Journal of Applied Psychology
1990, Vol. 75, No. 5, 547-560

0021-9010/90/$00.75

Assessing Method Variance in Multitrait–Multimethod Matrices: The Case of Self-Reported Affect and Perceptions at Work

Richard P. Bagozzi and Youjae Yi
School of Business Administration, University of Michigan

Spector (1987) concluded that there was little evidence of method variance in multitrait-multimethod data from 10 studies of self-reported affect and perceptions at work, but Williams, Cote, and Buckley (1989) concluded that method variance was prevalent. We extended these studies by examining several important but often neglected issues in assessing method variance. We describe a direct-product model that can represent multiplicative method effects and propose that model assumptions, individual parameters, and diagnostic indicators, as well as overall model fits, be carefully examined. Our reanalyses indicate that method variance in these studies is more prevalent than Spector concluded but less prevalent than Williams et al. asserted. We also found that methods can have multiplicative effects, supporting the claim made by Campbell and O'Connell (1967, 1982).

Researchers have often shown a substantial interest in assessing method variance with multitrait–multimethod (MTMM) matrices (e.g., Campbell & Fiske, 1959; Schmitt, Coyle, & Saari, 1977). As an artifact of measurement, method variance can bias results when researchers investigate relations among constructs measured with the common method. Because method variance provides a potential threat to the validity of empirical findings, it seems important to assess the extent to which method variance is problematic in typical research settings.

Spector (1987) addressed this issue by examining a series of MTMM matrices in research on self-reported affect and perceptions of work. Following the classic procedure proposed by Campbell and Fiske (1959), Spector assessed the amount of method effects by comparing the correlations of different traits measured with the same method (i.e., monomethod correlations) and the correlations among different traits across methods (i.e., heteromethod correlations). The monomethod correlations were not significantly different from the heteromethod correlations, and Spector concluded that there was little evidence of method variance.

Williams, Cote, and Buckley (1989) recently noted a number of limitations of this analytic procedure. As summarized by Schmitt and Stults (1986) and Widaman (1985), these limitations are (a) the lack of quantifiable criteria, (b) the inability to account for different reliability, and (c) the implicit assumptions underlying the procedure, especially the requirement of uncorrelated methods. It should be emphasized here that these limitations concern Campbell and Fiske's (1959) analytic procedures (i.e., interpretation of correlations); their core ideas (i.e., the use of multitrait–multimethod data and convergent and discriminant validity) are sound. Williams et al. reanalyzed the same data that Spector (1987) analyzed by using chi-square difference tests and variance partitioning with confirmatory factor analyses (CFA). Their analyses indicated that method variance is present and accounts for substantial variance in the measures originally examined by Spector.

Williams et al.'s (1989) study was executed carefully with a powerful CFA approach. However, their findings are inconclusive because their procedure has several limitations. First, because their tests examined only overall effects of method factors, they failed to provide information about method effects on individual measures. Suppose, for example, that the chi-square difference test indicates significant method effects in an MTMM model with 10 measures. This omnibus test does not identify how many and which of the measures are significantly affected by the methods. For instance, a global test based on model fits can indicate significant method variance when only 1 of 10 measures is affected by the method factor. Although Williams et al. partitioned the variance into trait, method, and error variance at the scale level, they did not test the significance of the method variance either at the scale level or at the individual item level. Thus, Williams et al.'s study does not give diagnostic information for conclusions about individual measures in the MTMM matrix.

Second, Williams et al. (1989) examined the chi-square goodness-of-fit test and the normed-fit index (NFI; Bentler & Bonett, 1980) but ignored other indicators, such as the adjusted goodness-of-fit index (AGFI), root-mean-square residual (rmr), standardized residuals, and improper estimates, which can provide useful information about model fit. For example, the chi-square test is sensitive to sample size and could possibly point to a satisfactory fit because of a lack of statistical power (e.g., Satorra & Saris, 1985). Likewise, when many trait and method factors are introduced into an MTMM model, a satisfactory chi-square may arise simply as a result of overfitting.

We would like to thank three anonymous reviewers for their helpful comments on the previous versions of this article. The financial assistance of the University of Michigan's School of Business Administration is also gratefully acknowledged.

Correspondence concerning this article should be addressed to Richard P. Bagozzi or Youjae Yi, School of Business Administration, University of Michigan, Ann Arbor, Michigan 48109-1234.

One should evaluate a structural model by using all the measures of the overall degree of fit as well as information on individual parameters provided in any particular application.

Finally, the CFA approach taken by Williams et al. (1989) is based on the assumption that variation in measures will be a linear combination of traits, methods, and error. That is, methods are presumed to have additive effects on measures in the CFA model. This may be a reasonable assumption and, in fact, can be tested as a hypothesis in any particular MTMM analysis. However, in certain contexts, traits and methods may interact in the determination of measure variation. Campbell and O'Connell (1967) went so far as to suggest that such an interaction is "quite general in nature" (p. 421). The multiplicative relation occurs 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). If methods do have multiplicative effects, then the CFA model is inappropriate for examining method effects. By using CFA models only, Williams et al. assumed that methods had linear effects for all of the data examined by Spector (1987) and ignored the possibility of multiplicative method effects.

In fact, Williams et al. (1989) seem to have confused trait–method interactions with trait–method correlations. For example, Williams et al. (1989, p. 463) asserted that their "analysis assumes that Trait X Method interactions do not exist (zero correlation among trait and method factors)." Also, their justification for the assumption, based on Widaman's (1985, p. 7) argument that correlations among trait and method factors "present both logical and empirical estimation problems of great magnitude," is misleading. The degree of association among traits and methods may be independent of the interaction between traits and methods, if any.

One purpose of the present article was to investigate these important, but often ignored, issues. A second purpose was to correct any erroneous conclusions about method variance in and construct validity of the data examined by Spector (1987) and Williams et al. (1989). To do so, we incorporated the aforementioned issues into our analyses. The third purpose was to show that the analysis and interpretation of MTMM data are not straightforward endeavors but require a careful, detailed consideration of many criteria for model specification, goodness of fit, and other statistical findings.

Evaluation of Method Variance and Model Fit

The general form of the CFA model for MTMM data can be expressed with two sets of equations (e.g., Jöreskog, 1974; Werts, Jöreskog, & Linn, 1972):

$$\mathbf{y} = [\Lambda_T \Lambda_M] \begin{bmatrix} \eta_T \\ \eta_M \end{bmatrix} + \epsilon, \text{ and} \tag{1}$$

$$\Sigma = \Lambda_T \Psi_T \Lambda_T' + \Lambda_M \Psi_M \Lambda_M' + \theta, \tag{2}$$

where $\mathbf{y}$ is a vector of $r \times s$ measures for r traits by s methods; Λ_T and Λ_M are factor-loading matrices for traits and methods, respectively (defined in the next sentence); η_T and η_M are vectors of r traits and s method factors, respectively; ϵ is a vector of residuals for $\mathbf{y}$; Σ is the implied variance–covariance matrix for $\mathbf{y}$; Ψ_T and Ψ_M are correlation matrices for traits and methods, respectively; θ is the vector of unique variances for ϵ; $\Lambda_T = [\Lambda_1, \Lambda_2, \cdots, \Lambda_r]'$; Λ_i is a diagonal matrix with factor loadings corresponding to the measures of the ith trait; and

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

where λ_j is a vector of factor loadings corresponding to the measures obtained by the jth method. Application of the CFA model to MTMM data permits one to partition variance into trait, method, and random error. These reside, respectively, in the squared factor loadings for Λ_T and Λ_M and in θ.

Four CFA models can be tested and compared to yield meaningful tests of hypotheses about method and trait factors (Widaman, 1985): Model 1 hypothesizes that only unique variances are free (i.e., it is the null model); Model 2 hypothesizes that variation in measures can be explained completely by traits plus random error (i.e., it is the trait-only model); Model 3 hypothesizes that variation in measures can be explained completely by methods plus random error (i.e., it is the method-only model); and Model 4 hypothesizes that variation in measures can be explained completely by traits, methods, and random error (i.e., it is the trait–method model).

Model 4 is, in fact, the hypothesis implied by Equations 1 and 2. Models 1 though 3 are special cases formed by constraining certain parameters of Model 4. Notice that the null model is nested in both the method-only and trait-only models and that the method-only and trait-only models are nested in the trait–method model. Consequently, chi-square difference tests can be used to test whether trait, method, or trait and method factors are present. For example, a test of method variance is provided by comparing Models 1 and 3 as well as Models 2 and 4.

