“Construct validity in cross-cultural management research: classical test theory and latent trait theory approaches”

AUTHORS	Debi P. Mishra

ARTICLE INFO	Debi P. Mishra (2013). Construct validity in cross-cultural management research: classical test theory and latent trait theory approaches. *Problems and Perspectives in Management*, *11*(1)
RELEASED ON	Monday, 22 April 2013
JOURNAL	"Problems and Perspectives in Management"
FOUNDER	LLC “Consulting Publishing Company “Business Perspectives”

NUMBER OF REFERENCES	NUMBER OF FIGURES	NUMBER OF TABLES
0	0	0

Debi P. Mishra (USA)

Construct validity in cross-cultural management research: classical test theory and latent trait theory approaches

Abstract

As businesses become increasingly global, the field of management stands to benefit from theories, best practices, tools, and techniques that can be used in different cultures. By providing guidelines for theory development and testing, this type of research can facilitate the generalizability and validity of management theories and concepts. Cross-cultural research can also be used by scholars and policy makers to better understand the comparative implications of theories that have originated in unicultural settings. Despite its importance, there is a paucity of research on the use of appropriate tools and techniques for measuring and comparing constructs across cultures. To address this gap, this paper highlights the importance of investigating conceptual, functional, and measurement equivalence of constructs as a prerequisite for cross-cultural comparisons. This study also discusses how two measurement approaches, i.e., *classical test theory* (CTT) and *item response theory* (IRT) can be used in conjunction to gainfully investigate equivalence. The use of CTT and IRT models is illustrated via an empirical investigation of the Supplier Reputation Display (SRD) construct by analyzing data collected from US and Canadian automotive service managers. Implications of this research for management theory and practice, and the scope for further research are also discussed.

Key words: cross-cultural measurement, item response theory, classical test theory, measurement equivalence.
JEL Classification: M10, M16, M19, C1, C8.

Introduction

Cross-cultural management research is important for a variety of reasons. First, with increasing globalization, there is a need for developing theoretical models and best practices that can be used in different settings (Adler, 1983; Eisenberg et al., 2013; Reeb and Mahmoud, 2012). Second, academicians (Clark, 1990; Moseley, 2013) have noted that most management research has been conducted in the United States, with assumptions pertinent to this country. This skewness in research emphasis can be addressed through theory testing across multiple cultural situations (Monroe, 1990; Runyan et al., 2012). Third, government institutions as well as multinationals can benefit from cross-cultural research, especially in their policy decisions in economic and social matters (Dholakia, 1987; Jackson, 2012; Scott, 1984). Finally, cross-cultural research can help researchers better appreciate the comparative implications of theories developed in other nations (Dant and Barnes, 1988; Jackson, 2012).

Despite the importance of cross-cultural research and the availability of a large body of literature in the social sciences on measurement issues (Drasgow and Kanfer, 1985; Freitag and Bauer, 2013; Taylor and Bowen, 2012; Tomyn et al., 2013), mainstream management scholars have paid little systematic attention to the critical issue of cross-cultural equivalence (see Durvasula et al., 1993; Singh, 1995; for notable exceptions).

The issue of cross-cultural equivalence is an issue of *validity*. Unless scholars can confidently assert that they have used equivalent measures of concepts, it is not possible to achieve scientific generalization. In other words, an item may be a valid measure of a construct within one culture, but it may not be suitable in another setting. For example, the cost of a product (construct) can be validly measured in one country using local price (scale item). However, local price may not be a direct basis for comparing the price of a product across countries because of different exchange rates. Hence, studies lacking equivalence are no different than attempts at cross-cultural generalization from a uni-cultural perspective.

In light of the preceding observations, the objectives of this research are *four-fold*: (1) to systematically appraise extant social science research and delineate criteria for achieving equivalence; (2) to ascertain the extent to which issues of equivalence have been paid attention to in the management field; (3) to introduce and describe two promising measurement theories, i.e., *classical test theory*, and *latent trait theory* for investigating measurement equivalence; and (4) to empirically investigate cross-cultural equivalence of the "Supplier Reputation Display" (SRD) construct (Mishra, 1998) by analyzing data collected from automotive service managers in the United States and Canada.

The remainder of this paper is organized as follows. First, the criteria for cross-cultural equivalence are outlined. Next, a set of cross-cultural studies are evaluated for equivalence. In light of this appraisal, the classical test theory and latent trait theory

models for investigating measurement equivalence are described. This is followed by a discussion of the Supplier Reputation Display construct. In the penultimate section, cross-cultural equivalence of the SRD construct is empirically investigated using CTT and LTT approaches. Finally, results of the empirical analysis together with directions for future research are described. I begin by outlining the criteria for cross-cultural equivalence.

1. Cross-cultural equivalence

Equivalence of *concepts* and *measures* is a prerequisite for meaningfully comparing theoretical relations across cultures (Hui and Triandis, 1985). Concepts fall into two broad categories. On the one hand, there are *molar* or *universal* concepts like "the psychic unity of mankind" (Kroeber, 1948) which may not be specified further through empirical referents. Substantive relations among these concepts can therefore be directly compared across cultures. On the other hand, and more commonplace, concepts are operationalized through empirical referents (an operationalized concept is a 'construct'; Kerlinger, 1978).

In order to meaningfully compare constructs across cultures, a researcher has to establish equivalence at the *conceptual, functional, and measurement* levels (Van de Vijier and Poortinga, 1982).

1.1. Conceptual equivalence. Conceptual equivalence implies that a construct from one culture should have an equal meaning in the other (Freitag and Bauer, 2013; Hui and Triandis, 1985). For instance, as noted by Dant and Barnes (1988), in the 1980's, the construct of *brand loyalty* in the US lacked conceptual equivalence in the erstwhile USSR where the *number* of the manufacturing factory was commonly used as a *brand* proxy. Other examples of concepts lacking conceptual equivalence across cultures are *cognitive consistency* (Green and White, 1976) and *perceived risk* (Hoover, Green and Saegert, 1978).

1.2. Functional equivalence. Functional equivalence entails similar antecedent-construct-consequent relations in different cultures (Hui and Triandis, 1985). For example, if *brand loyalty* and *number loyalty* (measured by the *frequency of purchase*) reduce risk (have similar consequences), functional equivalence of the "loyalty" construct is attained.

1.3. Measurement equivalence. Measurement equivalence is achieved when relations between observed scores and latent traits are identical across relevant groups. More specifically, individuals with the same standing on a latent trait, but sampled from different cultures, should have the same observed score on a scale measuring that trait (Drasgow and Kanfer, 1985; Hox et al., 2012; Singh, 1995). In other words, a numerical value on the scale refers to the same degree, intensity, or magnitude of the construct regardless of the culture. Measurement equivalence can be investigated only after conceptual and functional equivalences have been established.

