Debi P. Mishra (USA), Somali Ghosh (USA)

Detecting item bias in latent construct between group comparisons: an illustrative example using multi-sample covariance structural equations modeling

Abstract

Analyzing, responding to, and managing group differences in consumer behavior are critical for the successful formulation and execution of strategy. By understanding such distinctions, managers can design and deliver differentiated product offerings and promotions tailored to the needs of specific demographic, socioeconomic, and psychographic segments in the marketplace. From a methodology standpoint, comparing group differences involves two broad approaches. First, qualitative techniques such as interview protocols and ethnographies can generate unique insights about both within and between group phenomena. Second, managers and marketers typically use statistical tools such as *t*-tests and multivariate analysis of variance (MANOVA) techniques to obtain quantitative estimates of group differences. While MANOVA techniques are useful, they can also distort true group differences especially when latent constructs (e.g., satisfaction) are involved. In particular, MANOVA models for latent constructs will yield optimal results only if measures possess identical psychometric properties (e.g., item to construct relationships) across groups. However, the notion of psychometric equivalence is rarely tested in practice. Hence, biased items and measurement artifacts can confound true between group differences in MANOVA models. This paper discusses how item bias can be explicitly accounted for and controlled while making between group comparisons using MANOVA. Specifically, it describes how the multi-sample covariance structural equations modeling (SEM) method can provide researchers with a better basis to assess and address item bias. We illustrate the multi-sample SEM method by analyzing empirical data on perceptual latent constructs (*performance ambiguity* and *input uncertainty*) collected from automotive repair managers. The results show that by using the SEM and MANOVA techniques in tandem, researchers can control the deleterious effects of item bias and obtain robust and meaningful estimates of between group differences. The implications of our study for future research are also discussed.

Keywords: group differences, MANOVA, structural equations, service uncertainty.
JEL Classification: C12, C18, C30, C39, C65, Z00, Z19.

Introduction

The central objective of modern business management is to understand and satisfy customer needs and wants. To achieve this goal, firms typically deploy strategies that involve: (a) identifying and studying distinct homogeneous customer groups or segments; (b) designing product and service offerings tailored to the needs of customers; and (c) deploying targeted promotions and communications program to influence appropriate behaviors. Not surprisingly, there is a rich and growing body of academic research, tools, and techniques devoted to the topic of group customer behavior (Griskevicius and Kenrick, 2013; Morgan et al., 2013). For example, researchers use tools such as in-depth interviews, focus groups, immersion studies, survey methodologies, and ethnography to better understand myriad aspects of group behavior (Batinic and Appel, 2013; Cayla and Arnould, 2013; Oritz et al., 2013).

In recent years, the emphasis on studying *within* group behavior is being supplemented by an emerging perspective that highlights the importance of isolating, identifying, and studying *between* group differences in greater detail. Notice that the focus on between group differences has always been a central part of management tasks involving segmentation, differentiation, and positioning. However, there is renewed emphasis on appraising between group distinctions given the increasing multi-ethnicity and rapid globalization of today's marketplace. In particular, marketers have to devote considerable energy toward creating differentiated product offerings for a number of distinct racial, ethic, and religious groups in traditionally homogeneous domestic markets. For example, McDonald's, which historically offered a standardized menu of products in North America, today follows a strategy of "leading with ethic insights" (Helm, 2010). By understanding differences across ethnic groups, the company has "discovered how dramatically minority tastes can influence mainstream preferences" (Helm, 2010, p. 1). Similarly, Wal-Mart has selectively modified its standardized retail approach by introducing variations such as the 'neighborhood store' targeted to the Hispanic community (Flores, 2011). In sum, the ability of firms to efficiently conceptualize, understand, and measure between group differences is critical to their marketing prowess and success.

Given the importance of studying group differences in consumer behavior, accurate measurement of differential effects is a significant task for managers. However, as we subsequently show, most consumer group differences are measured by using simple methodologies that might not yield optimal results.

Hence, managers should focus on selecting methodologies that are most appropriate for gauging group differences.

Typically, group differences are measured via simple *t*-tests or multivariate analysis of variance (MANOVA) designs. The general *k* sample MANOVA effectively reduces to the common analysis of variance (ANOVA) method when comparing *two* samples. Intuitively, MANOVA computes the ratio of between to within group variances, and a statistically significant ratio (high value of *F* or Wilks Lambda) indicates meaningful group differences. The straightforward logic of MANOVA designs and the availability of easy to use statistical software have led to its widespread use for measuring group differences in marketing and other social sciences.

Despite the popularity of MANOVA, precise and meaningful measurement of group differences can be compromised under certain conditions. In particular, one important assumption of MANOVA is that items being compared across groups should be measured in an equivalent metric. For example, if an item such as average income is compared across two countries without accounting for exchange rate conversions, the observed difference is meaningless. In general, such item bias can be eliminated by suitably transforming variables into a common metric before undertaking group comparisons.

Controlling item bias is possible when an acceptable and accurate transformation method (e.g., currency exchange rate) is available to convert variables from one scale to another. For relatively tangible items such as height, weight, and temperature, item bias can be completely eliminated using appropriate conversions (e.g., Metric to British and vice versa). In contrast, when items represent perceptual or latent constructs (satisfaction, service quality), achieving meaningful conversions is not a straightforward exercise.

A critical question facing researchers is whether differences (or similarities) across groups on latent variables are confounded by measurement artifacts or not. In other words, assuming that a set of items validly measures a construct in one particular group, is it a sufficient condition for meaningful comparison across groups? Unfortunately, as we address in the following sections, within group validity of items is a necessary but not sufficient condition for meaningful between group comparisons. For instance, there may be biased items which provide valid measurement within a group but have the potential to confound across group comparisons.

Given the importance of item bias and its potential to distort the results of MANOVA designs, the objective of this paper is to delineate and demonstrate an approach for detecting bias before undertaking group difference tests. Specifically, we outline a step-wise procedure using multi-group structural equations modeling (SEM) for detecting and accounting for item bias when comparing latent constructs. In particular, we demonstrate the use of our approach by analyzing data regarding the perceptual constructs of performance ambiguity and input uncertainty which are present in the interface between service firms and customers. Our empirical results show that the SEM approach can be gainfully employed by researchers to account for item bias in group comparisons.

An appraisal of item bias has two key implications for management scholars. First, by deleting or suitably modifying biased items, comparison of latent constructs becomes meaningful. For instance, two groups may have the same score on a latent construct, but the presence of biased items may cause their observed scores to differ. In such a situation, detecting and modifying biased items will facilitate meaningful group comparisons via subsequent MANOVA designs. Second, with the increasing globalization of markets, there will be a renewed emphasis on the part of marketers to compare latent constructs across cultures using U.S. scale items as a starting point. Hence, the detection of item bias must necessarily precede any attempt at comparing latent constructs using MANOVA and related designs.