The chi-square difference test is an omnibus test that indicates whether or not measures are significantly affected by methods (or traits). In many cases, however, one may wish to determine how many and which of the measures are responsible for the global significance. Moreover, the method effects should be meaningful and interpretable (e.g., Browne, 1984). For this purpose, an inspection of the loadings linking method factors to individual measures is useful because loadings in Λ_M represent method-related variance for each measure (Widaman, 1985). Thus, an examination of loadings for method factors will provide useful information as to how often the method effects are significant at the individual item level and whether or not they are meaningful and interpretable.

Multiplicative Effects of Methods

It has been suggested that method factors may interact with trait factors in a multiplicative way (e.g., Campbell & O'Connell, 1967, 1982; Schmitt & Stults, 1986). That is, the higher the relationship between traits, the higher the method effects. Swain (1975) proposed the following direct-product model (DPM) to represent the multiplicative interaction of traits and methods in an MTMM analysis:

$$\Sigma = \Sigma_M \otimes \Sigma_T, \tag{3}$$

where Σ is the covariance matrix of the observed variables, Σ_M and Σ_T are method and trait covariance matrices, respectively, 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, this model has several limitations. It does not allow for measurement errors or different scales for different variables, which can limit the applicability of the model in many MTMM studies. Browne (1984, 1989) therefore extended the DPM to overcome these limitations (see also Cudeck, 1988) as follows:

$$\Sigma = \mathbf{Z}(\mathbf{P}_M \otimes \mathbf{P}_T + \mathbf{E})\mathbf{Z}, \quad (4)$$

where $\mathbf{Z}$ is a nonnegative definite diagonal matrix of scale constants, some of which are set equal to unity to achieve identification; $\mathbf{P}_M$ and $\mathbf{P}_T$ are nonnegative definite method and trait correlation matrices, respectively, whose elements are particular multiplicative components of common score correlations (i.e., correlations corrected for attenuation); and $\mathbf{E}$ is a diagonal matrix of nonnegative unique variances.

The DPM in Equation 4 is called the *heteroscedastic error* model and is equivalent to a three-mode factor analysis model (Bentler & Lee, 1978; Bloxom, 1968; Tucker, 1966) with some constraints (Browne, 1984). It can be seen that Equation 4 decomposes test scores into true scores plus error-score components. Under Equation 4 the correlation matrix corrected for attenuation has a direct-product structure,

$$\mathbf{P}_C = \mathbf{P}_M \otimes \mathbf{P}_T, \quad (5)$$

where $\mathbf{P}_C$ is the disattenuated correlation matrix with a typical element $\rho(T_iM_k, T_jM_l)$; $\mathbf{P}_M$ is the latent method correlation matrix with a typical element $\rho(M_k, M_l)$; and $\mathbf{P}_T$ is the latent trait correlation matrix with a typical element $\rho(T_i, T_j)$. From the definition of a right direct product, one can then see that a typical element of Equation 5 is

$$\rho(T_iM_k, T_jM_l) = \rho(T_i, T_j)\rho(M_k, M_l). \quad (6)$$

Notice that this equation assumes a multiplicative structure for true scores or common scores in the factor analysis sense, rather than for observed scores. Browne (1985) has developed a program, MUTMUM, to estimate the parameters in the DPM, but the program has not been widely used, perhaps because of its limited distribution.

Wothke and Browne (1990) have recently shown that the DPM can be reformulated as a linear model, allowing researchers to estimate the model by using the widely available LISREL program. Specifically, Equation 4 can be written as a second order confirmatory factor analysis model as follows:

$$\Sigma = \Lambda\Gamma\Phi\Gamma'\Lambda', \quad (7)$$

where $\Lambda = \mathbf{Z}$, Γ is the partitioned matrix

$$\Gamma = (\mathbf{C}_M \otimes \mathbf{I}_t | \mathbf{I}_{mt}) = (\Gamma_1 | \Gamma_2), \quad (8)$$

$\mathbf{C}_M$ is a square, lower triangular matrix chosen such that $\mathbf{P}_M = \mathbf{C}_M\mathbf{C}_M'$, $\mathbf{I}_t$ and $\mathbf{I}_{mt}$ are identity matrices, and

$$\Phi = \left(\begin{array}{c|c} \mathbf{I}_M \otimes \mathbf{P}_t & 0 \\ \hline 0 & \mathbf{E} \end{array}\right) = \left(\begin{array}{c|c} \Phi_1 & 0 \\ \hline 0 & \Phi_2 \end{array}\right). \quad (9)$$

The DPM can be easily restricted to suitable submodels. One useful version of the model, a composite error model, is defined by the additional restriction

$$\mathbf{E} = \mathbf{E}_M \otimes \mathbf{E}_T, \quad (10)$$

with $\mathbf{E}_M$ and $\mathbf{E}_T$ diagonal. By using the fact that any symmetric, nonnegative definite matrix can be expressed as the product of a square matrix and its transpose (e.g., Searle, 1982), this restriction can be rewritten as follows:

$$\mathbf{E} = (\mathbf{E}_M^{1/2} \otimes \mathbf{I}_t)(\mathbf{I}_m \otimes \mathbf{E}_T)(\mathbf{E}_M^{1/2} \otimes \mathbf{I}_t)' = \Gamma_2\Phi_2\Gamma_2'. \quad (11)$$

Several restrictions are needed for the identification of the DPM. First, one equality constraint per method is required for identification of scale factor estimates (Wothke & Browne, 1990). This restriction will fix the scale of the component scores. For instance, one may select a trait and set all its scale factors in $\mathbf{Z}$ equal to unity. Alternatively, one can constrain all diagonal elements of $\mathbf{C}_M$ to unity. The two types of restriction may be suitably combined. Another restriction is required to fix the scale of the error components because $(a\mathbf{E}_M) \otimes (b\mathbf{E}_T)$ equals $\mathbf{E}_M \otimes \mathbf{E}_T$ for any a equal to $1/b$. This may be achieved by fixing one element in either $\mathbf{E}_M$ or $\mathbf{E}_T$ at unity.

$\mathbf{P}_T$ is directly estimated in the model, and standard errors of its elements will be available from the LISREL solution. In contrast, the estimate of $\mathbf{P}_M$ is obtained by rescaling $\mathbf{C}_M\mathbf{C}_M'$ into a correlation matrix, and standard errors are not available from the LISREL output. However, one can obtain the standard errors for the method correlations by employing an alternative parameterization in which $\mathbf{P}_M$ (rather than $\mathbf{P}_T$) is directly estimated. Under the multinormality assumption, the model fit can be evaluated by using the maximum-likelihood chi-square statistic, computed as

$$\chi^2 = (N-1)[\ln|\Sigma| - \ln|\mathbf{S}| + \mathrm{trace}(\mathbf{S}\Sigma^{-1}) - rs], \quad (12)$$

where N is the sample size, and r and s are the number of traits and methods, respectively. The corresponding degrees of freedom are computed as $[rs(rs+1)/2] - k$, where k is the number of free parameters to be estimated in the model.

Campbell and Fiske's (1959) original criteria for convergent and discriminant validity have the following direct interpretations in the DPM (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 DPM holds.

These interpretations follow from the DPM specification. Recall that $\rho(T_iM_k, T_jM_l)$, a typical element of $\mathbf{P}_C$, denotes the disattenuated correlation between the ith trait measured with the kth method and the jth trait measured with the lth method. From Equation 6, $\rho(T_iM_k, T_jM_l)$ equals $\rho(T_i, T_j)$ $\rho(M_k, M_l)$. Campbell and Fiske's (1959) criterion for convergent validity is that the monotrait–heteromethod correlations should be substantially greater than zero. When the monotrait–heteromethod correlation $\rho(T_iM_k, T_iM_l)$ is examined under the DPM, it can be seen that

$$\rho(T_iM_k, T_iM_l) = \rho(T_i, T_i)\rho(M_k, M_l) = \rho(M_k, M_l). \quad (13)$$

That is, the monotrait–heteromethod correlations are equal to method correlations under the DPM. As a consequence, convergent validity is achieved when method correlations are large and positive.