Utilizing the criteria given above, I examined a few articles on cross-cultural marketing research from the major marketing journals. I chose the sub-discipline of marketing because: (1) researchers have made great strides in developing measures of latent constructs in this field; and (2) recent research focuses on theory testing across diverse cultural settings. The results of the exploratory review show mixed support for *conceptual* and *functional* equivalence, and virtually no support for *measurement* equivalence. For instance, Shimp and Sharma (1987), in validating the CETSCALE, noted that for respondents in all the groups, ethnocentrism led to a feeling of belongingness. Many studies have also carried out factor analysis of items for two or more groups and concluded that cross-national differences for constructs existed. More recent research has addressed some shortcomings primarily in the areas of conceptual and functional equivalence (Singh, 1995). However, considered as a whole, measurement equivalence has been primarily assessed by comparing factor pattern matrices in the relevant cultures. Furthermore, studies have also utilized variance partitioning (MANOVA and ANOVA) and other multivariate techniques (e.g., regression).

Overall, the measurement techniques used in extant cross-cultural marketing research appear to be flawed for a number of reasons. For instance, comparison of individual factor patterns is a *necessary* but not *sufficient* condition for equivalence (Drasgow and Kanfer, 1985). Although Meredith (1964), using Lawley's (1943) selection theorem has proved that there are powerful reasons to expect equal factor pattern matrices across groups, a researcher has to ensure that co-variation in the observed variables is indeed caused by identical constructs. A stricter test is to estimate the parameters (population values) of the pattern matrices as if the variables came from a single population. Such an approach, developed by Joreskog (1971) and subsequently modified by Bagozzi (1983) and by Drasgow and Kanfer (1985),

is known as the *simultaneous investigation of factor analysis in several populations* (SIFASP).

MANOVA models appear inadequate for investigating metric equivalence because the technique assumes scalar equivalence which requires constructs to be measured with an identical metric.

In summary, it appears that researchers have not paid adequate attention to equivalence criteria before making comparisons. Hence, cross-cultural management research stands to benefit from systematic procedures to establish sequential equivalence (i.e., conceptual → functional → measurement) of constructs.

Against the preceding observations, I empirically investigate equivalence of the Supplier Reputation Display (SRD) construct in a cross-cultural setting. The following section describes two promising models, i.e., classical test theory and latent trait theory, which address the critical issue of measurement equivalence.

2. Classical test theory (CTT)

In classical test theory, researchers seek a relationship between the observed score of an individual and his or her true (unobservable or latent) score on the latent construct. Mathematically, this relationship is expressed as,

$$x = t + e, \quad (1)$$

where x is the observed score, t is the true score, and e is the measurement error (random and systematic). For a situation where a set of items is used to measure a trait (θ), the common factor model, in light of (1) is,

$$x_i = \lambda_i + e_i, \quad (2)$$

where i refers to the i^{th} item, λ_i refers to the loading of the i^{th} item on the latent (unobservable) trait θ, and e_i is the error variance (systematic and random) for the i^{th} item. Under the assumption that variables are mutually uncorrelated, and that θ is standardized, the variances and co-variances of items can be partitioned into three sources, i.e., (a) valid variance of λ_i^2; (b) unique variance attributable to a particular item; and (c) random error variance. This partitioning of an item's variance is based on the principle of local independence (Lord and Novick, 1978).

2.1. Simultaneous investigation of factor analysis in several populations (SIFASP). Measurement equivalence in two different groups for a set of items (loading on a particular trait) is attained when the corresponding parameters (population values) in the two groups are equal. In other words,

$$\lambda_i 1 = \lambda_i 2, \quad (3)$$

$$e_{si} 1 = e_{si} 2, \quad (4)$$

$$e_{ri} 1 = e_{ri} 2, \quad (5)$$

where θ refers to an item's loading on its latent trait (θ), e_s refers to the specific (unique) variance, and e_r refers to random variance. Furthermore, the subscript i refers to the i^{th} item out of a set of n items loading on a trait, whereas 1 and 2 refer to the two groups under consideration. Note that these equations can be easily extended to k groups. Although measurement equivalence is achieved for the unidimensional case stated above, the multidimensional case imposes another restriction. In other words, if a set of indicators is measuring a set of constructs, the factor co-variance in the two groups should be equal. Mathematically,

$$\phi_{(ab)} 1 = \phi_{(ab)} 2, \quad (6)$$

where 1 and 2 refer to the two groups, and a, b refer to the constructs.

One important aspect of the SIFASP model is that stepwise constraining is possible. We can first constrain factor loadings across multiple groups to be equal and check for adequacy of fit. In the next step, we can impose additional measurement constraints (i.e., equal error variances), and check for the improvement in fit. Utilizing this successive constraining procedure, the degree of measurement equivalence can be determined. We can, therefore, ascertain whether items are being measured equivalently only to the extent of their valid variances or additional components also (i.e., specific and random variances, and factor co-variances). Note however, that we naturally expect equal factor loading matrices across groups where identical constructs are being studied (Meredith, 1964). This implies that the first step in the SIFASP procedure is to constrain factor loadings in the groups to be equal. Successive constraining can be done by fixing the other components of variance to be equal across groups.

3. Latent trait theory (LTT)

Latent Trait Theory, also called Item Response Theory (IRT) (see Lord, 1952) seeks a relationship between a person's standing on a construct (latent trait) and the probability of his or her responding positively to an item which is measuring the trait. In other words, if a person has a high standing on a latent trait (customer

satisfaction), and if he/she is administered an item in the "customer satisfaction" domain, his/her probability of checking a correct response will be high. Note that the trait level is theoretically unbounded. Hence, on the X axis, it will range from positive infinity to negative infinity. The Y axis however is bounded (0, 1). This relationship will therefore be non-linear in θ, p (θ refers to an estimate of the trait, p is the probability of a correct response). Furthermore, the curve will be asymptotic to the line $Y = 1$. An example of the mathematical relationship between θ and p is the three parameter logistic model (Birnbaum, 1968) for dichotomous data,

$$p(\theta) = \frac{c_i + (1 - c_i)}{1 + e^{\{-Da_i(\theta - b_i)\}}}$$

where, θ refers to a numerical estimate of a person's latent trait, D is a constant, and a_i, b_i, and c_i are parameters of the model which are described below.

