JOURNAL OF CONSUMER PSYCHOLOGY, *2*(2), 143–170
Copyright © 1993, Lawrence Erlbaum Associates, Inc.

Multitrait–Multimethod Matrices in Consumer Research: Critique and New Developments

Richard P. Bagozzi
School of Business Administration
The University of Michigan

Youjae Yi
School of Management
Seoul National University

In this article we address problems with recently proposed methods for examining construct validity and we introduce alternatives. With respect to the first-order confirmatory factor analysis model, shortcomings in the application of the method to multitrait–multimethod data are considered. The correlated uniqueness model is then presented as an alternative when ill-defined solutions arise in first-order confirmatory factor analyses or when method and trait variance are confounded. Procedures are then developed for investigating important hypotheses in the application of the direct product model. These procedures were not considered in previous treatments. Finally, panel models are shown to be adaptable to certain investigations of construct validity.

Measurement error is frequently recognized as a problem by many researchers because it can have serious confounding influences on empirical research (e.g., Cote & Buckley, 1987; Peter, 1981). Random error tends to attenuate the observed correlations among variables and may yield misleading conclusions. Systematic error such as method variance may also bias results by inflating the observed correlations among variables measured with the common method. It is thus important to validate measures prior to theory testing. Such construct validation can be done with multitrait–multimethod (MTMM) matrices (Campbell & Fiske, 1959).

Requests for reprints should be sent to Richard P. Bagozzi, School of Business Administration, University of Michigan, Ann Arbor, MI 48109.

Bagozzi and Yi (1991) examined three alternative procedures for analyzing MTMM matrices: the classic Campbell and Fiske (1959) procedure, the conventional first-order confirmatory factor analysis (CFA) model (e.g., Cote & Buckley, 1987; Jöreskog, 1981; Widaman, 1985), and the direct product (DP) model (e.g., Browne, 1984; Lastovicka, Murry, & Joachimsthaler, 1990; Wothke & Browne, 1990). The latter two procedures were offered as alternatives to the classic approach which has serious limitations.

This article begins with a critique of the conventional CFA model as applied to MTMM matrices and introduces an alternative (i.e., the correlated uniqueness model) which overcomes problems encountered by Bagozzi and Yi (1991). Next, hypotheses not considered by Bagozzi and Yi (1991), but important to the assessment of construct validity, are developed for the DP model. Procedures for testing these hypotheses are then illustrated. A more flexible and less cumbersome software package is used than that employed by Bagozzi and Yi (1991). Finally, a new approach to construct validity is described that applies to cases where as few as two traits are measured either by two or more methods or by two or more indicators per trait from a single method. The analysis of MTMMs by structural equation procedures, in contrast, requires either (a) three traits and three methods, (b) four traits and two methods, or (c) two traits and four methods for the achievement of identification in the general case.

FIRST-ORDER CFA

Background

To understand the limitations of the first-order CFA model, it is useful to begin with the criteria proposed by Campbell and Fiske (1959). Campbell and Fiske argued that construct validity can be ascertained by examining *convergent validity* and *discriminant validity.* Convergent validity is the degree to which multiple attempts to measure the same concept are in agreement. Measures of the same trait, no matter how derived, should be highly correlated if they validly measure a common construct. Discriminant validity is the degree to which measures of different concepts are distinct. That is, if two or more concepts are unique, valid measures of each should not covary too highly.

Convergent validity obtains, according to their criteria, when the monotrait–heteromethod coefficients are statistically significant and sufficiently large. The monotrait–heteromethod coefficients represent correlations between measures of the same trait by different methods and are sometimes termed validity coefficients. Establishment of convergent validity provides evidence that multiple measures of a construct obtained by multiple methods potentially indicate the same underlying construct (Peter, 1981). If the monotrait–heteromethod correlations were nonsignificant or too low in magnitude, there is little basis to argue that the measures tap the same construct, and consideration of discriminant validity is not warranted.

However, if convergent validity is demonstrated, this only provides minimal evidence for the construct validity of measures of a construct. It is possible that the measures also reflect other constructs 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 monotrait–heteromethod coefficients should be higher than their corresponding heterotrait–heteromethod coefficients. In other words, efforts to measure the same trait by different methods should yield higher correlations than efforts to measure different traits by different methods.

A second discriminant validity criterion specifies that the monotrait–heteromethod coefficients should be higher than their corresponding heterotrait–monomethod coefficients. Efforts to measure the same construct by different methods should produce higher correlations than efforts to measure different constructs by the same method. The final criterion for discriminant validity is that the pattern of correlations among traits should be the same in the monomethod and heteromethod blocks. When this criterion holds, correlations among traits will be independent of methods. But when it fails, trait correlations will be differentially impacted by methods.

Campbell and Fiske (1959) hoped that their procedure would provide a disentangling of trait and method effects. However, the assumptions of their approach are so restrictive, and the information provided so limited, as to make its use treacherous (Peter, 1981; Widaman, 1985). The key assumptions are the following: Traits and methods are assumed uncorrelated, methods affect all traits equally, methods are orthogonal, and measures are assumed equally reliable (Campbell & Fiske, 1959; Schmitt & Stults, 1986). Information not supplied by the approach, but essential to the interpretation of construct validity, includes statistical tests of the comparisons just noted and the amount of variation in measures due to traits, methods, and error.

First-Order Trait–Method Model

It was against this backdrop that researchers offered the first-order CFA model (Cote & Buckley, 1987; Widaman, 1985). The hope was that valid insights would be provided into the amounts of measure variance due to traits, methods, and error, and the degree to which convergent and discriminant validity are achieved. Bagozzi and Yi (1991) adapted Widaman's (1985) perspective in their article.

The CFA model does exhibit a number of advantages over the Campbell and Fiske (1959) procedure. It allows methods to correlate freely and affect measures to different degrees. It provides various measures of fit for an overall model, as well as estimates and tests of significance of convergent and discriminant validity. It gives a partitioning of variance into trait, method, and error components (Cote & Buckley, 1987). And by placing restrictions on the covariances among traits or among methods, it even permits estimation of trait–method correlations (Kumar & Dillon, 1992).

Nevertheless, it is important to understand the limitations of the CFA model. One assumption inherent in the model is that the error terms contain both specific and error variance (see Anderson, 1985). The consequences of this are 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). Kumar and Dillon (1990) proposed a model which overcomes this drawback of the standard CFA model by separating specific and error variance (see also Anderson, 1985). In the general case, their model requires three or more items from each of three or more methods on each of three or more traits and thus demands at least three times as many measures as the traditional MTMM analysis. Another characteristic making the model proposed by Kumar and Dillon (1990) limited in practical utility is the likelihood that attempts to apply the model will result in overfitting. So many parameters are fit to this model that failures to converge and improper solutions will occur frequently. In this article, focus is placed on the more common MTMM context where each of three or more traits is indicated by three or more methods.

A little-known limitation of the first-order CFA model is that the partitioning of variance into trait and method components does not, in general, yield method-free and trait-free interpretations (Kumar & Dillon, 1992). This is because the individual factor loadings take different values corresponding to the distinct trait–method pairings. For example, factor loadings concerning a trait vary across methods, and the corresponding variation cannot be attributed solely to the trait factor. Because each factor loading is specific to the particular trait–method combination, the associated variation is not really trait free or method free.

