This is the published version:


Available from Deakin Research Online:

http://hdl.handle.net/10536/DRO/DU:30018040

Reproduced with the kind permissions of the copyright owner.

Copyright: 2008, The Authors
The purpose of this paper is to present an empirical analysis of complex sample data with regard to the biasing effect of nonindependence of observations on standard error parameter estimates. In a two-factor confirmatory factor analysis model, using real data, we show how the bias in standard errors can be derived when the nonindependence is ignored. We demonstrate that the standard error bias produced by the nonindependence of observations can be considerable and we briefly discuss solutions to overcome the problem.

Keywords: Complex Sample Survey, Repeated Measures, Multilevel Structural Equation Modelling, Standard Error.

Track: Marketing Research and Research Methodology
1. Introduction

There has been an increasing interest in recent years in analysing models fitting data collected from longitudinal surveys or more generally surveys involving repeated measures that use complex sample designs (CSDs). A complex sampling design typically refers to data acquired by stratification (often on the basis of geography), cluster sampling and/or sampling with an unequal probability selection (Kish, 1965; Lee & Forthofer, 2005; Skinner, Holt, & Smith, 1989). Complex survey data are also referred to as multilevel or hierarchical data. Such designs, however, render the statistical analysis more complicated since the observations are not independent and identically distributed (i.i.d.). The conventional methods for estimating standard errors (SEs) included in most statistical software packages rely on a simple random samples (SRS) design (Lee & Forthofer, 2005). Whilst independence of sample constituents is normally assumed, it is hardly ever realised in the procedures of practical survey work (Kish & Frankel, 1974). The assumption of independent observations, while not entirely essential for the estimation of parameters (e.g., path coefficients), is central for the SEs of those parameters (Frenkel & Frenkel, 1977; Lee & Forthofer, 2005; Skinner et al., 1989). Since conventional estimation of SEs assumes that the correlation of the errors across individuals is zero, a researcher using clustered data may underestimate the SE. Failure to use the appropriate statistical analysis leads, therefore, to an increased probability of committing Type I errors (erroneously rejecting the null hypothesis), underestimation of SEs and misleadingly positive test results (Kish & Frankel, 1974; Scariano & Davenport, 1987; Stapleton, 2006).

In a Monte Carlo simulation study, Julian (2001) reports that covariance models that do not account for the multilevel structure of the data present estimation problems in the \( \chi^2 \) statistic, parameter estimates and SEs (when the intraclass correlations are >.05). The author states that, in such a case, the effect of disregarding the data dependence can no longer be overlooked. Furthermore, when the group/member ratio decreases, the consequence of not accounting for the multilevel data structure will become more severe on the quality of estimation (Julian, 2001). Other researchers such as Shackman (2001) and Maas & Hox (2005) reveal however that in multilevel modelling the concern is not so much the intraclass correlation \( \rho \) (or Rho), as it is the design effect (deff) which Shackman classifies as a correction that should be used to determine sample size. The Monte Carlo method is “an empirical method for evaluating statistics” (Paxton, Curran, Bollen, Kirby, & Chen, 2001, p. 289), and facilitates the investigator to examine the performance of a given test statistic across a number of random samples (Rubenstein, 1981). However, Monte Carlo studies have also been the object of various criticisms since they may lack of strong theory guiding the design and analysis of the simulation and because of their questionable external validity and generalisability. Without strong theory, simulation studies are frequently considered “to be akin to randomly looking for a needle in a haystack” (Paxton et al., 2001, p. 290).

In this paper, an analysis of real data is used to explore whether the theory proposed by Julian (2001), Shackman (2005) and Maas & Hox (1998) holds in an applied field such as marketing. As far as we know, there is no empirical research in marketing that has investigated the extent of the problems resulting from the application of nonhierarchical covariance structure modelling techniques which assume i.i.d. to complex sample data. We argue that it is an empirical question whether or not interdependence between multilevel measures needs to be accommodated in models. We use cross-sectional data in the form of repeated measurements of the same respondents across different brand ratings on a set of items. The items reflect two different dimensions (i.e., brand investments and brand quality).
that are part of the brand signalling framework proposed by Erdem & Swait (1998). Since our research question is not hierarchical or multilevel in nature, we employ aggregated analysis of complex survey data as suggested by Stapleton (2006), benefiting from the analysis of complex designs included in LISREL 8 (Jöreskog & Sörbom, 2006). In doing so, our research differentiates from the work offered by Julian (2001) for two reasons since: (a) whilst Julian (2001) used synthetic computer generated data, we offer an analysis based on real field data and (b) Julian (2001) used a disaggregated analysis (with a Multi Trait Multi Method [MTMM] model) whereas we propose an aggregated analysis of complex survey data. The consequences of disregarding the dependencies among observations intrinsic to the multilevel data for their estimated SEs were appraised in a two-factor confirmatory factor analysis (CFA) model.

