Expert Systems with Applications 39 (2012) 4026–4031

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

A simulation study using EFA and CFA programs based the impact of missing data on test dimensionality

Shin-Feng Chen [a], Shuyi Wang [b], Chen-Yuan Chen [c,d,e,*]

[a] Department of Education, National Pingtung University of Education, No. 4-18, Ming Shen Rd., Pingtung 90003, Taiwan
[b] Department of Measurement, Statistics and Evaluation, University of Maryland, College Park, MD 20742, USA
[c] Department and Graduate School of Computer Science, National Pingtung University of Education, No. 4-18, Ming Shen Rd., Pingtung 90003, Taiwan
[d] Global Earth Observation and Data Analysis Center (GEODAC), National Cheng Kung University, No 1, Ta-Hsueh Road, Tainan 701, Taiwan
[e] Department of Information Management, National Kaohsiung First University of Science and Technology, 2 Jhuoyue Rd. Nanzih, Kaohsiung 811, Taiwan

ARTICLE INFO

Keywords:
Data imputation
Test dimensionality
Confirmatory factor analysis
Exploratory factor analysis
Statistics package for social science

ABSTRACT

This study examines the impact of missing rates and data imputation methods on test dimensionality. We consider how missing rate levels (10%, 20%, 30%, and 50%) and the six missed data imputation methods (Listwise, Serial Mean, Linear Interpolation, Linear Trend, EM, and Regression) affect the structure of a test. A simulation study is conducted using the SPSS 15.0 EFA and CFA programs. The EFA results for the six methods are similar, and all results obtained two factors. The CFA results also fit the hypothesized two factor structure model for all six methods. However, we observed that the EM method fits the EFA results relatively well. When the percentage of missing data is less than 20%, the impact of the imputation methods on test dimensionality is not statistically significant. The Serial Mean and Linear Trend methods are suggested for use when the percentage of missing data is greater than 30%.

1. Objectives or purpose

Missing data often occur in large scale surveys or testing (Chen & Chen, 2010a; Lin & Chen, 2011). Some previous reports revealed its influence on scientific observation (Trabia, Renno, & Moustafa, 2008; Chen, 2009a, 2009b; Chen, 2010a, 2010c, 2010d; Chen, 2010b, 2010c, 2010d; Chen, 2011a, 2011b, 2011c, 2011d, 2011e, 2011f). The significance of the analysis depends heavily on the accuracy of the dataset. Therefore, the issue of missing data must be addressed since ignoring this problem can cause bias in the models being evaluated and lead to inaccurate conclusions.

Assessing the test structure is also an important issue for the development, evaluation and maintenance of large-scale tests (Tate, 2003). One important reason is that an assessment may provide empirical support for the content and cognitive process aspects for test validity. An assessment could provide evidence of validity based on the internal structure. This is attained by assessing the test structure. This type of evidence includes relationships among items within and across sub-scales, and consistency between sub-scale relationships and understanding regarding the construct or content domain. Evidence based on the internal structure is normally correlational, including item-scale relationships between both the scale the item is assigned to and the other scales in the test, as well as scale-to-scale relationships. Methods that are designed to study groups of relationships, such as factor analysis, are also commonly used.

Therefore, the purposes of this study are two-fold: (1) to investigate the effects of the missing data on the test structure; and (2) to examine the test structure (dimensionality of a test) by using different missing data imputation methods.

2. Missing data

The missing item values can be classified with respect to their underlying missing data mechanisms. As described by Little and Rubin (1987), these mechanisms include missing completely at random (MCAR), missing at random (MAR), and missing not at random (NMAR).

(1) Missing completely at random (MCAR): $f_{\psi,\theta}(R) = f_{\psi}(R|Y_{obs}, Y_{mis})$: Data are considered to be MCAR if missingness occurs only by chance. When data are considered to be MCAR, a person's missing value for a variable is independent of that person's values for other observable variables in the model. For example, suppose a survey is designed to measure attendance at school. The survey is administered to a group of students but an item is accidentally skipped by an individual due to some random interruption. This item would be considered MCAR. Planned missingness, such as when certain sets of items are given to one sample and other sets are given to another sample, would also result in MCAR missingness (Peugh & Enders, 2004).

* Corresponding author.
E-mail address: cyc@mail.npue.edu.tw (C.-Y. Chen).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.09.085

Table 1
Research design.