3.1. LTT parameters. Three parameters characterize a LTT model. These are: (1) a_i or the *discrimination* parameter; (2) b_i or the *difficulty* parameter; and (3) c_i or the *guessing* parameter. The a_i parameter is the slope of the ICC at the inflexion point. Higher values of a_i imply a steep curve which easily discriminates among individuals within a narrow range of θ. The b_i parameter is the value along the θ continuum at which the probability of a positive response is 50 per cent. Lower values of b_i will shift the ICC toward the left, making the item easier. The guessing parameter c_i implies that persons with infinitely low levels of the trait may still guess a favorable response. In this situation, the ICC will always have a positive intercept and never pass through the origin. However, some models (i.e., Samejima's graded model, 1969) assume that $c_i = 0$. In such cases, the lower asymptote of the ICC is zero.

The above ideas are captured by the item characteristic curves depicted in Figure 1. As can be seen from this figure, an item designed to measure a latent trait (θ) exhibits differential performance across two different cultures (1 and 2). First, as expected, the item achieves discrimination in both cultures because respondents possessing higher abilities on the latent trait (θ) perform better. When the ICC is a flat horizontal line, it achieves no discrimination. In the present situation, in culture 1, the ICC for the item (a = +2) is steeper than in culture 2 (a = +1.4) indicating better discrimination. On the other hand, in culture 1, respondents find the item to be easier (b = 0.29) than in culture 2 (b = 1.14) because even at low ability (θ) levels, its relative performance is better.

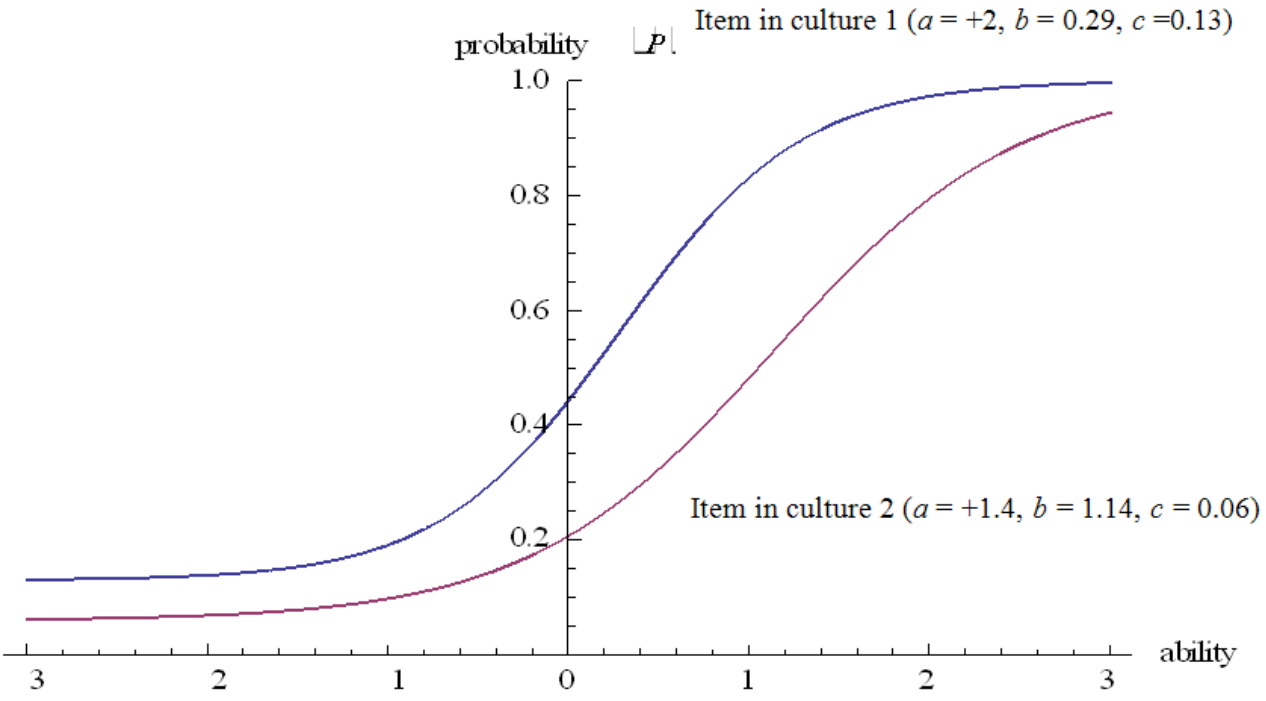

Fig. 1. Comparison of item characteristic curves

3.2. LTT assumptions and properties. The main assumption of the LTT model is that items measure a uni-dimensional construct. Furthermore, when a small number of items are used per construct, IRT parameter estimates can be obtained using the marginal maximum likelihood approach (Bock and Liebermann, 1970). LTT is characterized by item parameter invariance (parameters do not depend upon the trait being measured; Guion and Ironson, 1983), while in classical test theory, item statistics are sample dependent.

3.3. LTT and cross-cultural research. If an item exhibits measurement equivaence in two cultures, the corresponding ICC's are expected to be identical. If the ICC's differ, measurement equivalence is not attained. In such a situation, biased items may have to be modified or deleted from further analysis. LTT therefore provides an avenue for estimating the degree of bias.

3.4. The two parameter graded model. The discussion thus far has considered models for dichotomous (e.g., Yes/No) data. For a Likert type situation, the graded model given by Samejima (1969) is appropriate. The parameters of this model are easily estimated using the MULTILOG software (Thissen, 1990).

For Likert type data (a five point scale), there are 4 curves (one for each response category) per ICC. The first corresponds to the probability of checking 2 or higher (on a five point scale), the second 3 or higher, and so on. Note that the probability of checking 1 or higher on the scale is always unity. The graded model is mathematically depicted as,

$$p(\theta) = \frac{1}{1 + e^{\{-Da_i(\theta - b_i)\}}}$$

where θ refers to a numerical estimate of a person's latent trait, D is a constant, and a_i, b_i are parameters of the model described earlier.

4. Advantages of LTT over CTT

LTT affords several advantages over CTT. First, LTT parameters are sample invariant, while CTT parameters are sample specific. Second, LTT operates at a more micro level than CTT. This is because in LTT a response category is the unit of analysis while in CTT an item is the unit of analysis. Third, LTT is more general than CTT. While CTT assumes that observed variables are linearly related to traits, LTT makes no such assumption. Finally, for investigating measurement equivalence, LTT techniques are more sophisticated than CTT. For instance, the extent and degree of bias at the micro level of an item can be easily detected using IRT. More specifically, we may be able to detect bias across the theta continuum and response categories. Such an approach makes the measurement of bias more accurate and renders construct validation more meaningful.

