ELSEVIER

Journal of Operations Management 16 (1998) 387–405

JOURNAL OF OPERATIONS MANAGEMENT

The empirical assessment of construct validity

Scott W. O'Leary-Kelly [a,b,*], Robert J. Vokurka [a,b]

[a] *Computer Information Systems and Quantitative Analysis, University of Arkansas, Fayetteville, AR 72701, USA*
[b] *Dept. of Engineering Technology and Industrial Distribution, Texas A&M University, College Station, TX 77843, USA*

Abstract

This paper provides an in-depth review of the different methods available for assessing the construct validity of measures used in empirical research. Construct validity pertains to the degree to which the measure of a construct sufficiently measures the intended concept (e.g., is free of measurement error) and has been shown to be a necessary component of the research process. In order to illustrate the steps required to establish construct validity, we drew upon empirical research in the operations management area of manufacturing flexibility.

Keywords: Construct validity; Research methodology; Review of empirical assessment

1. Introduction

Traditional operations management (OM) research is often directed at solving a highly specific problem and has historically relied heavily on mathematical theory and modeling techniques. This emphasis has contributed to the general criticism that there is a serious gap between research that is academic in its orientation and research that is considered to be beneficial by industry practitioners (Buffa, 1981; Hax, 1981; Miller and Graham, 1981). Empirical research that is field based and makes use of data drawn from industry practice can help bridge this gap between academics and industry.

In the past decade, there has been a dramatic increase in this type of field based empirical research covering a broad array of OM topics. However, a substantial portion of this research is lacking a strong conceptual and methodological base (Flynn et al., 1990). In particular, the methodological issue of construct validity is generally ignored. In comparison, construct validity has been routinely assessed in fields like marketing and organizational behavior. Construct validity has been defined by Schwab (1980) as ''representing the correspondence between a construct (conceptual definition of a variable) and the operational procedure to measure or manipulate that construct'' (p. 5). Construct validity is an important part of construct validation, a multistep process for assessing the adequacy of measures (Schwab, 1980). As noted by Schwab (1980), ''construct validation is a necessary and major element in the research process'' (p. 33).

As with other disciplines, empirical research in OM is about examining the relationships between relevant variables. The ability to correctly identify significant relationships among variables depends on our ability to adequately measure the variables. Studies in which the measures are flawed (e.g., by excessive amounts of random error, or measurement-specific error, or method bias) can lead to erroneous conclusions. For instance, all measures (objective

* Corresponding author.

0272-6963/98/$19.00
PII S0272-6963(98)00020-5

Table 1
Construct validity assessment: manufacturing flexibility studies

Study	Unidimensional assessment	Reliability assessment	Convergent validity	Discriminant validity	Flexibility measurement	
					Type	Item scale
Schroeder et al. (1986)					Perceptual	MI
Swamidass and Newell (1987)		Cronbach's			Perceptual	MI
De Meyer et al. (1989)					Perceptual	SI
Fiegenbaum and Karnani (1991)					Objective	SI
Hörte et al. (1991)					Perceptual	SI
Dixon (1992)	EFA				Objective	MI
Gupta and Somers (1992)	EFA	Cronbach's	MTMM	MTMM	Perceptual	MI
Tunälv (1992)					Perceptual	SI
Das et al. (1993)					Objective	SI
Parthasarthy and Sethi (1993)		Cronbach's			Objective	SI
Ettlie and Penner-Hahn (1994)					Objective	SI
Upton (1995)					Objective	SI
Ward et al. (1995)	EFA	Cronbach's			Perceptual	MI
Suarez et al. (1996)	EFA				Objective	MI, SI

SI—Single Item measures.
MI—Multiple Item measures.

and perceptual) contain some degree of random error. Measures that contain excessive random error can attenuate the statistical results and lead to false acceptance of a null hypothesis (Nunnally, 1978). Construct validity involves the assessment of the degree to which a measure correctly measures its targeted variable. Studies that utilize empirical measures yet fail to adequately assess the construct validity of the measures are open to criticism.

Given the importance of construct validity to the research process, it is critical that we have a clear understanding of the methods used in its assessment. The purpose of this paper is to provide an in-depth review of the different methods of empirically assessing construct validity and to compare traditional methods with more recently developed methodologies. The aim of this paper is to provide researchers with critical information necessary to make informed decisions regarding the most appropriate methods for assessing construct validity.

To provide context to our discussion of construct validation, we drew upon empirical research in the OM area of manufacturing flexibility. This body of research is discussed only as an example to clarify the arguments presented about construct validity (i.e., this is *not* intended to be a review of manufacturing flexibility). We chose the manufacturing flexibility research because of its centrality in the OM literature. Empirical research in this area spans the last decade, the period over which empirical OM research has evolved, and it encompasses a wide range of empirical methodologies. It should be noted that our choice does not imply that research in this area is more or less valid than in other areas of empirical OM research (e.g., manufacturing strategy, just-in-time, total quality management). Rather, we use the manufacturing flexibility construct because it typifies the types and quality of empirical research being conducted across most areas of OM.

Manufacturing flexibility corresponds to the ability of a manufacturing firm to respond to changes in its market environments. We examined the empirical research on manufacturing flexibility, utilizing the ABI computerized index to located relevant articles dating back to the year 1986. This index covers most business journals (including most OM journals). In addition, we searched the references of all relevant articles obtained from the ABI search and included those articles that were related to manufacturing flexibility. Table 1 lists the 14 studies identified in our search.

2. Construct validity—empirical assessment

Construct validation is a multifaceted process that is comprised of three basic steps outlined in Fig. 1. The first step requires the identification of a group of measurement items (empirical indicators) which are thought to measure the construct. It is necessary to demonstrate that the empirical indicators are logically, as well as theoretically, connected to the construct (Nunnally, 1978; Pedhazur and Schmelkin, 1991). This step is commonly referred to as *content* validity in the literature (Nunnally, 1978; Carmines and Zeller, 1979; Kerlinger, 1986; Pedhazur and Schmelkin, 1991). The second step establishes the degree to which the empirical indicators measure the construct (the process of establishing the construct validity of a measure) (Schwab, 1980). This step, which requires a series of empirical tests that examine the measurement properties of the indicators, is the central focus of this paper. The final step in-

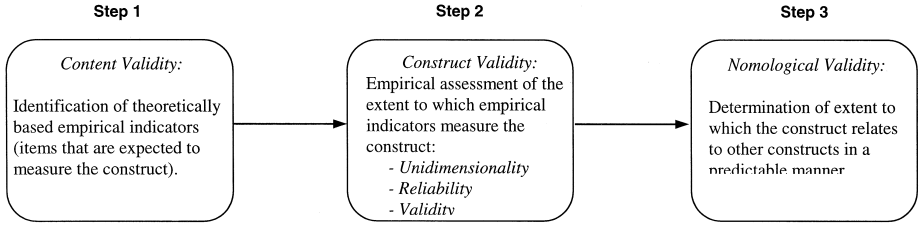

Fig. 1. Construct validation process.

volves the determination of the extent to which a construct relates to other constructs in a predictable manner, which is essentially hypothesis testing. This third step has been referred to as establishing *nomological* validity (Bagozzi, 1980; Venkatraman, 1989), or *substantive* validity (Schwab, 1980).

In the field of OM, many studies undertake the first step in construct validation. That is, empirical indicators that are thought to measure the constructs are chosen. However, many researchers then move directly to hypothesis testing, the third step, without ever assessing construct validity (e.g., Upton, 1995; Ettlie and Penner-Hahn, 1994; Das et al., 1993). Indeed, our review of manufacturing flexibility studies revealed that many studies failed to address the issues involved in construct validation (see Table 1). By failing to assess construct validity, we are in effect ignoring the many corrupting elements embedded in measures (e.g., measurement error, informant bias). As will be detailed in the remainder of the paper, this can seriously jeopardize the conclusions drawn in a study.