The first criterion for discriminant validity is that the monotrait–heteromethod correlations, $\rho(T_iM_k, T_iM_l)$, should be greater than the corresponding heterotrait–heteromethod correlations, $\rho(T_iM_k, T_jM_l)$ and $\rho(T_jM_k, T_iM_l)$, for $i \neq j$. One can see that

$$\begin{aligned}\rho(T_iM_k, T_jM_l)/\rho(T_iM_k, T_iM_l) &= \rho(T_jM_k, T_iM_l)/\rho(T_iM_k, T_iM_l) \quad (14)\\ &= \rho(T_i, T_j).\end{aligned}$$

That is, the ratios of a monotrait–heteromethod correlation to the heterotrait–heteromethod correlations become trait correlations under the DPM. Thus, the first criterion for discriminant validity is met when trait correlations are less than unity. The second criterion is that the monotrait–heteromethod correlations, $\rho(T_iM_k, T_iM_l)$, should be higher than the corresponding heterotrait–monomethod correlations, $\rho(T_iM_k, T_jM_k)$ and $\rho(T_iM_l, T_jM_l)$. From Equation 6, it can be seen that

$$\begin{aligned}\rho(T_iM_k, T_jM_k)/\rho(T_iM_k, T_iM_l) &= \rho(T_iM_l, T_jM_l)/\rho(T_iM_k, T_iM_l) \quad (15)\\ &= \rho(T_i, T_j)/\rho(M_k, M_l).\end{aligned}$$

That is, the ratios of monotrait–heteromethod correlations to heterotrait–monomethod correlations become the ratios of trait correlations to method correlations under the DPM. As a consequence, this criterion is met when the method correlations are greater than the trait correlations. The final criterion is that all matrices of intertrait correlations should have the same pattern whichever methods are used. This criterion is met whenever the DPM holds because the ratio

$$\rho(T_iM_k, T_jM_l)/\rho(T_mM_k, T_nM_l) = \rho(T_i, T_j)/\rho(T_m, T_n) \quad (16)$$

has the same value for any M_k and M_l.

The DPM 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. An important question then arises: What processes underlie the multiplicative effects of method factors?

One view might be called *differential augmentation* (Campbell & O'Connell, 1967, 1982). In this view, multiplicative effects are a functional 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-irrevelant) variance to the observed relationships. That is, sharing a method is expected to augment or increase the correlations between two measures above the true relationship; halo effects and response sets provide evidence for such method bias. However, not all relationships are likely to be equally exaggerated by shared method. Only relationships that are large enough to get noticed are more likely to be exaggerated. Campbell and O'Connell (1967, pp. 421–422) provide an example of such effects when ratings (e.g., self-ratings and peer ratings) are used as methods. Each rater might have an implicit theory (expectations) about the relationships (co-occurrence) of certain traits, which will lead to a rater-specific bias. In such cases, the stronger the true associations between traits, the more likely they are to be noticed and exaggerated, thus producing the multiplicative-method effect pattern. In summary, method factors augment or exaggerate the observed correlations differently, depending on the level of true trait relationships.

The differential attenuation perspective (Campbell & O'Connell, 1967, 1982) offers another explanation for multiplicative-method effects. A conceptual basis for this view is that using different methods will attenuate the trait relationships better represented when method is held constant rather than varied. Not sharing a method attenuates the true relationship so that it appears to be less than it should be; that is, methods are seen as diluting trait relationships rather than adding irrelevant systematic variance. From the differential attenuation perspective, 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. It is often found in longitudinal studies that correlations are lower for longer time lapses than shorter lapses, following a so-called autoregressive process. In this process, a high correlation between two traits will be more attenuated over time than a low correlation (for more details, see 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 (here, occasions). The traditional concept of attenuation as due to the unreliability of measures shows a multiplicative pattern because high correlations are more attenuated by unreliability than low ones. See Campbell and O'Connell (1982) for a detailed discussion of these two explanations for multiplicative effects.

In summary, in the CFA model and the DPM, different functional forms are hypothesized for trait and method effects: in the former, additive effects are assumed, whereas in the latter, multiplicative effects are assumed. In principle, the two models constitute alternative explanations for MTMM data. Specifically, the effects of a method are hypothesized to be constant in the CFA model. In contrast, method effects are hypothesized to vary with the level of trait correlations in the DPM. Although Campbell and O'Connell (1967, 1982) implied that trait and method interactions are the rule rather than the exception, it might be better in any specific case to examine which (additive or multiplicative) model is more appropriate. Ideally, one should have substantive expectations about the method effects prior to selecting a model. If no prior expectations are available, the researcher should test both models to discover which process is at work.

Method

For each of the 10 studies (11 data sets) examined originally by Spector (1987) and later by Williams et al. (1989), four CFA models (Models 1 to 4) were fitted by following the procedures suggested by Widaman (1985; see also Williams et al., 1989). Figure 1 provides an example of the full CFA model (i.e., Model 4) for MTMM data with three traits and three methods. It can be seen that Models 1 through 3 are derived from Model 4 by constraining certain parameters.

The effects of method factors were examined in two ways. First, we compared the hierarchically nested models to determine whether the introduction of method factors improved the fit of the model. Specifically, Model 1 (null) was compared with Model 3 (method only), and Model 2 (trait only) was compared with Model 4 (trait and method). Second, we examined the specific effects of method factors by examining the statistical significance of the individual-method factor loadings. For each measure, the method factor loading indicates the effect of the method factor, and the square of the loading indicates the percentage of variance due to the method factor (Widaman, 1985). Thus, the significance of factor loadings was examined to determine whether the method variance was significant.

As noted earlier, we also tested the possibility that Trait × Method interactions exist. In this regard, the DPMs were fitted on the basis of the procedures proposed by Wothke and Browne (1990). Because the causal diagram for the LISREL operationalization is quite cumbersome (e.g., there are 27 latent variables alone for the smallest model with three traits and three methods), we have not provided a figure. However, the Appendix contains the input program needed to perform a DPM analysis of the data found in Gillet and Schwab (1975).

Marsh and Hocevar (1988) recently introduced a new, powerful approach to MTMM analyses that uses hierarchical confirmatory factor analysis. As an alternative to CFA, this procedure explicitly partitions total variance into additive components for random measurement error, variance specific to each trait–method combination, variance common to a trait across methods, and variance common to a method across traits. Unlike the CFA model or the DPM, however, the approach requires multiple measures for each trait–method combination. Because none of the data sets reported in the studies considered satisfied this requirement, we did not use this approach in our study.

All statistical analyses were performed with the LISREL 7 program (Jöreskog & Sörbom, 1989), given the widespread use of LISREL among researchers (e.g., Bagozzi, 1980; Widaman 1985). LISREL 7 provides several advantages over earlier versions (e.g., LISREL 6); for example, the correct formula for the asymptotic variances of the residuals is used, and an error in the computation of the AGFI has been corrected. Throughout our analyses, the models were evaluated with multiple indicators of goodness of fit. These indicators included (a) chi-square tests, (b) AGFI, (c) rmr, (d) the number of large standardized residuals, and (e) the number of improper estimates (these will be discussed later).

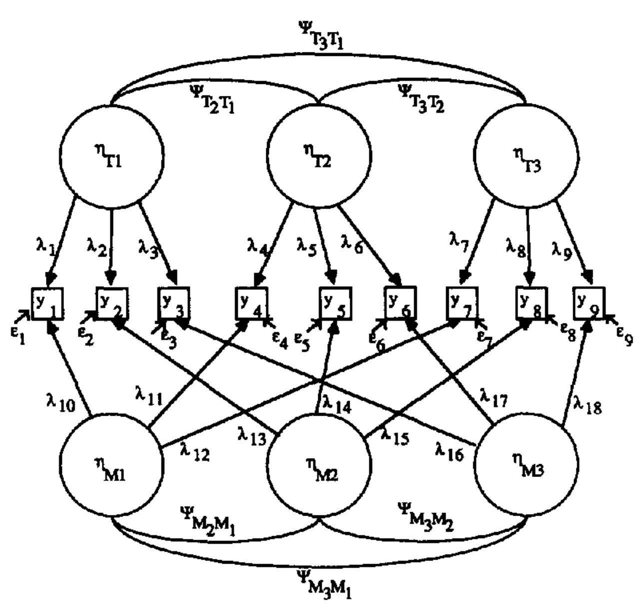

Figure 1. An illustration of the confirmatory factor analysis model for multitrait–multimethod designs with three traits and three methods.