5. Plan of analysis and criteria for evaluating results

This section details the plan of analysis. First, the supplier reputation display construct (Mishra, 1998) is described. Next, the SRD measurement model is discussed. Finally, the steps for conducting SIFASP and LTT analyses, as well as the criteria for evaluating the results are laid out. We begin with a discussion of the SRD construct.

6. The supplier reputation display (SRD) construct

As described by Mishra (1998), "In asymmetric marketing relationships, sellers typically possess more information about the object of an exchange than buyers. To ameliorate customers' evaluation problems, sellers use signals to promise the delivery of a certain level of quality to the market. An important signal in asymmetric markets is the manner in which sellers display their reputation to customers" (p. 123). More specifically, the SRD construct is described in the following manner.

"Reputation display involves disclosing information to the market about a firm's past and future quality orientation. For example, firms may undertake investments in physical surroundings to signal to customers the presence of irreversible sunk assets which may be expropriated if quality deteriorates in the future. Likewise, a firm's past conduct may be endorsed by third-parties, and this certification could act as a signal of future quality" (Mishra, 1998, p. 127).

7. Facets, dimensions, and items of the SRD construct

The conceptualization of the SRD construct is depicted in Figure 2. The corresponding scale items are depicted in Table A1 (see Appendix). As may be seen from Figure 2, SRD is a second order factor comprising three first order dimensions, i.e., certification, specific investments, and advertising intensity which are described below.

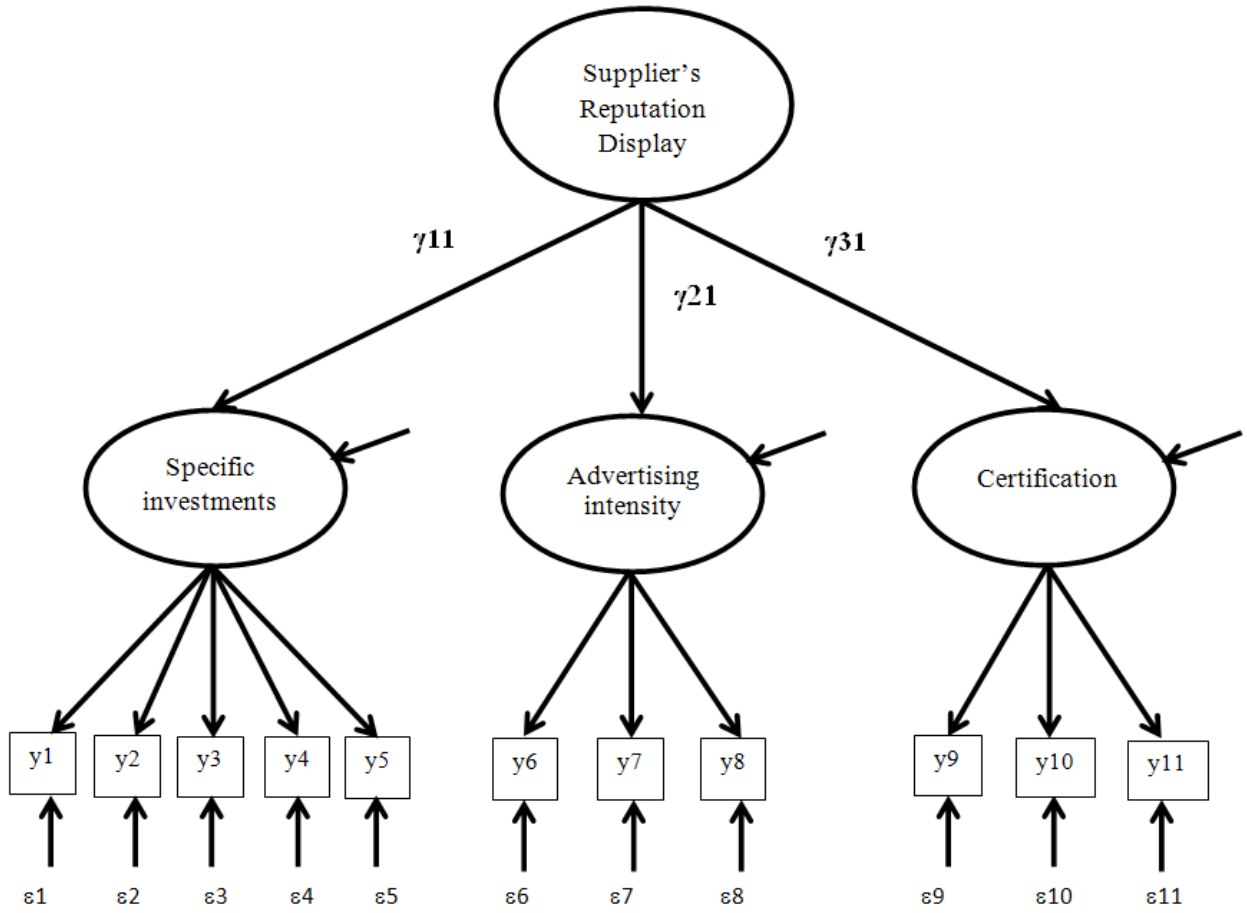

Fig. 2. Empirical representation of the Supplier's Reputation Display (SRD) construct

7.1. Certification. Research from a number of different disciplines (accounting, economics, education, finance, organizational theory) suggests that the display of certification can play a major role in reducing information asymmetry for a firm's target audience (Mishra, 2006; Mishra, 2000a). The economics literature (Akerlof, 1970; Grossman, 1981; Png and Reitman, 1995) suggests how certification display can lower customers' information acquisition costs (Stigler, 1961) and assure buyers of a minimum level of quality *ex-ante*. For example, many automotive repair shop display their "AAA" affiliation in order to reduce customers' information asymmetry. The efficacy of certification as a reputation signal is enhanced only when it is prominently displayed. In the parlance of economists, when information asymmetry is high, certification display creates a "separating equilibrium" across multiple firms. Hence, certification serves as a signal of reputation that customers rely upon to make rational assumptions about future intentions and behavior of firms.

7.2. Specific investments. In asymmetric markets, firms may signal their reputation via investments in firm specific assets whose sole purpose is to serve as bonding mechanisms (Klein and Leffler, 1981; Rubin, 1990; Rashid, 1988). Specific assets like investments in expensive upholstery and ornate surroundings do not yield any direct consumer benefit. By definition, these assets are non-salvageable because they find very little use outside the focal relationship between a firm and its customers. Investments in specific assets send a strong signal to customers that a firm does not intend to cheat in the short run because of the sunk nature of investments. This strategy also deters dishonest firms from entering the market. Specifically, potential entrants need to undertake substantial investments in fixed assets before they can reap the benefits of price premiums.