The remainder of this paper is organized as follows. First, we discuss the concept of item bias and depict the procedure for detecting bias using multi-group structural equations modeling. Second, we describe the conceptual framework pertaining to the perceptual constructs of performance ambiguity and input uncertainty, and hypothesize expected mean differences across groups (categories) of automotive repair services. In the third section, we delineate the data collection procedure and provide evidence of construct validity for the perceptual constructs. The fourth section depicts the initial results of a MANOVA test conducted to uncover mean differences across service categories. In the following section, we implement the proposed SEM approach by analyzing data regarding the perceptual constructs of performance ambiguity and input uncertainty. In the penultimate section, we show how detection and elimination of item bias enhances the accuracy of MANOVA designs. We conclude by discussing the implications of our research.

1. The concept of item bias

An item is biased if it does not measure the same latent construct in an equivalent way between two groups even if it validly measures the construct within a group. For example, the cost of a product (construct) can be validly measured in each group using the local price (scale item). However, local price may not provide a direct basis for comparison because of different exchange rates between countries. The item "local price" is therefore a biased item. The presence of item bias implies that groups cannot be meaningfully and validly compared without first transforming items to a common scale.

Transformation of scale items and removal of item bias is relatively straightforward in the physical sciences where concepts can be observed and accurately measured. For example, heat (concept) can be measured with a Celsius or a Fahrenheit thermometer (scale) in two situations and compared using a suitable transformation (from Celsius to Fahrenheit and vice versa). Unfortunately, such simple linear transformations are impossible to undertake when measuring latent (unobservable) constructs that do not possess unique measures.

Strictly speaking, latent constructs cannot be directly compared across groups on mean scores. For instance, parameter values (population estimates) linking the scale item (x) with the latent construct (ξ) may differ across groups, thereby rendering direct comparison of scores meaningless. As an example, assume that males and females actually have the same standing on a latent construct such as role conflict (ξ). However, a particular measure of role conflict (x) may be related to the latent construct in the female group such that $x = \xi + \delta$, while for males the corresponding relationship might be $x = 0.7\xi + \delta$ (δ is the error term). In such a situation, even if the two groups have the same standing on a latent construct, the mean score of the role conflict measure will be higher for females because x is a biased item. Meaningful comparison of latent constructs is therefore not possible in the presence of item bias.

As a first step toward ensuring meaningful and valid comparisons, measures should demonstrate identical parameter estimates across groups. This can be achieved if items have the same factor loadings and error variances across groups. Note that according to principles of psychometric theory, there are two sources of variance for the measure of a latent construct: (a) the true variance (squared loading), and (b) error variance (random and specific). If these variance components for an item are identical across groups, the item is not biased. Furthermore, equality of variance components is a more stringent requirement than the traditional factor matching procedure typically used in the management literature. For example, studies by Weinberger and Spotts (1989) and Seymor and Lessne (1984) have administered the same set of items to two (or more) groups and then compared the resulting factor patterns based on significant loadings. In other words, even if item i in group 1 loads 0.8 on a factor (ξ), and loads 0.6 on the same factor (ξ) in group 2, a numerical comparison of scales is meaningless even though a similar factor structure is suggested. In sum, meaningful comparison is possible only if the items are not biased.

Based on the preceding discussion, item bias can be detected by constraining the loadings and error variances for items to be equal across groups and estimating parameters simultaneously. The detection of bias is easily accomplished using the multi-group confirmatory structural equations modeling approach pioneered by Joreskog (1971). Briefly, parameter values (i.e., loadings and error variances) are estimated for the two (or more) groups as if the data came from a single population. In the next step, loadings in the two groups are constrained to be equal and the goodness of fit indices are inspected. If the model fits the data, the next step is to constrain the variances across the two groups to be equal and inspect the goodness of fit. If the model does not fit the data, items causing bias (misfit) can be detected using the Lagrange Multiplier test (L-M test) and then targeted for remedial action. A stepwise procedure for implementing the multi-sample SEM approach is depicted in the Appendix. We implement this procedure with a view to detecting and removing item bias while measuring group differences for the perceptual constructs of input uncertainty and performance ambiguity. The next section provides the conceptual discussion for these perceptual constructs.

2. Conceptual framework and research hypotheses

This section discusses the rationale for expecting differences in the values of uncertainty constructs across different groups (categories) of automotive repair services. First, we describe the concepts of performance ambiguity and input uncertainty in detail, followed by specific hypotheses regarding construct validity (H1) and group differences (H2).

Uncertainty is the difference between the amount of information required to perform a task and the degree of information already possessed by an organization (Galbraith, 1977; Larsson and Bowen, 1989). Past studies (Burns and Stalker, 1961; Jurkovich, 1974;

Lawrence and Lorsch, 1967) suggest a two-step approach for dealing with uncertainty. First, firms scan their environments to isolate salient dimensions of uncertainty (Tosi and Slocum, 1984). Second, organizations tackle differential information by adopting specific internal governance structures (Mishra, Heide, and Cort, 1998), or by providing additional information to exchange partners (Jones, 1987; Larsson and Bowen, 1989).

Consider the dimensions of uncertainty first. Typically, research in this area represents a movement from complex and diffuse con-ceptualizations of the environment toward more parsimonious and simpler facets. Mills and Margulies (1980) made the first formal attempt to identify conceptual dimensions of the client-service firm interface. These authors suggested that seven independent facets described the service organi-zation-customer interface, i.e., (a) *information*, involving quality, quantity, and confidentiality, (b) *decision*, or the nature and importance of employee decisions, (c) *time*, or the contact duration between employee and client, (d) *problem awareness*, indicating the extent to which clients are knowledgeable about problems and whether they can evaluate services, (e) *transferability*, or whether employees are substitutable or not, (f) *power* involving notions of dependency between employees and clients, and (g) *attachment* implying conflict potential between employees and clients.

Though multidimensional conceptualizations of the Mills and Margulies (1980) type have merit, the lack of mutual exclusivity among uncertainty dimensions undermines their use (Snyder, Cox and Jesse, 1982). For example, there might be a complex non-linear relationship between the time and information dimensions in the Mills and Margulies (1980) typology. Specifically, in certain situations, customers may spend a lot of time interacting with a service provider (e.g., a restaurant waiter) without exchanging any critical information. On the other hand, customers might exchange critical and sensitive information in a short span of time while interacting with the provider (e.g., a physician). Hence, treating time and information as independent dimensions is perhaps not justified. It is therefore possible that a more macro concept of uncertainty might subsume subordinate dimensions like time and information.