Missingness% Methods	10%	20%	30%	50%
List	Dimensionality (EFA/CFA)	Dimensionality (EFA/CFA)	Dimensionality (EFA/CFA)	Dimensionality (EFA/CFA)
Mean	Dimensionality (EFA/CFA)	Dimensionality (EFA/CFA)	Dimensionality (EFA/CFA)	Dimensionality (EFA/CFA)
Lint	Dimensionality (EFA/CFA)	Dimensionality (EFA/CFA)	Dimensionality (EFA/CFA)	Dimensionality (EFA/CFA)
Trend	Dimensionality (EFA/CFA)	Dimensionality (EFA/CFA)	Dimensionality (EFA/CFA)	Dimensionality (EFA/CFA)
EM	Dimensionality (EFA/CFA)	Dimensionality (EFA/CFA)	Dimensionality (EFA/CFA)	Dimensionality (EFA/CFA)
Regression	Dimensionality (EFA/CFA)	Dimensionality (EFA/CFA)	Dimensionality (EFA/CFA)	Dimensionality (EFA/CFA)

Table 2
Data generation structure and CFA results.

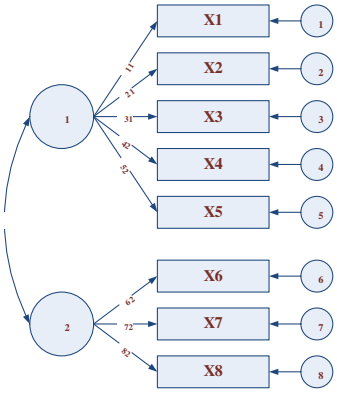
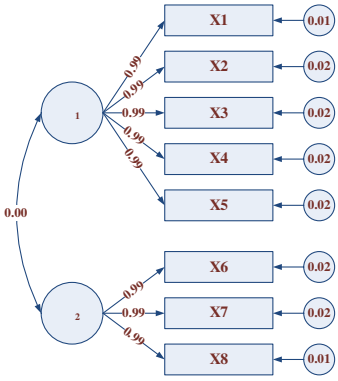
Data generation structure	Data generation (CFA results)	
		SRMR:0.0030 Eigen Value:0.61652/0.37055(98.707) Chi-Square/DF:105.625/19 Pr>Chi-Square:0.000 NFI:0.995 NNFI:0.994 CFI:0.996 GFI:0.973 SRMR:0.0030 AGFI:0.949 RMSEA:0.068 AIC:139.625 BIC:223.056

Table 3
EFA data generation results.

	Min	Max	Mean	SD	Factor1	Factor2
X1	1	7	4.04	1.050	0.994	
X2	1	7	4.04	1.048	0.992	
X3	1	7	4.04	1.056	0.993	
X4	1	7	4.04	1.050	0.993	
X5	1	7	4.04	1.042	0.993	
X6	1	7	4.05	1.038		0.994
X7	1	7	4.05	1.031		0.993
X8	1	7	4.05	1.030		0.995
0.98707					$\lambda_1 = 0.61652$	$\lambda_2 = 0.37055$

(2) Missing at random (MAR): $f_\psi(R|Y_{obs}, Y_{mis}) = f_\psi(R|Y_{obs})$: Data are considered MAR if the missingness does not depend upon the missing data itself but also on some characteristic of the observed data. An example of this is accidentally omitting an answer on a questionnaire.

(3) Missing not at random (NMAR) or non-ignorable missing: NMAR refers to data that is missing for a specific reason. An example of this is if a question on a questionnaire has been skipped deliberately by the participant (e.g., participants with higher incomes tend to skip income related questions). This data must be replaced or the entire case deleted before running any analyses.

Several approaches are used to handle such missing values: (1) the complete observed vectors method (listwise or pairwise); (2) Serial Mean; (3) Linear Interpolation (LINT); (4) Linear Trend; (5) EM; and (6) regression method.

(1) Listwise: The listwise method uses only cases that contain complete responses to all observations and removes cases for which some or all data are missing. The complete cases are assumed to be representative of the original sample of cases (Pigott, 1994; Furlow, Foulodi, Gagne, & Whittakar, 2007). However, a disadvantage of this method is the remaining sample might be biased when large amounts of data are missing.

Table 4
EFA results for the six missing imputation methods under 10%, 20%, 30% and 50% missingness.(For interpretation of the references to color in this Table, the reader is referred to the web version of this article.)

Method/missing%	10%	20%	30%	50%
List	98.6200	98.7190	88.3430	N/A
Mean	91.6520	83.1780	76.8590	62.7520
Lint	88.4240	77.1950	68.8610	54.5030
Trend	91.6510	83.1460	76.8390	62.6660
EM	98.8310	99.0240	99.1600	99.1430
Regression	98.5040	96.4870	96.312	81.8250