In general, the CFA model cannot disentangle the source of variation when the sources are highly correlated. 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 increase, trait and method variance will be confounded. For example, a general trait factor may underlie traits so that traits are highly correlated and that substantial variance in measures is primarily due to traits, while methods are relatively distinct. In such circumstances, application of the first-order CFA model can misleadingly yield highly correlated methods accounting for much variation in measures (Marsh, 1989). However, a good fitting first-order 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. In such cases, correlations among method factors represent the convergence of the general trait factor across methods, rather than true relationships among methods. Because most applications of the MTMM analysis involve substantially correlated traits and/or methods, the interpretation of the results from a first-order CFA is likely to be misleading in practice.

Yet another problem with the traditional first-order CFA model is the all too frequent occurrence of ill-defined solutions (Wothke, 1984, 1987). Ill-defined solutions include:

> 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 of values (e.g., negative variance estimates called Heywood cases), or standard errors of parameter estimates that are excessively large. (Marsh, 1989, p. 339)

In the analyses reported by Bagozzi and Yi (1991), for example, all solutions included either negative error variances (albeit nonsignificant), correlations greater than 1, or method factor loadings opposite in sign to that predicted by theory. Findings such as these bring into question the interpretation of the first-order CFA model as a proper representation of some MTMM data.

CORRELATED UNIQUENESS MODEL

As a remedy to the ill-defined solution problem, Marsh (1989) proposed a new model which he termed the correlated uniqueness (CU) model. Figure 1 presents the CU model for the three-trait, three-method case. Trait factors are represented identically to that found in the first-order CFA model, but instead of method factors, all uniquenesses corresponding to measures derived from

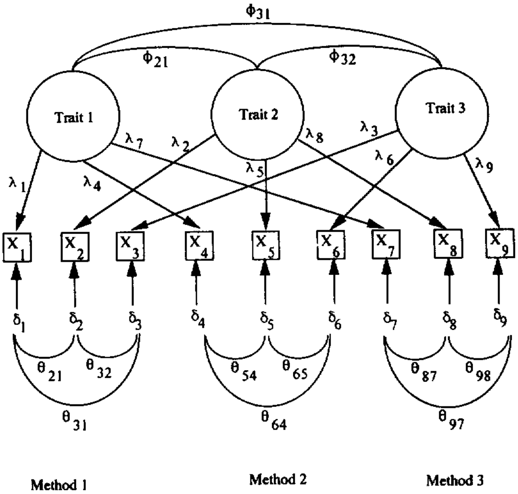

FIGURE 1 The correlated uniqueness model.

the same method are allowed to correlate freely. Whereas the first-order CFA model assumes that each method factor is unidimensional and follows a congeneric-like structure, the CU model represents the effects of methods as correlations among pairs of error terms. The CU model reduces considerably the possibility for ill-defined solutions. Marsh and Bailey (1991) compared the general CFA model and the CU model by analyzing 255 MTMM matrices constructed from real data (Study 1) and 180 MTMM matrices from simulated data (Study 2). The CU model converged to proper solutions for 99% in Study 1 and for 96% in Study 2, whereas the general CFA model showed only 24% and 22%, respectively. Even when both models converged to proper solutions, parameter estimates from the CU model were more accurate and precise than those from the general CFA model.

The CU model also allows for tests of convergent and discriminant validity. Specifically, trait factor loadings can be scrutinized for assessing convergent validity (i.e., the agreement among measures of the same trait by different methods), because the trait factor loading reflects the degree to which the observed measure is determined by the trait factor. Also, the trait correlations can be examined for assessing discriminant validity.

Furthermore, unlike the first-order CFA model, the CU model does not suffer from the potential ambiguity in interpreting the correlated method factors. It has been noted that the first-order CFA model could produce ostensibly correlated method factors which in fact reflect general trait effects across methods instead of true method effects (Marsh, 1989). In contrast, correlated uniquenesses representing method effects in the CU model would not be affected by trait variance across methods, because they are allowed only within each method, not across methods. Thus, the substantive interpretation of method factors can be less ambiguous under the CU model than under the first-order CFA model.

To illustrate and make comparisons to the findings noted in Bagozzi and Yi (1991), the CU model was applied to the same data sets examined by Bagozzi and Yi (i.e., Arora, 1982; Foxman, Tansuhaj, & Ekstrom, 1989; Menezes & Elbert, 1979; Seymour & Lessne, 1984). All analyses were performed with EQS (Bentler, 1989). The first thing to note is that, based on the chi-square goodness-of-fit criterion, all CU models except that for the data in Seymour and Lessne (1984), fit satisfactorily: $\chi^2(15) = 21.70, p \approx .12$ (Arora, 1982); $\chi^2(15) = 16.25, p \approx .37$ (Foxman et al., 1989); $\chi^2(15) = 14.02, p \approx .52$ (Menezes & Elbert, 1979); $\chi^2(15) = 27.51, p \approx .02$ (Seymour & Lessne, 1984).[1] Table 1 summarize the results of the parameter estimates for the CU model. It can be

[1] It is possible to examine other criteria related to goodness-of-fit such as so-called "practical relevance" indices. But this is not done in this article for purposes of simplicity and because use of such indices does not change the points made and the conclusions drawn therefrom in this article.

TABLE 1
Parameter Estimates for Correlated Uniqueness Models

Method–Trait	*Factor Loadings*	*Correlated Uniqueness*			*Factor Correlations*		
Arora (1982)							
Semantic Differential							
Situational involvement (SI)	.84 (.08)	.21 (.06)			1.00		
Enduring involvement (EI)	.90 (.08)	.09 (.04)	.18 (.04)		.34 (.10)	1.00	
Response involvement (RI)	.69 (.09)	.16 (.06)	.12 (.05)	.53 (.10)	.21 (.12)	.36 (.10)	1.00
Likert							
Situational involvement (SI)	.76 (.08)	.49 (.09)					
Enduring involvement (EI)	.84 (.08)	.09 (.05)	.27 (.05)				
Response involvement (RI)	.72 (.09)	.35 (.07)	.10 (.05)	.52 (.10)			
Stapel							
Situational involvement (SI)	.81 (.09)	.36 (.08)					
Enduring involvement (EI)	.94 (.08)	−.01 (.04)	.13 (.04)				
Response involvement (RI)	.86 (.09)	.01 (.05)	.10 (.04)	.25 (.09)			
Foxman et al. (1989)							
Father							
Suggest price range (SPR)	.36 (.10)	.88 (.11)			1.00		
Shop with parents (SP)	.15 (.10)	.36 (.08)	.98 (.11)		.78 (.12)	1.00	
Suggest stores (SS)	.21 (.10)	.31 (.08)	.49 (.09)	.96 (.11)	.64 (.12)	.73 (.11)	1.00
Mother							
Suggest price range (SPR)	.51 (.12)	.74 (.13)					
Shop with parents (SP)	.66 (.15)	.02 (.10)	.57 (.18)				
Suggest stores (SS)	.56 (.15)	.18 (.09)	.19 (.12)	.68 (.16)			
Child							
Suggest price range (SPR)	.68 (.13)	.53 (.16)					
Shop with parents (SP)	.55 (.13)	.18 (.12)	.70 (.14)				
Suggest stores (SS)	.52 (.14)	.22 (.10)	.33 (.11)	.73 (.15)			
Menezes & Elbert (1979)							
Likert							
Appearance (A)	.85 (.05)	.28 (.04)			1.00		
Products (P)	.85 (.05)	.08 (.03)	.27 (.04)		.76 (.03)	1.00	
Prices (Pr)	.90 (.05)	.07 (.02)	.04 (.02)	.18 (.03)	.46 (.06)	.49 (.06)	1.00