7.3. Advertising intensity. The role of advertising in reputation display has received attention in the economics literature (Klein and Leffler, 1981; Milgrom and Roberts, 1986; Nelson, 1970, 1974, 1978). Specifically, researchers question whether advertisements can provide tangible evidence of product quality in asymmetric markets. On the other hand, the level of advertising expenditure may signal a firm's reputation for providing quality services and products.

In a series of articles, Nelson (1970, 1974, 1978) argues that advertisements for credence goods do not emphasize tangible product related information. On

the other hand, customers appear to associate the level and intensity of advertising (or expenses) with quality. According to Nelson (1970, 1974, 1978), high quality firms whose products satisfy many customers are expected to generate more repeat sales than low quality firms. Consequently, such high quality firms are more profitable and have the wherewithal to expend resources in the market through heavy advertising which customers can associate with quality.

8. Criteria and steps for the SIFASP analysis (CTT approach)

The analysis starts with the variance-covariance matrices since it contains more information than the correlation matrix (which has unity in its diagonals). In addition, correlation matrices are typically scaled to have a mean of zero. In this situation, it makes little sense to compare groups with the same mean. The stepwise procedure for analysis is depicted in Figure 3 (see Appendix).

9. Criteria and steps for LTT analysis

As a first step, unidimensionality of the items is assessed. For the present study, items can be considered approximately uni-dimensional. However, approximate unidimensionality can be assessed by submitting items to an exploratory factor analysis (EFA) procedure. Next, parameter estimates for the biased items (identified by CTT) are computed using MULTILOG (Thissen, 1990). In the third step, ICC's for individual items in the two groups are plotted and compared. A precise assessment of bias is possible by linking metrics (Lord, 1952). A number of multivariate tests at this stage may indicate the effect of deleting biased items on substantive relationships. More specifically, substantive relations in one culture may change if biased items are deleted. On the other hand, substantive relations may remain unaffected even after deleting biased items. To this extent, it makes sense to pool the data and then make stronger substantive inferences. The stepwise procedure for the LTT analysis is depicted in Figure 4 (see Appendix).

10. Results

The sampling and data collection procedure is described in detail in Mishra (1998). A parallel survey using the scale items administered to automotive repair service managers was administered to a sample of Canadian automotive repair establishments. The CFA analysis was carried out using the program EQS Bentler (1980). The major advantage of EQS over LISREL is that it automatically imposes bounds on parameter estimates. Specifically, negative error variances (Heywood cases) do not affect the solution, as the parameter estimate is constrained at zero (lower bound).

10.1. Confirmatory factor analysis of SRD items for the USA and Canada samples. For both samples, *Convergent validity* of the indicators is suggested by (a), all hypothesized loadings being positive, and (b), *t* values being > 2. *Discriminant validity* of the three SRD facets is established through sequential χ^2 difference tests for 1 *df* by fixing correlations between facets pairwise to 1 (Howell, 1986; Mishra, 2000b). In the present analysis, all χ^2 difference tests are significant indicating that discriminant validity for the three facets of the SRD construct is achieved. Finally, the goodness-of-fit indices are acceptable, suggesting a good fit of the model to the data.

The nested multi-group confirmatory factor analysis suggests that the measurement model of equal factor loadings does not hold across the USA and Canada samples. In addition, the Lagrange Multiplier (LM) test indicates that the *specific investment* dimension contributes to measurement nonequivalence (hence bias) in the two groups.

10.2. IRT analysis. Unidimensionality of the items was assessed using exploratory factor analysis. Next, all steps depicted in Figure 4 were followed to estimate IRT parameters for the US and Canadian samples. Parameter estimates for the US and Canadian items loading on the specific investment dimensions appear in Table 1.

Table 1. Item response theory parameter estimates for advertising intensity items

US sample						*Canada sample*					
Advertising intensity item	*a*	*b2*	*b3*	*b4*	*b5*	Advertising intensity item	*a*	*b2*	*b3*	*b4*	*b5*
We spend significant amounts of money for advertising our services	1.60	-3.21	-1.71	-1.90	0.91	We spend significant amounts of money for advertising our services	2.62	-5.11	-1.56	-1.16	0.73
We advertise our services on a very regular basis	1.71	-3.72	-1.92	-0.91	0.87	We advertise our service on a very regular basis	1.82	-6.31	-2.63	-1.15	1.13
We usually undertake large scale advertising for promoting our service	1.61	-2.91	-1.91	-0.97	0.85	We usually undertake large scale advertising for promoting our service	7.62	-1.62	-1.22	-0.45	1.09

Notes: Parameter Estimates are based on Samejima's two parameter graded response model for polytomous data. All parameter estimates are significant. Parameter estimates are estimated using MULTILOG

11. Discussion

Utilizing the CTT approach, the Lagrange Multiplier (L-M) test suggested that items comprising the specific investment dimension were biased. However, CTT does not indicate the degree of this bias. At best, it gives an idea about the item being biased in a global sense.

For the *advertising intensity* dimension, the second response category of the "we spend significant amounts of money for advertising our services" item is biased against US managers at very low trait levels. This means that for this response category, scores will be lower for the US respondents as compared with their Canadian counterparts. Interestingly, for the other response categories, the a parameter for the Canadian managers is higher than those for the US managers. This suggests that the item discriminates better for Canadian managers than it does for American managers. Similar comparisons are possible for the remaining items. Thus, IRT offers a powerful avenue for focusing on an item at the micro level.

Given that we are able to focus on an item at the micro level, there are a lot of possibilities for tackling bias in the context of cross-cultural management research. Specifically, if bias extends for a narrow range of the theta continuum, and only for a few response categories, retaining the item will not distort substantive results. However, if the bias is uniform (across a wider range), deletion may be required. Against this backdrop, the individual strengths of the IRT and CTT models become evident. In the CTT model, location of item bias is possible via the L-M test. In the present analysis, the L-M test indicated that *advertising intensity* items are biased in a macro sense. On the other hand, the IRT analysis takes a more microscopic approach and suggests that bias for the first *advertising intensity* item may not be severe, and that the item can be retained for further analysis.

The findings of this research have to be considered against certain limitations. First, IRT analysis was not carried out for all the items comprising the facets. Such an analysis would have provided an estimate of bias across the whole scale. Second, visual inspection of ICC's is only a crude way of determining bias. There are a number of sophisticated techniques which can be used to precisely estimate bias (see Berk, 1982; for a review).

