Expert Systems with Applications 39 (2012) 10456–10464

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Improving measurement invariance assessments in survey research with missing data by novel artificial neural networks

Liang Ting Tsai, Chih-Chien Yang *

Cognitive NeuroMetrics Laboratory, Graduate Institute of Educational Measurement & Statistics, National Taichung University of Education, 140 MingSheng Road, Taichung 403, Taiwan

ARTICLE INFO

Keywords:
Measurement invariance
LVQ-ESW
Missing at random
Artificial neural networks
Sampling weights

ABSTRACT

This study proposes the learning vector quantization estimated stratum weight (LVQ-ESW) method to interpolate missing group membership and weights in identifying the accuracy of measurement invariance (MI) in a stratified sampling survey. Survey data is rife with missing information, such as gender and race, which is critical for identifying MI, and in ensuring that conclusions from large-scale testing campaigns are accurate. In the current study, simulations were conducted to examine the accuracy and consistency of MI detection using multiple-group confirmatory factor analysis (MG-CFA) to compare different approaches for interpolating missing information. The results of the computerized simulations showed that the proposed method outperformed traditional methods, such as List-wise deletion, in terms of accurately and stably identifying MI. The implications for interpolating missing group membership and weights for survey research are discussed.

© 2012 Elsevier Ltd. All rights reserved.

1. Introduction

Researchers who use the same questionnaires or tests across multiple groups assume that these tests measure the same properties in each group (MacDonald, 1999; Millsap & Kwok, 2004). In other words, the questionnaires or tests must provide measurement invariance (MI) across different groups. When MI holds, the observed test score(s) reflect true diversity in the construct measured between the groups. However, if MI is violated, it is difficult to accurately interpret differences in the observed score (Baure, 2005; Millsap & Kwok, 2004; Raju, Laffitte, & Byrne, 2002). Establishing MI thus has become an essential condition of test construction and research analysis. Nevertheless, the weighting effects of stratified sampling procedures play an important role in the accuracy of MI identification in large scale testing campaigns (Yang & Tsai, 2007, 2008). Even though the application of sampling weights in test analysis is critical in determining the accuracy of MI in a stratified sampling survey, this procedure is often neglected (Yang & Tsai, 2007, 2008). In some cases, group membership information is missing, and this makes the calculation of sampling weights impossible. This raises the question: can researchers simply ignore or delete data with missing group membership?

Many factors can cause systematic differences in the random sampling of groups in many experimental designs. These factors may include missing data, non-responses, or some other unexpected factor. For example, many people leave certain sensitive background questions blank, such as gender, age or ethnicity, when taking anonymous personality tests or questionnaires. However, these missing values encode sampling information that is used to calculate group weights. Deleting or removing any data means a loss of information that can seriously bias conclusions, particularly when researchers attempt to identify MI. Therefore, a proper process to interpolate missing values is required, which could thereby avoid the removal or exclusion of these data, together with other variables of interest (Chen, Tsai, & Yang, 2010).

To solve the problem of missing group membership data in calculating weights and identifying MI, the current study proposes an interpolation algorithm: learning vector quantization estimated stratum weights (LVQ-ESW) (Chen et al., 2010). The LVQ-ESW algorithm uses individuals with complete responses on the independent variable(s) to interpolate missing data on group memberships and weights. Firstly, LVQ (learning vector quantization)—that is, the supervised learning network of an artificial neural network (ANN)—is adopted to find the independent variables that are matched with group membership information, and to analyze their pattern of inter-relationship. Details about the theoretical and mathematical specifications of LVQ algorithms are discussed in the section entitled "The LVQ-ESW Algorithm". Next, these relationships are used to assess possible group memberships and weights for responses on the independent variable that have missing information on group membership. Finally, researchers can interpolate these estimated group memberships and weights into an MI identification process that considers weight values.

* Corresponding author. Tel.: +886 4 2218 3523.
E-mail address: noahyang@ntcu.edu.tw (C.-C. Yang).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.02.048

To ensure the accuracy and consistency of the LVQ-ESW in MI detection, this study used the multiple-group confirmatory factor analysis (MG-CFA) model to identify MI. This study also manipulated levels of missing data rates, sampling sizes, numbers of indicators, percentages of violated MI indicators (PMI), and patterns of invariance to explore the accuracy and consistency of LVQ-ESW performances in calculating unknown group information.

2. Establishing measurement invariance

The identification of MI includes a series of detection processes: (1) invariance of covariance across groups; (2) configural invariance; (3) weak factorial invariance; (4) strong factorial invariance; and (5) strict factorial invariance (Meredith, 1993; Reise, Widaman, & Pugh, 1993; Vandenberg & Lance, 2000). First, the null hypothesis of the invariance of subpopulation covariance tests whether the covariance matrices are equal across all groups. If the invariance of population covariance matrices is rejected, one can continue with the following detection processes with the assumption that the model parameters may be invariant. In the second step, configural invariance requires that the pattern of fixed and free factor loadings is the same across groups with no equality constraints. Next, weak factorial invariance tests whether the factor loadings are equivalent across groups. Fourth, if weak factorial invariance holds, strong factorial invariance assesses whether the intercepts across groups are the same. In the final step, strict factorial invariance tests whether there is invariance in measurement error across groups, if the previous invariances are achieved.

Many approaches have been developed to assess the invariances mentioned above, including the multiple indicators, multiple causes method (MIMIC; Finch, 2005; Oort, 1998) and the multiple group confirmatory factor analysis (MG-CFA; Meade, Johnson, & Braddly, 2008; Millsap & Kwok, 2004; Vandenberg & Lance, 2000). The MG-CFA technique has become the most commonly used method for detecting MI, and is superior to the MIMIC method because it can better accommodate the processes of MI detection outlined above (Meade et al., 2008; Raju et al., 2002; Vandenberg & Lance, 2000). In this study, data was simulated with a lack of MI, and was analyzed with MG-CFA testing for MI.