Recognizing the importance of carefully delineating the domain of uncertainty facing a service organization, Argote (1982) called for a movement "from diffuse characterizations of an organization's environment or task to more precise descriptions of uncertainty characterizing a particular element of an organization's task environment" (p. 422). Hence, later studies have tended to focus on more fundamental dimensions of customer uncertainty. Specifically, researchers (Argote, 1982; Bowen and Jones, 1986; Jones, 1987; Jones, 1990; Larsson and Bowen, 1989; Mills and Margulies, 1980; Siehl, Bowen, and Pearson, 1992) have conceptualized the service organization-customer interface along *two* main dimensions: (a) *performance ambiguity*, or customers' inability to assess quality *ex-ante*, and (b) *input uncertainty* or the complex demands that customers place on service organizations. These dimensions have received empirical validation in a recent study by Mishra (2013).

2.1. Performance ambiguity and input uncertainty. According to Bowen and Jones (1986), "performance ambiguity arises when any dimension of an exchange makes it difficult for one party to evaluate the performance of the other" (p. 431). In a similar vein, Jones (1990) notes that "performance ambiguity is particularly prevalent when the goods or services being purchased are intrinsically complex, and their quality can only be really evaluated after purchase and use" (p. 24). In general, performance ambiguity represents a form of information asymmetry or a situation where one party to the exchange possesses more information than the other party (Rao and Bergen, 1992).

Performance ambiguity creates uncertainty for firms because additional information has to be provided to customers for completing a transaction. It may be noted that customers possess limited ability to evaluate quality prior to purchase since services possess experience (Nelson, 1970) and credence (Darby and Karni, 1973) properties. While experience attributes can be evaluated only after use (e.g., "Was the restaurant meal delicious?"), credence properties cannot be evaluated by customers even after consumption (e.g., "Did I really need that much of automotive repairs?"). Faced with such ambiguity, customers may postpone their buying decision, or may even switch to firms which provide additional information to customers.

While performance ambiguity is related to customers' lack of information about a service, input uncertainty is related to organizations' lack of information about the nature of customer demands. As Larsson and Bowen (1986), note, input uncer-tainty is "the organization's incomplete information about what, where, when, and how customer input is going to be processed to produce desired outcomes" (p. 217). The concept of input uncertainty is similar to the notion of variability in raw materials (Perrow, 1967; Woodward, 1970). Variability is described by

Overton, Schneck, and Hazlett (1977) as "differences between patients in the degree of criticalness of their health problems and the frequency with which emergency situations may occur" (p. 205). The general implication of input uncertainty is that firms may lack information to design and incorporate standardized processes in the service production and delivery system.

In sum, it appears that *two* dimensions of uncertainty characterize the firm-customer interface. First, since a service is intangible, customers experience *performance ambiguity*, or difficulty in evaluating a service prior to consumption. Second, firms have to deal with *input uncertainty* or disturbances created by customer presence and participation during the production and delivery of a service.

It may be noted that the two dimensional categorization of the customer-firm interface also incorporates the defining characteristics of services. For instance, *intangibility*, which is defined as a condition where services "cannot be seen, felt, tasted, or touched in the same manner in which goods can be sensed" (Zeithaml, Parasuraman, and Berry, 1985, p. 33) is closest in meaning to customer performance ambiguity. This view is endorsed by Siehl, Bowen, and Pearson (1992) who explicitly define intangibility as performance ambiguity. The other characteristics of services, i.e., *inseparability of production and consumption* (Booms and Nyquist, 1981; Bateson, 1989), *heterogeneity*, and *perishability* (Berry, 1980; Booms and Bitner, 1980) are related to input uncertainty. Figure 1 depicts the uncertainty dimensions for a service firm.

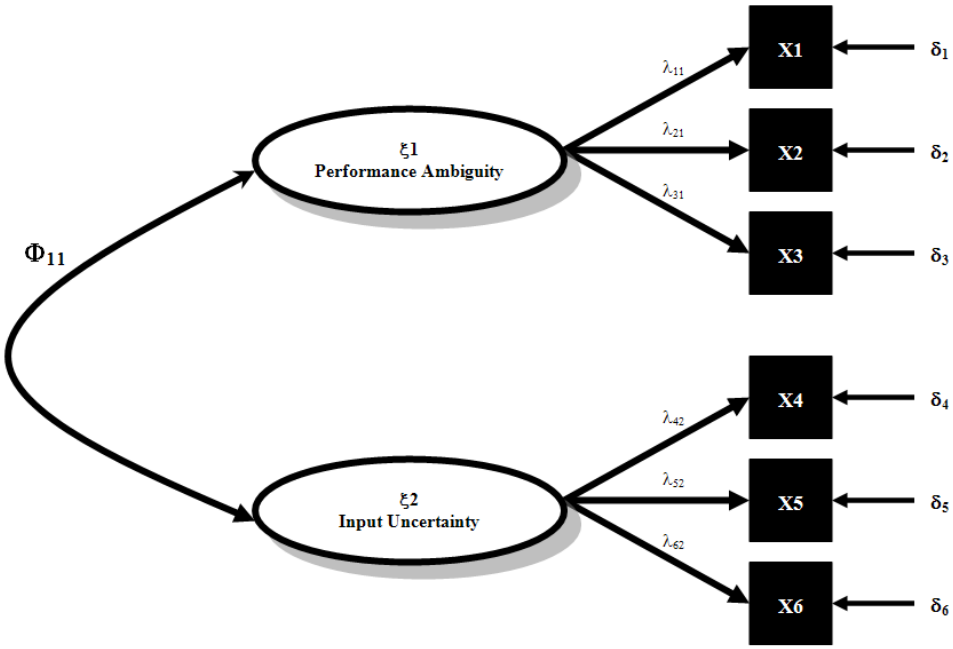

Fig. 1. Conceptual model of performance ambiguity and input uncertainty

Note: X represents measured variables; δ represents error in measured variables; ξ represents latent constructs; ϕ represents inter-factor correlation; λ represents factor loadings.

2.2. Hypotheses. In view of the above discussion, the following hypotheses are offered for empirical testing. Please note than our central objective is to investigate H2, i.e., if bias indeed influences measured differences for the three categories of automotive repair services varying along the continuum of information asymmetry.

H1: Uncertainty in the service organization-customer interface can be described by the two distinct dimensions of (i) performance ambiguity, and (ii) input uncertainty.

H2: Performance ambiguity and input uncertainty will differ across service categories (groups) possessing different levels of information asymmetry. For instance, uncertainty will be lowest for standardized oil change automotive repair service, at an intermediate level for brake service, and highest for transmission service.