Conclusion and directions for future research

The objectives of this research were four-fold, i.e., (1) to appraise the social sciences literature and delineate the criteria for achieving equivalence in construct measurement; (2) to ascertain the extent to which equivalence issues have been addressed in the management field; (3) to describe two competing measurement models (i.e., IRT and CTT) for conducting cross-cultural comparisons; and (4) to empirically implement the IRT and CTT approaches by investigating cross cultural equivalence of the "Supplier's Reputation Display" (SRD) construct using data collected from managers in the United States and Canada.

A systematic review of the management literature suggests that the notion of equivalence has received scant attention from researchers. More importantly, a number of research studies conveniently ignore the concept of equivalence, and aim at cross-cultural generalization from a unicultural standpoint. In many cases, the unit of analysis (nation state, culture, system, ethnic group, society) has not been specified a priori. This calls for a more rigorous implementation of the concept of equivalence while carrying out cross-cultural management research.

Both measurement models tested here (IRT and CTT) have their individual strengths. Specifically, the IRT model provides a microscopic investigation of item bias while CTT tackles the problem at a macro level. Hence, when CTT and LTT models are combined, they yield additional insights into the location and nature of item bias. Such a plural approach facilitates more in-depth investigation of item bias and holds considerable promise for theory development and testing in cross cultural settings.

Finally, the social sciences (particularly psychology) have used item response theory for quite some time now by focusing on equivalence with respect to test scores. By utilizing IRT approaches, we may be in a better position to validate and generalize latent constructs and theories in the management field.

References

1. Adler, N.J. (1983), "Cross-cultural management research: The ostrich and the Trend", *Academy of Management Review*, 8 (2), pp. 226-232.
2. Akerlof, George A. (1970). "The Market for 'Lemons': Quality under Uncertainty and the Market Mechanism", *Quarterly Journal of Economics*, 84 (August), pp. 488-500.
3. Bagozzi, Richard P. (1983). "A Holistic Method for Modeling Consumer Response to Innovation", *Operations Research*, 31, p. 1.

4. Bentler, P. (1980). "Significance Tests and Goodness of fit in the Analysis of Covariance Structures", *Psychological Bulletin*, 56, pp. 81-105.
5. Berk, R.A. (1982). *Handbook of methods for detecting test bias*, Johns Hopkins University Press.
6. Bock, R.D. and M. Lieberman (1970). "Fitting a Response Model for n Dichotomously Scored Items", *Psychometrika*, 35, pp. 179-197.
7. Clark, Terry (1990). "International Marketing and National Character: A Review and Proposal for an Integrative Theory", *Journal of Marketing*, October.
8. Dant, Rajiv P. and James H. Barnes (1988). "Methodological Concerns in Cross-Cultural Research", in J.N. Sheth (Ed.), *Research in Marketing*, Greenwich, Conn: Jai Press.
9. Dholakia, N. (1987). "Industrial Policy, Competitiveness, and the Restructuring of World Markets", in J.N. Sheth (Ed.), *Advances in Marketing and Public Policy*, Greenwich, Conn: Jai Press.
10. Drasgow, Fritz and Ruth Kanfer (1985). "Equivalence of Psychological Measurement in Heterogeneous Populations", *Journal of Applied Psychology*, 4, pp. 662-680.
11. Durvasula, Srinivas, J. Craig Andrews, Steven Lysonski and Richard G. Netemeyer (1993). "Assessing the Cross-National Applicability of Consumer Behavior Models: A Model of Attitude toward Advertising in General", *Journal of Consumer Research*, 19 (March), pp. 626-636.
12. Eisenberg, J., Lee, H.J., Brück, F., Brenner, B., Claes, M.T., Mironski, J. & Bell, R. (2013). "Can Business Schools Make Students Culturally Competent? Effects of Cross-Cultural Management Courses on Cultural Intelligence", *Academy of Management Learning & Education*.
13. Freitag, M. & Bauer, P.C. (2013). "Testing for Measurement Equivalence in Surveys Dimensions of Social Trust across Cultural Contexts. *Public Opinion Quarterly*, 77 (S1), pp. 24-44.
14. Green, R. and P. White (1976). "Methodological Considerations in Cross-National Consumer Research", *Journal of International Business Studies*, 7. (Fall/Winter).
15. Grossman, Sanford J. (1981). "The Informational Role of Warranties and Private Disclosure about Product Quality", *Journal of Law and Economics*, 24, December, pp. 461-483.
16. Guion, R.M. and G.H. Ironson (1983). "Latent Trait Theory for Organizational Research", *Organizational Behavior and Human Performance*, 31, pp. 54-87.
17. Hoover, R.J., R. Green and J. Saegert (1978). "Cross-National Study of Perceived Risk", *Journal of Marketing* (July), pp. 102-108.
18. Howell, Roy D. (1986). "Analysis of Covariance Structures", *Journal of Marketing Research*, Fall.
19. Hox, J.J., Leeuw, E.D., Brinkhuis, M.J. & Ooms, J. (2012). "Multigroup and Multilevel Approaches to Measurement Equivalence", *Methods, Theories, and Empirical Applications in the Social Sciences*, pp. 91-96.
20. Hui, C.H. and Harry C. Triandis (1985). "Measurement in Cross-Cultural Psychology: A Review and Comparison of Strategies", *Journal of Cross-Cultural Psychology*, 16 (2), pp. 131-152.
21. Jackson, T. (2012). "Cross-cultural management and the informal economy in sub-Saharan Africa: implications for organization, employment and skills development", *The International Journal of Human Resource Management*, 23(14), pp. 2901-2916.
22. Joreskog, K.L. (1971). "Simultaneous Factor Analysis in Several Populations", *Psychometrika*, 36, pp. 409-426.
23. Kerlinger, F. (1978). *Foundations of Behavioral Research*, Holt Reinhart and Winston.
24. Klein, Benjamin and Keith B. Leffler (1981). "The Role of Market Forces in Assuring Contractual Performance", *Journal of Political Economy*, 89 (4), pp. 615-41.
25. Kroeber, A.L. (1948). *Anthropology: Race, Language, Culture, Psychology, Prehistory*, Harcourt.
26. Lawley, D.N. (1943-44). "A Note on Karl Pearson's Selection Formula", Proceedings of the Royal Society of Edinburgh (Section A), 62, pp. 28-30.
27. Lord, F.M. (1952). "A Theory of Test Scores", *Psychometric Monograph*, 7.
28. Lord, F.M. and M.R. Novick (1978). *Statistical Theories of Mental Test Scores*, Reading, MA: Addison-Wesley.
29. Meredith, W. (1964). "Notes on Factorial Invariance", *Psychometrika*, 29, pp. 177-185.
30. Milgrom P. and J. Roberts (1986). "Price and Advertising Signals of Product Quality", *Journal of Political Economy*, 94, pp. 796-821.
31. Mishra, Debi P. (1998). "The Conceptualization and Measurement of Supplier's Reputation Display in Asymmetric Marketing Relationships", *Journal of Market Focused Management*, 3, pp. 123-150.
32. Mishra, Debi P. (2000a). "Interdisciplinary Contributions in Retail Service Delivery: Review and Future Directions"*, Journal of Retailing and Consumer Services*, 7, pp. 101-118.
33. Mishra, Debi P. (2000b). "An Empirical Assessment of Measurement Error in Health-Care Survey Research", *Journal of Business Research*, 48, pp. 193-205.
34. Mishra, Debi P. (2006). "The Role of Certification in Service Relationships: Theory and Empirical Evidence", *Journal of Retailing and Consumer Services*, 13 (4), pp. 81-96.
35. Monroe, Kent B. (1990). "Editorial", *Journal of Consumer Research*, June.
36. Moseley, J.L. (2013). Understanding Cross-Cultural Management. *Performance Improvement*, 52(1), pp. 43-45.
37. Nelson, Philip (1978). "Advertising as Information Once More", in *Issues in Advertising*: The Economics of Persuasion, D.C. Tuerck (Ed.), Washington: American Enterprise Institute.
38. Nelson, Phillip (1974). "Advertising as Information", *Journal of Political Economy*, 81 (4), July-August, pp. 729-754.