TABLE 1 *(Continued)*

Method–Trait	*Factor Loadings*	*Correlated Uniqueness*			*Factor Correlations*		
Semantic Differential							
Appearance (A)	.92 (.05)	.16 (.03)					
Products (P)	.84 (.05)	.05 (.02)	.30 (.04)				
Prices (Pr)	.88 (.05)	.00 (.02)	.06 (.02)	.20 (.03)			
Stapel							
Appearance (A)	.81 (.05)	.35 (.04)					
Products (P)	.84 (.05)	.08 (.03)	.30 (.04)				
Prices (Pr)	.88 (.05)	.00 (.02)	.03 (.02)	.24 (.03)			
Seymour & Lessne (1984)							
Likert							
Involvement (I)	.83 (.09)	.30 (.06)			1.00		
Power (P)	.95 (.08)	−.01 (.03)	.10 (.03)		.20 (.11)	1.00	
Interpersonal Need (IN)	.82 (.09)	.04 (.04)	−.01 (.03)	.32 (.07)	.73 (.06)	.61 (.08)	1.00
Mixed Scales							
Involvement (I)	.83 (.09)	.37 (.07)					
Power (P)	.96 (.08)	−.12 (.03)	.14 (.03)				
Interpersonal Need (IN)	.85 (.09)	.19 (.05)	−.12 (.03)	.30 (.07)			
Graphic Rating							
Involvement (I)	.95 (.08)	.12 (.06)					
Power (P)	.92 (.08)	.06 (.03)	.19 (.04)				
Interpersonal Need (IN)	.50 (.10)	−.07 (.05)	−.02 (.04)	.73 (.12)			

Note. Standard errors in parentheses.

seen that, in contrast to the findings reported by Bagozzi and Yi (1991) for the CFA model, no ill-defined solutions exist.

It is informative to examine the parameter estimates for each set of data. The trait factor loadings for the data of Arora (1982) are generally quite high, revealing considerable variation due to traits. These results indicate achievement of strong convergent validity. Trait correlations are nonsignificant to low, pointing to achievement of discriminant validity. Six of nine correlated uniquenesses are significant but are relatively low in magnitude, except for situational and response involvement measured by Likert scales. The differentially correlated uniquenesses—some significant, some nonsignificant—show that method effects are present, but that the assumption of unidimensional effects made by the CFA model is untenable.

In the analyses of the Foxman et al. (1989) data, the trait factor loadings are rather low for the father factor and mediocre for the mother and child factors, thus showing that variation due to traits is small. Six of nine uniqueness correlations are significant. Of these, the three associated with the father factor generally have large values. Again the assumption of unidimensional method factors does not hold. The correlations among traits are all lower than unity but quite high, suggesting achievement of weak discriminant validity.

For the data in Menezes and Elbert (1979), trait factor loadings are very high (ranging between .81 and .92), pointing to strong convergence among measures. Although six of nine correlated uniquenesses reach significance, they are generally low in magnitude (ranging from .04 to .08), suggesting rather weak method effects. Here, too, the assumption of unidimensional method factors is not met. Finally, traits are moderately to highly correlated.

The overall fit of the CU model was poor for the data in Seymour and Lessne (1984), suggesting that the CU model might be inappropriate. Thus, caution is in order for interpreting the related findings. Trait variance is generally high, except for interpersonal need measured with the graphic rating method where the variance is 25%. The correlation between involvement and power is low ($r = .20$), but the correlations are rather high between involvement and interpersonal need ($r = .73$), and between power and interpersonal need ($r = .61$). Only four of nine correlated uniquenesses are significant. However, among those obtained by the mixed scales, two are negative and significant, whereas the third is positive and significant. As there is no apparent theoretical reason to expect these conflicting findings, it is likely that the correlated uniquenesses constitute "wastebasket parameters" (Browne, 1984, p. 7) in that they enhance the overall fit of the model but do so at the expense of providing no substantive interpretation.

In sum, the CU model overcomes three drawbacks with the first-order CFA model: (a) the tendency to yield ill-defined solutions, (b) confounding of method variance with trait variance (when this is due to common trait variation across methods and traits are highly correlated), and (c) the false belief that variation can be partitioned into trait-free and method-free components.

Nevertheless, the CU model has limitations of its own. First, the model implicitly assumes that the effects of one method are uncorrelated with those of others.[2] The appropriateness of the CU model would thus depend on the plausibility of this assumption. The CU model would be most appropriate when maximally different methods are employed, which is desirable for assessing validity (Campbell & Fiske, 1959). When this assumption is not met, however, the CU model may be inappropriate, which should be indicated by the poor fit of the model.

Second, the error terms in the model still confound random error with measure specificity and make it difficult to distinguish trait variance from trait plus method variance (e.g., Bagozzi, Yi, & Phillips, 1991). The inclusion of separate correlated uniquenesses for each pair of measures from a common

[2] A CFA model with uncorrelated methods is a special case of the CU model in which the effects of each method are assumed to be unidimensional.

method overcomes the restrictive assumption that methods have unidimensional effects. But in so doing, parsimony is lost and the interpretation of each correlated uniqueness is made difficult. When a unidimensional method factor is found to hold, it is frequently reasonable to interpret its effects as systematic error due to the method. However, when some correlated uniquenesses are significant, others nonsignificant, or when some are positive and others negative, it may be difficult to explain the source of the differing patterns of influence. Despite these limitations, the CU model does provide information on convergent and discriminant validity, and thus can be useful in the analysis of construct validity where the classic Campbell and Fiske (1959) criteria and the first-order CFA model fail to apply.

THE DP MODEL

The DP model postulates that the effects of methods and traits are multiplicative, rather than additive. By reanalyzing four data sets from consumer research, Bagozzi and Yi (1991) found that application of the DP model gave adequate fit in two of four cases. However, their findings are inconclusive because their procedure has some limitations. The following critique and extension of the use of the DP model is conducted in this regard.

Two shortcomings of the application of the DP model by Bagozzi and Yi (1991) are the following. First, Bagozzi and Yi (1991) used LISREL (Jöreskog & Sörbom, 1989) to carry out their DP model analyses.[3] Although LISREL can be used in this regard, it is quite cumbersome to employ because it requires an extensive reparameterization of the DP model to accommodate multiplicative effects and other peculiarities of the model. For example, there are 27 latent variables alone for the model with three traits and three methods. Likewise, it is not possible to simultaneously estimate trait and method correlations or to impose equality constraints on parameter estimates under a LISREL specification. Second, and more important, Bagozzi and Yi (1991) limited their examination of the DP model to overall tests of significance and visual inspections of the disattenuated trait and method correlation matrices to ascertain convergent and discriminant validity. Tests of specific hypotheses regarding trait and method effects and more formal tests of construct validity are desirable.

In the analyses to follow, the MUTMUM program (Browne, 1991) was

[3] The DP model can also be parameterized and estimated by using the EQS program. In illustrating the estimation of the DP model via EQS, Bentler, Poon, and Lee (1988) showed that the LISREL results agreed with the EQS solutions. However, the EQS approach is not considered further because it suffers from the same problems facing the LISREL approach.

used to investigate the DP model for each of the data sets just noted.[4] The MUTMUM program is less cumbersome than LISREL, provides standard errors for both trait and method correlations (a particular 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 matrices. MUTMUM can be used to advantage to test important pattern hypotheses neglected by Bagozzi and Yi (1991), as developed next.