3. Research design

The present study used a mail survey to gather data about different theoretical concepts. A mail survey was used because data pertaining to focal constructs like performance ambiguity and input uncertainty are unlikely to be found in secondary data sources like company annual reports. Customer service managers in automotive service outlets provided data about the different variables. The survey was administered to customer service managers in North America who

were employed by different automotive service repair firms as identified by relevant "Standard Industry Classification" (SIC) codes.

3.1. Survey context. Three categories of firms providing specialized automotive services were chosen, i.e., (1) lubrication services (SIC code 7549-03), (2) brake services (SIC code 7539-14), and (3) transmission repair services (SIC code 7537-01). These categories were selected based upon Cook and Campbell's (1979) guidelines on survey design. Specifically, we expected the concepts and phenomenon being investigated like performance ambiguity to naturally occur in the automotive service setting. Furthermore, it was hoped that the present context would provide adequate variation and co-variation among the theoretical concepts of interest.

A careful consideration of the literature suggests that constructs like "performance ambiguity" are widely present in the automotive service sector. For instance, a number of studies have noted that customers experience considerable ambiguity while evaluating automotive services (Andaleeb and Basu, 1994; Heskett, Sasser, and Hart, 1990; Hubbard 1998; Schleisinger, 1992). As a consequence of this performance ambiguity, automotive service customers are more likely to be significant complainers (Day and Bodur, 1978).

The automotive service sector also provides evidence of suppliers' use of signals like certification and customer bonds. For example, Biehal (1983, p. 90) notes that automotive repair firms stress certification of mechanics (mechanic's competence) to better communicate with customers. Likewise, Grove, Fisk, and Bitner (1992) document how automotive service establishments might post customer bonds by investing in their physical surroundings.

As noted earlier, three different service categories (transmission service, brake service, and lubrication service) were chosen to maximize variability among concepts. The extant literature (Iacobucci, 1992) provides evidence of variability across automotive service categories for constructs such as performance ambiguity. For instance, Iacobucci's (1992) study found that customers experienced greater performance ambiguity for general car repair service as opposed to brake relining. Furthermore, customers relied on more quality cues as performance ambiguity increased (Iacobucci, 1992). In a similar vein, Biehal (1983) noted that customers who desired preventive and scheduled maintenance service like oil changes perceived less ambiguity than customers who demanded more complex services like general repair. Given the preceding theoretical literature and the results of a pilot study that was conducted, we expected that on a continuum of performance ambiguity, lubrication service would be at the low end (Biehal, 1983), while car repair service like transmission service would occupy the high end. We expected brake service to occupy an intermediate point on the continuum (Iacobucci, 1992). Hence, the three sub-categories of automotive repair service were chosen to maximize variability among theoretical constructs in the model.

3.2. Sampling frame. Three different national mailing lists corresponding to transmission service, brake service, and lubrication service were obtained from a commercial list broker. Each list provided the name and address of companies, selected demographic data and the name and telephone number of a prospective key informant for each company. The purchased list for each category consisted of 3042 names.

Given the time and resources needed to conduct a census of each list, and keeping in mind the overall sample size requirements for obtaining statistical results, the initial list for each category was trimmed in a systematic way. First, duplication (of key informant names and company names) was removed after visual inspection. Removal of duplicate names from each list resulted in 2892 names for the transmission category, 2916 names for the brake service category, and 2800 names for the lubrication service category.

3.3. Questionnaire administration. Before the questionnaire was administered, efforts were made to identify key informants within each company, using the procedures recommended by Heide and John (1990) and Kumar, Stern, and Anderson (1993). Questionnaires were mailed to all firms that agreed to participate in the survey. The questionnaire packet consisted of a cover letter, a pre-paid envelope, and the questionnaire itself. In order to motivate firms to respond, they were offered an incentive in the form of a report that summarized the results of the study. Five weeks after the initial mailing, a reminder postcard was sent to all non-respondents.

3.4. Response rate and assessment of non-response bias. The response rate obtained in this study is 22.6% (287 completed questionnaires). It may be noted that although the response rate appears to be low, it is consistent with acceptable response rates reported in research employing similar research designs (Morgan and Hunt, 1994; Rao and Mahi, 2003). All subsequent empirical analysis is based upon these 287 responses.

To investigate whether non response bias was an issue in the present study, an extrapolation method (Armstrong and Overton, 1976) was adopted. The basic logic of this approach is that late respondents will most likely approximate the characteristics of non respondents. Statistically, a *t*-test comparing the means of key demographic variables across the early and late responding groups should not be significant.

To investigate non-response bias, total responses were divided into two groups on the basis of a cut-off date designated as the completion of the fourth week after the last survey was mailed. Based on our results, we could not reject the null hypothesis of no mean differences across the early and late responding groups for "number of employees" (t = = 1.43, p = 0.153) and "annual sales volume" (t = = 1.36, p = 0.174).

3.5. Measures. The measures and their item reliabilities are depicted in Table 1.

Performance ambiguity refers to difficulties faced by customers in evaluating a service. Three items were used to measure this construct, based on ones previously developed by Jones (1987).

Input uncertainty refers to the degree to which organizations experience variability in service demands made by customers. Three items were used based on an earlier study by Jones (1987).

Table 1. Scale items and reliability

Construct	Scale items	Format	Reliability
Performance ambiguity	♦ Customers have to assume that they are getting good service from us because there is no other way they can tell. ♦ It would be very time consuming for a customer to check up on how well a mechanic is performing his or her job. ♦ Customers can easily determine the amount of service that is needed by them (reverse coded).	7-point Likert scale with "strongly disagree" and "strongly agree" as anchors.	0.80
Input uncertainty	♦ Mechanics usually encounter the same problems in their day-to-day work. ♦ Customers often come up with problems that mechanics have never encountered before. ♦ The work performed by mechanics does not vary a lot for one customer to another.	7-point Likert scale with "strongly disagree" and "strongly agree" as anchors.	0.66

4. Examination of construct validity

To assess the construct validity of measures, items were first submitted to a common factor analysis. A two-factor model pertaining to the two constructs in the study was estimated. As per accepted practice (Norusis, 1991), the maximum likelihood procedure was used to extract a range of factor solutions. Specifically, the two factor structure was successively re-specified as a 2, 3, and 4 factor model respectively. By submitting items to a range of factor solutions, one can "pick" the best one by a relative comparison of estimates (such as the overall root mean square of off diagonal elements in the reproduced correlation matrix) across various models (Norusis, 1991).