The second step, establishing construct validity, involves the empirical assessment of the adequacy of a measure and requires that three essential components be established: unidimensionality, reliability and validity. Unidimensionality involves establishing that a set of empirical indicators relates to one and only one construct. According to Bagozzi (1980), ''It is a matter of logical and empirical necessity that a variable be unidimensional'' and a multidimensional measure (comprised of indicators related to more than one construct) ''cannot, by definition, be considered a variable and hence must not be treated as such in one's theory'' (p. 126). Treatment of multidimensional measures as if they were unidimensional can lead to false conclusions. For example, when a measure of one variable improperly includes empirical indicators that are related to another variable, we are in a sense combining two variables (A and B) to form a new variable (C). Serious problems arise regarding the interpretation of association between C and other variables, in that it is impossible to determine whether the association is a function of the underlying variable(s) A and/or B. This obviously has the potential to create false associations between two variables and therefore taint the findings of a study.

The reliability and validity components pertain to the degree of measurement error. It is important to note that all measures reflect not only the construct they are intended to measure, but also measurement error (Carmines and Zeller, 1979). Measurement error can be partitioned into random error and systematic error. Random error occurs in an unpredictable manner, such that it creates random variances across repeated measures. Random error is inversely related to the reliability of a measure and can lead to incorrect results. For instance, random error can attenuate the results of statistical tests (Bagozzi et al., 1991; Bollen, 1989). On the other hand, studies have shown that, under certain circumstances, random error can inflate parameter estimates and lead to false indications of significant relationships between variables (Bollen, 1989).

Similarly, systematic errors, errors that vary in a consistent manner, are negatively associated with the validity of a measure. A common analogy used to explain systematic errors involves archery. If an archer continually misses the bulls-eye, but consistently places her arrows in the upper left corner, then some form of systematic error has affected her targeting. Likewise, if the quality of a product is consistently overstated by a machine operator, then the quality data will have a systematic error. The larger the systematic error, the less valid the measure. Systematic errors pertain to problems such as key-informant bias (method variance). The use of key-informants is a common method for obtaining data in most empirical OM studies. Typically, the informants are not randomly chosen; rather, certain individuals are targeted because they have specific qualifications related to the research topic. For example, most OM studies rely on manufacturing executives or personnel for reporting information used in their research. However, it has been shown that informant's position within an organization can systematically affect his responses, thereby creating biased measures (Kumar et al., 1993; Seidler, 1974). For example, studies have shown that the views of CEOs may systematically vary from those of lower level executives because differences in their organizational positions influence their interpretation of events (Golden, 1992; Hambrick, 1981).

Sections 2.1, 2.2 and 2.3 examine in detail the three components (unidimensionality, reliability, and

validity) required to establish the construct validity of a measure. Each section provides a detailed discussion of the empirical methods currently available for assessing the different components of construct validity, a general explanation of each methodology, and a discussion of the strengths and weaknesses of the different methods.

2.1. Unidimensionality

Gerbing and Anderson (1988) state, ''unidimensionality refers to the existence of a single trait or construct underlying a set of measures (or empirical indicators)'' (p. 186, text in parentheses added). There are two implicit conditions for establishing unidimensionality. First, an empirical indicator must be significantly associated with an underlying latent variable (i.e., the empirical representation of a construct) and, second, it can be associated with one and only one latent variable (Hair et al., 1992; Phillips and Bagozzi, 1986; Anderson and Gerbing, 1982). A measure must satisfy both of these conditions in order to be considered unidimensional. For example, an empirical indicator of the variable volume flexibility must be related to only volume flexibility and not other variables (e.g., delivery flexibility or process flexibility).

There are two common methods for assessing the unidimensionality of a measure: exploratory factor analysis and confirmatory factor analysis (Pedhazur and Schmelkin, 1991). Loosely speaking, factor analysis examines the linear association among empirical indicators as they relate to the underlying latent variable(s) (Kim and Mueller, 1978). Under exploratory factor analysis, the associations between empirical indicators and latent variables are not prespecified, whereas in confirmatory factor analysis the associations are specified.

2.1.1. Exploratory factor analysis method

In general terms, exploratory factor analysis (EFA) is an analytic method used to condense a group of empirical indicators into a smaller set of composite factors (latent variables) with a minimum loss of information (Hair et al., 1992). EFA is a method commonly used to explore data in search of unidimensional latent variables. Although this method has some serious drawbacks, which will be discussed below, it was used in all manufacturing flexibility studies in which unidimensionality was assessed (see Table 1).

Perhaps the most effective means of explaining EFA is through a diagrammatic representation of a general EFA model. In Fig. 2a, the ξ's are termed *common* factors, reflecting the fact that their common effects are shared across all the empirical indicators (X's) to varying degrees. The common factors (ξ's) correspond to the latent variables and are a linear combination of all the empirical indicators included in the analysis (Hair et al., 1992). The model depicted in Fig. 2a has two latent variables (ξ_1 and ξ_2), each comprised of a linear combination of the five empirical indicators (X_1–X_5). The λ_{ji}'s in the model represent *factor loadings*, the correlation between the jth latent variable and the ith empirical indicator. Squaring the factor loading (λ) yields the percentage of variance in the empirical indicator (X_i) that is explained by the latent variable (ξ_j) (Hair et al., 1992). The δ_i's in Fig. 2a are termed *unique* factors (or errors) and characterize the variance that is unique to each empirical indicator. The unique variance is comprised of both random and

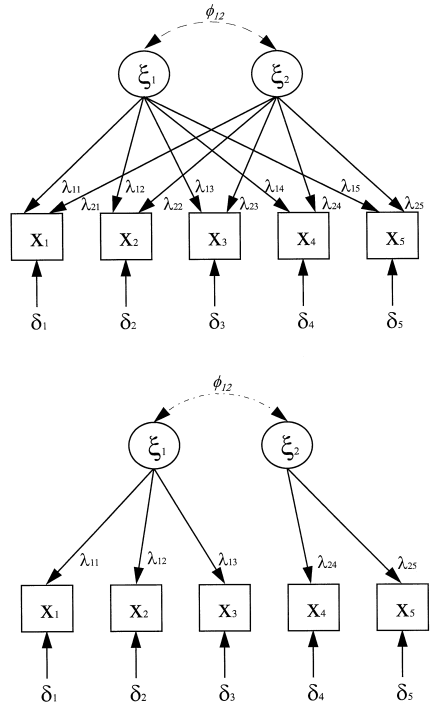

Fig. 2. Unidimentionality–measurement models.

specific measurement error which is not shared by the other empirical indicators. Finally, the double-headed dashed curve between the two latent variables represents the correlation (ϕ) between latent variables. In EFA, it is all or nothing with regard to allowing the latent variables to be free to correlate. That is, with some EFA techniques, the latent variables are not permitted to correlate (orthogonal techniques), whereas with others, all variables are free to correlate (oblique techniques) (Bollen, 1989).

Regarding unidimensionality, the primary purpose of EFA is to identify which empirical indicators are strongly linked to a particular latent variable (the strength of the 'link' is determined by the size of the factor loading). This process (identifying significant empirical indicators) is usually centered around one of two basic activities. One involves the 'pure' exploration of a group of theoretically-based empirical indicators for their linkage to an underlying latent variable or variables (Schwab, 1980). This involves studies in which there is little or no existing evidence that links a set of empirical indicators to a specific latent variable. Under these circumstances, EFA is applied to a set of empirical indicators in order to identify those indicators that form a unidimensional factor (latent variable), thereby providing the basis for a measure.

EFA is also used to establish that a group of empirical indicators are unidimensional with regard to a predefined latent variable (Schwab, 1980). As in the study by Ward et al. (1995), empirical indicators that were theoretically related to each of the different latent variables were specified a priori and unidimensionality was assessed via EFA. For example, they specified that indicators that measured manufacturing lead-time and setup/changeover time would be associated with flexibility, whereas indicators measuring the rate at which products become outdated and the rate of innovation of new products would be associated with market dynamism. EFA relies on rules of thumb regarding the size of factor loadings (λ's) to demonstrate unidimensionality.