The MUTMUM program was applied to the four data sets just described, giving the following results for goodness-of-fit indices: $\chi^2(25) = 52.38$, $p < .005$ (Arora, 1982); $\chi^2(25) = 30.04$, $p \approx .22$ (Foxman et al., 1989); $\chi^2(25) = 29.80$, $p \approx .23$ (Menezes & Elbert, 1979); $\chi^2(25) = 79.97$, $p < .001$ (Seymour & Lessne, 1984). These differ somewhat from those reported by Bagozzi and Yi (1991), which were based on LISREL. The differences are apparently due to empirical underidentification problems in the LISREL analyses (Bagozzi & Yi, 1991, p. 435). No such problems arose in the MUTMUM analyses. This points to still another potential advantage of the use of MUTMUM, although little is known about the sensitivity of LISREL to empirical underidentification problems when the DP model is investigated.

Table 2 presents the parameter estimates for the DP model analyses. It is interesting to apply Browne's (1984) criteria for convergent and discriminant validity to these results. The criterion for convergent validity is satisfied when all method correlations are large. As shown in Table 2, this criterion holds for all data sets, except for Foxman et al. (1989) where two method correlations are rather small ($r_{M1M2} = .31$ and $r_{M1M3} = .18$) and therefore bring into question the achievement of convergent validity.

The first criterion for discriminant validity states that the correlations among traits should be lower than 1 in absolute terms. All four data sets meet this requirement. The second criterion for discriminant validity stipulates that every method correlation should be greater than all trait correlations. Table 2 reveals that this criterion is met for all data sets, except for Foxman et al. (1989) where, in fact, all methods are correlated at levels lower than all trait correlations. The first criterion for discriminant validity will be satisfied whenever the DP model satisfactorily fits the data. By this requirement, the data in Foxman et al. (1989) and Menezes and Elbert (1979) satisfy the criterion, whereas the data in Arora (1982) and Seymour and Lessne (1984) do not. In sum, it might be concluded that convergent and discriminant validity are not achieved for the data in Foxman et al. (1989), but are achieved for the data in

[4] The MUTMUM program is available from Michael W. Browne, Department of Psychology, Ohio State University, Columbus, OH 43210. The program has also been employed to carry out analyses in consumer research (e.g., Lastovicka et al., 1990).

TABLE 2
Findings for the Direct Product Model

				Trait Correlations					*Method Correlations*				
Study	*Measures*	*Communalities*	*Error*	T_1	*SE*	T_2	*SE*	T_3	M_1	*SE*	M_2	*SE*	M_3
Arora (1982)	(T_1M_1)	.99	.01	1					1				
	(T_2M_1)	1.00	.00	.33	.06	1			.68	.04	1		
	(T_3M_1)	.99	.01	.49	.06	.39	.06	1	.82	.03	.75	.03	1
	(T_1M_2)	1.00	.00										
	(T_2M_2)	1.00	.00										
	(T_3M_2)	1.00	.00										
	(T_1M_3)	.84	.16										
	(T_2M_3)	1.00	.00										
	(T_3M_3)	.89	.11										
Foxman et al. (1989)	(T_1M_1)	.77	.23	1					1				
	(T_2M_1)	.81	.19	.63	.10	1			.31	.09	1		
	(T_3M_1)	.79	.21	.63	.10	.79	.09	1	.18	.08	.53	.09	1
	(T_1M_2)	.71	.29										
	(T_2M_2)	.75	.25										
	(T_3M_2)	.73	.27										
	(T_1M_3)	.82	.18										
	(T_2M_3)	.85	.15										
	(T_3M_3)	.83	.17										

TABLE 2 *(Continued)*

				Trait Correlations					*Method Correlations*				
Study	*Measures*	*Communalities*	*Error*	T_1	*SE*	T_2	*SE*	T_3	M_1	*SE*	M_2	*SE*	M_3
Menezes & Elbert (1979)	(T_1M_1)	.92	.08	1					1				
	(T_2M_1)	.91	.09	.77	.03	1			.90	.02	1		
	(T_3M_1)	.94	.06	.44	.05	.48	.05	1	.89	.02	.90	.02	1
	(T_1M_2)	.92	.08										
	(T_2M_2)	.90	.10										
	(T_3M_2)	.94	.06										
	(T_1M_3)	.88	.12										
	(T_2M_3)	.86	.14										
	(T_3M_3)	.92	.08										
Seymour & Lessne (1984)	(T_1M_1)	.96	.04	1					1				
	(T_2M_1)	.96	.04	−.07	.09	1			.83	.03	1		
	(T_3M_1)	.80	.20	.78	.05	.30	.09	1	1	.00	.95	.03	1
	(T_1M_2)	.98	.02										
	(T_2M_2)	.98	.02										
	(T_3M_2)	.89	.11										
	(T_1M_3)	.87	.13										
	(T_2M_3)	.86	.14										
	(T_3M_3)	.54	.46										

Note. T_iM_j refers to the appropriate trait–method pairing.

Menezes and Elbert (1979) on the basis of the DP model results. No conclusions can be drawn for construct validity for the other data sets on the basis of the DP model findings.

Note that construct validity was assessed based on the visual inspection of estimates. It is desirable to examine specific hypotheses concerning construct validity, reliability, trait effects, and method effects more formally (Bagozzi & Yi, 1992). Table 3 presents the results for certain hypotheses of interest, where only the data in Menezes and Elbert (1979) are investigated for illustrative purposes. The first set of comparisons in the table focuses on the equivalence of traits and tests whether each trait correlation is lower than 1 in an absolute sense. This provides a formal test of the first discriminant validity criterion. A comparison of the model hypothesizing that all traits are perfectly correlated (i.e., $\rho_{t21} = \rho_{t32} = \rho_{t31} = 1$) to the baseline DP model shows that one must reject this hypothesis—$\chi_d^2(5) = 625.98, p < .001$. Indeed as shown in Table 2, each trait correlation is significantly lower than 1.

The bottom of Table 3 shows the results for tests of whether two or more methods are equivalent. These tests are not related to construct validity per se, but are useful for discovering redundancy in methods. For example, a researcher in the early stages of a program of research might wish to discover which methods from a set used in a pretest are distinct in order to avoid unnecessary duplication in a subsequent study. Comparison of the hypothesis that all methods are perfectly correlated (i.e., $\rho_{M21} = \rho_{M32} = \rho_{M31} = 1$) to the baseline DP model shows that the hypothesis of equivalent methods should be

TABLE 3
Tests of Hypotheses of the Direct Product Model for the Data of Menezes & Elbert (1979)

Model	*χ^2 Goodness-of-Fit*	*χ^2 Difference Test*
Baseline	$\chi^2(25) = 29.80, p \approx .23$	
All traits equivalent		
$\rho_{t21} = \rho_{t32} = \rho_{t31} = 1$	$\chi^2(30) = 655.78, p < .001$	$\chi_d^2(5) = 625.98, p < .001$
Traits 1 and 2 equivalent	$\chi^2(28) = 182.93, p < .001$	$\chi_d^2(3) = 153.13, p < .001$
$\rho_{t21} = 1$		
Traits 2 and 3 equivalent	$\chi^2(28) = 498.77, p < .001$	$\chi_d^2(3) = 468.97, p < .001$
$\rho_{t32} = 1$		
Traits 1 and 3 equivalent	$\chi^2(28) = 539.83, p < .001$	$\chi_d^2(3) = 510.03, p < .001$
$\rho_{t31} = 1$		
All methods equivalent		
$\rho_{m21} = \rho_{m32} = \rho_{m31} = 1$	$\rho^2(30) = 86.31, p < .001$	$\chi_d^2(5) = 56.51, p < .001$
Methods 1 and 2 equivalent		
$\rho_{m21} = 1$	$\chi^2(28) = 61.90, p < .001$	$\chi_d^2(3) = 32.10, p < .001$
Methods 2 and 3 equivalent	$\chi^2(28) = 58.48, p < .001$	$\chi_d^2(3) = 28.68, p < .001$
$\rho_{m32} = 1$		
Methods 1 and 3 equivalent	$\chi^2(28) = 64.45, p < .001$	$\chi_d^2(3) = 34.65, p < .001$
$\rho_{m31} = 1$		

rejected—$\chi_d^2(5) = 56.51, p < .001$. The remaining comparisons reveal that, in fact, each pair of methods is significantly correlated at a level lower than 1.