In the present analysis, the lowest residuals were observed for the hypothesized two factor structure. Furthermore, the scree plot which depicts the relationship between Eigen values and the number of factors was also inspected. A sharp break corresponding to the second factor suggests that the two-factor model is adequate.

Next, all factor loadings in the pattern matrix of the two-factor solution were inspected. As Stevens (1992) suggests, the cut-off for identifying a significant loading is determined by sample size as well as overall alpha (Stevens, 1992, p. 383). Given the present sample size, and using an alpha value of 0.05, a cut-off level of 0.36 was used to identify variables to be retained for further analysis. At this point, the Cronbach alpha (Cronbach, 1951) value for each scale was computed and items exhibiting poor item-to-total correlations were dropped. Only those items exhibiting alpha values of 0.6 or greater were retained (Peterson, 1994).

5. Test of hypotheses

To empirically test various hypotheses, Latent Variable Structural Equations (LVSE) modeling was adopted. LVSE appears to be particularly well suited for analyzing causal structures like the one in the present study, for a number of reasons. First, this approach enables researchers to estimate the strength of relationships among *latent* variables in a model. Second, LVSE permits the *simultaneous estimation* of relationships among latent and observable variables. Finally, the LVSE procedure can be used by researchers to estimate the effect of *measurement error* in latent and observable variables on parameter estimates as well as on overall model fit.

The goodness-of-fit for various LVSE models was assessed by using multiple criteria. Specifically, an average off diagonal standardized residuals (AOSR) value of 0.06 or lower was used as evidence of good

fit (Bagozzi and Yi, 1988). Furthermore, a Comparative Fit Index (CFI) value of 0.9 or higher was used for assessing the degree of overall fit, as recommended by Bentler (1990). Convergent validity was determined by inspecting the parameter estimates of each restricted model. Specifically, large (> 0.4), positive, and statistically significant estimates (t > 2) indicated that loadings were not trivially different from zero. Finally, discriminant validity was assessed by restricting factor intercorrelations pair wise to 1 and then computing a χ^2 difference statistic with 1 degree of freedom (Bagozzi and Yi, 1988). A significant χ^2 difference test provides evidence of discriminant validity.

5.1. Test for hypothesis 1 (H1). To investigate H1, or whether performance ambiguity and input uncertainty represented distinct dimensions in the firm-customer interface, a two-factor measurement model depicted in Figure 1 was estimated. The various parameter estimates pertaining to this model are depicted in Table 2.

As can be seen from Table 2, all measures load on their hypothesized factors and estimates are positive and significant, providing evidence of convergent validity (Bagozzi and Yi, 1988). For instance, estimates range from .40 (t = 5.51) for one item measuring the input uncertainty construct to .91 (t = = 10.34) for another item measuring the same construct.

To investigate discriminant validity between performance ambiguity and input uncertainty, a nested estimation procedure was adopted (Howell, 1987). First, a baseline model (M0) was estimated by allowing performance ambiguity and input uncertainty to be correlated. Next, the chi-square (χ^2) and degrees of freedom (df) were calculated for a restricted model (M1) in which the correlation between performance ambiguity and input uncertainty was fixed to unity. The χ^2 difference between M0 and M1 ($\chi^2_{M1} - \chi^2_{M0}$) for 1 degree of freedom provides a statistical test of the null hypothesis that M0 and M1 represent the same model. Hence, rejection of the preceding null hypothesis implies that M1 and M0 differ. Stated differently, a significant χ^2 difference test indicates that the two constructs in question are distinct because fixing a correlation to unity (as in M1) does not improve the goodness-of-fit of the baseline model. In the present study, χ^2_{M1} = 34.44 (df_{M0} = 8) and χ^2_{M0} = 161.91 (df_{M1} = 9). Hence, the χ^2 difference test is significant ($\Delta\chi^2$ = 127.47, df = 1, p < 0.001), implying that measures of performance ambiguity and input uncertainty are distinct.

Finally, overall goodness of fit estimates for the two-factor measurement model suggest that the hypothesized factor structure reproduces observed correlations within sampling error. For instance, both the average off-diagonal squared residual (0.06) and CFI (.95) values indicate good fit. Hence, H1, which hypothesizes that performance ambiguity and input uncertainty are distinct dimensions, is supported.

Table 2. Parameter estimates of construct validity

(Construct)	Estimate[a]	T[b]
Performance ambiguity (ξ_1)		
λ_{11}	.81	14.29
λ_{21}	.82	14.45
λ_{31}	.67	11.60
Input uncertainty (ξ_2)		
λ_{42}	.70	8.96
λ_{52}	.91	10.34
λ_{62}	.40	5.51
Overall goodness of fit indices		
χ^2 = 34.44, df = 8, p = <0.001		
AOSR[c] = 0.06		
CFI[d] = 0.95		

Notes: [a] Standardized factor loadings computed via EQS using the iteratively re-weighted generalized squares method. [b] Tests the null hypothesis that the parameter is zero. [c] Precentage of average off-diagonal squared residuals in the reproduced correlation matriz. [d] Comparative Fit Index (Bentler, 1990).

5.2. Test for hypothesis 2 (H2). To test H2, a multivariate analysis of variance (MANOVA) model was estimated with the three service categories (lubrication, brake, transmission) as treatment variables, and firm-client interface constructs (performance ambiguity, input uncertainty) as dependent factors. Recall that the three service categories were selected because of expected variation in the focal constructs of performance ambiguity and input uncertainty. Specifically, input uncertainty and

performance ambiguity are expected to be highest in the transmission category, lowest in the lubrication category, and of intermediate value in the brake category. Hence, if post-hoc analysis corresponds to this pattern of variation, it further supports the hypothesized two factor conceptualization.

Turning to the MANOVA results, the null hypothesis (H_0: The population mean vectors of dependent variables are equal across the three sub-groups) cannot be rejected because both the Pillai's Trace and Wilks Lambda values are non-significant. In other words, the results of this analysis show that the expected hypothesized group differences across the three different service categories is not supported.

5.3. Multigroup SEM analysis and additional tests of hypothesis 2 (H2). Recall that hypothesis 2 suggests expected variation across service categories. However, the above MANOVA results suggest that the hypothesis is unsupported. Recall from an earlier discussion that one important reason for not observing hypothesized group differences could be due to the presence of item bias. To investigate if item bias is an issue in the present analysis, we conducted a multi-group structural equations modeling exercise that we describe below.