39. Nelson, Phillip (1970). "Information and Consumer Behavior", *Journal of Political Economy*, 72 (March-April), pp. 311-329.
40. Png, I.P.L. and David Reitman (1995). "Why are Some Products Branded and Others Not?", *Journal of Law and Economics*, 38 (1), pp. 207-224.
41. Rashid, Salim (1988). "Quality in Contestable Markets: A Historical Problem?", *The Quarterly Journal of Economics*, February, pp. 246-249.
42. Reeb, D., Sakakibara, M. & Mahmood, I.P. (2012). "From the Editors: Endogeneity in International Business Research", *Journal of International Business Studies*, 43 (3), pp. 211-218.
43. Rubin, Paul (1990). *Managing Business Transactions*, New York: John Wiley.
44. Runyan, R.C., Ge, B., Dong, B. & Swinney, J.L. (2012). "Entrepreneurial Orientation in Cross-Cultural Research: Assessing Measurement Invariance in the Construct", *Entrepreneurship Theory and Practice*.
45. Samejima, Fumiko (1969). "Estimation of Latent Ability Using a Response of Graded Scores", Monograph, 17, *Psychometrika*.
46. Scott, B.R. (1984). "National Strategy for US Competitiveness", *Harvard Business Review*, 62, (March-April), pp. 77-91.
47. Shimp, Terence A. and Subhash Sharma (1987). "Consumer Ethnocentrism: Construction and Validation of the CETSCALE", *Journal of Marketing Research*, 24, pp. 280-289.
48. Singh, J. (1995). "Measurement issues in cross-national research", *Journal of International Business Studies*, 26(3), pp. 597-619.
49. Stigler, G (1961). "The Economics of Information", *Journal of Political Economy*, 69 (June), pp. 213-225.
50. Taylor, C.R. & Bowen, C.L. (2012). "Best practices for cross-cultural advertising research: Are the rules being followed? In *Handbook of Research on International Advertising*, Edward Elgar Publishing, pp. 3-19.
51. Thissen, David (1990). *Multiple, Categorical Item Analysis and Test Scoring Using Item Response Theory,* Mooresville, IN: Scientific Software.
52. Tomyn, A.J., Tyszkiewicz, M.D.F. & Norrish, J.M. (2013). "The Psychometric Equivalence of the Personal Wellbeing Index School-Children for Indigenous and Non-Indigenous Australian Adolescents", *Journal of Happiness Studies*, pp. 1-14.
53. Van de Vijier, F.R.J. and Y.H. Poortinga (1982). "Cross-Cultural Generalization and Univesality", *Journal of Cross-Cultural Psychology*, 13, pp. 387-408.

Appendix

Table A1. Scale items and reliability for the supplier reputation display construct

Construct	Scale items	Format	Reliability*
Specific Investments	♦ We have undertaken significant investments in the decor of our surroundings ♦ From time to time we undertake extensive investments in the interior and exterior modeling of our buildings ♦ We have spent significant amounts of money in designing and displaying signs in our buildings ♦ We have undertaken significant investments in our facilities dedicated to the needs of our customers ♦ If this particular location closed down it would be very difficult for us to recover the investments we have made in the decor of our buildings	7 point Likert scale with "strongly disagree" and "strongly agree" as anchors	0.75, 0.77
Certification	♦ Awards and recognition that we have received for our service ♦ Signs which depict the training and qualifications of our mechanics ♦ Membership in professional organizations (like ASE or AAA)	7 point Likert scale with "not prominently displayed" and "prominently displayed" as anchors	0.88, 0.77
Advertising Intensity	♦ We spend significant amounts of money for advertising our services ♦ We advertise our service on a very regular basis ♦ We usually undertake large scale advertising for promoting our service	7 point Likert scale with "strongly disagree" and "strongly agree" as anchors	0.86, 0.82

Note: Refers to alpha values in the US and Canadian samples respectively.