3. Missing group membership and sampling weights

Since large scale survey databases usually cover a wide array of topics, the original designers are often limited to using broad or frequently used structures in designing sampling and weighting strategies. For example, survey drafters in large cross-country research projects, such as TIMSS (Trends in International Mathematics and Science Study; Martin, 2004) and PISA (Programme for International Students Assessment; OECD, 2005), usually use standard survey structures to frame their sampling designs and to calculate the sampling weights. However, whether or not they accurately reflect demographics within each country remains a highly controversial topic. In addition, it is also debatable whether or not it is possible to distinguish differences in various groups under different combinations. Secondary analysts who use these databases usually require different population structures for their research purposes. While the total population may be identical, completely different clusterings or meanings are possible. For example, gender information might be missing under a sampling design that is grouped by geographical regions. In this case, problems occur when secondary researchers adopt large scale multinational databases to identity MI across gender. These problems require proper correction. Chen et al. (2010) suggests the LVQ-ESW as a technique to allow researchers to regroup some variables from a database, so as to satisfy their own research perspectives. This method is effective in solving frequently encountered, but often neglected, practical problems in different research objectives. However, Chen et al. (2010) only focuses on applying the LVQ-ESW to CFA (confirmatory factor analysis) parameter estimation. The accuracy and stability of the LVQ-ESW in assessing MI in a stratified sampling survey has not been fully established.

4. The LVQ-ESW algorithm

To reach the objective of stable and accurate assessment of MI, this study used the LVQ-ESW to infer missing group memberships and sampling weights for the subjects. These subjects had complete responses on the independent variables but missed information on group membership. Chen et al. (2010) conducted a simulation study to assess the stability and accuracy of using the LVQ-ESW. Unlike other methods for dealing with missing group information (e.g., List-wise deletion), they found that the LVQ-ESW produced more accurate estimations for inferring the population parameters in CFA models. Therefore, this paper adopted the LVQ-ESW to determine the accuracy of MI in a stratified sampling survey.

The basic LVQ algorithm includes learning and recalling processes. The main function of the LVQ is to categorize information and to make predictions about missing information through learning. LVQ is one of the competitive learning networks (Liu, Zuo, Zeng, Vroman, & Rabenasolo, 2010; Luengo, García, & Herrera, 2009), such that it treats the parameters as a kind of "learning," and uses the information gathered from the learning process to infer the category to which the missing information most likely belongs (Chen et al., 2010; Er, Yumusak, & Temurtas, 2010; Liu et al., 2010; Tsai & Yang, 2004). The first step in interpolating missing information for subjects is to establish linkages between group memberships and the independent variables in the sample that have complete group membership and independent variable data. These linkages can be established via the LVQ algorithm. This step is learning process. With these linkages established, the next step is recalling process. Researchers can use these linkages with the target subjects' complete independent variable data to estimate their missing group memberships and weights.

The LVQ framework consists of input, hidden, and output layers. The input layer conveys the observation data—that is, the learning paradigm vectors—to the network. The hidden layer, which interacts with the input vector, represents the parameter calculation and optimization process, and is connected to the input layer through a reference matrix. The hidden layer is partially connected to the output layer, and the connected reference vector is fixed at 1. The output layer then generates a category from the input vector.

In order to obtain a better classification performance, the learning procedure is a very important stage (Yang & Yang, 2002). The simplest LVQ learning process is as follows (Chen et al., 2010; Liu et al., 2010; Luengo et al., 2009; Sagiroglu & Pham, 2000; Su & Chou, 2006; Tsai & Yang, 2004; White, 1989; Yang & Yang, 2002):

Step 1: Initialize the reference vector Z_i of neuron i.

Step 2: Input a learning paradigm vector Y and a corresponding category to the network.

Step 3: Calculate the Euclidean distance between Y and Z_i, where Y_i and Z_{ij} are the jth elements of Y and Z_i, respectively.

$$D_i = \|Z_i - Y\| = \sqrt{\sum_j (Z_{ij} - Y_j)^2} \tag{1}$$

Step 4: Update the reference vector Z_i that is closest to the input vector. The neuron with the smallest Euclidean distance calculated from Step 3 is called the winner, and only this winner is permitted to modify the reference vector. If the classification of the winner matches the category of the training vector, the reference vector moves toward the input vector. The formula for this movement is as follows (Tsai & Yang, 2004; Yang & Yang, 2002).

$$z_i^{(t+1)} = z_i^{(t)} + \gamma^{(t)} \times h_i \times (Y^{(t)} - z_i^{(t)}) \tag{2}$$

Otherwise, the reference vector moves away from the input vector as follows:

$$z_i^{(t+1)} = z_i^{(t)} - \gamma^{(t)} \times h_i \times (Y^{(t)} - z_i^{(t)}) \tag{3}$$

where γ is the learning coefficient. The variable h_i equals 1 if the ith neuron is the winner, and 0 otherwise. For better convergence, the learning coefficient must decrease monotonically with time t (Khuwaja & Abu-Rezq, 2003; Tsai & Yang, 2004). This usually occurs by multiplying a constant k, making the winner move closer to the learning categories. The constant values range from 0 to 1. This process is called moving convergence.

$$\gamma^{(t)} = k \times \gamma^{(t-1)} \tag{4}$$

Step 5: Return to Step 2 with a new learning input vector and repeat the process until all learning vectors are used or the difference, $z_i^{(t+1)} - z_i^{(t)}$, converges to the stopping criterion.

After completing the network learning process, the next procedure is the recall process. The researcher can use the optimum parameter values and related reference matrix obtained from the learning process to predict which group membership subjects belong to. The LVQ algorithm provides subjects with an estimated group membership after the recalling process. The researcher can simply calculate the group weights and interpolate the group membership and weights for the subjects when MI is identified.

To obtain better classification performance when using the LVQ algorithm to estimate missing group memberships, many parameters must be identified, including the learning coefficient, number of hidden layers, number of learning paradigms, etcetera. The improper specification of these parameters will negatively influence classification accuracy and learning time (Tsai & Yang, 2004; Yang & Yang, 2002). This study adopted the same parameter definitions as Chen et al. (2010): the learning coefficient was 0.2 and the constant k was 0.9. The number of learning paradigms was defined by adopting all the subjects with known group membership and repeating the learning process until all paradigms were used.