We closely followed the multi group procedure depicted in the Appendix. The results of this analysis show that measurement equivalence no longer holds when the constraints of equal error variance are added to those of equal factor loadings ($\Delta\chi^2$ = 32.4, df = 7, p < 0.005). In other words, automatically fixing the error variances to be equal across groups has probably led to bias, which in turn, has compromised group differences. To determine which item(s) contribute to cross-group bias, the L-M test for each constraint was inspected. The results imply that one of the performance ambiguity items titled "It would be very time consuming for a customer to check up on how well a mechanic is performing his or her job" is a source of bias when comparing means across the two groups (χ^2 = = 19.812, p < 0.0005).

In the next step we re-estimated the MANOVA model by deleting the above performance ambiguity item. The results of this new MANOVA analysis depicted in Table 2 show variance along expected lines. Specifically, both Pillai's Trace (Statistic = = 0.67, F = 3.6, p = 0.00) and Wilks Λ (Statistic = = .37, F = 5.12, p = 0.00), are significant, implying that both performance ambiguity and input uncertainty vary across the three service categories. Second, post-hoc univariate t tests comparing performance ambiguity and input uncertainty across service categories are significant and in the expected direction. Taken together, results of the MANOVA and CFA models provide strong support for H1 and H2. In summary, the presence of item bias affected the initial multi-group comparisons. However, after eliminating bias, group differences in uncertainty across the three service categories appear to be in the right direction.

Table 3. MANOVA results

	Treatments			Univariate post hoc comparisons[a]			
	Transmission (N = 82)	Brake (N = 90)	Lube (N = 115)	Transmission vs. Brake		Transmission vs. Lube	
Dependent variables	Means			t	p	t	p
Performance ambiguity	5.46	4.27	2.54	3.75	.001	4.88	.000
Input uncertainly	4.17	3.06	2.19	3.52	.002	4.39	.000
Multivariate statistics[b]							
Statistic	Value		F		p		
Pillai's trace	0.68		3.6		.002		
Wilks Λ	.37		4.36		.000		

Notes: [a] Tests the null hypothesis that there is no difference in mean values for dependent variables across treatment groups. [b] Tests the null hypothesis that there is no difference in population mean vectors for dependent variables across treatment groups.

Discussion and conclusion

The aim of this paper has been three-fold: (a) to discuss how item bias can confound group comparisons for latent variables, (b) to delineate a hybrid approach using MANOVA and multi-group structural equations modeling to identify and eliminate bias, and (c) to apply the suggested approach to a real life setting and investigate if item bias can confound group differences for perceptual constructs in the service firm-customer boundary.

We implemented the above procedure by studying the concepts of input uncertainty and performance ambiguity that characterize the service firm-customer interface. In brief, service firms face unpredictable customer demands (input uncertainty) and also react to customers' inherent uncertainty about services (performance ambiguity). There is unequivocal theoretical evidence to suggest that different types of services will exhibit different levels of performance ambiguity and input uncertainty. For example,

consider automotive repair service firms that range from outlets providing simple standardized oil changes to more complex units that undertake diagnosis and cure tasks (e.g., engine transmission). The straightforward expectation is that group differences for the latent constructs of uncertainty across service categories will be significant and meaningful. In other words, the lowest uncertainty will be observed for the oil change group, while the transmission category is expected to possess the highest score.

As we have described, our initial empirical finding runs counter to the expected variation in uncertainty across service categories. However, after implementing a multi-group structural equations modeling procedure, we could identify and eliminate one item that potentially caused bias across groups. Elimination of this biased item produced the expected hypothesized results. In other words, the combined use of MANOVA and structural equations modeling provides an efficient approach for investigating group differences. These results imply that if true uncertainty differences indeed exist, but managers cannot uncover them, a standardized (one suit fits all) strategy may be deployed across all service categories, when in reality, a customized approach is required. This potential mismatch in interventional approaches might lead to poor organizational and customer outcomes.

A significant development in the current world economy has been the emergence of distinct institutions such as emerging economies, multiethnic customers, and newer forms of organizations. A key implication of this shift is that managers are keen to study how such forms compare with established models. By understanding group differences, managers will be in a better position to formulate and execute effective strategy. For example, today the mushrooming of on-line firms that conduct global e-commerce is well accepted. However, very little is known about how customer service expectations and quality compare between bricks and mortar and online shopping services. The key to understanding these differences is to use robust methodological tools for group comparisons. Our study provides a discussion of how extant comparison approaches can be significantly enhanced by considering the deleterious effects of item bias in confounding true group differences. We hope that the combined MANOVA and structural equations approach will go a long way in uncovering true group differences for managers.

From a managerial standpoint, detection of item bias is particularly relevant in today's multiethnic society where observed group differences may be caused by the nature and wording of a particular item than due to true differences on latent constructs and variables. If managers erroneously conclude that groups differ on a latent construct when in reality they do not, resources spent for targeting a particular group would be misdirected. Conversely, the presence of item bias may lead managers to conclude that groups do not differ on latent constructs when in reality they do. In such a situation, some market segments will not be identified or targeted, resulting in lost opportunity for increased market share. In any case, bias is a measurement artifact that confounds group difference measurements. By identifying and tackling bias, managers may be better equipped to understand the needs of diverse ethnic groups in society, thereby ensuring the optimal allocation of resources. For these reasons, we advocate increased attention to the issue of item bias measurement.

References

1. Andaleeb, Syed S., and Amiya K. Basu (1994). Technical Complexity and Consumer Knowledge as Moderators of Service Quality Evaluation in Automobile Service Industry, *Journal of Retailing*, Vol. 70 (4), pp. 367-382.
2. Argote, Linda (1982). Input Uncertainty and Organizational Coordination in Hospital Emergency Units, *Administrative Science Quarterly*, Vol. 27, pp. 420-434.
3. Armstrong, J.S., and T. Overton (1976). Estimating Nonresponse Bias in Mail Surveys, *Journal of Marketing Research*, Vol. 14 (August), pp. 396-402.
4. Bagozzi, Richard P., and Youjae Yi (1988). On The Evaluation of Structural Equation Models, *Journal of the Academy of Marketing Science*, Vol. 16 (Spring), pp. 74-79.
5. Bateson, John E. (1989). *Managing Services Marketing*, London: Dryden Press.
6. Batinic, B. & Appel, M. (2013). Mass communication, social influence, and consumer behavior: Two field experiments, *Journal of Applied Social Psychology*, Vol. 43 (7), pp. 1353-1368.
7. Bentler, Peter M. (1990). Comparative Fit Indices in Structural Models, *Psychological Bulletin*, Vol. 107, pp. 238-246.
8. Berry, Leonard L. (1980). Service Marketing is Different, *Business*, Vol. 30 (May-June), pp. 24-29.
9. Biehal, Gabriel J. (1983). Consumers' Prior Experiences and Perceptions in Auto Repair Choice, *Journal of Marketing*, Vol. 47, Summer, pp. 82-91.
10. Booms, B.H. and M.J. Bitner (1980). New management tools for the successful tourism manager, *Annals of Tourism Research*, Vol. 7 (3), pp. 337-352.