Results

The results of the four CFA models discussed earlier are presented in Table 1. The descriptions of the four data sets are different from those provided by Spector (1987) and Williams et al. (1989). Specifically, the sample size is 111 (not 302) and 723 (not 941) for Alderfer (1967) and Sims, Szilagyi, and Keller (1976), respectively, and the number of traits is 4 rather than 5 for both Dunham, Smith, and Blackburn (1977) and Gillet and Schwab (1975). We made these corrections after closely inspecting the data given in the original articles. For example, only four traits are common across methods in Dunham et al. (1977, p. 429), contrary to the description given by Williams et al. (1989, p. 465). Although convergence failures were common when both method and trait factors were included in CFA models, we were nevertheless able, by judicious choice of starting values, to achieve satisfactory solutions in most instances.

The first thing to notice is that the CFA model with traits and methods fit 10 of the 11 MTMM data sets quite well, as shown in the last column in Table 1. This conclusion is based on an interpretation of the chi-square goodness-of-fit tests alone. Later we scrutinize additional goodness-of-fit measures and other diagnostic criteria that make this interpretation problematic.

The chi-square difference tests based on comparison of two sets of nested models are presented in Table 2. The comparison of Model 1 and Model 3 showed that the introduction of method factors significantly dropped the chi-square value in each data set, indicating that meaningful improvements over the null models were achieved (see the first column in Table 2). The comparison of Model 2 and Model 4 also showed that the introduction of method factors provided significant improvements over the trait-only models for all data sets except Spector (1985). Notice that the two chi-square difference statistics are generally quite different in their magnitude (e.g., 1,096.8 vs 53.1 for Meier, 1984) although they have the same degrees of freedom. It is thus possible that two chi-square difference tests can lead to different conclusions. Indeed, for Spector's (1985) data, the comparison of Model 1 with Model 3 yielded a significant chi-square difference test, whereas the comparison of Model 2 with Model 4 yielded a nonsignificant chi-square difference test.

Thus, a question arises: Which chi-square difference test should be used to assess method effects? We believe that the test should be based on the comparison of Model 2 with Model 4 for two reasons. The first reason stems from the definition of method variance. Campbell and Fiske (1959) defined method variance as variance attributable to measurement method rather than to the constructs of interest. This definition suggests that method variance refers to the variance that cannot be explained by traits but that is explained by methods. Second, the baseline model should be chosen from the set of meaningful models that researchers already accept as valid (Sobel & Bohrn-

Table 1
Results of Confirmatory Factor Analysis of Multitrait–Multimethod Matrices Applied to Data Summarized in Spector (1987)

Study	Model 1		Model 2		Model 3		Model 4	
	χ^2	*df*	χ^2	*df*	χ^2	*df*	χ^2	*df*
Alderfer, 1967 ($n = 111, T = 5, M = 2$)	260.8*	45	46.2*	25	119.2*	34	12.9	14
Dunham, Smith, & Blackburn, 1977 ($n = 622, T = 4, M = 4$)	6,799.7*	120	1,850.1*	98	2,659.6*	98	258.0*	76
Gillet & Schwab, 1975 ($n = 273, T = 4, M = 2$)	710.1*	28	39.2*	14	259.4*	19	6.1	5
Johnson, Smith, & Tucker, 1982 ($n = 100, T = 5, M = 2$)	528.3*	45	77.1*	25	204.5*	34	10.4	14
McCabe, Dalessio, Briga, & Sasaki, 1980 (Study 1) ($n = 82, T = 5, M = 2$)	774.8*	45	50.5*	25	435.0*	34	6.5	14
McCabe et al., 1980 (Study 2) ($n = 82, T = 5, M = 2$)	926.2*	45	34.6	25	376.9*	34	13.9	14
Soutar & Weaver, 1982 ($n = 242, T = 5, M = 2$)	1,264.2*	45	48.7*	25	199.8*	34	13.0	14
Spector, 1985 ($n = 102, T = 5, M = 2$)	485.7*	45	25.6	25	218.4*	34	10.4	14
Pierce & Dunham, 1978 ($n = 155, T = 4, M = 2$)	682.9*	28	58.9*	14	214.5*	19	2.6	5
Sims, Szilagyi, & Keller, 1976 ($n = 723, T = 4, M = 2$)	1,082.4*	28	78.6*	14	259.7*	19	11.5	6
Meier, 1984 ($n = 320, T = 3, M = 3$)	1,682.3*	36	64.8*	24	585.4*	24	11.7	12

Note. Model 1 is the null model; Model 2 is the trait-only model; Model 3 is the method-only model; and Method 4 is the trait–method model. T = number of trait factors; and M = number of method factors.
* $p < .05$.

stedt, 1985). In most MTMM matrices, the measures are not selected at random but are systematically chosen to capture traits. These considerations imply that the trait-only model should be a baseline model for the chi-square difference tests. In summary, the chi-square different tests suggest that method variance is significant in 10 of 11 cases.

Next we examined the statistical significance of method factor loadings for individual measures. If a loading is greater than twice the value of its standard error, then it is judged to differ from zero. Because the method factor loading reflects the degree to which the observed measure is determined by the method factor, this test indicates whether or not the variance due to the method factor is significant. These results are summarized in the third column of Table 2.

Table 2
Results of Assessment of Method Variance in Confirmatory Factor Analyses

Study	Model 1 vs. Model 3		Model 2 vs. Model 4		No. of significant method factor loadings	No. of inconsistent method factor loadings
	χ^2	*df*	χ^2	*df*		
Alderfer, 1967	141.6*	11	33.2*	11	0/10	0
Dunham, Smith, & Blackburn, 1977	4140.1*	22	1,592.0*	22	10/10	0
Gillet & Schwab, 1975	450.7*	9	33.1*	9	0/8	0
Johnson, Smith, & Tucker, 1982	323.8*	11	66.7*	11	5/10	0
McCabe, Dalessio, Briga, & Sasaki, 1980 (Study 1)	339.8*	11	44.0*	11	0/10	0
McCabe et al., 1980 (Study 2)	549.3*	11	20.7*	11	0/10	0
Soutar & Weaver, 1982	1064.4*	11	35.7*	11	10/10	0
Spector, 1985	267.3*	11	15.2	11	1/10	0
Pierce & Dunham, 1978	468.4*	9	56.4*	9	0/8	0
Sims, Szilagyi, & Keller, 1976	822.7*	9	67.1*	9	7/8	2
Meier, 1984	1096.8*	12	53.1*	12	5/9	1

Note. Model 1 is the null model; Model 2 is the trait-only model; Model 3 is the method-only model; and Method 4 is the trait–method model.
* $p < .05$.

None of the method factor loadings were significant for five data sets (i.e., Alderfer, 1967; Gillet & Schwab, 1975; McCabe et al., 1980, Study 1 & Study 2; Pierce & Dunham, 1978). For four data sets (Meier, 1984; Johnson et al., 1982; Sims et al., 1976; Spector, 1985), method factors were found to have significant effects for some of the measures employed in each study. In two studies (Dunham et al., 1977; Soutar & Weaver, 1982), all the method effects were significant. Overall, in 5 out of 11 data sets, half or more of the measures showed significant effects of method factors. Across these five studies, 78% (37 out of 47) of the method factor loadings were significant.

We also examined the possibility that the nonsignificance of method loadings might be due to empirical underidentification or overfitting (Kenny, 1979). For example, empirical underidentification can increase the standard errors of parameter estimates (Rindskopf, 1984), which may lead to the nonsignificance of method factor loadings. However, an inspection of the parameter estimates and their standard errors gave no indication of empirical underidentification or overfitting.[1]

Another important consideration in assessing method effects is the interpretability of parameter estimates. Parameter estimates are inconsistent if they are highly unlikely or contradict what would be expected on the basis of theoretical or methodological reasoning. For example, if method effects yield both positive and negative loadings on the same method factor, they are typically uninterpretable.[2] Browne (1984, p. 7) called these *wastebasket parameters* to indicate that they are introduced to achieve satisfactory goodness of fit but do not have a substantive interpretation (see also Kenny, 1979, p. 154). In our reanalyses, the results showed that Sims et al. (1976) and Meier (1974) had 2 and 1 inconsistent loadings for method factors, respectively, rendering method factors in these studies difficult to interpret. However, across these two studies, 14 of 17 method factor loadings were consistent, and 12 of 17 were significant.