5. Method

5.1. Design

The primary purpose of this study is to evaluate the stability and accuracy of applying the LVQ-ESW method to detect MI. This study defined a total population of $N = 100{,}000$ and divided this population into two groups of $N_1 = 80{,}000$ and $N_2 = 20{,}000$. Both groups followed the same one-factor CFA model, but differed in terms of factor loadings, with both of these showing a lack of MI across the two groups. Five study variables were manipulated: sample size (200, 500, or 1000); number of total indicators (8 or 16); percentage of violated MI indicators (0%, 25%, or 50%); missing data rates (5%, 10%, or 15%); and pattern of invariance (lower, balanced, and no-difference). Designs capturing the different variable combinations for the two groups are presented in Table 1.

The total sample sizes of 200, 500, and 1000 were further divided into two groups (group 1: n_1, and group 2:n_2) and taken separately from the two groups (N_1 and N_2) to represent small, medium, and large sample sizes. These values were derived from simulation studies that recommended sample sizes from 150 to 200 or above, to adequately detect MI under a variety of circumstances (Meade & Lautenschlager, 2004; Meade et al., 2008).

In the other experimental designs, 8 and 16 represent the number of indicators simulated in shorter and longer surveys, respectively. Each survey was based on a one-factor CFA model rather than a multi-factor model. The values of 0%, 25%, and 50% were adopted to show the percentage of violated MI across groups. From these percentages, numbers of different factor loadings were simulated for 0, 2, and 4 indicators in the shorter survey and 0, 4, and 8 indicators in the longer survey. Except for the indicators denoting a lack of invariance, all the others were designed as reference indicators. The magnitude of the difference in factor loadings between group 1 and group 2 was 0.25 for all conditions containing indicators lacking MI. Indicator intercepts and measurement error invariance were not manipulated in this study. Intercepts were set to zero and measurement error was generated from $N(0,\ 1)$ for all conditions.

From these variable settings, indicator factor loadings can differ across the two groups in three different patterns – lower, balanced, and no-difference. The lower pattern reflects a state where the indicators denoting a lack of MI were set to favor group 1. In other words, group 1 was set to have a higher factor loading than group 2. When the lack of invariance was in a balanced pattern, half of the loadings were set to be higher in group 1, whereas the other half was set to be higher in group 2. Finally, the no-difference pattern exhibits no difference in factor loadings for both groups.

This study will report the results of two simulation studies. In Study 1, the proportionate sampling ($R = 1$) of the two groups was designed by the ratio ($R = w_1/w_2$) of sampling weights from the two groups ($w_1 = N_1/n_1, w_2 = N_2/n_2$). This design attempts to use proportionate sampling to avoid the problem of weighting. In Study 2, the ratio was equal 2 ($R = 2$) to represent the disproportionate sampling of the two groups, thereby reflecting under- or over-sampling conditions. For example, the sampling conditions set parameters of $R = 2$ and a sample size of 500, for group 1, with 333 ($n_1 = 333$) subjects drawn at random with a sampling weight of 240.24 $\left(w_1 = \frac{N_1=80{,}000}{n_1=333} = 240.24\right)$. For group 2, 167 ($n_2 = 167$) subjects were drawn at random, with a sampling weight of 119.76 $\left(w_2 = \frac{N_2=20{,}000}{n_2=167} = 119.76\right)$. In this example, the sampling for group 1 and group 2 reflects under- and over-sampling, respectively. Table 2 presents all the sampling conditions in this study.

To assess the robustness of the LVQ-ESW, three proportions of subjects with missing group membership and weights (5%, 10%, and 15%) were focused on group 2. Therefore, in the example mentioned above, 5% of the combined sample size of 500 had missing information, as 25 subjects in group 2 had missing group memberships and weights, with the remaining 142 subjects having complete information. Similarly, there were 117 and 92 complete subjects after taking into account 10% and 15% missing data rates respectively. These missing data rates were adopted and revised from Ender and Bandalos (2001), Ender and Peugh (2004) and Chen et al. (2010). Table 2 presents other specific sampling conditions.

In this study, all the missing data were designed to occur completely in group 2; and to be missing at random (MAR) (Chen et al., 2010; Little & Rubin, 1987; Little & Schenker, 1994). According to the methodologies of the missing data analysis, only when MAR holds, may researchers interpolate or evaluate the missing data with complete data to ensure correct statistical inferences for various population characteristics (Ender & Peugh, 2004; Little & Rubin, 1987; Little & Schenker, 1994).

5.2. Process of the simulation studies

The same steps were used for both the proportionate and disproportionate simulation studies. First, the subjects in both groups were generated with the Mplus 4.12 software (Muthén & Muthén, 1998–2006) following the framework of the MG-CFA model and the factor loadings as seen in Table 1. Second, the three sample sizes were selected randomly from the two individual groups, with missing group membership for those subjects removed according to the missing proportions as seen in Table 2. The authors wrote a Matlab 7.1 computer program for this sampling procedure. Third,

Table 1
Factor loadings used to generate two groups artificial observations.

	0%		25%				50%				0%		25%				50%			
	No		Lower		Balanced		Lower		Balanced		No		Lower		Balanced		Lower		Balanced	
	G1	G2	G1	G2	G1	G2	G1	G2	G1	G2	G1	G2	G1	G2	G1	G2	G1	G2	G1	G2
λ_1	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
λ_2	.6	.6	.6	.35	.6	.85	.6	.35	.6	.85	.6	.6	.6	.35	.6	.85	.6	.35	.6	.85
λ_3	.6	.6	.6	.35	.6	.35	.6	.35	.6	.85	.6	.6	.6	.35	.6	.85	.6	.35	.6	.85
λ_4	.6	.6	.6	.6	.6	.6	.6	.35	.6	.35	.6	.6	.6	.35	.6	.35	.6	.35	.6	.85
λ_5	.6	.6	.6	.6	.6	.6	.6	.35	.6	.35	.6	.6	.6	.35	.6	.35	.6	.35	.6	.85
λ_6	.6	.6	.6	.6	.6	.6	.6	.6	.6	.6	.6	.6	.6	.6	.6	.6	.6	.35	.6	.35
λ_7	.6	.6	.6	.6	.6	.6	.6	.6	.6	.6	.6	.6	.6	.6	.6	.6	.6	.35	.6	.35
λ_8	.6	.6	.6	.6	.6	.6	.6	.6	.6	.6	.6	.6	.6	.6	.6	.6	.6	.35	.6	.35
λ_9											.6	.6	.6	.6	.6	.6	.6	.35	.6	.35
λ_{10}											.6	.6	.6	.6	.6	.6	.6	.6	.6	.6
λ_{10}											.6	.6	.6	.6	.6	.6	.6	.6	.6	.6
λ_{12}											.6	.6	.6	.6	.6	.6	.6	.6	.6	.6
λ_{13}											.6	.6	.6	.6	.6	.6	.6	.6	.6	.6
λ_{14}											.6	.6	.6	.6	.6	.6	.6	.6	.6	.6
λ_{15}											.6	.6	.6	.6	.6	.6	.6	.6	.6	.6
λ_{16}											.6	.6	.6	.6	.6	.6	.6	.6	.6	.6