11. Booms, Bernard H., and Jody Nyquist (1981). Analyzing the Customer/Firm Communication Component of the Service Marketing Mix, in *Marketing of Services*, J.H. Donnelly and W.R. George (eds.), Chicago: American Marketing Association, pp. 172-177.
12. Bowen, David E. and Gareth R. Jones (1986). Transaction-Cost Analysis of Service Organization-Customer Exchange, *Academy of Management Review*, Vol. 11 (2), pp. 428-441.
13. Burns, T. and G. Stalker (1961). *The Management of Innovation*, London: Tavistok.
14. Cayla, J. and Arnould, E. (2013). Ethnographic Stories for Market Learning, *Journal of Marketing*, Vol. 77 (4), pp. 1-16.
15. Cook, T.D. and D.T. Campbell (1979). *Quasi-experimentation: Design & Analysis Issues For Field Settings*, Boston: Houghton Mifflin.
16. Cronbach, L.J. (1951). Coefficient Alpha and the Internal Structure of Tests, *Psychometrika*, Vol. 6, pp. 297-334.
17. Darby, M.R. & E. Karni (1973). Free Competition and the Optimal Amount of Fraud, *Journal of Law and Economics*, Vol. 16 (April), pp. 67-86.
18. Day, R.L., and M. Bodur (1978). Consumer Response to Dissatisfaction with Services and Intangibles, In *Advances in Consumer Research*, H. Keith Hunt (ed.), Ann Arbor: MI, Association for Consumer Research.
19. Flores, Aleen B. (2011). Walmart's New Approach in Socorro Launched, *El Paso Times*, July 20 (accessed online at http://www.elpasotimes.com/news/ci_18312763?source=most_viewed).
20. Galbraith, J. (1977). *Organizational Design*, Reading MA: Addison-Wesley.
21. Griskevicius, V. and Kenrick, D.T. (2013). Fundamental Motives for Why We Buy: How Evolutionary Needs Influence Consumer Behavior, *Journal of Consumer Psychology*.
22. Grove, Stephen J., Raymond P. Fisk, and Mary Jo Bitner (1992). Dramatizing the Service Experience: A Managerial Approach, *Advances in Services Marketing and Management: Research and Practice*, Vol. 2, Teresa A. Swartz, David E. Bower, and Stephen W. Brown (eds.), Greenwich, CT: JAI Press.
23. Heide, Jan B. and George John (1990). Alliances in Industrial Purchasing: The Determinants of Joint Action in Buyer-Supplier Relationships, *Journal of Marketing Research*, Vol. 27 (February), pp. 24-36.
24. Helm, Burt (2010). Ethic Marketing: McDonald's Is Lovin' It, *Bloomberg Businessweek*, July 08 (accessed online at: http://www.businessweek.com/printer/articles/50558-ethnic-marketing-mcdonalds-is-lovin-it).
25. Heskett, James L., W. Earl Sasser, Jr., and Christopher W.L. Hart (1990). *Service Breakthroughs: Changing the Rules of the Game*, New York: The Free Press; p. 107.
26. Howell, Roy D. (1987). Covariance Structure Modeling and Measurement Issues: A Note on Interrelations Among a Channel Entity's Power Sources, *Journal of Marketing Research*, Vol. 24 (February), pp. 119-126.
27. Hubbard, Thomas N. (1998). An Empirical Examination of Moral Hazard in the Vehicle Inspection Market, *RAND Journal of Economics*, Vol. 29 (2), Summer, pp. 406-426.
28. Iacobucci, Dawn (1992). An Empirical Investigation of Some Basic Tenets in Services: Goods-Services Continua, *Advances in Services Marketing and Management*: *Research and Practice*, Vol 1, Teresa A. Swartz., David E. Bowen, and Stephen W. Brown (eds.), Greenwich, CT: JAI Press, pp. 23-52.
29. Jones, Gareth R (1987). Organization-Client Transactions and Organizational Governance Structures, *Academy of Management Journal*, Vol. 30 (2), pp. 197-218.
30. Jones, Gareth R. (1990). Governing Customer-Service Organizational Exchange, *Journal of Business Research*, Vol. 20, pp. 23-29.
31. Joreskog, K.G. (1971). Simultaneous Factor Analysis in Several Populations, *Psychometrika*, Vol. 36, pp. 409-426.
32. Jurkovich, R. (1974). A Core Typology of Organizational Environments, *Administrative Science Quarterly*, Vol. 19, pp. 380-394.
33. Kumar, Nirmalya, Louis W. Stern, and James C. Anderson (1993). Conducting Interorganizational Research Using Key Informants, *Academy of Management Journal*, Vol. 36 (6), pp. 1633-51.
34. Larsson, Rikard and David E. Bowen (1989). Organization and Customer: Managing Design and Coordination of Services, *Academy of Management Review*, Vol. 14 (2), pp. 213-233.
35. Lawrence, Paul R., and Jay W. Lorsch (1967). Organization and Environment, Boston, Mass: *Harvard Business School*.
36. Mills, Peter K., and N. Margulies (1980). Toward a Core Typology of Service Organizations, *Academy of Management Review*, Vol. 5, pp. 255-265.
37. Mishra (2013). Firms' Strategic Response to Service Uncertainty: An Empirical Signaling Study, *Australasian Marketing Journal*, Vol. 21, pp. 187-197.
38. Mishra, D.P., J.B. Heide and S.G. Cort (1998). Levels of agency relationships in service delivery: theory and empirical evidence, *Journal of Marketing Research*, Vol. 35 (3), pp. 277-295.
39. Morgan, O.A., J.C. Whitehead, W.L., Huth, G.S. Martin, and R. Sjolander, (2013). A Split-Sample Revealed and Stated Preference Demand Model to Examine Homogenous Subgroup Consumer Behavior Responses to Information and Food Safety Technology Treatments, *Environmental and Resource Economics*, pp. 1-19.
40. Morgan, Robert M. and Shelby Hunt (1994). The Commitment-Trust Theory of Relationship Marketing, *Journal of Marketing*, Vol. 58, pp. 20-38.
41. Nelson, Phillip (1970). Information and Consumer Behavior, *Journal of Political Economy*, Vol. 81 (4), July-August, pp. 729-754.
42. Norusis, Marija J. (1991). *SPSS Base System User's Guides*, Chicago: SPSS Inc.