To summarize, in five data sets (i.e., Dunham et al., 1977; Johnson et al., 1982; Meier 1984; Sims et al., 1976; Soutar & Weaver, 1982), most method factors (78%) showed statistically significant effects. These results suggest that method variance can often be significant, consistent with the findings of Williams et al. (1989). In the other six data sets, however, many of the individual method loadings showed statistical nonsignificance. Spector (1987) concluded that there was little evidence of method variance (i.e., in 1 of 10 studies), whereas Williams et al. concluded that there was strong evidence of method variance (i.e., in 9 of 11 data sets). Thus, our findings suggest a conclusion somewhere between Spector's and Williams et al.'s.

So far, we have implicitly assumed that the CFA model is adequate for analyzing all the data sets. However, the assumed structure of the trait and method effects should be tested with various goodness-of-fit indicators for the CFA model. A number of diagnostics for each data set are summarized in Table 3. The chi-square test, AGFI, and RMR are overall measures of fit in the sense that they express the discrepancy between the variance–covariance matrix implied by the hypothesized model and the observed variance–covariance matrix.

We used two additional criteria to evaluate the CFA solutions. Standardized residuals are formed by taking the residuals from the observed and implied variance–covariance matrices and dividing these residuals by their asymptotic standard errors. According to Jöreskog and Sörbom (1989, p. 32), "each standardized residual can be interpreted as a standard normal deviate and considered 'large' if it exceeds the value 2.58 in absolute value." Standardized residuals can be obtained as an option on LISREL 7. The presence of large standardized residuals indicates that a significant amount of variance remains unexplained and that the model may be misspecified.

The second procedure we used was an examination of each parameter for the presence of improper estimates. An improper estimate is one that is either illogical or outside the range of conventional acceptability. Negative error variances, correlations greater than 1.00, and standardized factor loadings greater than 1.00 are examples. Because improper estimates often result from model misspecifications (e.g., van Driel, 1978), they provide useful information about the adequacy of a model. Thus, the presence of large standardized residuals or improper estimates would indicate that the hypothesized model is not appropriate for the given data set.

Only statistically significant anomalies were considered improper estimates in this study. For example, nonsignificant negative error variances were not counted as improper estimates because they could occur as the result of sampling errors. Similarly, we counted as inconsistent estimates only those that were statistically significant and opposite in sign to that expected. Nevertheless, it should be acknowledged that the presence of a significant number of nonsignificant error variances could point to model misspecification if no theoretical or methodological reason can be offered to explain their occurrence.

Applying the above criteria to our analyses, we obtained the summary of results shown in Table 3. Notice first that the Dunham et al. (1977) analysis yielded an unsatisfactory goodness of fit based on the chi-square test, $\chi^2(76, N = 622) = 258.0$, $p < .001$, and 91 large standardized residuals. The data of Gillet and Schwab (1975) yielded a satisfactory chi-square statistic, $\chi^2(5, N = 273) = 6.1$, $p > .28$, but revealed 6 large standardized residuals. All studies showed satisfactory AGFIs (.87–.97), with the possible exception of McCabe et al.'s (1980) Study 2, and all studies yielded satisfactory rmrs (.01–.08), with the possible exception of Dunham et al. (1977). No improper estimate was found for any of the data sets. On balance, we have reason to accept the CFA solutions for 9 of the 11 data sets.

---

[1] We thank two anonymous reviewers for suggesting these ideas. A third reviewer pointed out that the difference in conclusions for the test of overall versus individual method effects may reflect the properties of the chi-square statistic and its degrees of freedom. For example, a chi-square value of 3 with 1 degree of freedom is not significant. But if this chi-square value were tested together with another value (an omnibus chi-square test of 6 with 2 degrees of freedom), it would be considered significant.

[2] A reviewer suggested that negative as well as positive loadings on the method factor could occur if organizational informants were used as methods. Presumably this could reflect systematic biases from multiple informants in a manner leading to opposite effects on each measurement. For example, if two informants from each of a set of organizations provided information on each measurement in an MTMM design, and the responses were systematically influenced in opposite ways (e.g., due to differences in knowledge, position in hierarchy, vested interest, power, etc.), negative and positive loadings could occur across items on a method factor.

Table 3
Summary of Goodness-of-Fit Measures for Confirmatory Factor Analysis Models

Study	χ^2	*df*	*p*	AGFI	rmr	No. of large standardized residuals	No. of improper estimates
Alderfer, 1967	12.9	14	.53	.91	.06	0	0
Dunham, Smith, & Blackburn, 1977	258.0	76	.00	.92	.08	91	0
Gillet & Schwab, 1975	6.1	5	.29	.96	.02	6	0
Johnson, Smith, & Tucker, 1982	10.4	14	.73	.92	.03	0	0
McCabe, Dalessio, Briga, & Sasaki, 1980 (Study 1)	6.5	14	.95	.94	.02	0	0
McCabe et al., 1980 (Study 2)	13.9	14	.46	.87	.02	0	0
Soutar & Weaver, 1982	13.0	14	.53	.96	.02	0	0
Spector, 1985	10.4	14	.73	.93	.03	0	0
Pierce & Dunham, 1978	2.6	5	.76	.97	.01	0	0
Sims, Szilagyi, & Keller, 1976	11.5	6	.07	.98	.01	0	0
Meier, 1984	11.7	12	.47	.97	.02	0	0

Note. AGFI = adjusted goodness-of-fit index; rmr = root-mean-square residual.

However, the analyses of the Dunham et al. (1977) and Gillet and Schwab (1975) studies yielded unsatisfactory goodness-of-fit indexes. Therefore, we rejected the hypothesis of linear, additive effects for methods implied by the CFA for these two data sets.

Next we examined whether the DPM is a viable alternative, especially for the two data sets not fitting the CFA pattern. A summary of the findings for the DPM applied to each data set is presented in Table 4. On the basis of the standard goodness-of-fit indicators, the DPM appears to fit the data of Gillet and Schwab (1975) and possibly Spector (1985). For example, the chi-square tests indicated an acceptable fit for these data sets, $\chi^2(16, N=273)=17.6, p>.30$; and $\chi^2(28, N=102)=40.9, p>.06$, respectively. However, the Spector (1985) data revealed 6 large standardized residuals, suggesting a model specification error. In fact, an inspection of the standardized residuals revealed the presence of large values in 8 of the 11 DPM analyses. In addition, 4 improper estimates were found in the Alderfer (1967) analysis. Finally, one error message arose in the analysis of McCabe et al.'s (1980, Study 2) data, suggesting that one parameter was unidentified. Because the parameter in question was in fact theoretically identified, it is likely that the message refers to empirical underidentification (Dillon, Kumar, & Mulani, 1987; Kenny, 1979; Rindskopf, 1984). In short, when all the goodness-of-fit indicators and diagnostics were taken into account, the evidence supported the DPM for the data of Gillet and Schwab (1975) but not for any of the remaining 10 data sets. The individual parameter estimates for the DPM analysis of the data in Gillet and Schwab (1975) are presented in Table 5.

The parameter estimates in Table 5 reveal some useful information about the properties of trait and method factors. The trait correlation matrix $\mathbf{P}_T$ can be easily retrieved from Φ_1. As expected, the elements of $\mathbf{P}_T$, the trait correlation coefficients corrected for attenuation, are larger than the original (observed) correlation coefficients (reported by Gillet & Schwab, 1985). For instance, the disattenuated correlation between promotion and pay is .63, which is higher than the corresponding raw correlations (ranging from .34 to .55). However, $\mathbf{P}_T$ reflects trends in the correlations among the observed measures. That is, the promotion–pay correlation is relatively large, whereas the promotion–coworkers and pay–coworkers correlations are small, consistent with the pattern in the original MTMM data. The elements of $\mathbf{P}_M$ can be similarly examined. Using the findings in Table 5, we found that correlation between the two methods to be .79.