Note: no: no-difference pattern; lower: all lower pattern; balanced: balanced pattern; G1: group 1; G2: group 2.

Table 2
Design for the simulation study.

Total sampling sizes (n)	R	Projected stratum sizes		Sample sizes by missing rates		
				5%	10%	15%
		n_1	n_2	n_2	n_2	n_2
200	2	133	67	57	47	37
	1	160	40	30	20	10
500	2	333	167	142	117	92
	1	400	100	75	50	25
1000	2	667	333	283	233	183
	1	800	200	150	100	50

three methods were used to handle subjects with missing group membership. The first method was List-wise deletion (LWD), which deletes subjects with missing group membership, and uses the remaining complete subjects to infer the accuracy of detection rates for MI identification. The second method was the weighting-adjustment class (WAC; see, e.g., Lohr, 2010), which deletes the subjects with missing group membership and then reweights the membership based on the remaining complete observations. In the previous example, for 5% of the sample with missing group membership information in group 2, the new sampling weight for the 142 subjects with complete information was 140.85 $\left(w_2 = \frac{N_2=20.000}{n_2=142} = 140.85\right)$. The third method used to interpolate the missing group membership and weights was the LVQ-ESW. In this case, the LVQ was used to classify the 25 subjects with missing group membership. If the LVQ inferred the subject with missing data to group 1, the LVQ-ESW recorded the weights to be 240.24, but 119.76 if the LVQ inferred that subject to group 2. Following these three methods of handling subjects with missing group membership and weights, the datasets were used to assess MI.

Two nested models were used to identify a lack of MI. In Model 1, the baseline model, the datasets for group 1 and group 2 were analyzed simultaneously, but the factor loadings for the indicator were allowed to vary across groups. Model 2, the constrained model, tested the factor loadings across groups by examining a model that was identical to the baseline model, apart from the factor loadings being controlled to be equal across groups (Meade & Lautenschlager, 2004). Both models were estimated by the robust maximum likelihood estimator (MLR) using Mplus 4.21 (Muthén & Muthén, 1998–2006). With this estimator, the individual model x^2 values are scaled, and cannot be used for nested model comparisons because the difference between two scaled x^2 values is not distributed as x^2. To compare models, the Satorra-Bentler scaled chi-square difference test (TRd) was then used to evaluate the significance of the increment in fit for the nested model (Satorra & Bentler, 2001). The formulas for this test are as follows (http://www.statmodel.com/chidiff.shtml):

$$TRd = (T_0 \times C_0 - T_1 \times C_1)/cd \tag{5}$$

$$cd = (d_0 \times C_0 - d_1 \times C_1)/(d_0 - d_1) \tag{6}$$

where T_0 and T_1 are MLR chi-square values for the nested and comparison model, respectively, df_0 is the degrees of freedom in the nested model, C_0 is the scaling correction factor for the nested model, df_1 is the degrees of freedom in the comparison model, and C_1 is the scaling correction factor for the comparison model. An alpha level of .05 was used to determine statistical significance. A particular example of the MLR estimation method and the adjusted chi-square difference test can be found on the Mplus website, Muthén & Muthén (1998–2006) and Woods (2009).

To confirm the stability of the experimental results in this study, 200 replications were conducted to achieve a reasonable estimation in each condition. Each replication involved the three step processes: the sampling of two individual groups, interpolating the missing group membership and weights, and identifying MI. The true-positive (TP) rates of the List-wise deletion, the WAC and LVQ-ESW methods were assessed by the proportion of times MI was correctly identified in 200 replications under both lower and balanced patterns. Conversely, the false-positive (FP) rates of identifying MI under the no-difference pattern were assessed by the proportion of times that MI was mistakenly identified in 200 replications.

6. Results for Study 1: proportionate sampling

6.1. No difference pattern

Fig. 1 displays the FP rate profiles for the no-difference pattern of MI indicators under proportionate sampling. The numerical results are shown in Appendix A. When questionnaires did not have any violated indicators, the FP rates ranged from 0.03 to 0.07, with a mean of 0.054 for List-wise deletion, 0.03–0.07, with a mean of 0.053 for WAC, and 0.04–0.07, with a mean of 0.056 for LVQ-ESW. The FP rates were very close to the expected value of 0.05, even when 15% of the sampling subjects had missing group membership. Thus, it appears that all three methods were able to yield well-controlled FP rates when questionnaires did not contain any violated indicators under proportionate sampling.