A number of other hypotheses might also be examined (Bagozzi & Yi, 1992). 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 trait correlations constrained to be zero to the baseline DP model. Formal comparisons could be made as well between trait and method 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 just outlined 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 done by comparing the baseline model to a model fixing the diagonal matrix of errors corresponding to methods to unity.

It can be noted that the DP model did not fit the Arora (1982) data, whereas the CU model did fit. This result suggests that the effects of methods and traits might be additive, as opposed to multiplicative, for the Arora data. However, the goodness of fit cannot be used as the sole criterion in selecting a model. For example, the data in Foxman et al. (1989) and Menezes and Elbert (1979) fit both models well. These results indicate that alternative models can fit the same MTMM data. In such cases, it is meaningful to ask if any theoretical or methodological reasons can be brought to bear for deciding between the two models for the two data sets at hand.

In fact, there is reason to believe that traits and methods interact according to the DP model structure for the data in Foxman et al. (1989) and Menezes and Elbert (1979). An intuitive description of the DP model proceeds as follows. The DP model posits a functional interaction between the true level of trait correlation and the magnitude of method effects. Traits and methods interact in the sense that sharing a method exaggerates the correlations between highly correlated traits relative to traits that are less correlated. That is, not all relationships are equally exaggerated by sharing a method, but relationships that are large enough to get noticed are more likely to be inflated (Campbell & O'Connell, 1967).

In Foxman et al. (1989), three members of each family in the sample (father, mother, child) provided estimates of the child's purchasing influence in three areas: suggesting a price range, going shopping with parents when looking for a product for family use, and suggesting stores. Each member might have an implicit theory (expectations) about the relationships of certain traits, which will lead to a member-specific (method) bias. In this case, the stronger the true associations between traits are, the more likely they are to be noticed and exaggerated, thus producing an interaction between traits and methods.

Likewise in Menezes and Elbert (1979), business students were asked to use similar scales (Likert, semantic differential, and Stapel scales) to rate retail store image on appearance, products, and price. As these three image characteristics naturally tend to covary in the market place and therefore lead to an expectation of highly correlated traits, an assumption likely to be held by respondents as well, it is anticipated that the stronger the true correlations among traits, the greater the exaggeration by respondents. This, too, is consistent with a multiplicative trait–method pattern. In sum, the data in Foxman et al. (1989) and Menezes and Elbert (1979) are more consistent with a multiplicative interpretation of trait and method interactions than additive effects.

CONSTRUCT VALIDATION BY USE OF PANEL MODELS

Two drawbacks with the CFA and CU models are the following. First, the smallest combinations of traits and methods needed to implement these models are either: (a) three traits and three methods, (b) four traits and two methods, or (c) two traits and four methods. It is possible to use fewer traits and/or methods, but this requires the imposition of restrictions on the parameter space which may not be valid or reasonable in typical data collection contexts. Hence, the CFA and CU models can be difficult to implement in practice. Second, in both models, random error is confounded with specific error (Bagozzi et al., 1991). The CU model shares this drawback as well. This can result in biased estimates of reliability and convergent validity and underestimate or obscure method effects.

An alternative way to examine construct validity in certain situations is by use of panel models. To show this, Figure 2 represents a two-period model for two constructs, η_1 and η_2, measured by three and two indicators, respectively. This model provides the following information. The temporal stabilities of η_1 and η_2 are shown by β_{31} and β_{42}, respectively, which reflect corrections for random and specific errors, as developed next. Discriminant validity between measures of η_1 and η_2 can be assessed by inspection of ψ_{21} and ψ_{43}. The former is the correlation between the two latent variables measured at Time 1; the latter is the partial covariation between the two latent variables measured at

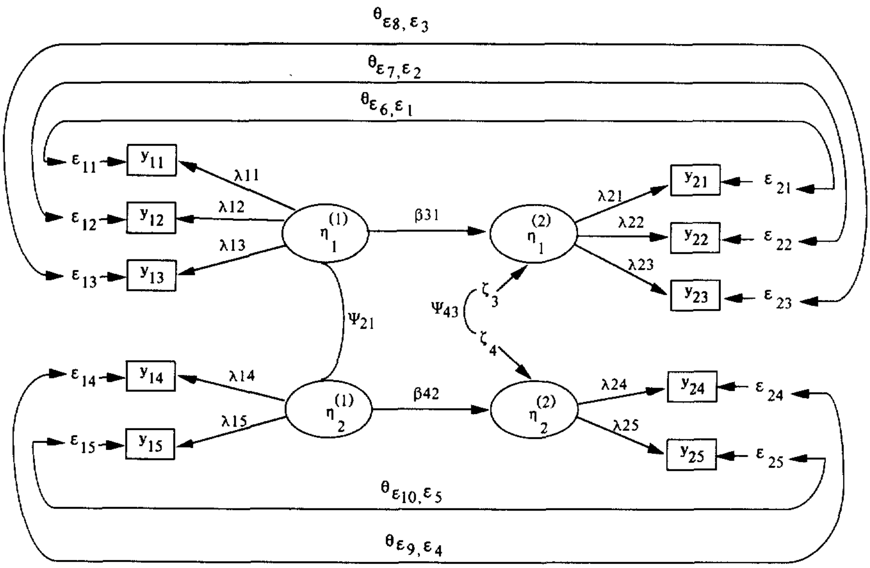

FIGURE 2 Structural equation models for examining reliability, construct validity, and stability of measures of two constructs.

Time 2 (i.e., the covariation after partialling out variance due to dependence on the same measures at time 1).[5] The estimates for ψ_{21} and ψ_{43} also reflect corrections for random and specific errors.

The achievement of convergent validity is implied when the overall fit of the model shown in Figure 2 holds and the factor loadings are high and significant. To show how the model provides separate estimates for random and specific errors, one can write the measurement equations for the first measure of η_1 at both points in time as follows:

$$y_{11} = \lambda_{11}\,\eta_1^{(1)} + \varepsilon_{11}$$

$$y_{21} = \lambda_{21}\,\eta_1^{(2)} + \varepsilon_{21},$$

[5] It is possible to modify the panel model shown in Figure 2 to include the cross-lagged influences of η_1 at Time 1 on η_2 at Time 2 and η_2 at Time 1 on η_1 at Time 2. Because this is straightforward and does not alter the general conclusions developed, nothing more is said about the cross-lagged path effects. The researcher should check for such possibilities in applications of such panel models as discussed herein. For the data in Table 5, the goodness-of-fit for the panel model of Figure 2, with the additional specification of β_{32} and β_{41} free, gave $\chi^2(27) = 20.40$. Compared to the baseline model, this yields $\chi^2_d(2) = 2.50, p < .25$ and thus the hypothesis that $\beta_{32} = \beta_{41} = 0$ can not be rejected. The estimates for the cross-lagged paths were $\beta_{32} = -.17(.20)$ and $\beta_{41} = .27\ (.19)$.

where y_{11} and y_{21} are the observed scores obtained by the same measure at Times 1 and 2, respectively. Rewriting these equations to express separate effects for random error and specific variance on the measures gives

$$y_{11} = \lambda_{11}\,\eta_1^{(1)} + s_{p1} + e_{11}$$

$$y_{21} = \lambda_{21}\,\eta_1^{(2)} + s_{p1} + e_{21},$$

where $\mathrm{s}_{\mathrm{p}1}$ is the specific error and the e_{11} and e_{21} are the random error components of y_{11} and y_{21}, respectively. Because the same measure (item) is used repeatedly across time, each measure might have systematic influence on the observed scores. This might occur because a measure has a specific meaning other than the underlying trait or there are memory effects. In such cases, it is necessary to postulate measure-specific method factors by allowing correlated errors for the same measures (Sörbom, 1975). But Figure 2 implies that Var (s_{p1}) = Cov $(\varepsilon_{11}, \varepsilon_{21})$ = $\theta\varepsilon_6\varepsilon_1$. That is, specific variance can be obtained by serially correlated errors.