Several issues confront researchers who use the EFA technique. One of these involves the criteria for assessing whether or not an empirical indicator is significantly related to an underlying latent variable. Specifically, it is not clear how large the factor loading should be in order to be regarded as significant. This is a fairly subjective matter, with no substantive criteria for making these judgments (Hair et al., 1992; Kerlinger, 1986). As previously noted, there are 'rules of thumb' employed for judging the significance of an indicator. For example, Hair et al. (1992) suggests that factor loadings as small as 0.30 are considered significant. However, these authors caution that, depending on such factors as sample size, unique variance, and the order in which the latent variable is extracted, the size of the factor loading required for 'significance' could be higher or lower. Subjectivity aside, a minimum factor loading of 0.30 seems to be the agreed upon rule of thumb (e.g., Hair et al., 1992; Pedhazur and Schmelkin, 1991; Kerlinger, 1986; Carmines and Zeller, 1979).

Although EFA can be a valuable tool for assessing unidimensionality, it is abundantly clear that EFA should not be used in the absence of a theoretically based rationale (Pedhazur and Schmelkin, 1991; Kerlinger, 1986; Carmines and Zeller, 1979; Nunnally, 1978). That is, even though the results from EFA may indicate that several empirical indicators are significantly related to a latent variable, there must be a logical reason for utilizing these indicators. Otherwise, we are engaging in what is commonly referred to as 'data snooping'. That is, analyzing the data in the absence of theoretical guidance, in which case the results are almost always sample specific and cannot be generalized to other circumstances.

Another issue with EFA is whether to analyze the empirical indicators as a homogeneous or heterogeneous group. Should the EFA be conducted simultaneously across multiple latent variables, (e.g., Gupta and Somers, 1992) or should each theorized latent variable be assessed separately (e.g., Suarez et al., 1996; Ward et al., 1995; Dixon, 1992)? By analyzing each latent variable separately, this provides only the capability for determining whether the empirical indicators are significantly associated with the latent variable. This method does not, however, establish that the indicators are unrelated to other latent variables included in the study, also a necessary requirement for establishing unidimensionality. Potential problems arise in that indicators that are reflective of more than one latent variable could create false associations between the latent variables, thereby seriously compromising the findings of a study.

With EFA, an important element affecting the stability of factor analytic results is sample size, in that the larger the sample the more stable the results obtained from EFA (Pedhazur and Schmelkin, 1991). As is often the case, having too small a sample may require that latent variables be analyzed separately (e.g., Dixon, 1992). In general, EFA should not be conducted on samples with fewer than 50 observations (samples of 100 or more are preferable) (Hair et al., 1992). Also, depending on the reference, as a rule of thumb there should be between four (Hair et al., 1992) and ten (Kerlinger, 1986) observations for each empirical indicator (or variable) included in the analysis.

In a related issue, when analyzing a heterogeneous set of empirical indicators, the researcher is faced with the decision of letting either all or none of the latent variables be free to correlate. This is particularly troublesome for studies involving latent variables where there is a theoretical basis for correlation to exist between only specific variables. For example, a recent study by Suarez et al. (1996) reported that there is empirical evidence that some manufacturing flexibility variables (e.g., mix flexibility and new product flexibility) are significantly correlated, while others are not (e.g., mix flexibility and volume flexibility). These types of situations create a dilemma regarding which EFA method to use to assess unidimensionality. One possible solution would be to conduct a separate analysis using an oblique EFA method (a method that allows latent variables to correlate) for those variables which are theorized to be correlated, and an orthogonal method (a method in which there is no correlation between any latent variable) for the others. However, this solution does not assess all indicators simultaneously and therefore fails to thoroughly investigate whether an indicator is associated with one or more latent variables.

In addition, there is a great deal of subjectivity regarding the criteria for determining the number of latent variables that are extracted for the analysis (Hair et al., 1992). This tends to be a greater problem for studies in which there is little certainty about the underlying structures contained in the indicators. In the absence of a quantitative technique for determining the number of latent variables, there are several subjective methods that are widely utilized for this task (e.g., latent root criterion, screen test, percentage of variance). [1] Several EFA techniques allow the number of latent variables (factors) that are to be extracted to be fixed a priori. This is more appropriate for studies that analyze unidimensionality based on a predetermined number of latent variables.

2.1.2. Confirmatory factor analysis method

Unlike the conventional techniques involving EFA, confirmatory factor analysis (CFA) contains inferential statistics that allow for hypothesis testing regarding the unidimensionality of a set of measures. This leads to a stricter and more objective interpretation of unidimensionality than does EFA and, therefore, it often produces different conclusions about the unidimensionality of a measure (Gerbing and Anderson, 1988). Unlike EFA, the use of CFA requires the researcher to specify the CFA-model prior to analyzing the data; that is, the latent variables and their associated empirical indicators have to be specified a priori. Referring to Fig. 2b, this is accomplished by restricting the empirical indicators to load on specific latent variables (ξ's) and to designate which latent variables are allowed to correlate. In doing so, CFA provides enhanced control for assessing unidimensionality and is more in line with the overall process of construct validation. That is, the CFA-model builds directly on step one of the multistep process of construct validation, which describes the linkages of the empirical indicators to the latent variables based on theoretical justification.

Some authors favor EFA instead of CFA, arguing that the former is more appropriate for early stages of theory development when there may not be a clear theoretical basis for the inclusion/exclusion of various indicators in a measure (e.g., Dixon, 1992; Gupta and Somers, 1992). For example, Gupta and Somers (1992) developed a set of unidimensional measures using EFA. However, they had already demonstrated a theoretical basis for why certain empirical indicators should be associated with specific flexibility variables, making CFA appropriate for their re-

[1] While it is beyond the scope of this paper to describe these methods, the interested reader should consult Hair et al. (1992) for a detailed description of these methods.

search. The use of EFA rather than CFA in this study resulted in a predictable outcome—the generation of fewer factors than there are underlying latent variables (Hunter and Gerbing, 1982). This is particularly true for variables that are correlated (a likely occurrence among different dimensions of manufacturing flexibility); under this condition, EFA tends to combine these variables into a single factor (latent variable) (Hunter and Gerbing, 1982).

CFA addresses other problems associated with EFA. For instance, unlike the subjective criteria for EFA, in CFA the significance of the factor loading (λ_{ji}) can be tested using a statistical *t*-test. In addition, CFA gives researchers the advantage of being able to evaluate the overall acceptability of the measurement model in terms of the model s fit to the data, using a χ^2 test. This is advantageous because it is possible for the indicators to be significantly related to their respective latent variables, even when the overall fit of the model is not statistically acceptable. This would indicate that there are serious problems with the way the current measures are specified in the model (e.g., indicators could be associated with more than one variable not specified in the model). This overall test of the measurement model is not possible in EFA.

In addition, CFA provides the capability to simultaneously test the unidimensionality of a set of empirical indicators that are comprised of both correlated and uncorrelated latent variables. This allows a more strict test of unidimensionality than in EFA. Additionally, CFA provides a means for directly assessing whether latent variables are correlated as stipulated in the model. This is accomplished by examining both the overall model fit (e.g., using a χ^2 statistic), as well as the significance of each of the correlations.

Although CFA is a superior technique for assessing unidimensionality, it is rarely used in OM research. As stated earlier, no studies involving manufacturing flexibility have used it. Given the advantages of CFA over EFA, as outlined above, it should be the method of choice for future OM studies that require the assessment of unidimensionality.

Section 2.2 describes a second component that is important to the establishment of the construct validity of a measure—reliability. Specifically, we discuss and critique the different methods that are currently available for assessing the reliability of a measure.

2.2. Reliability

Reliability pertains to the consistency or stability of a measure and is inversely related to the degree to which a measure is contaminated by random error (Bollen, 1989; Carmines and Zeller, 1979). Regardless of whether we are working with subjective or objective measures, random error is always present to a certain degree and in some cases can constitute a major problem that can jeopardized the validity of research findings. Therefore, it is incumbent upon researchers to assess and report the reliability of their measures. As important as this is, this basic step of construct validation has been ignored in the majority of manufacturing flexibility studies to date. In fact, of the fourteen studies involving manufacturing flexibility, less than a third of them assessed reliability (see Table 1). Many of the methods for empirically assessing reliability are based on classical test theory, therefore a brief review of this area is provided.