6.2. Lower and balanced pattern

Fig. 2 also shows the detection rate profiles for the three methods. The numerical results are also shown in Appendix A. However, the detection rates represent the true positive rates under the lower and balanced patterns. The left side of Fig. 2 displays the detection rates of factor loading invariance for the lower pattern. The detection rates for the three methods increased as the sample sizes increased. Apart from a ceiling effect with large sampling sizes, the LVQ-ESW method outperformed both the List-wise deletion and WAC methods. As anticipated, although the detection rates for

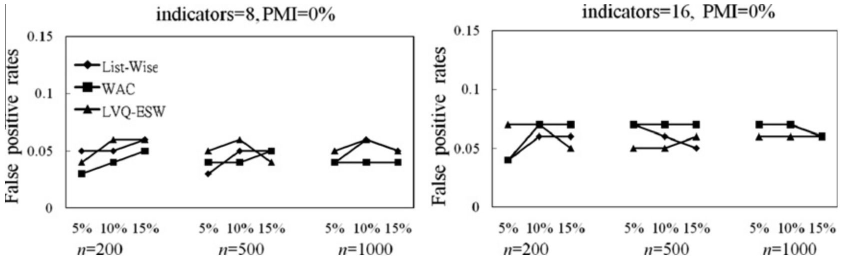

Fig. 1. False positive rates of measurement invariance detection under proportionate sampling.

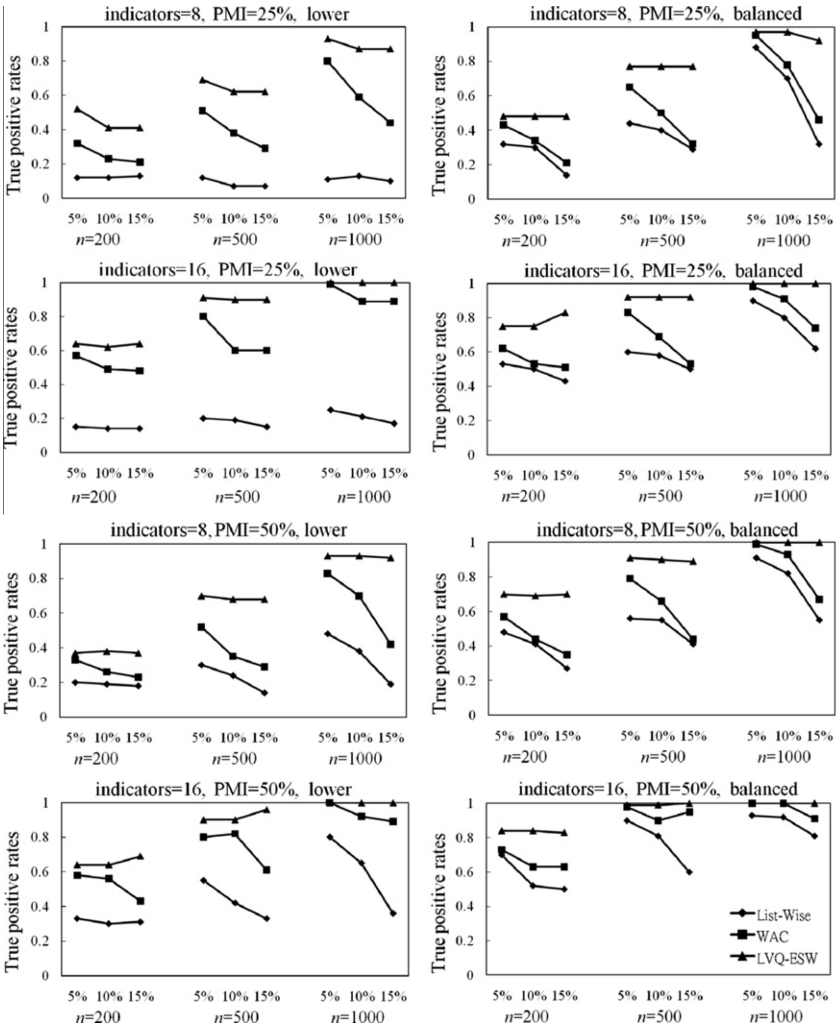

Fig. 2. True positive rates of measurement invariance detection under proportionate sampling.

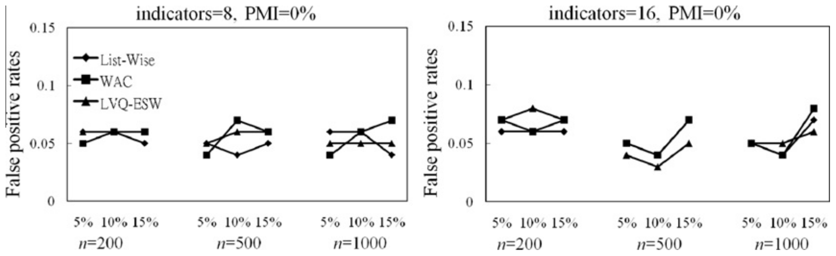

Fig. 3. False positive rates of measurement invariance detection under disproportionate sampling.

List-wise deletion and WAC were greatly influenced by the rate of missing data, with poorer detection as the rate of missing data increased, the LVQ-ESW outperformed the other two methods by being robust against any rate of missing data. Under almost all the conditions, varied across the three methods, more indicators in the survey offered higher detection rates. For instance, under the condition with 10% missing data and a sample size of 1000, and 8 indicators, the List-wise deletion, WAC, and LVQ-ESW methods reported detection rates of 0.13, 0.59, and 0.87, respectively. However, for 16 indicators, the detection rates were 0.21, 0.89, and 1.00, respectively. Finally, for almost all conditions, higher PMI yielded higher detection rates, except for small sample sizes ($n = 200$). In summary, the LVQ-ESW method produced higher rates of MI identification compared to both List-wise deletion and WAC.

The right side of Fig. 2 shows the detection rates for balanced-pattern factor loadings. Under this pattern, detection rates were higher than those under the lower pattern, consistent with the conclusions of Meade and Lautenschlager (2004). Specifically, the pattern of effects under the balanced pattern was similar to that described above for the lower pattern.

7. Results for Study 2: disproportionate sampling

7.1. No difference pattern

Fig. 3 indicates that the FP positive rates for the no difference pattern of MI indicators under disproportionate sampling were all close to 0.05 and similar for all three methods. Across all conditions that questionnaires did not have any violated indicators, the FP rates ranged from 0.04 to 0.07, with a mean of 0.054 for List-wise deletion, from 0.04 to 0.08, with a mean of 0.058 for WAC, and from 0.03 to 0.08, with a mean of 0.056 for LVQ-ESW (see Appendix B). In short, all three methods for dealing with missing group membership and weights under disproportionate sampling seemed not to be affected by sample size, percentage of missing MI indicators, and missing data rates at all.

