preferably three times, as many measures as standard procedures.
Hierarchical confirmatory factor analysis model (e.g., Marsh & Hocevar 1988)	Random error and measure specificity estimated separately.	Assumes constant proportions for measure k for the ratios of (a) trait variance to method variance (b) trait variance to measure specificity (c) method variance to measure specificity. Requires at least twice, and preferably three times, as many measures as standard procedures.
First-order multiple-informant, multiple indicator model (e.g., Kumar & Dillon 1990)	Random error and measure specificity estimated separately. Avoids assumptions on ratios made by second-order and hierarchical confirmatory factor analysis models.	Required at least twice, and preferably three times, as many measures as standard procedures. Likelihood of overfitting high (i.e., failures to converge or ill-defined solutions are likely).
Panel models with multitraits (e.g., Bagozzi & Heatherton 1992)	Random error and measure specificity estimated separately. Temporal stability and true reliability can be estimated. Applies to as few as two traits and three methods.	Needs at least two points in time for each measure.

of the traits under investigation, the properties of the methods used to measure traits, the correspondence of traits and methods to the underlying rationale of the model under consideration, the assumptions of the model and its sensitivity to their violation, and the kind of information desired (e.g., partitioning of variance into components). Properly used, the various procedures provide numerous insights into construct validity.

REFERENCES

Anderson, J. C. (1987). An approach for confirmatory measurment and structural equation modeling of organizational properties. *Management Science,* **33,** 525–541.

Bagozzi, R. P. (1991). Further thoughts on the validity of measures of elation, gladness, and joy. *Journal of Personality and Social Psychology,* **61,** 98–104.

Bagozzi, R. P., & Heatherton, T. F. (1991). *Further evidence on the psychometric properties of the state self-esteem scale.* Unpublished manuscript, the University of Michigan.

Bagozzi, R. P., & Yi, Y. (1990). Assessing method variance in multitrait-multimethod matrices: The case of self-reported affect and perceptions at work. *Journal of Applied Psychology,* **75,** 547–560

Bagozzi, R. P., & Yi, Y. (1991). Multitrait-multimethod matrices in consumer research. *Journal of Consumer Research,* **17,** 426–439.

Bagozzi, R. P., & Yi, Y. (1992a). Testing hypotheses about methods and traits in the direct product model for the MTMM matrix. *Applied Psychological Measurement* (in press).

Bagozzi, R. P. & Yi, Y. (1992b). Multitrait–multimethod matrices in consumer research: Critique and new developments. *Journal of Consumer Psychology* (in press).

Bagozzi, R. P., Yi, Y., & Phillips, L. W. (1991). Assessing construct validity in organizational research. *Administrative Science Quarterly,* **36,** 421–458.

Bentler, P. M. (1989). *EQS: Structural equations program manual* [Computer program manual]. Los Angeles: BMDP Statistical Software.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin,* **107,** 238–246.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin,* **88,** 588–606.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

Boruch, R. F., & Wolins, L. (1970). A procedure for estimation of trait, method, and error variance attributable to a measure. *Educational and Psychological Measurement,* **30,** 547–574.

Browne, M. W. (1984). The decomposition of multitrait-multimethod matrices. *British Journal of Mathematical and Statistical Psychology,* **37,** 1–21.

Browne, M. W. (1989). Relationships between an additive model and a multiplicative model for multitrait-multimethod matrices. In R. Coppi & S. Bolasco (Eds.), *Multiway data analysis* (pp. 507–520). Amsterdam: North-Holland.

Browne, M. W. (1990). *MUTMUM PC user's guide.* Unpublished manuscript. Department of Statistics, University of South Africa, Pretoria.

Campbell, D. T. (1955). The informant in quantitative research. *American Journal of Sociology,* **60,** 339–342.

Campbell, D. T., & Fiske, D. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin,* **56,** 81–105.

Campbell, D. T., & O'Connell, E. J. (1967). Method factors in multitrait–multimethod matrices: Multiplicative rather than additive? *Multivariate Behavioral Research,* **2,** 409–426.

Campbell, D. T., & O'Connell, E. J. (1982). Methods as diluting trait relationships rather than adding irrelevant systematic variance. In D. Brinberg & L. Kidder (Eds.), *Forms of validity in research* (pp. 93–111). San Francisco: Jossey-Bass.