It should be pointed out that the reliability of each measure of η_1 can be estimated as 1 − Var (e_{ij}), and the composite reliabilities can be computed as functions of these. Unlike the reliabilities of measures under the CFA, CU, and DP models, which are based on ε_{ij}, the reliabilities for the panel model in Figure 2 are true reliabilities. Reliabilities computed under the CFA, CU and DP models will generally underestimate the true reliabilities because the error terms of measures will contain both random and specific components.

As an illustration, the panel model shown in Figure 2 was applied to the data from a panel study of consumer attitudes toward using coupons (see Bagozzi, Baumgartner, & Yi, 1992). Female staff members at a major university participated in the study; the sample size was 151. Three measures of attitudes and two of subjective norms were selected from two consecutive points in time (i.e., 1 week). The attitude toward using coupons for shopping in the supermarket was assessed with three semantic differential scales: pleasant/unpleasant, good/bad, and favorable/unfavorable. Subjective norms were measured with two items: "Most people who are important to me think I definitely should/definitely should not use coupons for shopping in the supermarket during the coming week," and "Most people who are important to me probably consider my use of coupons to be wise/foolish."[6]

Before presenting the results for the panel model, it is informative to examine the measurement models separately for Times 1 and 2. Figure 3 shows the contemporaneous measurement model, and the findings are contained in Table 4. The chi-square values demonstrate satisfactory fits at Time 1, $\chi^2(4) = 4.68$, $p \approx .32$, and Time 2, $\chi^2(4) = 2.63$, $p \approx .62$. These results imply that the measures of attitudes and subjective norms converge on separate, unidimen-

[6] Further information on the panel data can be obtained from the authors on request.

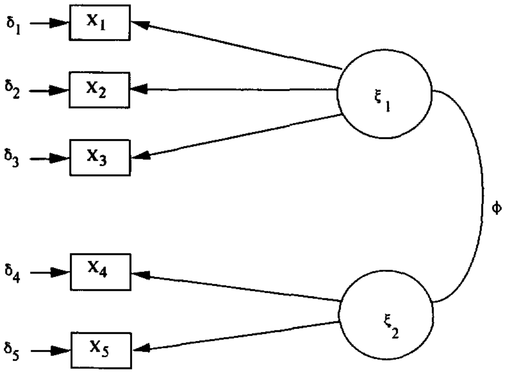

FIGURE 3 Contemporaneous measurement model for two constructs.

sional factors. Indeed, the factor loadings are high and significant in each case. To test whether the correlation between factors is significantly lower than 1, the fit of the model of Figure 3 must be compared to the fit of the model with phi fixed at unity. At Time 1, $\chi^2_d(1) = 9.95, p < .001$; at Time 2, $\chi^2_d(1) = 29.56, p < .001$. Thus, the hypothesis can be rejected that the factors are perfectly correlated. However, the factor correlations are rather large ($\phi = .86$, $SE = .05$, at Time 1; $\phi = .75$, $SE = .06$, at Time 2). Error variances are low to moderately high. The composite reliabilities are high for both attitudes and subjective norms at both points in time. All reliabilities based on the contemporaneous model, however, are likely to be lower than the true reliabilities because the error terms are likely to contain specific variance (cf. Anderson, 1985). Comparisons are made with reliabilities computed from the panel models where specific variance is estimated and used to assess reliability.

The panel model of Figure 2 was applied to the data. The model fits well: $\chi^2(26) = 22.16, p \approx .68$. Table 5 presents the parameter estimates where it can be seen that factor loadings are high and significant. These results indicate achievement of convergent validity. In addition, error variances are low to moderately high. Also shown in the table are estimates of specific variance that are unique and different from the variance explained by traits. Although low in value, all specific variances are significant.

Given that the panel model in Figure 2 is satisfactory, it is desirable to test whether the factor loadings are equal across time. Equal factor loadings imply that the reliabilities of measures are identical across time. If the null hypothesis of equal across-time factor loadings is rejected, the measures are not equally reliable across time and the estimated stability of the constructs (as reflected by β^{31} and β^{42} in Figure 2) would be affected by both true stability of constructs and changes in reliability. If the null hypothesis cannot be rejected, the stability of the measures can be interpreted in an unambiguous way. The test, which must be performed on the covariance matrices (Cudek, 1989), is conducted as follows. Equality constraints are first imposed on the factor loadings for each

TABLE 4
Findings for Contemporaneous Measurement Models (See Figure 3)

	Goodness-of-Fit			*Factor Loadings*				*Factor Correlation*		*Error Variances*				*Reliability*			
														Individual		*Composite*	
	χ^2	*df*	*p*	*Aact*	*SE*	*SN*	*SE*	ϕ	*SE*	*Aact*	*SE*	*SN*	*SE*	*Aact*	*SN*	*Aact*	*SN*
Time 1	4.68	4	.32	.75	.07			.86	.05	.43	.06			.56		.81	.78
				.82	.07					.33	.06			.67			
				.73	.08					.47	.07			.53			
						.69	.08					.52	.07		.48		
						.90	.07					.19	.07		.81		
Time 2	2.63	4	.62	.86	.07			.75	.06	.27	.04			.74		.90	.78
				.87	.07					.24	.04			.76			
				.86	.07					.26	.04			.74			
						.75	.08					.44	.08		.56		
						.85	.08					.27	.08		.72		

Note. Aact = attitude; SN = subjective norm.

TABLE 5
Estimates for Measurement Parameters in Panel Model (See Figure 2)

	Factor Loadings													
	Time 1				*Time 2*						*Error Variances*			
Measures	*Aact*	*SE*	*SN*	*SE*	*Aact*	*SE*	*SN*	*SE*	*Specific Variance*	*SE*	*Time 1*	*SE*	*Time 2*	*SE*
Attitude														
Aact 1	.77	.07			.86	.07			.09	.04	.45	.06	.26	.04
Aact 2	.78	.07			.88	.07			.12	.04	.34	.05	.23	.04
Aact 3	.75	.07			.85	.07			.11	.04	.45	.06	.27	.04
Subjective norm														
SN1			.70	.12			.71	.12	.17	.05	.53	.07	.49	.07
SN2			.88	.12			.90	.12	.10	.05	.22	.06	.19	.07

respective pair of measures, one pair at a time, to arrive at chi-square values for the models hypothesizing invariance of measurement instruments over time. Next, each of the goodness-of-fit values so obtained is compared to the chi-square value for the model with factor loadings unconstrained. The differences in chi-square values with one degree of freedom provide tests of the equality of factor loadings over time.