2.2.1. Classical test theory and measurement error

The basic equation associated with test theory is

$$o_i = \tau_i + e_i \quad (1)$$

where, parameter o_i is the observation of the ith indicator (or actual 'test' score) of the measure, τ_i is the true-score, and e_i is the random error term of the ith indicator (Pedhazur and Schmelkin, 1991; Bollen, 1989). Several assumptions are made regarding measured variables: (1) covariance between τ_i and e_i is zero: $\mathrm{COV}(\tau_i, e_i) = 0$; (2) the mean of the random errors (e_i) is zero: $E(e_i) = 0$; (3) the covariance between τ_i from the ith indicator and e_j from the jth indicator of the same variable is zero: $\mathrm{COV}(\tau_i, e_j) = 0$; and (4) the covariance between the errors of two different indicators of the same variable is zero: $\mathrm{COV}(e_i, e_j) = 0$ (Pedhazur and Schmelkin, 1991; Bollen, 1989; Carmines and Zeller, 1979).

In test theory there are three general types of observed measures: *parallel*, *τ-equivalent*, and *congeneric* (Bollen, 1989). The different types of measures are most easily described in the context of two different indicators (or measures), o_i and o_j of the same variable.

(1) *Parallel measures*. In addition to the general assumptions stated above (regarding classical test theory), two different indicators (measures) are parallel if they have equivalent true-scores (i.e., τ_i equals (τ_j), and the errors terms have equal variances (i.e., $\mathrm{VAR}(e_i) = \mathrm{VAR}(e_j)$). The assumptions made concerning parallel measures are the most restrictive and consequently are rarely satisfied in empirical studies (Pedhazur and Schmelkin, 1991; Bollen, 1989).

(2) *τ-Equivalent measures*. The assumption of equal error variances across indicators is relaxed for τ-equivalent measures. That is, the only assumption

Table 2
Summary of reliability methods

Reliability methods	Assumptions	Advantages	Disadvantages
Test–retest: involves measuring a variable at two points in time (t and $t+1$) using the same scale and sample	Variable are stable	Appropriate for measures comprised of single indicators	Not appropriate for variables that are not stable over time
	Measures are parallel		Perceptual based measures susceptible to carryover effects Costs of administering two surveys can be prohibitive
Alternative forms: involves measuring a variable at time (t) using one measure and again at time ($t+1$) using a different measure. Both utilize the same sample	Variables are stable over time	Appropriate for measures comprised of single indicators	Not appropriate for variables that are not stable over time
	Measures are parallel	Less susceptible to carryover effects than Test–retest method	Costs of administering two surveys can be prohibitive Requires development of two unique measures
Cronbach's α: involves deriving an index, which ranges from 0 to 1, based on the correlations of the indicators that comprise the measure	Measures are τ-equivalent	Assumption of τ-equivalent measures is a less restrictive assumption than parallel measures	Underestimates reliability of measures that are not τ-equivalent
	Measures are comprised of multiple indicators	Requires only a single sample Virtually no chance of carryover effects Reliability of a measure may be improved by increasing the number of indicators	Requires multiple indicators for a measure
WLJ composite reliability: involves confirmatory factor analysis to derive a composite index, which also ranges from 0 to 1	Measures are congeneric	Assumption of congeneric measures is the least restrictive	Underestimates reliability of measures that are not congeneric
	Measures are comprised of multiple indicators	Requires only a single sample Virtually no chance of carryover effects It provides the capability to directly test assumption of congeneric measures	Requires multiple indicators for a measure

is that the true-scores are identical across different indicators (i.e., $\tau_i = \tau_j$) (Pedhazur and Schmelkin, 1991; Bollen, 1989).

(3) *Congeneric measures*. Congeneric measures are the least restrictive type of measures. The only assumption is that the true scores of the indicators are perfectly correlated. That is, error variances and true scores may be unequal (Pedhazur and Schmelkin, 1991; Bollen, 1989).

There are several methods for assessing the reliability of a measure. The type of empirical test that should be used depends, in part, on the assumptions regarding the type of measure. Each method of reliability assessment has some unique drawbacks that should be considered. In the remainder of this section, we review and critique some of the more common methods for assessing reliability (see Table 2).

2.2.2. Test–retest method

Conceptually, the test–retest method involves taking a measurement at two different points in time (e.g., t and $t+1$), using the same set of indicators and sample group. The correlation coefficient between the two sets of test scores (o_i) is used as an estimate of reliability (Pedhazur and Schmelkin, 1991). Use of this method requires the more restrictive assumption of parallel measures (Bollen, 1989). In addition, this method assumes that the true scores (τ's) are stable over time—that is, the phenomenon being measured remains constant and the covariance between errors is zero ($\mathrm{COV}(e_t, e_{t+1}) = 0$) (Bollen, 1989). The test–retest method of reliability assessment has several drawbacks.

First, the test–retest method assumes that measures are stable over time. That is, if we could measure the true score of a variable at two different points in time (t and $t+1$), the measurements would be the same. However, most variables are not completely stable over time. For example, given the dynamic nature of many manufacturers, it is highly probable that the conditions surrounding manufacturing flexibility would be changing over time, in which case the stability assumption would not be appropriate. In general, the greater the time period between tests, the more likely it is that this assumption will be violated (Carmines and Zeller, 1979). The overall effect of unstable true scores is that the reliability of the measures is underestimated (Bollen, 1989; Carmines and Zeller, 1979).

Second, temporal stability affects created by response tendencies, such as carry-over effects and social desirability, can lead to overestimation of reliability when assessed using the test–retest method. For example, problems regarding carry-over effects (sometimes referred to as memory effects) have been associated with the test–retest method (Pedhazur and Schmelkin, 1991; Carmines and Zeller, 1979). It has been shown that surveying individuals twice using the same measures has a high potential for biasing perceptual-based measures. That is, responses given the first time tend to influence responses given the second time (Pedhazur and Schmelkin, 1991; Nunnally, 1978). These types of measures are fairly common in manufacturing flexibility research and in OM research in general. For example, half of the manufacturing flexibility studies utilized perceptual measures of flexibility.

Another potential bias inherent in the test–retest method is the tendency of respondents to represent themselves in a positive manner - commonly referred to as the social-desirability bias. If respondents consistently rate themselves in a socially desirable manner, the can artificially increase the stability of the measure, because a high reliability coefficient for measures taken at two different points in time may be a result of the respondents' desires to present themselves (or their companies) in a positive light rather than the constancy of the measurement instrument itself. In general, the test–retest method of reliability assessment should be avoided for perceptual-based measures, given that carry-over and social desirability effects can have the undesirable result of inflating reliability estimates (Nunnally, 1978).

Finally, there are practical problems associated with the test–retest method. In the case of manufacturing flexibility studies, as well as most other OM empirical studies, getting executives to take the time to fill out a single questionnaire can be difficult at best; however, convincing executives to fill out the same questionnaire a second time presents an enormous challenge. Another practical concern involves the substantial cost of conducting a second survey. Although the simplicity of the test–retest method makes it very appealing, numerous experts have strongly discouraged use of this technique for esti-

mating reliability due to its many potential problems (e.g., Pedhazur and Schmelkin, 1991; Bollen, 1989; Nunnally, 1978).

2.2.3. Alternative forms method

Another technique for estimating reliability is the alternative forms method (also referred to as equivalent forms). This method is very similar to the test–retest method, however the alternative forms method involves two different measures (o_1 and o_2) of the same variable (e.g., use of two different measurement instruments) at times t and $t+1$, where

$$o_1 = \tau_t + e_t \quad (2)$$

$$o_2 = \tau_{t+1} + e_{t+1} \quad (3)$$

o_1 is a measure at time t, and o_2 is a different measure at time $t+1$. The alternative forms method makes the same assumptions as the test–retest method: the measures are parallel, τ_t equals τ_{t+1}, and the errors (e_t and e_{t+1}) do not covary (Pedhazur and Schmelkin, 1991; Bollen, 1989). Under these assumptions, the correlation between the two different measures (o_1 and o_2) equals the reliability of both measures (Pedhazur and Schmelkin, 1991; Bollen, 1989).

In manufacturing flexibility research, the alternative forms method would require that two different measurement instruments be developed to measure the same construct. For example, Ward et al. (1995) and Swamidass and Newell (1987) have developed two different measures of manufacturing flexibility. The alternative forms method of assessing reliability would involve administering one measure of manufacturing flexibility (e.g., the Ward et. al. measure) at time t, and the other measure (e.g., Swamidass and Newell's) at time $t+1$. The correlation of the two different measures would be used as an estimate of the reliability of the measures.