7.2. Lower and balanced pattern

Fig. 4 shows the detection rate profiles for the List-wise deletion, WAC, and LVQ-ESW methods under disproportionate sampling. The numerical results are shown in Appendix B. The left and right sides of Fig. 4 display the detection rates for the lower and balanced patterns respectively. Under the same conditions, the detection rates for the three methods of disproportionate sampling were higher than those under proportionate sampling, except when the detection rate encountered the ceiling effect, or when sample sizes were small ($n = 200$). For instance, under the condition with a rate of 10% missing data, 25% PMI, a balanced pattern, and a sample sizes of 500, the List-wise deletion, WAC, and LVQ-ESW methods achieved detection rates of 0.40, 0.50, and 0.77 respectively for proportionate sampling, and 0.57, 0.80, and 0.87 respectively for disproportionate sampling. This was expected, because the sample sizes for group 2 (n_2) under disproportionate sampling were greater than their counterparts under proportionate sampling for the same sample sizes (n).

With respect to the overall trends listed above, the detection rates under disproportionate sampling were very similar to those under proportionate sampling, when the List-wise deletion, WAC, and LVQ-ESW methods were used for the lower and balanced patterns. For these three methods, the detection rates increased as the sampling sizes increased. The LVQ-ESW method outperformed the List-wise deletion and WAC methods except for when the detection rates reached ceiling because of large sample sizes.

8. Conclusions and recommendations

This article proposes a learning vector quantization estimated missing stratum weights (LVQ-ESW) method to infer the accuracy of detecting measurement invariance (MI) in a stratified sampling survey. The results of computerized numerical simulations showed that the LVQ-ESW method was more accurate, stable, and reliable than List-wise deletion and WAC when accurately detecting MI. For instance, across a variety of conditions, the LVQ-ESW method produced a higher detection rate than the other two methods. This method is convenient for interpolating important information for calculating group weights because it uses the LVQ algorithm to estimate the possible group membership of the samples. The superiority of this approach lies in its ability to categorize information and to make robust predictions about missing information.

Specific to the variable manipulated in the simulations, for proportionate sampling, almost all conditions under the lower and balanced patterns showed that the greater the sampling size, and the longer the survey, the higher the detection rates were for List-wise deletion and WAC. Similarly, higher PMI resulted in higher MI detection rates, except when detection rates reached the ceiling. These tendencies could also be found in the conditions with disproportionate sampling. In this case, WAC clearly outperformed List-wise deletion, because the sum of weights for the WAC method was equal to the population size. The LVQ-ESW method, unlike List-wise deletion and WAC, produced higher detection rates in identifying MI, because the efficiency and stability of the LVQ algorithm as a predictive tool was able to estimate the group memberships and weights correctly. However, the analytic sample size used in the LVQ-ESW method was larger than the List-wise deletion and WAC methods, and this may have contributed to the higher detection rates.

Sampling weights play an important role in the statistical analysis of population inferences with disproportionate sampling or missing data in survey research. Therefore, researchers adopt proportionate sampling techniques to avoid the problem of differential weighting in statistical analyses. However, it is fairly common for respondents to refuse particular questionnaires, or to leave certain questions blank. This poses a problem for researchers, because the people who refuse to answer questionnaires are typical examples of people who carry systematic “MAR” information. This study attempted to simulate this phenomenon of loss

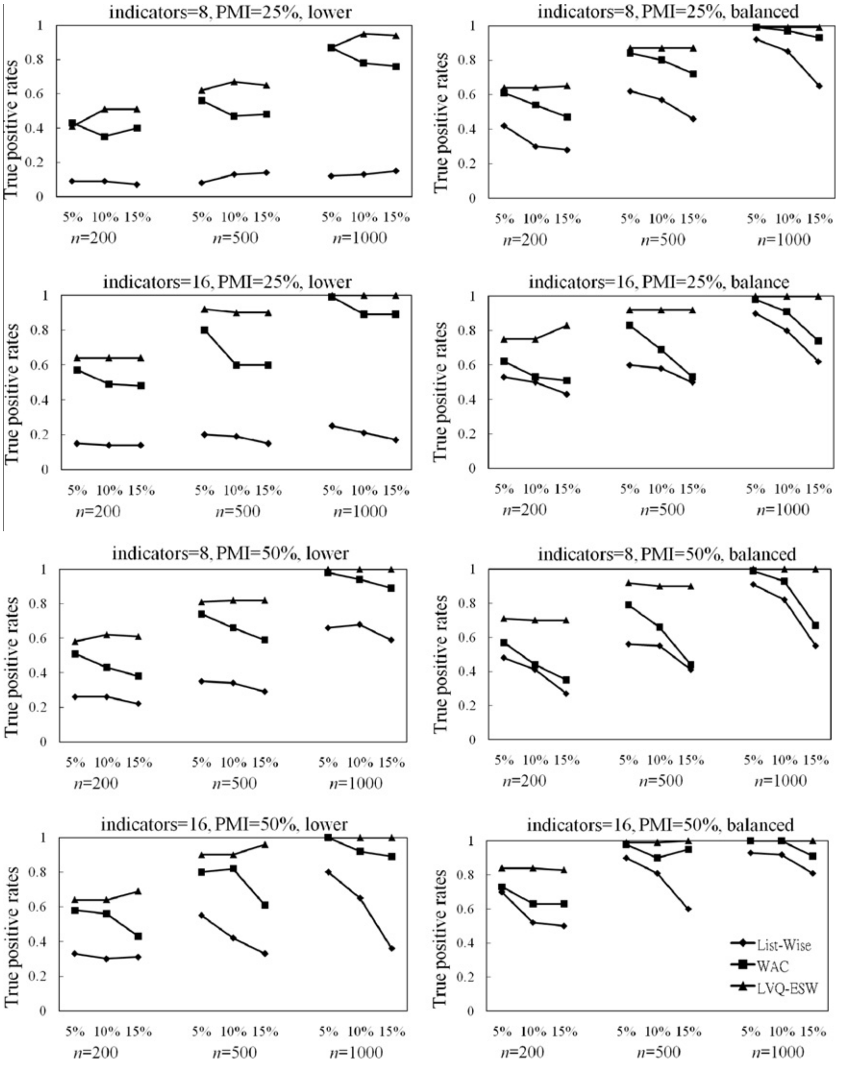

Fig. 4. True positive rates of measurement invariance detection under disproportionate sampling.

of systematic MAR by focusing all of the missing group information in the same group. The results of the simulation showed that, regardless of whether sampling was disproportionate or whether the missing data was systematic, detection rates for identifying MI were lower when missing group membership and weights were not properly interpolated.