The chi-square difference tests indicate that the factor loadings for measures of attitudes and for measures of subjective norms are equal over time. For example, the restricted model with the invariance constraint for the first measure of attitudes (i.e., $\lambda_{11} = \lambda_{21}$) gives the following fit: $\chi^2(27) = 22.80$, $p \approx .70$, and the chi-square differences is not significant, $\chi^2_d(1) = .64$, $p > .45$. In fact, all the chi-square differences are not significant (all ps $> .45$), suggesting that the measures are equally reliable across time. Therefore, estimates of stability and tests of hypotheses can be interpreted in an unambiguous way.

Before presenting the results for discriminant validity, residual covariances, and stability, it is interesting to compare estimates of reliability between contemporaneous and panel models. For example, the reliability of Subjective Norm 1 at Time 1 is .48 for contemporaneous models and .64 for panel models. It is found that the reliabilities generally increased quite a bit from contemporaneous models to panel models; the average reliability estimate is .66 for contemporaneous models and .78 for panel models. This is a consequence of the estimates in the panel model, taking into account measure specificity.

Now that a satisfactory goodness-of-fit of the panel model has been established and convergent validity and factorial invariance demonstrated, it is meaningful to examine discriminant validity between the measures of attitudes and subjective norms (Figure 2). The standardized correlations among factors are very high: $\psi_{21}^{\text{std}} = .90$ and $\psi_{43}^{\text{std}} = .70$. The residual covariance between attitudes ($\eta_1^{(2)}$) and subjective norms ($\eta_2^{(2)}$) at Time 2 is very small and in fact is nonsignificant: .05(.04). This suggests that for the data at hand no omitted variables exist, and the standardized residual variances can be interpreted as the amount of variance unexplained in attitudes and subjective norms. If the residual covariance had been significant, the explained variance would have to be interpreted as pseudo R^2 values. The explained variances in attitudes and subjective norms are $R^2 = .66$ and $R^2 = .83$, respectively.

Finally, it is interesting to examine the stability of attitudes and subjective norms. The stability coefficients are quite high: $\beta_{31}^{\text{std}} = .81$ and $\beta_{42}^{\text{std}} = .91$. It should be pointed out that the stability coefficients have been corrected for random and specific errors. When measure specificity is not taken into account in panel models, stability coefficients will be inflated by an amount proportional to the stability of the specific components in the measures. For comparison, the stability coefficients were estimated for the panel model of Figure 2 where no provisions were made for measure specificity. This gives $\beta_{31}^{std} = .85$ and $\beta_{42}^{std} = .97$. Thus, a failure to model measure specificity does indeed result in inflated stability coefficients. For the particular data at hand, the inflation is relatively minor and does not affect substantive conclusions. However, in data sets with higher levels of measure specificity, failure to account for such systematic biases can affect substantive conclusions. A hypothesis of interest in some studies might be a formal test of the equality of the stability coefficients between constructs (i.e., $\beta_{31} = \beta_{42}$). The goodness-of-fit for the model with this constraint was, $\chi^2(30) = 23.01, p \approx .82$. Comparing this to the baseline model gives, $\chi_d^2(1) = .11, p > .74$. Therefore the hypothesis of equal stability cannot be rejected.

DISCUSSION

It is premature to give rigid guidelines for conducting construct validation research. The conceptual criteria for construct validity are complex and in need of further development. The procedures used for analyzing MTMM matrices are in a state of flux and require better integration. Nevertheless, to the extent that convergent and discriminant validity are useful operationalizations of important aspects of construct validity, it is crucial to evaluate the procedures currently used in consumer research.

This article began with the premise that the procedures introduced by Bagozzi and Yi (1991) have several shortcomings. The first-order CFA model is often overparameterized and typically produces ill-defined solutions (Marsh & Bailey, 1991; Wothke, 1987). However, unlike the DP model, it is one of the few procedures giving unique estimates of trait, method, and error components. Therefore, if a researcher wants to test an underlying model additive in trait, method, and error components, and to obtain parameter estimates showing the relative contributions, the CFA model is a viable alternative.

If ill-defined solutions result for the first-order CFA model, a likely occurrence, and if one is confident that an additive structure characterizes the underlying model, one can try the CU model in which method effects are represented as correlated uniquenesses. If the CU model is found acceptable, this might suggest that the correlated uniquenesses associated with each method cannot be explained in terms of a single method factor.

One important advantage of the CU model is that it seldom leads to ill-defined solutions that plague MTMM studies. In the analyses of the four data sets reported by Bagozzi and Yi (1991), all CFA results included some ill-defined solutions. However, when the CU model was used to analyze the same data, no ill-defined solutions occurred (see also Marsh & Bailey, 1991). This result may happen because the CU model does not make the restrictive assumption of congeneric-like method effects. In addition, the CU model provides a more accurate interpretation of trait and method effects by eliminating the possibility that the so-called method factors actually reflect a general trait effect instead of, or in addition to, method effects (Marsh, 1989).[7] It also gives information on convergent and discriminant validity. On the other hand, the major drawbacks of the CU model are that methods are assumed to be uncorrelated, that the interpretation of correlated uniquenesses can be ambiguous, that contributions of specific method effects are not estimated, and that, like the CFA model, random and specific errors are confounded.

If the CU model yields ill-defined solutions, or if the researcher has reason to believe that the data reflect multiplicative trait and method components with additive uniquenesses, the DP model can be applied. If one desires a global assessment of convergent and discriminant validity according to the logic implied by Campbell and Fiske's (1959) original criteria, the DP model is the only currently available procedure for doing this.

This article developed several useful hypotheses for the DP model not considered by Bagozzi and Yi (1991), and illustrated formal procedures for testing these hypotheses. For example, one can test whether two particular trait factors are equivalent or orthogonal. Although Bagozzi and Yi (1991) used omnibus tests for examining construct validity, they failed to offer tests of specific hypotheses regarding construct validity as well as individual traits and methods. This study also used a program more flexible, accurate, and easier to use than that employed by Bagozzi and Yi (1991). The program provides estimates and their standard errors for both method and trait correlations, allows constraints on the parameter space to avoid ill-defined solutions, and reduces empirical underidentification problems. Nevertheless, the DP model is not without limitations. For example, validity as represented in the method correlation matrix is difficult to interpret (especially as the correlations approach unity).

Other limitations can be mentioned with the CFA, CU, and DP models. First, these models cannot be implemented when the number of traits and methods is too low. For example, when two traits are investigated with two methods, neither the CFA nor the CU model can be used for construct valida-

[7] An example can be found in the MTMM data from Van Tuinen and Ramanaiah (1979). An application of the CFA model showed very highly correlated methods (98, .92, and .85), whereas the CU model showed that methods were distinct and that method effects were negligible. Thus, this example illustrates that an application of the CFA model could misleadingly yield highly correlated methods as a result of the convergence of a general trait factor across methods. In such cases, the CFA model often underestimates the correlations among traits. For the data in Van Tuinen and Ramanaiah (1979), the CFA model yielded correlations among traits ranging from − .32 to .26, which was contrary to expectations. The CU model, in contrast, revealed more theoretically consistent correlations ranging from .19 to .81. For further results and discussion on this issue, see Bagozzi (1993).

tion, thus making their use rather limited in practice. Second, random error is confounded with specific error in the CFA, CU, and DP models. This problem is potentially serious because it may result in biased estimates of reliability, convergent validity, and method effects.