An important purpose of MTMM analyses is to assess the construct validity of measures. Thus, we examined the convergent and discriminant validity for each data set by using both CFA models and DPMs. The convergent validity in the CFA model was first assessed by comparing hierarchically nested models (Schmitt & Stults, 1986; Widaman, 1985): Model 1 versus Model 2 and Model 3 versus Model 4 (see Table 1). The comparison of Model 1 with Model 2 resulted in a significant chi-square difference in all studies, suggesting that the addition of trait factors to a null model results in a better fit. The comparison of Model 3 and Model 4 also yielded a significant chi-square difference in all the studies, indicating that the addition of trait factors improved the model fit significantly.

We then examined the loadings in Λ_T to gain information on the degree of convergent validity. The loadings for trait factors indicated trait-related variation in the measures, and the extent of trait variation reflected the magnitude of shared variation for two or more measures on a common factor. Within the context of CFA, this variation has method and error variance removed from it. Trait variance thus yields a quantitative indicator of the degree of convergent validity. Convergent validity can be said to result when the trait factor loading on a measure of interest is statistically significant.

As shown in Table 6, convergent validity in this sense was achieved for most data sets. Overall, 81% of measures showed convergent validity across studies. Two data sets failed to achieve convergent validity, one data set revealed mixed results, and eight data sets achieved convergent validity (see Table 7 for a summary). We should stress that this conclusion is based on the statistical significance of trait factor loadings. It is possible that significant trait factor loadings can be low from a practical point of view.

We assessed discriminant validity by examining the correlations among traits and their standard errors under the CFA

Table 4
Summary of Goodness-of-Fit Measures for Direct Product Models

Study	χ^2	*df*	*p*	AGFI	rmr	No. of large standardized residuals	No. of improper estimates
Alderfer, 1967	50.1	28	.01	.82	.12	6	4
Dunham, Smith, & Blackburn, 1977	465.2	101	.00	.89	.08	46	0
Gillet & Schwab, 1975	17.6	16	.30	.96	.04	0	0
Johnson, Smith, & Tucker, 1982	44.3	28	.03	.85	.07	4	0
McCabe, Dalessio, Briga, & Sasaki, 1980 (Study 1)	52.7	28	.00	.80	.05	3	0
McCabe et al., 1980 (Study 2)	42.7	28	.04	.82	.02	0	0
Soutar & Weaver, 1982	53.8	28	.00	.92	.04	11	0
Spector, 1985	40.9	28	.06	.86	.07	6	0
Pierce & Dunham, 1978	33.3	16	.01	.89	.06	1	0
Sims, Szilagyi, & Keller, 1976	92.4	16	.00	.93	.05	15	0
Meier, 1984	49.3	25	.00	.94	.03	0	0

Note. AGFI = adjusted goodness-of-fit index; rmr = root-mean-square residual.

model. Discriminant validity among traits is achieved when an intertrait correlation is significantly different from 1.00 or when the chi-square difference test indicates that the two traits are not perfectly correlated (e.g., Schmitt & Stults, 1986; Widaman, 1985). Discriminant validity was established for all the measures of 7 data sets: Dunham et al. (1977), Johnson et al. (1982), McCabe et al. (1980, Study 1 and Study 2), Pierce and Dunham (1978), Spector (1985), and Sims et al. (1976). In contrast, discriminant validity was achieved 8 of 10 times for Alderfer (1967), 0 of 6 times for Gillet and Schwab (1975), 6 of 10 times for Soutar and Weaver (1982), and 2 of 3 times for Meier (1984). The third column of Table 7 summarizes the results of these analyses.

We also examined convergent and discriminant validity under the DPM, using the criteria described earlier. Because the method correlation (i.e., $r = .18$) was small in Alderfer (1967), we concluded that convergent validity was not achieved. Because the method correlations were relatively large in all other studies, we concluded that convergent validity was achieved. Specifically, the average method correlation was .64 for Dunham et al. (1977), .79 for Gillet and Schwab (1975), .77 for Johnson et al. (1982), .98 for McCabe et al. (1980, Studies 1 and 2), .94 for Soutar and Weaver (1982), .94 for Spector (1985), .83 for Pierce and Dunham (1978), .83 for Sims et al. (1976), and .94 for Meier (1984).

The test of discriminant validity can be illustrated with the data of Alderfer (1967). The correlations among traits ranged from −.02 to .03 and are significantly less than unity, satisfying the first criterion of discriminant validity. The method correlation was .18, which was larger than any of the trait correlations, satisfying the second criterion. The third requirement was also met. All the studies were evaluated in terms of these criteria, and as the final column of Table 7 shows, the criteria for discriminant validity were met for all studies except Dunham et al. (1977).[3]

Discussion

We investigated the nature of method effects (i.e., additive or multiplicative) by comparing two alternative models: CFA and DPM. To gain perspective into the issues, let us examine Table 7, which summarizes the conclusions we drew from the CFA and DPM analyses. The conclusions are based on a full interpretation of goodness-of-fit measures, parameter estimates, and the other diagnostics mentioned earlier (cf. Tables 3 and 4). When the model-fit criteria in the full sense are considered, neither the CFA nor DPM hypotheses fit the data of Dunham et al. (1977). The DPM, but not the CFA model, fits the data of Gillet and Schwab (1975).[4] In contrast, the CFA model, but not the DPM, fits all the other data sets. The results across studies tend to support the premise that MTMM data can be explained by either additive or multiplicative method effects, but not by both.[5] Only the data in Dunham et al.'s (1977) study failed to fit either model. One explanation for this lack of fit is methodological. Dunham et al. administered 41 scales to respondents, and each scale contained many items. The length of their survey might have induced fatigue and other biases, leading to the lack of a discernable structure in their data.

[3] No standard errors were available in McCabe, Dalessio, Briga, and Sasaki's (1980) Study 2 because of the empirical underidentification problem; as a result, the first criterion could not be tested statistically. However, the size of the correlations (.40–.76) suggests that they all are most likely lower than unity and thus probably satisfy the first criterion.

[4] A convergence failure occurred for the CFA model (with two method factors) in this data set, but an alternative model hypothesizing only one method factor yielded converging solutions with a satisfactory fit. The results suggest that overfitting might be a problem for the CFA model in this data set.

[5] The findings for the DPM applied to the data of Meier (1984) might be interpreted as supporting the model, thus implying that both the CFA model and DPM fit the data. However, although no standardized residuals were significant and all other parameter estimates pointed to an acceptable DPM fit, the chi-square test suggested a poor fit. Thus, the evidence is mixed for accepting the DPM. When it is not possible to differentiate between the CFA and DPM analyses for a particular data set on the basis of the criteria scrutinized herein, the researcher might wish to apply cross-validation and penalty functions (e.g., Cudeck & Browne, 1983). We acknowledge that it is possible for both the CFA model and DPM to fit a particular data set, although it is unlikely (cf. Bagozzi & Yi, in press).

Table 5
Parameter Estimates for Direct-Product-Model Analysis of Data in Gillet and Schwab (1975)

Trait–method combination	Z (Λ)	$E_M^{1/2} \otimes I_t$ (Γ_2)	$C_M \otimes I_t$ (Γ_1) 1	2	3	4	5	6	7	8	$I_m \otimes P_T$ (Φ_1) 1	2	3	4	5	6	7	8	$I_m \otimes E_T$ (Φ_2)
1. Promotion–MSQ	0.91	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00								0.18
2. Pay–MSQ	0.92	1.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.63	1.00							0.17
3. Coworkers–MSQ	0.91	1.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.20	0.17	1.00						0.23
4. Supervision–MSQ	0.97	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.49	0.36	0.37	1.00					0.05
5. Promotion–JDI	0.91	1.71	0.65	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00				0.18
6. Pay–JDI	0.93	1.71	0.00	0.65	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.63	1.00			0.17
7. Coworkers–JDI	0.87	1.71	0.00	0.00	0.65	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.20	0.17	1.00		0.23
8. Supervision–JDI	1.09	1.71	0.00	0.00	0.00	0.65	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.49	0.36	0.37	1.00	0.05

Note. MSQ = Minnesota Satisfaction Questionnaire; JDI = Job Descriptive Index.

If we had based our conclusions about model fit solely on the chi-square goodness-of-fit tests (cf. Tables 3 and 4), the CFA model would have been accepted for Gillet and Schwab (1975) and the DPM accepted for Spector (1985). Aside from overlooking the anomalies noted in the Results section, the use of only the chi-square goodness-of-fit test would thus have led to ambiguous and contradictory results for these data sets. We see the need for careful examination of individual parameter estimates, standardized residuals, and the additional diagnostics noted earlier.