In terms of practicality in applying the LVQ-ESW, apart from customized computer programs, artificial neural network computer packages (e.g., NeuroSolution or PCNeuro) also can run the LVQ-ESW. The increasing accessibility of the LVQ algorithm should see greater flexibility in experimental designs and better statistical control, for more reliable conclusions.

Appendix A. Detection rates of measurement invariance identification under proportionate sampling

Indicators	Percentage of a lack of MI (%)	Sampling size (n)	List-wise			WAC			LVQ-ESW		
			5%	10%	15%	5%	10%	15%	5%	10%	15%
8	0% (No-difference)	200	0.05	0.05	0.06	0.03	0.04	0.05	0.04	0.06	0.06
		500	0.03	0.05	0.05	0.04	0.04	0.05	0.05	0.06	0.04
		1000	0.04	0.06	0.05	0.04	0.04	0.04	0.05	0.06	0.05
	25% (Lower)	200	0.12	0.12	0.13	0.32	0.23	0.21	0.52	0.41	0.41
		500	0.12	0.07	0.07	0.51	0.38	0.29	0.69	0.62	0.62
		1000	0.11	0.13	0.10	0.80	0.59	0.44	0.93	0.87	0.87
	25% (Balanced)	200	0.32	0.30	0.14	0.43	0.34	0.21	0.48	0.48	0.48
		500	0.44	0.40	0.29	0.65	0.50	0.32	0.77	0.77	0.77
		1000	0.88	0.70	0.32	0.95	0.78	0.46	0.97	0.97	0.92

Appendix A (*continued*)

Indicators	Percentage of a lack of MI (%)	Sampling size (n)	List-wise			WAC			LVQ-ESW		
			5%	10%	15%	5%	10%	15%	5%	10%	15%
	50% (Lower)	200	0.20	0.19	0.18	0.33	0.26	0.23	0.37	0.37	0.37
		500	0.30	0.24	0.14	0.52	0.35	0.29	0.68	0.68	0.68
		1000	0.48	0.38	0.19	0.83	0.70	0.42	0.93	0.93	0.93
	50% (Balanced)	200	0.48	0.41	0.27	0.57	0.44	0.35	0.71	0.70	0.70
		500	0.56	0.55	0.41	0.79	0.66	0.44	0.92	0.90	0.90
		1000	0.91	0.82	0.55	0.99	0.93	0.67	1.00	1.00	1.00
16	0% (No-difference)	200	0.04	0.06	0.06	0.04	0.07	0.07	0.07	0.07	0.05
		500	0.07	0.06	0.05	0.07	0.07	0.07	0.05	0.05	0.06
		1000	0.07	0.07	0.06	0.07	0.07	0.06	0.06	0.06	0.06
	25% (Lower)	200	0.15	0.14	0.14	0.57	0.49	0.48	0.66	0.64	0.64
		500	0.20	0.19	0.15	0.80	0.60	0.60	0.92	0.90	0.90
		1000	0.25	0.21	0.17	0.99	0.89	0.89	1.00	1.00	1.00
	25% (Balanced)	200	0.53	0.50	0.43	0.62	0.53	0.51	0.75	0.75	0.83
		500	0.60	0.58	0.50	0.83	0.69	0.53	0.92	0.92	0.92
		1000	0.90	0.80	0.62	0.98	0.91	0.74	1.00	1.00	1.00
	50% (Lower)	200	0.33	0.30	0.31	0.58	0.56	0.43	0.64	0.64	0.69
		500	0.55	0.42	0.33	0.80	0.82	0.61	0.90	0.90	0.96
		1000	0.80	0.65	0.36	1.00	0.92	0.89	1.00	1.00	1.00
	50% (Balanced)	200	0.70	0.52	0.50	0.73	0.63	0.63	0.84	0.84	0.83
		500	0.90	0.81	0.60	0.98	0.90	0.95	0.99	0.99	1.00
		1000	0.93	0.92	0.81	1.00	1.00	0.91	1.00	1.00	1.00

Appendix B. Detection rates of measurement invariance identification under disproportionate sampling