In such circumstances, panel models may be used to examine construct validity. Panel models can simultaneously assess the temporal stability and reliability of constructs, while correcting for the separate influences of random and specific errors. This article has illustrated that when measure specificity is not taken into account, factor loadings and reliability coefficients are underestimated. Failure to model measure specificity results in overestimated stability coefficients. To the extent that measure specificity is substantial or the number of available traits and methods is small, panel models can be useful for making valid conclusions about construct validity and reliability.

A drawback with panel models is that they do not explicitly model method variance. However, one could model various types of method variance by modifying the model. For example, it is possible to hypothesize that all the responses are affected by sharing a common method of measurement (e.g., self-report). By introducing one method factor common to all measures, one can write the measurement equations for the first measure of η_1 as follows:

$$y_{11} = \lambda_{11}\,\eta_1^{(1)} + \lambda_{13}\,\eta_M + \varepsilon_{11}$$

$$y_{21} = \lambda_{21}\,\eta_1^{(2)} + \lambda_{23}\eta_M + \varepsilon_{21},$$

where η_M is the method factor shared by all measures.

Alternatively, it is possible to posit effects of multiple methods for each trait over time. For instance, each item may have its own systematic influence on the individual responses, as occurs when an item has a specific meaning other than the construct of interest. In such cases, it is necessary to postulate a method factor for each type of item (i.e., all traits measured with the same item at different waves are hypothesized to share method variance). It is thus possible to decompose an observed score into three components: (a) the latent variable common to all items (common factor), (b) the item-specific method factor (specific factor), and (c) the random measurement error (Raffalovich & Bohrnstedt, 1987).

$$y_{11} = \lambda_{11}\,\eta_1^{(1)} + \lambda_{13}\eta_{M1} + \varepsilon_{11}$$

$$y_{21} = \lambda_{21}\,\eta_1^{(2)} + \lambda_{23}\eta_{M1} + \varepsilon_{21},$$

where multiple method factors (e.g., η_{M1}, η_{M2}, η_{M3}) are allowed.

In estimating panel models with multiple method factors, one should consider identification of the parameters. In general, for all parameters to be identified one must have three or more indicators at three or more waves (Raffalovich & Bohrnstedt, 1987). When the number of indicators (I) or the number of periods (P) is smaller than 3, some constraints should be imposed to achieve identification. For example, if one has three or more measures at only two time periods, one must impose I independent constraints on the factor loadings, error variances, or factor correlations.

In sum, there may not be a single best model under all conditions for analyzing MTMM data. Rather, one should examine the appropriateness of alternative models in each situation. One can begin with the CFA model because of its desirable features and parsimonious structure. When the model yields ill-defined solutions, or when method factors are unlikely to be unidimensional, one can apply the CU model. If the CU model still yields ill-defined solutions, or if method effects are likely to be multiplicative, one can try the DP model and examine several hypotheses proposed in this article concerning construct validity. When the number of traits and methods is small and/or the uniqueness of each scale (or measure) is substantial, however, the panel model should be useful.

Although the four models are distinct in terms of functional forms, selection of a particular model may not be so clear. The fits of some models may be empirically indistinguishable. For example, the DP model is sometimes indistinguishable in fit from the CFA model (Bagozzi & Yi, 1991; Kumar & Dillon, 1992).[8] Furthermore, selection of a model should not be based solely on empirical considerations. One should also bring in theoretical and methodological considerations when deciding among the alternative approaches. In this regard, this article has presented the underlying assumptions, as well as strengths and weaknesses, of each model and examined how these models can be used in construct validation. A researcher should explicitly consider the nature of the assumptions underlying each model, examine the plausibility of these assumptions, select the most appropriate one in the given situation, and describe these assumptions clearly to the reader.

ACKNOWLEDGMENTS

We thank Dipankar Chakravarti and two anonymous reviewers for their helpful comments on an earlier version of this article.

[8] Kumar and Dillon (1992) noted that both the DP model and the covariance component analysis (CCA) model (Wothke, 1987) can also fit the same data. The CCA model can be used to analyze MTMM data in special instances and is discussed further in Kumar and Dillon (1992), where problems with its specification are considered. Because the CCA model has been thoroughly evaluated by Kumar and Dillon (1992), and at the same time is less viable than the other procedures, nothing more is said about it in this article.

REFERENCES

Anderson, J. C. (1985). A measurement model to assess measure-specific factors in multiple-informant research. *Journal of Marketing Research, 22,* 86–92.

Arora, R. (1982). Validation of an S-O-R model for situation, enduring, and response components of involvement. *Journal of Marketing Research, 19,* 505–516.

Bagozzi, R. P. (1993). Assessing construct validity in personality research: Applications to measures of self-esteem. *Journal of Research in Personality, 27,* 49–87.

Bagozzi, R. P., Baumgartner, H., & Yi, Y. (1992). State versus action orientation and the theory of reasoned action: An application to coupon usage. *Journal of Consumer Research, 18,* 505–518.

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

Bagozzi, R. P., & Yi, Y. (1992). Testing hypotheses about methods, traits, and communalities in the direct product model. *Applied Psychological Measurement, 16,* 373–380.

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.* Los Angeles: BMDP Statistical Software.

Bentler, P. M., Poon, W. Y., & Lee, S. Y. (1988). Generalized multimode latent variable models: Implementation by standard programs. *Computational Statistics and Data Analysis, 6,* 107–118.

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

Browne, M. W. (1991). *MUTMUM PC user's guide.* Columbus: Ohio State University.

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

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

Cote, J. A., & Buckley, M. R. (1987). Estimating trait, method, and error variance: Generalizing across 70 construct validation studies. *Journal of Marketing Research, 24,* 315–318.

Cudeck, R. (1989). Analysis of correlation matrices using covariance structure models. *Psychological Bulletin, 105,* 317–327.

Foxman, E. R., Tansuhaj, P. S., & Ekstrom, K. M. (1989). Family members' perceptions of adolescents' influence in family decision making. *Journal of Consumer Research, 15,* 482–491.

Jöreskog, K. G. (1981). Analysis of covariance structures. *Scandinavian Journal of Statistics, 8,* 65–92.

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

Kumar, A., & Dillon, W. R. (1990). On the use of confirmatory measurement models in the analysis of multiple-informant reports. *Journal of Marketing Research, 28,* 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.

Lastovicka, J. L., Murry, J. P., Jr., & Joachimsthaler, E. (1990). Evaluating the measurement validity of lifestyle typologies with qualitative measures and multiplicative factoring. *Journal of Marketing Research, 28,* 11–23.

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 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.

Menezes, D., & Elbert, N. F. (1979). Alternative semantic scaling formats for measuring store image. *Journal of Marketing Research, 16,* 80–87.

Peter, J. P. (1981). Construct validity: A review of basic issues and marketing practices. *Journal of Marketing Research, 18,* 133–145.

Raffalovich, L. E., & Bohrnstedt, G. W. (1987). Common, specific, and error variance components of factor models: Estimation with longitudinal data. *Sociological Methods & Research, 15,* 385–405.

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

Seymour, D., & Lessne, G. (1984). Spousal conflict arousal: Scale development. *Journal of Consumer Research, 11,* 810–821.

Sörbom, D. (1975). Detection of correlated errors in longitudinal data. *British Journal of Mathematical and Statistical Psychology, 28,* 138–151.

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

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

Wothke, W. (1984). *The estimation of trait and method components in multitrait–multimethod measurement.* Unpublished doctoral dissertation, University of Chicago.

Wothke, W. (1987, April). *Multivariate linear models of the multitrait–multimethod matrix.* Paper presented at the American Educational Research Association Annual Meetings, Washington, DC.

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.

Accepted by Dipankar Chakravarti.