When the evidence for convergent and discriminant validity is examined, the CFA and DPM conclusions can be quite different (see Table 7). Consider, for example, the analyses of the data in Gillet and Schwab (1975). Here CFA indicated a failure to achieve convergent and discriminant validity, whereas the DPM analysis led to a conclusion of satisfactory convergent and discriminant validity. Which conclusion is appropriate? Recall that the DPM gave a satisfactory fit to the data, whereas the CFA model was rejected. From this perspective, more credence should be given to the conclusion from the DPM analysis. This result suggests that one should investigate the structure of method effects before assessing construct validity.

To date, most analyses of MTMM matrices have been based on the assumption that effects of methods and traits are additive. As Campbell and O'Connell (1967, p. 424) argued, the assumptions underlying the additive models of factor analysis have been untested. Researchers have invariably used factor analysis as the criterion to which the data should fit, never vice versa. However, analytical procedures should be chosen on the basis of the fit of the models (e.g., CFA or DPM) to the data. That is, the models should be tested against data, rather than merely assumed. Our analyses revealed that the multiplicative effects are plausible for at least some data, suggesting that more attention should be given to the DPM in MTMM analyses (see also Bagozzi & Yi, in press).

One potential limitation of the DPM might be noted. The DPM criterion for convergent validity requires that the method correlations be substantial. This criterion is a composite indicator that implicitly takes into account the resultant convergence among multiple measures for each trait. The composite indicator does not identify the degree of convergent validity or point out which measure is satisfactory or not. The CFA criterion for convergent validity is based on the amount of trait variance for each measure and provides diagnostic information about which measures achieve convergent validity. Such diagnostic results can aid researchers in item selection for future research (cf. Anderson & Gerbing, 1988). It thus appears that the DPM is less informative than the CFA model with respect to convergent validity.

The primary focus of the current article was on the adequacy of MTMM models based on commonly accepted statistical criteria. We now consider the practical relevance of the findings, with particular emphasis on the nature of overfitting. For each of the data sets, we computed noncentralized NFIs for (a) the trait-only CFA model compared with the null model, (b) the trait–method CFA model compared with the null model, and (c) the trait–method CFA model compared with the trait-only CFA model. The first two indices give the proportion of total information accounted for by the trait-only and trait–method

Table 6
Results of Assessment of Convergent Validity in Confirmatory Factor Analyses

Study	Model 1 vs. Model 2 χ^2	df	Model 3 vs. Model 4 χ^2	df	No. of significant trait factor loadings	No. of inconsistent trait factor loadings
Alderfer, 1967	214.6	20	106.3	11	0/10	0
Dunham, Smith, & Blackburn, 1977	4,949.6	22	2,401.5	22	16/16	0
Gillet & Schwab, 1975	670.9	14	253.3	14	0/8	0
Johnson, Smith, & Tucker, 1982	451.2	20	194.1	20	10/10	0
McCabe, Dalessio, Briga, & Sasaki, 1980 (Study 1)	724.3	20	428.5	20	10/10	0
McCabe et al., 1980 (Study 2)	891.6	20	363.0	20	10/10	0
Soutar & Weaver, 1982	1,215.5	20	186.8	11	7/10	0
Spector, 1985	460.1	20	208.0	20	10/10	0
Pierce & Dunham, 1978	623.9	14	211.9	14	8/8	0
Sims, Szilagyi, & Keller, 1976	1,003.8	14	248.2	14	8/8	0
Meier, 1984	1,617.4	12	573.7	12	9/9	0

Note. Model 1 is the null model; Model 2 is the trait-only model; Model 3 is the method-only model; and Method 4 is the trait–method model. All chi-square values are significant ($p < .05$).

models, respectively, from a practical standpoint (e.g., Bentler & Bonett, 1980; Mulaik et al., 1989). The third index provides an indication of the gain in goodness-of-fit when going from the trait-only to trait–method CFA model. In all three indices, the appropriate degrees of freedom are subtracted from their respective chi-square values to yield noncentralized estimates. The noncentralized NFIs remove the bias in small samples of ordinary NFIs (e.g., McDonald & Marsh, 1990).

The findings for the application of these noncentralized NFIs to the 11 data sets examined in this article are presented in Table 8. Shown in the first column, the noncentralized NFIs for the trait-only model are quite large for most data sets. This suggests that trait factors explain a substantial amount of information in the 11 data sets. An inspection of the second column shows that both trait and method factors explain virtually all the information in the data. The increments in noncentralized NFIs going from the trait-only to the trait–method CFA model (see column 3 in Table 8) range from .01 to .24. Our analyses, based on commonly accepted statistical criteria, showed that method variance was significant for 5 of 11 data sets (see Table 2). In three of these five studies (i.e., Dunham et al., 1977; Johnson et al., 1982; Sims et al., 1976), the improvement in the noncentralized NFI value due to the addition of methods was larger than .05 (.24, .12, and .06, respectively). If the .05 value is used as a rule of thumb, one might conclude that method variance is significant in both statistical and practical senses for

Table 7
Qualitative Summary of Findings Across Studies

Study	Confirmatory factor analysis: Model fit	Convergent validity	Discriminant validity	Direct-product model: Model fit	Convergent validity	Discriminant validity
Alderfer, 1967	Accept	Fail	Mixed	Reject	Fail	Pass
Dunham, Smith, & Blackburn, 1977	Reject	Pass	Pass	Reject	Pass	Fail
Gillet & Schwab, 1975	Reject	Fail	Fail	Accept	Pass	Pass
Johnson, Smith, & Tucker, 1982	Accept	Pass	Pass	Reject	Pass	Pass
McCabe, Delessio, Briga, & Sasaki, 1980 (Study 1)	Accept	Pass	Pass	Reject	Pass	Pass
McCabe et al., 1980 (Study 2)	Accept	Pass	Pass	Accept	Pass	Pass
Soutar & Weaver, 1982	Accept	Mixed	Mixed	Reject	Pass	Pass
Spector, 1985	Accept	Pass	Pass	Reject	Pass	Pass
Pierce & Dunham, 1978	Accept	Pass	Pass	Reject	Pass	Pass
Sims, Szilagyi, & Keller, 1976	Accept	Pass	Pass	Reject	Pass	Pass
Meier, 1984	Accept	Pass	Mixed	Reject	Pass	Pass

Note. The qualitative conclusions summarized here should be interpreted, and tempered if necessary, in conjunction with the findings for goodness-of-fit of models and other diagnostics.

Table 8
Results for Testing Overfitting Hypothesis by Use of Noncentralized Normed-Fit-Index

Study	Model 2 vs. Model 1	Model 4 vs. Model 1	Model 4 vs. Model 2
Alderfer, 1967	0.90	1.00	0.11
Dunham, Smith, & Blackburn, 1977	0.74	0.97	0.24
Gillet & Schwab, 1975	0.96	1.00	0.04
Johnson, Smith, & Tucker, 1982	0.89	1.01	0.12
McCabe, Dalessio, Briga, & Sasaki, 1980 (Study 1)	0.97	1.01	0.05
McCabe et al., 1980 (Study 2)	0.99	1.00	0.01
Soutar & Weaver, 1982	0.98	1.00	0.02
Spector, 1985	1.00	1.01	0.01
Pierce & Dunham, 1978	0.93	1.00	0.07
Sims, Szilagyai, & Keller, 1976	0.94	0.99	0.06
Meier, 1984	0.98	1.00	0.02

Note. Model 1 is the null model; Model 2 is the trait-only model; and Model 4 is the trait–method model.

these data sets. In fact, three other studies with nonsignificant method loadings (Alderfer, 1967; McCabe et al., 1980, Study 1; Pierce & Dunham, 1978) had noncentralized NFI increments larger than .05 (.11, .05, and .07, respectively). On the basis of the noncentralized NFI values (as well as the chi-square tests), one could argue that method factors are important for these studies, though standards for what constitutes a significant increment in noncentralized NFIs are lacking.

Some caveats are in order. The findings of this investigation may be limited in their generalizability because they are based on empirical data sets in a selected research area (cf. Bagozzi & Yi, in press). Caution is also needed in comparing the results across the studies examined because they had different sample sizes. It has been shown that sample size has significant effects on goodness-of-fit indicators such as AGFI and rmr (La Du & Tanaka, 1989; Marsh, Balla, & McDonald, 1988). Moreover, the rules we suggest for assessing validity and practical significance are merely heuristics.