Indicators	Percentage of a lack of MI (%)	Sampling size (n)	List-wise			WAC			LVQ-ESW		
			5%	10%	15%	5%	10%	15%	5%	10%	15%
8	0% (No-difference)	200	0.06	0.06	0.05	0.05	0.06	0.06	0.06	0.06	0.06
		500	0.05	0.04	0.05	0.04	0.07	0.06	0.05	0.06	0.06
		1000	0.06	0.06	0.04	0.04	0.06	0.07	0.05	0.05	0.05
	25% (Lower)	200	0.09	0.09	0.07	0.43	0.35	0.40	0.41	0.51	0.51
		500	0.08	0.13	0.14	0.56	0.47	0.48	0.62	0.67	0.65
		1000	0.12	0.13	0.15	0.87	0.78	0.76	0.87	0.95	0.94
	25% (Balanced)	200	0.42	0.30	0.28	0.61	0.54	0.47	0.64	0.64	0.65
		500	0.62	0.57	0.46	0.84	0.80	0.72	0.87	0.87	0.87
		1000	0.92	0.85	0.65	0.99	0.97	0.93	0.99	0.99	0.99
	50% (Lower)	200	0.26	0.26	0.22	0.51	0.43	0.38	0.58	0.62	0.61
		500	0.35	0.34	0.29	0.74	0.66	0.59	0.81	0.82	0.82
		1000	0.66	0.68	0.59	0.98	0.94	0.89	1.00	1.00	1.00
	50% (Balanced)	200	0.59	0.42	0.35	0.78	0.71	0.60	0.82	0.82	0.81
		500	0.92	0.75	0.52	0.95	0.93	0.86	0.98	0.98	0.97
		1000	0.92	0.82	0.79	1.00	1.00	0.99	1.00	1.00	1.00
16	0% (No-difference)	200	0.06	0.06	0.06	0.07	0.06	0.07	0.07	0.08	0.07
		500	0.05	0.04	0.07	0.05	0.04	0.07	0.04	0.03	0.05
		1000	0.05	0.04	0.07	0.05	0.04	0.08	0.05	0.05	0.06
	25% (Lower)	200	0.15	0.12	0.13	0.78	0.70	0.70	0.81	0.82	0.82
		500	0.26	0.28	0.26	0.94	0.94	0.93	0.97	0.97	0.96
		1000	0.41	0.40	0.37	1.00	1.00	1.00	1.00	1.00	1.00
	25% (Balanced)	200	0.72	0.70	0.64	0.85	0.80	0.71	0.90	0.91	0.90
		500	0.88	0.85	0.70	0.97	0.95	0.88	0.99	0.99	0.99
		1000	0.94	0.97	0.90	1.00	1.00	1.00	1.00	1.00	1.00
	50% (Lower)	200	0.39	0.40	0.34	0.76	0.66	0.56	0.83	0.82	0.83
		500	0.67	0.65	0.55	0.96	0.93	0.86	0.97	0.97	0.97
		1000	0.91	0.88	0.82	1.00	1.00	1.00	1.00	1.00	1.00
	50% (Balanced)	200	0.80	0.72	0.65	0.96	0.93	0.93	0.98	0.98	0.97
		500	0.90	0.88	0.70	1.00	0.99	0.97	1.00	1.00	1.00
		1000	0.95	0.96	0.89	1.00	1.00	1.00	1.00	1.00	1.00

References

Baure, D. J. (2005). The role of nonlinear factor-to-indicator relationships in test of measurement equivalence. *Psychological Method, 10*(3), 305–316.

Chen, C. R., Tsai, L. T., & Yang, C. C. (2010). Supervised learning vector quantization for projecting missing weights of hierarchical neural networks. *WSEAS Transactions on Information Science and Applications, 6*(7), 799–808.

Ender, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling, 8*(3), 430–457.

Ender, C. K., & Peugh, J. L. (2004). Using an EM covariance matrix to estimate structural equation models with missing data: choosing an adjusted sample size to improve the accuracy of inferences. *Structural Equation Modeling, 11*(1), 1–19.

Er, o., Yumusak, N., & Temurtas, F. (2010). Chest diseases diagnosis using artificial neural networks. *Expert Systems with Applications, 37*(12), 7648–7655.

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement, 29*(4), 278–295.

Khuwaja, G. A., & Abu-Rezq, A. N. (2003). Data acquisition and recognition of fingerprints with LVQ. *International Journal of Computational Intelligence & Application, 3*(1), 65–88.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: John Wiley.

Little, R. J. A., & Schenker, N. (1994). Missing data. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 39–75). New York: Plenum.

Liu, J., Zuo, B., Zeng, X., Vroman, P., & Rabenasolo, B. (2010). Nonwoven uniformity identification using wavelet texture analysis and LVQ neural network. *Expert Systems with Applications, 37*(3), 2214–2246.

Lohr, S. (2010). *Sampling: Design and analysis* (2nd ed.). Pacific Grove, CA: Duxbury Press.

Luengo, J., García, S., & Herrera, F. (2009). A study on the use of statistical tests for experimentation with neural networks: Analysis of parametric test conditions and non-parametric tests. *Expert Systems with Applications, 36*(4), 7798–7808.

MacDonald, R. P. (1999). *Test theory: A unified perspective*. Mahwah, NJ: Erlbaum.

Martin, M. O. (2004). *TIMSS 2003 user guide for the international database*. Chestnut Hill, MA: Boston College.

Meade, A. M., Johnson, E. C., & Braddly, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology, 93*(3), 568–592.

Meade, A. W., & Lautenschlager, G. J. (2004). A Monte-Carlo study of confirmatory factor analytic tests of measurement equivalence/invariance. *Structural Equation Modeling, 11*(1), 60–72.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika, 58*(4), 525–543.

Millsap, R. E., & Kwok, O. M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods, 9*(1), 93–115.

Muthén, L. K., & Muthén, B. O. (1998–2006). Mplus user's guide (4th ed.). Los Angeles, CA: Muthén & Muthén.

OECD (2005). PISA 2003 technical report. Paris: OECD.

Oort, F. J. (1998). Simulation study of item bias detection with restricted factor analysis. *Structural Equation Modeling, 5*(2), 107–124.

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence. A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87*(3), 517–529.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*(3), 552–566.

Sagiroglu, S., & Pham, D. T. (2000). Neural network classification of defects in vector boards. *Journal of Process Mechanical Engineering Part B, 214*(3), 255–258.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66*(4), 507–514.

Su, C. T., & Chou, C. J. (2006). A neural network-based approach for statistical probability distribution recognition. *Quality Engineering, 18*, 293–297.

Tsai, L. T., & Yang, C. C. (2004). A simulation study on classified accuracy for learning vector quantization. *Journal of Research on Measurement & Statistics, 12*, 269–291.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4–69.

White, H. (1989). Some asymptotic results for learning in single hidden layer feedforward network models. *Journal of the American Statistical Associate, 84*(408), 1003–1013.

Woods, C. M. (2009). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research, 44*(1), 1–27.

Yang, C. C., & Tsai, L. T. (2007). Weighting effects of stratified sampling procedures on the accuracy of DIF identifications. In *Seventy-fourth international meeting of the psychometric society*.

Yang, C. C., & Tsai, L. T. (2008). Inferring measurement equivalence between Likert-type questionnaires under effects of sampling weights. *Chinese Journal of Psychology, 50*(3), 257–269.

Yang, M. S., & Yang, J. H. (2002). A fuzzy-soft learning vector quantization for control chart pattern recognition. *International Journal of Production Research, 40*(12), 2721–2731.