Like the test–retest method, the alternative forms method has several theoretical and practical drawbacks. These include the potential to underestimate reliability due to an unstable true score, where τ_t does not equal τ_{t+1}, as well as the difficulties and added cost of obtaining two sets of responses from the same company (Bollen, 1989). The alternative forms method has an additional disadvantage, in that it requires the construction of two different measures that are truly equal (i.e., both have equal true scores) (Carmines and Zeller, 1979). This is very problematic for studies that involve broadly defined constructs. For example, the construct of manufacturing flexibility encompasses many indicators covering a wide range of elements, such as 'the degree of R&D effort in a firm' (Swamidass and Newell, 1987) and 'the emphasis placed on reducing procurement lead-times' (Ward et al., 1995). Given the broad nature of the manufacturing flexibility construct, it would be very difficult to develop two different measures that have identical true scores (i.e., measure exactly the same construct).

Compared to the test–retest method, the alternative forms method does have advantages. First, the alternative forms method is less susceptible to carry-over effects because of the different scales used for each measure (Bollen, 1989). Likewise, because different versions of a measure are used, there is a reduced likelihood that the error terms (e_t and e_{t+1}) will be correlated—a condition that can lead to overestimation of reliability (Bollen, 1989).

2.2.4. Cronbach's α coefficient

The Cronbach α coefficient is one of the most popular methods for assessing reliability (Pedhazur and Schmelkin, 1991; Carmines and Zeller, 1979). In the manufacturing flexibility studies, all researchers that assessed reliability did so using Cronbach's α. The α coefficient can range from 0 to 1 (the higher the α the higher the reliability) and represents the estimated systematic variance (or true score) of a measure. The α coefficient is based on the correlations among the indicators that comprise a measure, with higher correlations among the indicators associated with higher α coefficients (Pedhazur and Schmelkin, 1991).

There is not a complete agreement on how large the α coefficient should be in order to be considered acceptable. For example, Nunnally (1978) has indicated that reliabilities below 0.70 are not acceptable. However, when confronted with lower α values, many studies still quote the earlier position taken by Nunnally (1967) that α values as low as 0.50 are acceptable for early stages of research. Still, others have argued that for broadly defined constructs, reliabilities as low as 0.40 are acceptable (Van de Venn

and Ferry, 1980). Clearly, though, higher levels of α engender greater confidence in the measure. However, factors such as the nature of the construct (narrowly vs. broadly defined), the stage of research (early vs. late), and the costs associated with the study (cost of developing a more reliable measure) may determine the degree of measurement error we are willing to tolerate.

The Cronbach α coefficient provides advantages over the previous two methods. First, unlike both the test–retest and alternative forms methods, it is based on the much less restrictive assumption that the indicators are τ-equivalent (Pedhazur and Schmelkin, 1991). Second, the Cronbach α method requires only a single sample, which essentially eliminates the chance of carry-over effects (Bollen, 1989). Finally, it has been shown that the α coefficient and, therefore, the reliability of a measure, can be increased by increasing the number of indicators comprising a measure (Pedhazur and Schmelkin, 1991; Carmines and Zeller, 1979). This provides researchers with a direct means for improving the reliability of their measures.

There are, however, a couple of drawbacks associated with the Cronbach α method. First, even though it is based on the less restrictive assumption of τ-equivalent measures, this is still problematic for studies involving congeneric measures. For these studies, Cronbach's α underestimates reliability and can therefore lead to the false conclusion that the measures are not reliable due to the underestimated (lower) α value (Bollen, 1989). Second, reliability estimates of single-item measures cannot be made using Cronbach's α method (Bollen, 1989). [2] In order to overcome this problem, multiple indicator measures must be used. However, this will increase the size of questionnaires in survey-based studies, which in turn can lead to a lower response rate. Although this is a problem, it is the authors' opinion that the benefits of being able to assess the reliability of the measures used in a study clearly outweigh the risk of a reduction in the response rate.

2.2.5. Werts, Linn, and Jöreskog (WLJ) composite reliability method

The Werts, Linn, and Jöreskog (WLJ) composite reliability method is increasingly used in other fields of research (e.g., marketing and strategy), however none of the manufacturing flexibility studies used this method. The WLJ method utilizes CFA to derive a composite reliability index, which is based on the proportion of variance attributable to only the latent variable (i.e., excluding measurement error) (Werts et al., 1974). Like the Cronbach α, the WLJ composite reliability index can range from 0 to 1, with a higher value indicating greater reliability.

The WLJ method has several advantages over the other methods used to assess reliability. First, the WLJ method is based on the least restrictive assumption that the measures need only be congeneric (true scores do not have to be equivalent, but must be perfectly correlated) (Bollen, 1989). Therefore, this method is applicable to a large set of measures, given that it can be used for parallel and τ-equivalent measures as well. Second, because this method incorporates CFA, it provides the ability for directly testing the assumption of congeneric measures. [3] Finally, like Cronbach's α, this method requires only a single sample, making it easier to implement than either the test–retest or alternative forms methods, and virtually eliminates the chance of carry-over effects.

The disadvantages associated with the WLJ method are similar to those of Cronbach's α. First, if the true scores of the measures are not perfectly correlated, which means they are not congeneric, then the WLJ method will underestimate reliability (Bollen, 1989). Second, like Cronbach's α, the WLJ method requires multiple indicators for each measure in order to estimate reliability. As previously stated, although increasing the number of indicators may reduce the response rate, the benefits of assessing reliability outweighs this risk.

Section 2.3 details the final components of the second stage in the construct validation process—

[2] In addition to the problems regarding Cronbach's α, the use of a single indicator is generally inadequate for capturing the complexity of most constructs.

[3] It is beyond the scope of this paper to explain this procedure, Jöreskog and Sörbom (1989) provide an excellent description accompanied by an example of testing the hypothesis of congeneric measures.

establishing the validity of a measure. Specifically, we explain and critique the two methods most often used for assessing the validity component.

2.3. Convergent and discriminant validity

The validity of a measure is the degree to which the variance in the measure is attributed to variations in the variable and not some other factor (e.g., method variance). Establishing the validity component of a measure involves two elements: convergent validity and discriminant validity (Campbell and Fiske, 1959). Convergent validity relates to the degree to which multiple methods of measuring a variable provide the same results. For example, if we measure manufacturing flexibility using different methods (e.g., data independently supplied by two different sources, such as a CEO and a manufacturing executive), to what degree does the data from the two methods converge (or covary)? The assumption is that if a measure is valid, it should yield the same results when utilized across different methods. If the results fail to converge, this suggests that the variations in the data may stem in part from the difference in methods, thereby raising doubt about the validity of the measure. Discriminant validity is the degree to which measures of different latent variables are unique. That is, in order for a measure to be valid, the variance in the measure should reflect only the variance attributable to its intended latent variable and not to other latent variables.

The use of measures that lack convergent and discriminant validity can lead to numerous problems in the interpretation of the results of a study. For example, the finding of a significant relationship between variables that lack convergent validity might be attributable to the method(s) used to measure the latent variables, not to any 'true' relationship between them (Fiske, 1982). Similarly, if we use measures of two latent variables, x and y, that fail to demonstrate discriminant validity, we cannot conclude that the measures are reflecting two unique constructs; in this case it would be inappropriate to analyze x and y as separate latent variables. Given the potentially serious problems posed by the lack of convergent and discriminant validity, it is alarming that only one of the fourteen manufacturing flexibility studies assessed this component of construct validity.

In Sections 2.3.1 and 2.3.2, we examine two of the most commonly used methods for assessing convergent and discriminant validity: the Multitrait–Multimethod Matrix Method, and the Confirmatory Factor Analysis Method. Once again, we review the techniques and discuss the strengths and weaknesses of each.