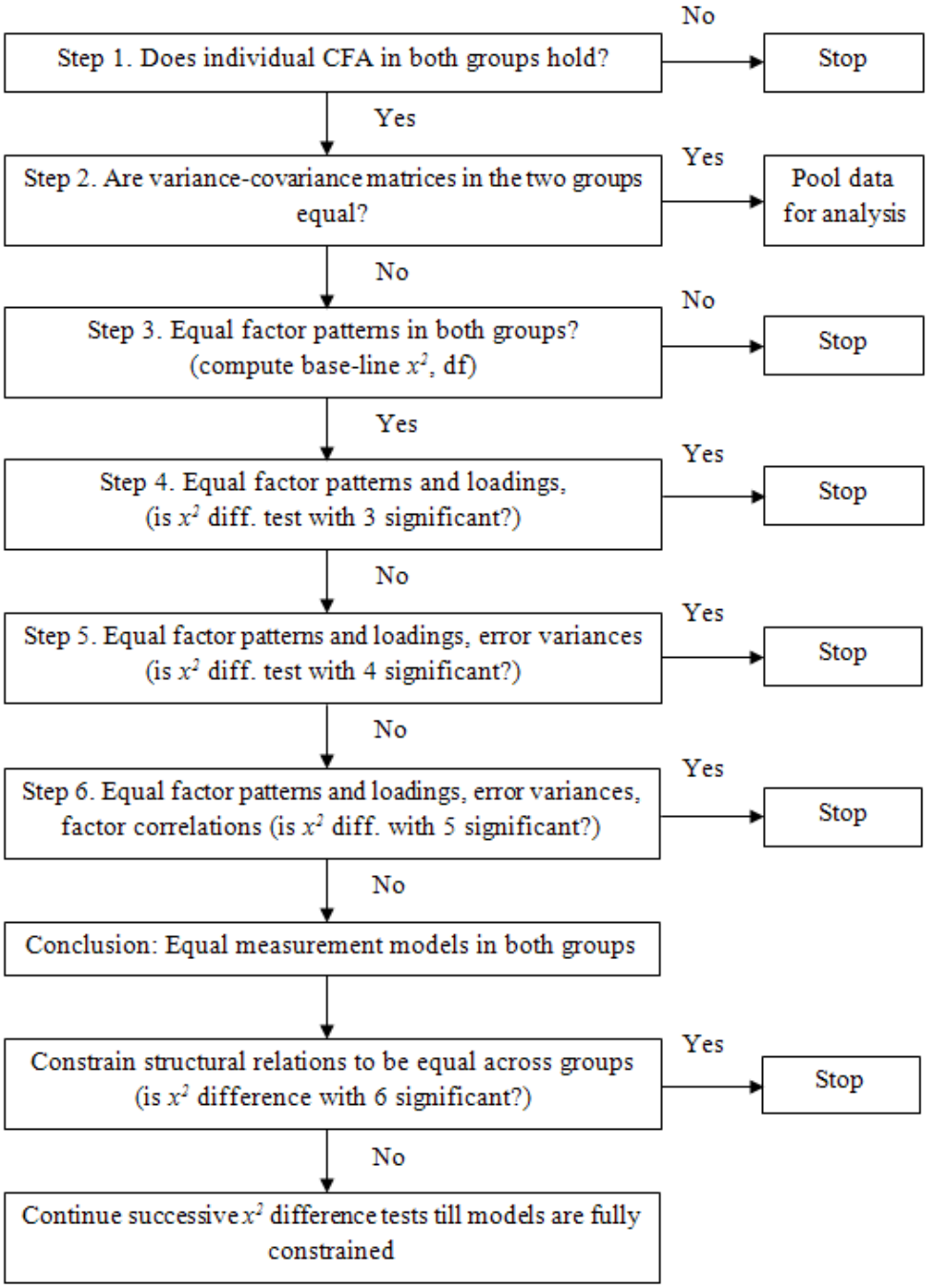

Fig. 3. Flowchart for CTT analysis

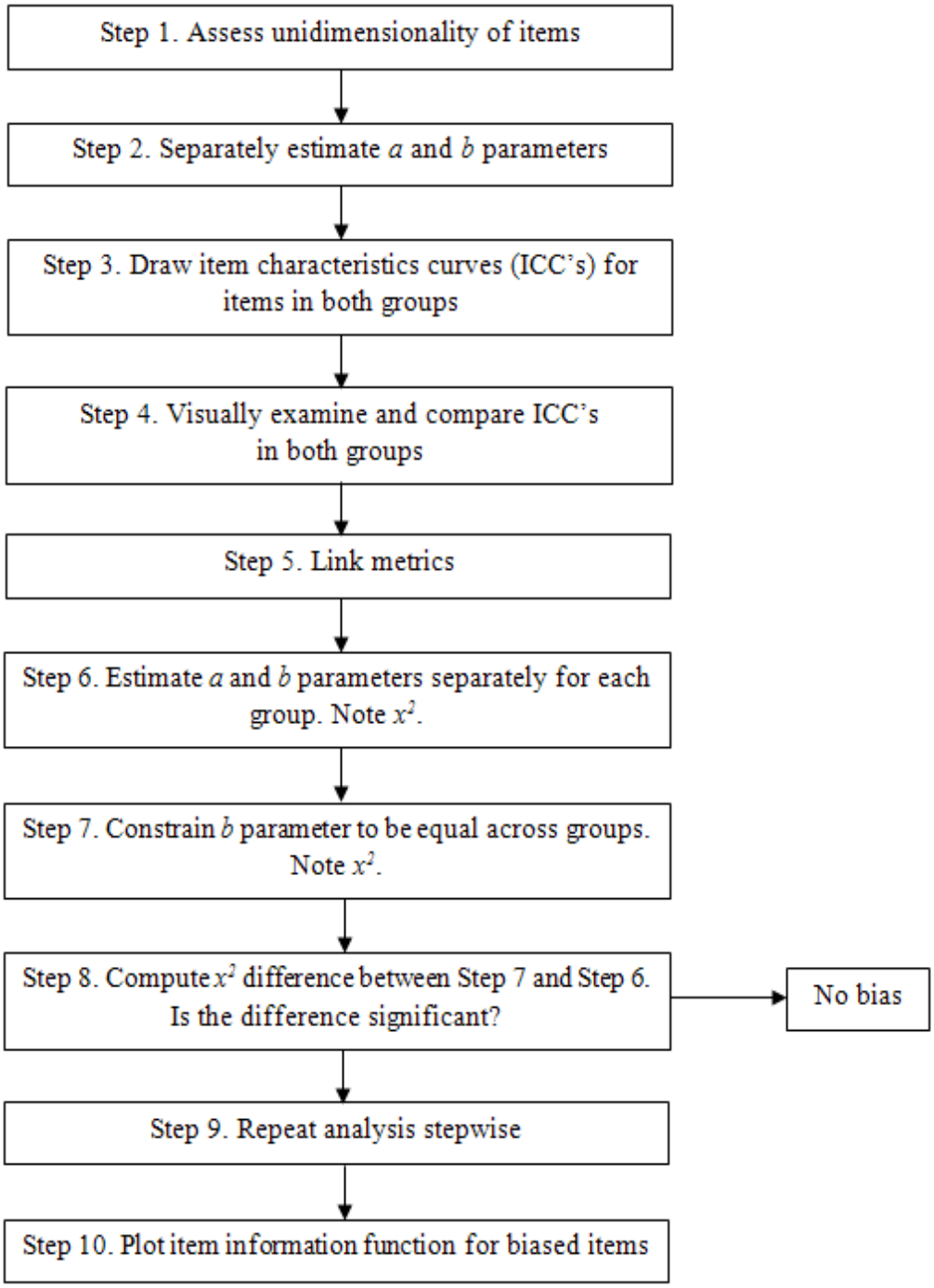

Fig. 4. Flowchart for IRT