Carver, C. S. (1989). How would multifaceted personality constructs be tested? Issues illustrated by self-monitoring, attributional style, and hardiness. *Journal of Personality and Social Psychology,* **56,** 577–585.

Comrey, A. L. (1970). *Manual: Comrey personality scales.* San Diego: Educational and Industrial Testing Service.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Chicago: Rand McNally.

Coopersmith, S. (1967). *The antecedents of self-esteem.* San Francisco: Freeman.

Crandall, R. (1973). The measurement of self-esteem and related constructs. In J. P. Robinson & P. R. Shaver (Eds.), *Measures of social psychological attitiudes* (pp. 45–168). Michigan: Institute for Social Research, the University of Michigan.

Cudeck, R. (1988). Multiplicative models and MTMM matrices. *Journal of Educational Statistics,* **13,** 131–147.

Cudeck, R. (1989). Analysis of correlation matrices using covariance structure models. *Psychological Bulletin,* **106,** 317–327.

Cudeck, R., & Browne, M. W. (1983). Cross-validation of covariance structures. *Multivariate Behavioral Research,* **18,** 147–167.

Dillon, W. R., Kumar, A., & Mulani, N. (1987). Offending estimates in covariance structure analysis: Comments on the causes of and solutions to Heywood cases. *Psychological Bulletin,* **101,** 126–135.

Eagly, A. H. (1967). Involvement as a determinant of response to favorable and unfavorable information. *Journal of Personality and Social Psychology,* **7** (3, Whole No. 643).

Fiske, D. W. (1982). Convergent-discriminant validation in measurments and research strategies. In D. Brinberg & L. H. Kidder (Eds.), *Forms of validity in research* (pp. 77–92). San Francisco: Jossey-Bass.

Fitts, W. (1965). *Manual: Tennessee self-concept scale.* Nashville: Counselor Recordings & Tests.

Fitts, W. H., Adams, J. L., Radford, G., Richard, W. C., Thomas, B. K., Thomas, M. M., & Thompson, W. (1971). *The self-concept and self-actualization* (Monograph 3). Nashville: Counselor Recordings & Tests.

Funder, D. C. (1989). Accuracy in personality judgment and the dancing bear. In D. M. Buss and N. Cantor (Eds.), *Personality psychology: Recent trends and emerging directions* (pp. 210–223).

Ganster, D. C., Hennessey, H. W., & Luthans, F. (1983). Social desirability response effects: Three alternative models. *Academy of Management Journal,* **26,** 321–331.

Golding, S. L., & Seidman, E. (1974). Analysis of multitrait-multimethod matrices: A two-step principal components procedure. *Multivariate Behavioral Research,* **9,** 479–496.

Hovland, C., & Janis, I. (Eds.) (1959). *Personality and persuasibility.* New Haven: Yale University Press.

Hubert, L. J., & Baker, F. B. (1979). A note on analyzing the multitrait-multimethod matrix: An application of a generalized proximity function comparison. *British Journal of Mathematical and Statistical Psychology,* **32,** 179–184.

Jackson, D. N. (1967). *Personality research from manual.* Goshen, NY: Research Psychologists Press.

Jackson, D. N. (1970). *The Jackson personality inventory.* Unpublished manuscript, University of Western Ontario, London.

Jackson, D. N. (1975). Multimethod factor analysis: A reformulation. *Multivariate Behavioral Research,* **10,** 259–275.

Jöreskog, K. G. (1974). Analyzing psychological data by structural analysis of covariance matrices. In R. C. Atkinson, D. H. Krantz, R. D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology* (Vol. 2, pp. 1–56). San Francisco: Freeman.

Jöreskog, K. G., & Sörbom, D. (1989). *LISREL 7—A guide to the program and applications*, 2nd ed., Chicago: SPSS.

Judd, C. M., Jessor, R., & Donovon, J. E. (1986). Structural equation models and personality research. *Journal of Personality*, **54,** 149–198.

Kenny, D. A. (1976). An empirical application of confirmatory factor analysis to the multitrait–multimethod matrix. *Journal of Experimental Social Psychology*, **12,** 247–252.

Kenny, D. A. (1979). *Correlation and casuality*. New York: Wiley.

Kumar, A., & Dillon, W. R. (1990). On the use of confirmatory measurment models in the analysis of multiple-informant reports. *Journal of Marketing Research*, **27,** 102–111.