2.3.1. Multitrait–multimethod matrix method

The multitrait–multimethod (MTMM) matrix method is the traditional way in which convergent and discriminant validity has been assessed. This method, developed by Campbell and Fiske (1959), suggests that convergent and discriminant validity be evaluated using the pattern and magnitude of correlations between the latent variables (measures of *traits* or constructs) and the different methods (e.g., different informants or different instruments) used to measure the constructs, thereby forming the MTMM matrix. As shown in Fig. 3, assessment is based on three sets of correlations. First, correlations between different constructs and the same method, referred to as heterotrait–monomethod (HTMM) correlations, are enclosed in the solid triangle. Second, correlations between different constructs and different methods, referred to as heterotrait–heteromethod (HTHM) correlations, are enclosed in the dashed triangle. Finally, correlations between the same construct as measured by different methods, monotrait–heteromethod (MTHM) correlations, are located on the diagonals of the matrix.

Based on the MTMM matrix method, convergent validity is evident when the correlations of the same construct (trait) measured by different methods

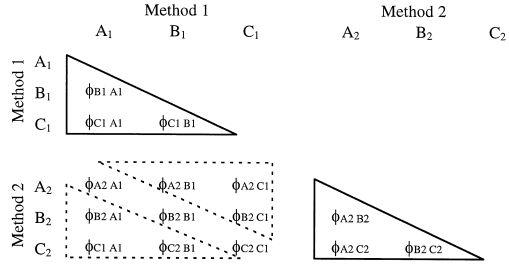

Fig. 3. MTMM matrix: three traits—two methods.

(MTHM correlations) are relatively large and significantly different from zero. Assessment of discriminant validity, on the other hand, involves three different criteria (Bagozzi et al., 1991; Pedhazur and Schmelkin, 1991). First, any one MTHM correlation must be significantly larger than any of the other correlations located on the same row and column. Second, a variable should correlate higher with the same variable measured by different methods (MTHM correlations), than with different variables measured by the same method (HTMM correlations). Third, the pattern of correlations between HTMM and HTHM correlation-triangles must be the same. Failure to meet these criteria would indicate that the measures are corrupted potentially by method bias (Bagozzi et al., 1991).

There are several limitations associated with the MTMM matrix method. One of the most frequently cited criticisms is that there are no precise standards for determining the degree to which the criteria outlined above are met. For example, if a HTMM correlation is greater than all but one of its related HTMM or HTHM correlations, does it fail to achieve discriminant validity? When are MTHM correlations large enough to represent convergent validity? Unfortunately, there are only relatively crude rules-of-thumb to help guide this decision and the adequacy of these are not agreed upon in the literature (Bagozzi et al., 1991; Farh et al., 1984).

Second, this method assumes that the traits are all equally influenced by the different methods used to measure them (Schmitt and Stults, 1986). For instance, if two methods are used to measure a latent variable, it is assumed that the degree of method effect (or bias) introduced into the measure by the different methods will be the same. Take, for example, the study by Gupta and Somers (1992) in which they assessed the convergent and discriminant validity of several manufacturing flexibility measures using the MTMM matrix method. They used two methods of collecting data, a phone interview and mail survey, in their study. Unlike a mail survey, there is a greater opportunity for a phone interviewer to influence the responses of the informant. This, in turn, has the potential to create differences in the magnitude of the method effect on the different measures, which would violate the assumption of equal effect. Unfortunately, the MTMM matrix method does not provide a means of evaluating this assumption. In addition, a recent study by Bagozzi et al. (1991) has shown that the significance and magnitude of method effects on each measure can vary across methods, a finding that calls into question this assumption. It would seem that the greater the difference in the types of methods used to measure the traits, the more likely that the traits will not be equally influenced.

Third, the MTMM matrix method also assumes that the methods factors are uncorrelated (Schmitt and Stults, 1986). Again, the results of the study by Bagozzi et al. (1991) suggest that it is not uncommon for methods factors to be correlated, calling into question this assumption. Finally, with the MTMM matrix method there is no way to separate out the variance in the measure that is attributable to the traits, vs. the methods, vs. random error. For example, using this method it is impossible to determine if high MTHM correlations (indicating convergent validity) are due to a high degree of shared method variance (an undesirable characteristic) or due to a high level of shared trait variance (a desirable characteristic).

Each of the above limitations greatly reduces the effectiveness and reliability of the MTMM matrix method. In light of these problems, there is growing agreement that a confirmatory factor analysis (CFA) approach is a superior method for assessing convergent and discriminant validity (Bagozzi et al., 1991; Pedhazur and Schmelkin, 1991; Schmitt and Stults, 1986; Farh et al., 1984).

2.3.2. Confirmatory factor analysis method

The CFA method for assessing convergent and discriminant validity involves a multistep procedure. Following the work of Widaman (1985) and Bagozzi et al. (1991), the CFA procedure for testing convergent and discriminant validity involves assessing a series of nested CFA-models and their individual parameter estimates. Assuming that the observed measures are multivariate normally distributed, the overall statistical acceptability of any CFA-model can be tested using the χ^2 statistic. In addition, comparisons of two nested models can be made in order to determine if one model is more appropriate (i.e., one model provides a statistically better fit to the data). Nested models are models that can be

formed by constraining specific parameters of another model. Given that the difference between two χ^2 values is distributed as a χ^2 with degrees of freedom equal to the difference between the degrees of freedom of the two models, a statistical comparison can be made between two nested models.

In order to simplify the discussion of the CFA procedure for testing convergent validity, we focus our discussion on the hypothetical example depicted in Fig. 4. It includes three latent variables (or traits) (ξ_{T1}, ξ_{T2}, and ξ_{T3}) and two different methods of measurement referred to as method factors (ξ_{M1} and ξ_{M2}). The use of CFA for assessing convergent validity requires at least two empirical measures (X's) for each latent variable. The latent variable factor loadings (λ_1, $\lambda_2, \ldots \lambda_6$) correspond to the variance attributed to the latent variables and the method factor loadings (λ_7, $\lambda_8, \ldots \lambda_{12}$) correspond to variance in the measures attributable to the different methods of measurement. The squared factor loadings (λ^2), in conjunction with the errors variables (δ's), account for the total variance in the data. The ϕ's represent the correlations among the latent variables and among the method factors.

The CFA method requires that the convergent validity of a set of measures be established first, followed by a test of discriminant validity (Bagozzi and Phillips, 1982). A general CFA procedure for establishing convergent validity is outlined in Fig. 5. The series of nested models necessary to test for convergent validity are described below (in relation to Fig. 4).

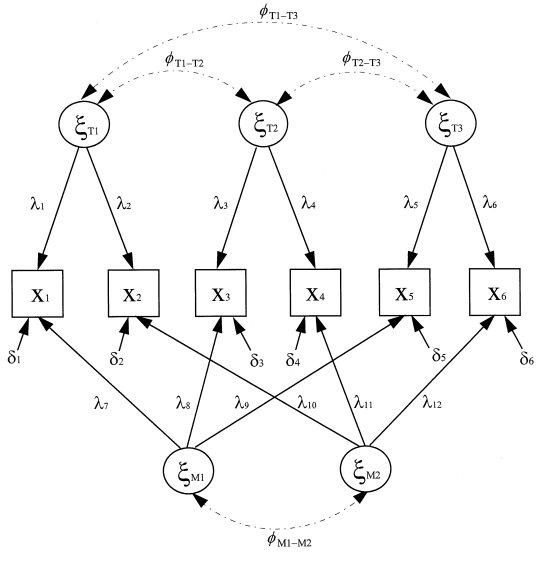

Fig. 4. Confirmationary factor analysis MTMM model.

M1: Null Model—all factor loadings (λ_1–λ_{12}) are constrained to zero. This model tests the hypothesis that the total variance in the indicators can be explained completely by the unique variance (or error).

M2: Trait Only Model—the factor loadings for the method factors (λ_7–λ_{12}) are constrained to zero. This model is used to test the hypothesis that the variance in the measures is due to only latent variable (trait) variance and random error.

M3: Trait–Method Model—the empirical measures (X's) are loaded on their associated latent variable and method factors (as depicted in Fig. 4). This model is used to test the hypothesis that the variance in the measures is due to both latent variable variance and method variance, as well as random error.