43. Ortiz, M.H., K.E. Reynolds and G.R. Franke (2013). Measuring Consumer Devotion: Antecedents and Consequences of Passionate Consumer Behavior, *The Journal of Marketing Theory and Practice*, Vol. 21(1), pp. 7-30.
44. Overton, P., R. Schneck and C.B. Hazlett (1977). An empirical study of the technology of nursing subunits, *Administrative Science Quarterly*, Vol. 22 (2), pp. 203-19.
45. Perrow, Charles (1967). A Framework for Comparative Analysis of Organizations, *American Sociological Review*, Vol. 32, pp. 194-208.
46. Peterson, Robert A. (1994). A meta-Analysis of Cronbach's Coefficient Alpha, *Journal of Consumer Research*, Vol. 21 (9), pp. 381-391.
47. Rao, Akshay R., and Mark E. Bergen (1992). Price Premiums as a Consequence of Buyers' Lack of Information, *Journal of Consumer Research*, Vol. 19 (3) (December), pp. 412-423.
48. Rao, Akshay R. and Humaira Mahi (2003). The Price of Launching a New Product: Emprical Evidence on Factors Affecting the Relative Magnitude of Slotting Allowances, *Marketing Science*, Vol. 22 (2), pp. 246-268.
49. Schleisinger, Leonard A. (1992). Automobile Dealer Sales and Service: Critical Incidents, *Harvard Business School Case* (Number 9-690-061).
50. Seymor D., and Gref Lessne (1984). Spousal Conflict Arousal: Scale Development, *Journal of Consumer Research*, Vol. 11 (3), pp. 810-821.
51. Siehl Caren., David E. Bowen., and Christine M. Pearson (1992). Service Encounters as Rites of Integration: An Information Processing Model, *Organization Science*, Vol. 3 (4), November, pp. 537-555.
52. Snyder, C.A., J.F. Cox., and R.R. Jesse (1982). A Dependent Demand Approach to Service Organization Planning and Control, *Academy of Management Review*, Vol. 7, pp. 455-466.
53. Stevens, J. (1992). *Applied Multivariate Statistics for the Social Sciences*, 2nd edition, New Jersey: Lawrence Erlbaum.
54. Tosi, H., and J.W. Slocum (1984). Contingency Theory: Some Suggested Directions, *Journal of Management*, Vol. 10, pp. 9-26.
55. Weinberger, M.G., and H.E. Spotts (1989). A Situational View of Information Content in TV Advertising in the US and UK, *Journal of Marketing*, Vol. 53 (1), pp. 89-94.
56. Woodward, Joan (1970). *Industrial Organization: Behavior and Control*, London: Oxford University Press.
57. Zeithaml, Valarie A., A. Parasuraman., and Leonard L. Berry (1985). Problems and Strategies in Services marketing, *Journal of Marketing*, Vol. 49 (Spring), pp. 33-46.

Appendix

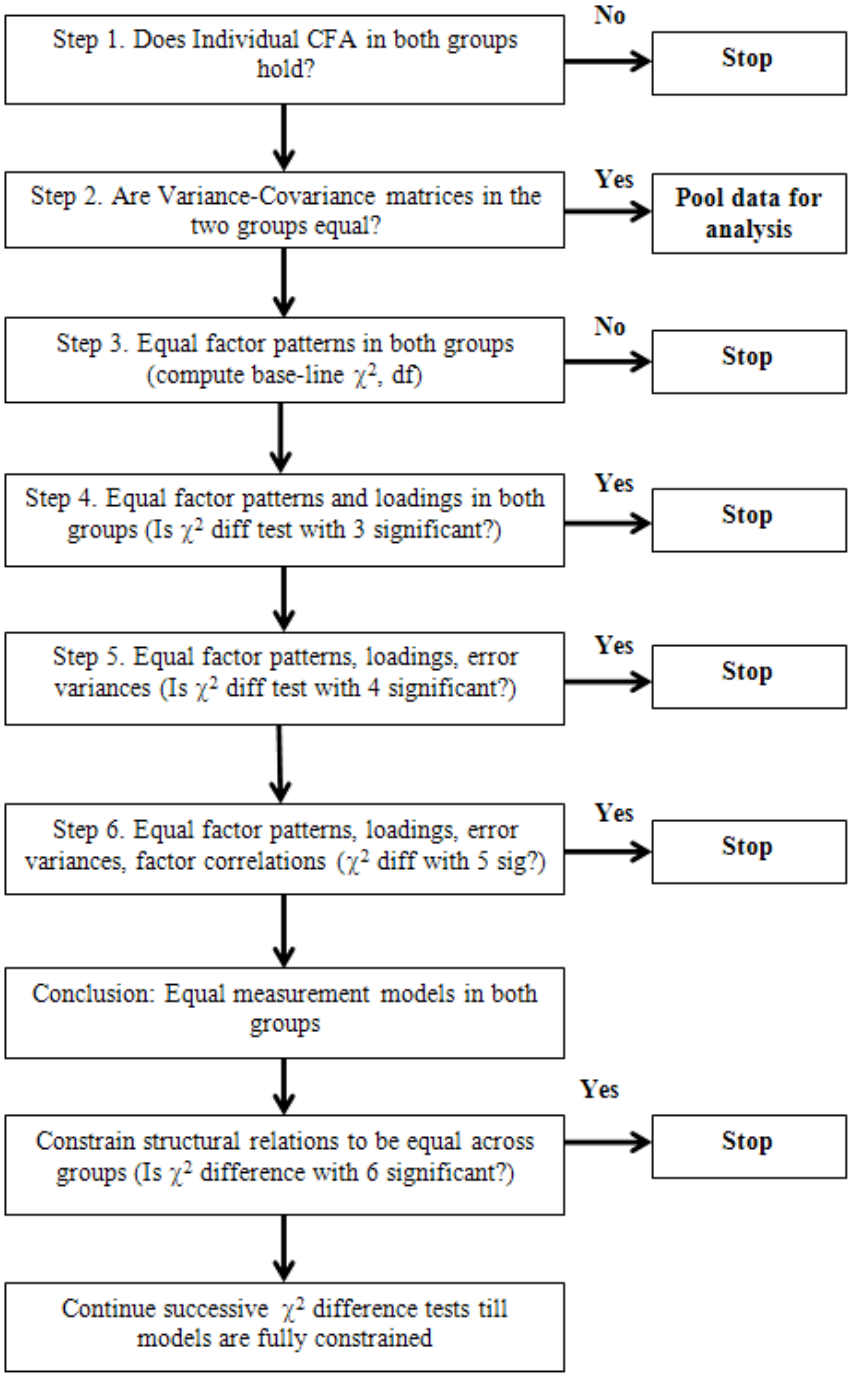

Fig. 1. Flow chart for multi-sample SEM analysis