Kumar, A., & Dillon, W. R. (1992). An integrative look at the use of additive and multiplicative covariance structure models in the analysis of MTMM data. *Journal of Marketing Research*, **29,** 51–64.

Levin, J., Montag, I., & Comrey, A. L. (1983). Comparison of multitrait–multimethod, factor, and smallest space analysis on personality scale data. *Psychological Reports*, **53,** 591–596.

Marsh, H. W. (1988). Multitrait-multimethod analyses. In J. P. Keeves (Ed.), *Educational research methodology, measurement and evaluation: An international handbook*. Oxford: Pergamon.

Marsh, H. W. (1989). Confirmatory factor analyses of multitrait-multimethod data: Many problems and a few solutions. *Applied Psychological Measurement*, **13,** 335–361.

Marsh, H. W., & Bailey, M. (1991). Confirmatory factor analyses of multitrait–multimethod data: A comparison of the behavior of alternative models. *Applied Psychological Measurement*, **15,** 47–70.

Marsh, H. W., & Hocevar, D. (1988). A new, more powerful approach to multitrait–multimethod analyses: Application of second-order confirmatory factor analysis. *Journal of Applied Psychology*, **73,** 107–117.

Maxwell, A. E. (1977). *Multivariate analysis in behavioral research*. London: Chapman & Hall.

McDonald, R. P., & Marsh, H. W. (1990). Choosing a multivariate model: Noncentrality and goodness of fit. *Psychological Bulletin*, **107,** 247–255.

Mulaik, S. A., James, L. R., Van Alstin, J., Bennett, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin*, **105,** 430–445.

Nicholls, J. G., Licht, B. G., & Pearl, R. A. (1982). Some dangers of using personality questionnaires to study personality. *Psychological Bulletin*, **92,** 572–580.

Ozer, D. J. (1989). Construct validity in personality assessment. In D. M. Buss and N. Cantor (Eds.), *Personality psychology: Recent trends and emerging directions* (pp. 224–234). New York: Springer-Verlag.

Paulhus, D. L. (1989). Socially desireable responding: Some new solutions to old problems. In D. M. Buss and N. Cantor (Eds.), *Personality psychology: Recent trends and emerging directions* (pp. 201–209). New York: Springer-Verlag.

Rindskopf, D. (1983). Parameterizing inequality constraints on unique variances in linear structural models. *Psychometrika*, **48,** 73–83.

Rindskopf, D. (1984). Structural equation models: Empirical identification, Heywood cases and related problems. *Sociological Methods & Research*, **13,** 109–119.

Rosenthal, R., & Rosnow, R. L. (Eds.) (1969). *Artifacts in behavioral research*. New York: Academic Press.

Schmitt, N., & Stults, D. N. (1986). Methodology review: Analysis of multitrait-multimethod matrices. *Applied Psychological Measurement,* **10,** 1–22.

Seidler, J. (1974). On using informants: A technique for collecting quantitative data and controlling measurement error in organizational analysis. *American Sociological Review,* **39,** 816–831.

Siegel, S. (1956). *Nonparametric statistics.* New York: McGraw–Hill.

Swain, A. J. (1975). *Analysis of parametric structure for variance matrices.* Unpublished doctoral dissertation, University of Adelaide, Australia.

Tucker, L. R., & Lewis, C. (1973). The reliability coefficient for maximum likelihood factor analysis. *Psychometrika,* **38,** 1–10.

Van Driel, O. P. (1978). On various causes of improper solutions of maximum likelihood factor analysis. *Psychometrika,* **43,** 225–243.

Van Tuinen, M., & Ramanaiah, N. V. (1979). A multimethod analysis of selected self-esteem measures. *Journal of Research in Personality,* **13,** 16–24.

Werts, C. E., & Linn, R. L. (1970). Path analysis: Psychological examples. *Psychological Bulletin,* **74,** 193–212.

Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait–multimethod data. *Applied Psychological Measurement,* **9,** 1–26.

Winkler, J. D., Kanouse, D. E., & Ware, J. E., Jr. (1982). Controlling for acquiescence response set in scale development. *Journal of Applied Psychology,* **67,** 555–561.

Wothke, W., & Browne, M. W. (1990). The direct product model for the MTMM matrix parameterized as a second order factor analysis model. *Psychometrica,* **55,** 255–262.