The steps in the CFA procedure for establishing convergent validity are illustrated in Fig. 5. The procedure begins with an examination of the fit of the null model (M1). M1 is considered the baseline model and is formed by constraining all factor loadings (λ's) to equal zero. Failure to reject M1 (using a χ^2 test) suggests that variance in the indicators is explained by unique (error) variance. This may indicate a lack of convergent validity and requires further assessment. The next step involves comparing the fit of M1 (null model) to M2 (trait only model) using a χ^2 difference test. If M2 is found to provide a better fit (i.e., a significant χ^2 difference test), this indicates that the addition of traits (or latent variables) improves the explanation of the variance in the data. It is then necessary to test goodness-of-fit of M2. If, however, M2 does not provide a better fit than M1, then a comparison between M1 and M3 is required. If M3 (trait–method) provides significant improvement over the goodness-of-fit of M1, then it is necessary to test the goodness-of-fit of M3. Alternatively, if M3 fails to provide a better fit than M1, it should be concluded that there is no convergent validity.

If the χ^2 test of fit for M1 is rejected, the next step is to examine the goodness-of-fit of the trait

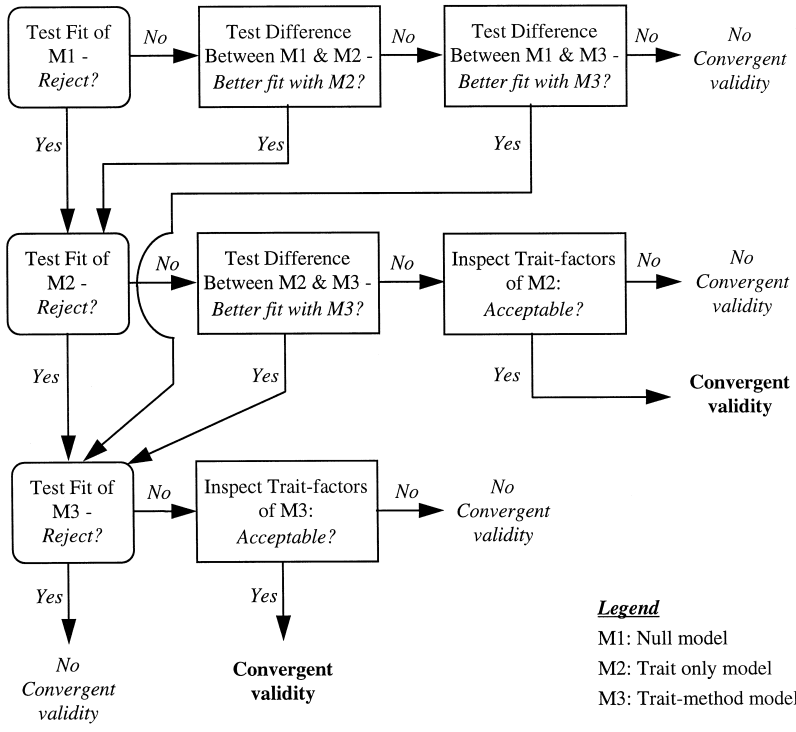

Fig. 5. Assessing convergent validity—CFA method.

only model (M2). Acceptance of M2 indicates that there is strong evidence that the variance in the empirical measures is explained by the latent variable variance and random error. In order to establish convergent validity, however, two additional criteria must be met. First, there must be no statistical difference between the fit of M2 (trait only) and M3 (trait–method), thereby demonstrating that the method factors do not provide meaningful improvements in explaining the variance in the data. Second, the individual parameters for M2 should be scrutinized. That is, each individual measure's factor loading should be statistically significant and the variance attributed to the latent variable (determined by the square of the factor loading) should be relatively large as compared to the error variance, although the term 'relatively large' has not been clearly defined to date (Bagozzi et al., 1991). Either the rejection of the trait only model (M2) or a significant difference between M2 and M3, requires the assessment of the overall goodness-of-fit of the trait–method model (M3).

If M3 provides a statistically acceptable fit to the data, then the latent variable factor loadings (λ) must be examined in order to establish convergent validity. Specifically, it must be shown that the latent variable factor loadings are statistically significant and account for a relative substantial proportion of the total variance in the data. Although there are no clear rules for what constitutes a 'substantial proportion of the total variance', an examination of the methods only model (constraining the factor loadings' of the latent variables to zero) provides some guidance. If either the methods only model does not yield an acceptable fit or the M3 model is significantly different from the methods only model, then it could be concluded that the latent variables account for a substantial amount of the variance in the data, over and above that of the methods factors.

Only when convergent validity is established is it acceptable to examine the discriminant validity of the measures (i.e., that they represent distinct latent variables). If it can be demonstrated that the variables are not perfectly correlated then discriminant

validity is established (Widaman, 1985). Two approaches have been used to demonstrate discriminant validity. One approach compares two CFA-models: one in which the correlation of a pair of latent variables is constrained to equal 1.0, and another in which the correlation is free to vary (Venkatraman, 1989). A significantly lower χ^2 value for the unconstrained model, as compared to the constrained model, provides support for discriminant validity. This approach requires separate comparisons for each pair of latent variables. The second approach simply involves testing that the correlations between the latent variables are statistically different from 1.0 (Bagozzi et al., 1991).

The confirmatory factor analysis (CFA) method of assessing convergent and discriminant validity is a more powerful tool and requires fewer assumptions than the traditional Campbell and Fiske MTMM matrix method. First, the CFA method provides a direct means for assessing the *degree* to which convergent and discriminant validity are achieved (Bagozzi et al., 1991; Bagozzi and Phillips, 1982). This is accomplish by examining the size and significance of factor loadings, as well as by examining the χ^2 goodness-of-fit of the overall model. Second, inherent in the CFA technique is the capability for method factors to influence the measures of the traits to varying degrees (Bagozzi et al., 1991; Bagozzi and Phillips, 1982). That is, the CFA technique does not require the strict assumption of equal method factor influence across all traits (variables). As discussed previously, it is highly likely that method factors influence the measures of traits to varying degrees. Finally, the information provided by CFA allows the variance to be partitioned into trait, method and error components (squared trait and method factor loading (λ^2's), and error variance (δ's)) (Bagozzi et al., 1991; Bagozzi and Phillips, 1982).

While CFA affords several advantages over the Campbell and Fiske MTMM matrix method, it has one potential drawback—the χ^2 statistic is directly influenced by the sample size (Tanaka, 1993). For example, a large sample size may cause the rejection of almost any model, even for models that explain most of the variance in the data. Alternatively, a small sample tends to produce non-significant χ^2 statistics, causing the acceptance of most models. In order to avoid these problems, it is best to include measures other than the χ^2 statistic for assessing of the goodness-of-fit for a model. For a thorough discussion of alternatives to the χ^2 statistic for evaluating model fit, see Bentler (1990), Jöreskog (1993) and Tanaka (1993).

3. Implications for researchers

This paper has reviewed methodologies available for the assessment of the different components of construct validity—unidimensionality, reliability, and convergent and discriminant validity—and has highlighted the advantages and disadvantages associated with each. Perhaps the most important conclusion we can draw is that CFA-based methodologies provide the most comprehensive method for assessing construct validity. Specifically, CFA is the only method that provides statistical tests for the assessment of construct validity (e.g., χ^2 test for the overall fit of the model, *t*-tests of the significance of the model parameters, such as factor loadings, error variances, factor correlations), allowing for a more objective means for establishing construct validity. In addition, CFA-based methodologies require fewer assumptions than the more traditional methods of assessment and, therefore, are more accommodating to most empirical data. Finally, CFA provides greater flexibility in explicitly incorporating and analyzing the different elements of measurement variance (e.g., trait, method, error), providing superior capabilities for determining the adequacy of the measures.

Because manufacturing flexibility studies provided the context for our construct validity discussion, it seems important to note the current state of this literature. Overall, we found a lack of attention to construct validity issues in research on manufacturing flexibility. For example, 50% of the studies did not assess any of the components related to construct validity. In fact, most studies (64%) used single item measures, making the assessment of construct validity impossible. Among those studies that did address some aspect of construct validity, none used the superior CFA-based methods. As mentioned earlier, the potential for confounded statistical results places a severe limitation on the conclusions drawn in studies that do not examine the construct validity of their measures. Given the importance of construct

validity in establishing the veracity of measurement, there is a clear need for greater attention to this part of the research process. In order to adequately assess construct validity, researchers should utilize both multiple empirical indicators for their measures and at least two methods for measuring their latent variables (e.g., multiple key-informants).