In summary, this research indicates that the assessment of method variance in MTMM analyses is a complex process involving a number of criteria. Our reanalyses of the data analyzed by Spector (1987) suggest that the conclusions stated by Williams et al. (1989) could have been an artifact of their analytic procedure, which was based solely on overall tests of fit. In the 10 studies on affect and perceptions at work, we found method variance to be sometimes significant, but not as prevalent as Williams et al. concluded. We also found that method effects were sometimes multiplicative (though much less prevalent than Campbell & O'Connell, 1967, 1982, suggested) rather than additive, so that the usual CFA model was inappropriate. Thus, it seems necessary for researchers to consider alternative models (i.e., the CFA model and DPM) when analyzing MTMM matrices.

Future research should be directed at determining the conditions under which each model is appropriate. Researchers might conduct simulation studies to compare the performance of alternative models (i.e., CFA model and DPM) over a range of relevant factors. Such studies ought to contribute to a better understanding of the consequences and implications of using the improper model when analyzing MTMM data. Researchers should examine MTMM data in other substantive areas as well to determine the generalizability of these findings.

References

Alderfer, C. (1967). Convergent and discriminant validation of satisfaction and desire measures by interviews and questionnaires. *Journal of Applied Psychology, 51,* 509–520.

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103,* 411–423.

Bagozzi, R. P. (1980). *Causal models in marketing.* New York: Wiley.

Bagozzi, R. P., & Yi, Y. (1990). On the analysis of multitrait–multimethod matrices in consumer research. *Journal of Consumer Research.*

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

Bentler, P., & Lee, S. (1978). Statistical aspects of a three-mode factor analysis model. *Psychometrika, 43,* 343–352.

Bloxom, B. (1968). A note on invariance in three-mode factor analysis. *Psychometrika, 33,* 347–350.

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

Browne, M. W. (1985). MUTMUM: *Decomposition of multitrait–multimethod matrices.* Unpublished manuscript, University of South Africa, Pretoria.

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.

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.

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

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.

Dunham, R., Smith, F., & Blackburn, R. (1977). Validation of the Index of Organizational Reactions with the JDI, the MSQ, and the Faces scales. *Academy of Management Journal, 20,* 420–432.

Gillet, B., & Schwab, D. (1975). Convergent and discriminant validities of corresponding Job Descriptive Index and Minnesota Satisfaction Questionnaire scales. *Journal of Applied Psychology, 60,* 313–317.

Johnson, S., Smith, P., & Tucker, S. (1982). Response format of the Job Descriptive Index: Assessment of reliability and validity by the multitrait–multimethod matrix. *Journal of Applied Psychology, 67,* 500–505.

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.

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

La Du, T. J., & Tanaka, J. S. (1989). Influence of sample size, estimation method, and model specification on goodness-of-fit assessments in structural equation models. *Journal of Applied Psychology, 74,* 625–635.

Marsh, H. W., Balla, J. R. & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103,* 391–410.

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

McCabe, D., Dalessio, A., Briga, J., & Sasaki, J. (1980). The convergent and discriminant validities between the IOR and the JDI: English and Spanish forms. *Academy of Management Journal, 23,* 778–786.

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

Meier, S. (1984). The construct validity of burnout. *Journal of Occupational Psychology, 57,* 211–219.

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

Pierce, J., & Dunham, R. (1978). The measurement of perceived job characteristics: The Job Diagnostic Survey versus the Job Characteristics Inventory. *Academy of Management Journal, 21,* 123–128.

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

Satorra, A., & Saris, W. E. (1985). Power of the likelihood ratio test in covariance structure analysis. *Psychometrika, 50,* 83–90.

Schmitt, N., Coyle, B. W., & Saari, B. (1977). A review and critique of analyses of multitrait–multimethod matrices. *Multivariate Behavioral Research, 12,* 447–478.

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

Searle, S. R. (1982). *Matrix algebra useful for statistics.* New York: Wiley.

Sims, H., Szilagyi, A., & Keller, R. (1976). The measurement of job characteristics. *Academy of Management Journal, 19,* 195–212.

Sobel, M. E., & Bohrnstedt, G. W. (1985). Use of null models in evaluating the fit of covariance structure models. In N. Tuma (Ed.), *Sociological methodology 1985* (pp. 152–178). San Francisco: Jossey-Bass.

Soutar, G., & Weaver, J. (1982). The measurement of shop-floor job satisfaction: The convergent and discriminant validity of the Worker Opinion Survey. *Journal of Occupational Psychology, 55,* 27–33.

Spector, P. (1985). Measurement of human service staff satisfaction: Development of the Job Satisfaction Survey. *American Journal of Community Psychology, 13,* 693–713.

Spector, P. (1987). Method variance as an artifact in self-reported affect and perceptions at work: Myth or significant problem? *Journal of Applied Psychology, 72,* 438–443.

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

Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika, 31,* 279–311.

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

Werts, C. E., Jöreskog, K. G., & Linn, R. L. (1972). A multitrait–multimethod model for studying growth. *Educational and Psychological Measurement, 32,* 655–678.

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

Williams, L. J., Cote, J. A., & Buckley, M. R. (1989). Lack of method variance in self-reported affect and perceptions at work: Reality or artifact? *Journal of Applied Psychology, 74,* 462–468.

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

(*Appendix follows on next page*)

Appendix

Input Program Specification of the Direct Product Model for the Data in Gillet and Schwab (1975)

```
DA NI=8 NO=273 MA=KM
[data]
MO NY=8 NE=8 NK=16 LY=DI, FR GA=FU, FR PH=SY, FR PS=ZE
BE=ZE, TE=ZE
ST.8 LY 1-LY 8
EQ LY 1 LY 5
PA GA
*
0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0
1 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0
0 1 0 0 0 1 0 0 0 0 0 0 0 1 0 0
0 0 1 0 0 0 1 0 0 0 0 0 0 0 1 0
0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 1
MA GA
*
1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0
0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0
0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0
1 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0
0 1 0 0 0 1 0 0 0 0 0 0 0 1 0 0
0 0 1 0 0 0 1 0 0 0 0 0 0 0 1 0
0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 1
EQ GA 5 1 GA 6 2 GA 7 3 GA 8 4
EQ GA 5 5 GA 6 6 GA 7 7 GA 8 8
FI GA 1 9 GA 2 10 GA 3 11 GA 4 12
EQ GA 5 13 GA 6 14 GA 7 15 GA 8 16
PA PH
*
0
1 0
1 1 0
1 1 1 0
0 0 0 0 0
0 0 0 0 1 0
0 0 0 0 1 1 0
0 0 0 0 1 1 1 0
0 0 0 0 0 0 0 0 1
0 0 0 0 0 0 0 0 0 1
0 0 0 0 0 0 0 0 0 0 1
0 0 0 0 0 0 0 0 0 0 0 1
0 0 0 0 0 0 0 0 0 0 0 0 1
0 0 0 0 0 0 0 0 0 0 0 0 0 1
0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
MA PH
*
1
.5 1
.5 .5 1
.5 .5 .5 1
0 0 0 0 1
0 0 0 0 .5 1
0 0 0 0 .5 .5 1
0 0 0 0 .5 .5 .5 1

0 0 0 0 0 0 0 0 .2
0 0 0 0 0 0 0 0 0 .2
0 0 0 0 0 0 0 0 0 0 .2
0 0 0 0 0 0 0 0 0 0 0 .2
0 0 0 0 0 0 0 0 0 0 0 0 .2
0 0 0 0 0 0 0 0 0 0 0 0 0 .2
0 0 0 0 0 0 0 0 0 0 0 0 0 0 .2
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 .2
EQ PH 2 1 PH 6 5
EQ PH 3 1 PH 7 5
EQ PH 4 1 PH 8 5
EQ PH 3 2 PH 7 6
EQ PH 4 2 PH 8 6
EQ PH 4 3 PH 8 7
EQ PH 9 9 PH 13 13
EQ PH 10 10 PH 14 14
EQ PH 11 11 PH 15 15
EQ PH 12 12 PH 16 16
OU NS SS TV RS ND=4 AD=5000
```

Received October 17, 1989
Revision received March 22, 1990
Accepted March 23, 1990 ■