Although we have focused on manufacturing flexibility in our discussion, we expect that the lack of consideration of construct validity issues may be a critical problem in other areas of OM research. Specifically, we expect that few studies have considered construct validity, and predict that most of those studies that have assessed some aspect of validity have used the less effective, traditional methods outlined in this paper. Given that the failure to adequately assess construct validity can greatly diminish the conclusions derived in a study, researchers must, in the future, pay closer attention to this important element of the research process.

References

Anderson, J.C., Gerbing, D.W., 1982. Some methods for respecifying measurement models to obtain unidimensional construct measurement. J. Marketing Res. 19, 453–460.

Bagozzi, R.P., 1980. Causal Modeling In Marketing. Wiley, New York.

Bagozzi, R.P., Phillips, L.W., 1982. Representing and testing organizational theories: a holistic construal. Administrative Sci. Q. 27, 459–489.

Bagozzi, R.P., Yi, Y., Phillips, L.W., 1991. Assessing construct validity in organizational research. Administrative Sci. Q. 36, 421–458.

Bentler, P.M., 1990. Comparative fit indexes in structural models. Psychol. Bull. 88, 588–606.

Bollen, K.A., 1989. Structural Equations with Latent Variables. Wiley, New York.

Buffa, E.S., 1981. Commentary on production/operations management: agenda for the '80s'. Decision Sciences 12 (4), 572–573.

Campbell, D.T., Fiske, D.W., 1959. Convergent and discriminant validation by the multitrait–multimethod matrix. Psychol. Bull. 58, 81–105.

Carmines, E.G., Zeller, R.A., 1979. Reliability and Validity Assessment, Sage University Paper series on Quantitative Applications in the Social Sciences, Series No. 07–017. Sage Publications, Beverly Hills.

Das, B.J., Chappell, W.G., Shughart, W.F., 1993. Demand fluctuations and firm heterogeneity. J. Ind. Econ. 41 (1), 51–60.

De Meyer, A., Nakane, J., Miller, J.G., Ferdows, K., 1989. Flexibility, the next competitive battle. Int. J. Operations Prod. Manage. 6 (4), 6–16.

Dixon, J.R., 1992. Measuring manufacturing flexibility: an empirical investigation. Eur. J. Operations Res. 60, 131–143.

Ettlie, J.E., Penner-Hahn, J.D., 1994. Flexibility ratios and manufacturing strategy. Manage. Sci. 40 (11), 1444–1454.

Farh, J.L., Hoffman, R.C., Hegarty, W.H., 1984. Assessing environmental scanning at the subunit level: a multitrait–multimethod analysis. Decision Sci. 15 (2), 197–219.

Fiegenbaum, A., Karnani, A., 1991. Output flexibility—a competitive advantage for small firms. Strategic Manage. J. 12, 101–114.

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

Flynn, B.B., Sakakibara, S., Schroeder, R.G., Bates, K.A., Flynn, E.J., 1990. Empirical research methods is operations management. J. Operations Manage. 9 (2), 250–284.

Gerbing, D.W., Anderson, J.C., 1988. An updated paradigm for scale development incorporating unidimensionality and its assessment. J. Marketing Res. 25 (5), 186–192.

Golden, B.R., 1992. The past is the past—or is it? The use of retrospective accounts as indicators of past strategy. Acad. Manage. J. 35, 848–860.

Gupta, Y.P., Somers, T.M., 1992. The measurement of manufacturing flexibility. Eur. J. Operations Res. 60, 166–182.

Hair, Jr., J.F., Anderson, R.E., Tatham, R.L., Black, W.C., 1992. Multivariate Data Analysis with Readings, 3rd edn. Macmillan, New York.

Hambrick, D.C., 1981. Strategic awareness within to management teams. Strategic Manage. J. 2, 263–279.

Hax, A.C., 1981. Commentary on production/operations management: agenda for the '80s'. Decision Sci. 12 (4), 574–577.

Hörte, S.A., Börjesson, S., Tunälv, C., 1991. A panel study of manufacturing strategies in Sweden. Int. J. Operations Prod. Manage. 11 (3), 135–144.

Hunter, J.E., Gerbing, D.W., 1982. Unidimensional measurement, second order factor analysis, and causal modeling. Res. Organizational Behav. 4, 267–320.

Jöreskog, K.G., 1993. Testing structural equation models. In: Bollen, K.A., Long, J.S. (Eds.), Common Problems/Proper Solutions. Sage Publications, Newbury Park, CA, pp. 294–316.

Jöreskog, K.G., Sörbom, D., 1989. LISREL 7: A Guide to the Program and Applications, 2nd edn. SPSS, Chicago.

Kerlinger, F.N., 1986. Foundations of Behavioral Research, 3rd edn. Holt, Rinehart and Winston, New York.

Kim, J., Mueller, C.W., 1978. Introduction to Factor Analysis: What it is and How To Do It. Sage University Paper series on Quantitative Applications in the Social Sciences, Series No. 07–013. Sage Publications, Beverly Hills.

Kumar, N., Stern, L.W., Anderson, J.C., 1993. Conducting interorganizational research using key informants. Acad. Manage. J. 36 (6), 1633–1651.

Miller, J.G., Graham, M.B.W., 1981. Production/operations management: agenda for the '80s'. Decision Sci. 12 (4), 547–571.

Nunnally, J.C., 1967. Psychometric Theory. McGraw-Hill, New York.

Nunnally, J.C., 1978. Psychometric Theory, 2nd edn. McGraw-Hill, New York.

Parthasarthy, R., Sethi, S.P., 1993. Relating strategy and structure to flexible automation: a test of fit and performance implications. Strategic Manage. J. 14, 529–549.

Pedhazur, E.J., Schmelkin, L.P., 1991. Measurement, Design, and Analysis: An Integrated Approach. Lawrence Erlbaum Associates, Publishers, Hillsdale, NJ.

Phillips, L.W., Bagozzi, R.P., 1986. On measuring organizational properties of distribution channels: methodological issues in the use of key informants. Res. Marketing 8, 313–369.

Schmitt, N., Stults, D.M., 1986. Methodology review: analysis of multitrait–multimethod matrices. Appl. Psychol. Measurement 10, 1–22.

Schroeder, R.G., Anderson, J.C., Cleveland, G., 1986. The content of manufacturing strategy: an empirical study. J. Operations Manage. 6 (1), 405–415.

Schwab, D.P., 1980. Construct validity in organizational behavior. Res. Organizational Behav. 2, 3–43.

Seidler, J., 1974. On using informants: a technique for collecting quantitative data and controlling for measurement error in organizational analysis. Am. Sociological Rev. 39, 816–831.

Suarez, F.F., Cusumano, M.A., Fine, C.H., 1996. An empirical study of manufacturing flexibility in printed circuit board assembly. Operations Res. 44 (1), 223–240.

Swamidass, P.M., Newell, W.T., 1987. Manufacturing strategy, environmental uncertainty and performance: a path analytic model. Manage. Sci. 33 (4), 509–524.

Tanaka, J.S., 1993. Multifaceted conceptions of fit in structural equations models. In: Bollen, K.A., Long, J.S. (Eds.), Common Problems/Proper Solutions. Sage Publications, Newbury Park, CA, pp. 11–39.

Tunälv, C., 1992. Manufacturing strategy—plans and business performance. Int. J. Operations Prod. Manage. 12 (3), 4–24.

Upton, D.M., 1995. Flexibility as process mobility: the management of plant capabilities for quick response manufacturing. J. Operations Manage. 13 (3–4), 205–224.

Van de Venn, A.H., Ferry, D.I., 1980. Measuring and Assessing Organizations. Wiley, New York.

Venkatraman, N., 1989. Strategic orientation of business enterprises: the construct, dimensionality, and measurement. Manage. Sci. 35 (8), 942–962.

Ward, P.T., Duray, R., Leong, G.K., Sum, C., 1995. Business environment, operations strategy, and performance: an empirical study of Singapore manufacturers. J. Operations Manage. 13, 99–115.

Werts, C.E., Linn, R.L., Jöreskog, K.G., 1974. Interclass reliability estimates: testing structural assumptions. Education Psychol. Measurement 34 (1), 25–33.

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