2. Complex designs and clustering
Repeated measures occur in many areas of research. It is not uncommon either in commercial or academic research to have repeated measures as a result of “stacking up” observations from different survey cycles. In the marketing discipline and particularly in the brand equity literature (e.g., Dillon, Madden, Kirmani, & Soumen, 2001; Erdem & Swait, 1998, 2004; Erdem, Swait, & Valenzuela, 2006) respondents are frequently asked to rate a set of items across different brand names. In cross-sectional data, repeated measures can be modelled as a two-level structure where measurement occasions are level one units and individual subjects are level two units. Thus, these repeated observations can be regarded as additional clusters within the larger design. However, for a set sample size, cluster designs are subject to larger standard errors. Because the calculation of the SEs entails the sample size in the denominator of the equation:

\[ SE(\bar{x}) = \frac{\sigma_x}{\sqrt{n}} \]  

when the observations are not independent, the effective sample size is not the mere number of cases in the dataset and an adjustment needs to be made in order to avoid the SEs being biased downwards. The difference in the accuracy of the estimates produced by a CSD relative to a SRS is known as the deff (Skinner et al., 1989). In samples characterised by clusters, the deff is the ratio of the actual variance, under the sampling method used, to the variance calculated under the assumption of SRS (Muthén & Satorra, 1995) and can be expressed as follows:

\[ deff = 1 + \rho(m-1) \]  

where deff is the design effect, \( \rho \) is the intraclass correlation for the variable under consideration and \( m \) is the size of the cluster. \( \rho \) exhibits the degree of correspondence within each group and Shrout & Fleiss (1979, p. 423) expressed it as follows:

\[ \rho = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_w^2} \]  

whereby \( \sigma_g^2 \) is the variability between cluster groups (or primary sample units, PSU) and \( \sigma_w^2 \) is the pulled variance within cluster groups. Hence \( \rho \) indicates the fraction of total variability \( \sigma_g^2 + \sigma_w^2 \) that can be attributed to the variability within cluster group \( \sigma_w^2 \). When data are
independent, ρ coefficients ought to be equal to zero. The scale of ρ hinges on the relation between the characteristics of the groups (i.e. structure, size and function) and the variable measured (Julian, 2001).

3. Application to brand signalling data

We follow previous work (Erdem & Swait, 1998) in signalling theory which develops an information economics perspective on the value (or equity) attributed to brands by consumers. Signalling theory explains how high quality firms can differentiate themselves from the lower quality ones. Based on this theory it has been argued that brand signalling inferences emerge from dissipative signals (Rao, Qu, & Reuekert, 1999), which derive from an ex ante expenditure comprising investment in building a reputation that could be lost, should the promised product quality not correspond to the actual quality delivered (Erdem & Swait, 1998). A brand incorporates and represents a firm’s past and present marketing mix strategy, activities and brand investments (Erdem & Swait, 1998). Firms spend resources on their brands to assure that promises are maintained. Firms make brand investments to exhibit commitment to their brands (Klein & Leffler, 1981). Brand investments contribute to credibility by signalling. When a branded product fails to fulfil the promise expressed in the brand signal, the brand compromises the expected returns on these brand investments as well as its reputation for delivering on its promises. The literature suggests that investments in brand strategies and activities (i.e. brand logo, sponsorship, or a powerful advertising campaign) are sunk costs that cannot be recuperated (Ippolito, 1990). If brands damage their credibility they cannot command the premium associated with their reputation and brand investment (Erdem & Swait, 1998).

To render the examination relatively simple and easy to interpret, we analysed, using CFA, the psychometric properties and association of only two constructs: brand investments and brand quality. Following signalling theory, this relationship indicates that the quality of a brand name relates to the investments made in such a brand. Both constructs are reflective latent variables in the brand signalling framework proposed by Erdem & Swait (1998) whereby each indicator was allowed to load exclusively on its respective latent variable. We used brand ratings obtained from 239 members of an online panel in Australia to assess the various dimensions of the signalling framework in the jeans (121 surveys) and digital cameras (118 surveys) product categories. Each respondent rated 8 brands; we therefore had a total of 1912 individual observations at the brand level (968 for jeans and 944 for digital cameras). We used scales similar to the ones used by Erdem & Swait (1998, 2004) and Erdem et al. (2006). Also, consistent with Erdem & Swait (1998, 2004) and Erdem et al. (2006) all items were measured on 9 point Likert scales. Following the same methodology used in Erdem & Swait (1998) and Erdem et al. (2006), data were stacked in one single matrix in order to produce a pooled model across different brands and the two product categories. The data were balanced, that is, there were the same number of brand observations per respondent. The data lent themselves to be modelled via complex sample analysis since they were structured in the form of repeated measurements of the same respondents across different brand ratings on a set of items.

4. Data analysis

We firstly assessed the two factor CFA pooled model making using of LISERL 8.8 (Jöreskog & Sörbom, 2006), employing design base adjustment of the likelihood ratio test (LRT) statistic which automatically rescales χ² value and estimates “robust” standard error (for a detailed estimation procedure see Scientific Software International Inc., 2005). All λ and Φ were left free to vary/covary, whilst the variance of each construct was set to 1.
The model converged very well resulting in $\chi^2=4.36$, df=4, P-value=0.35945, RMSEA=0.007 with $\Phi_{\text{inveq,quan}}=0.88$ (p < .001). A good fitting model allowed us to move to the next phase of the research study. In order to observe the effects of $\rho$ and $\text{deff}$ on the $SEs$ estimation, this study considered three respondent/repeated measures configurations for the multilevel data: (a) 239 respondents with 8 brand observations each, (b) 239 respondents with 5 brand observations each, and (c) 239 respondents with 3 brand observations each. The number and nature of brands chosen for this experiment was selected randomly from the available dataset and it was consistent across all respondents (i.e., all the brand names were constant across individuals). The number of brand observations per individual was manipulated with the expectation of having an incremental/decremental effect on $\rho$. On the other hand, as stated previously, we know that such a manipulation is, ceteris paribus, likely to influence $\text{deff}$ estimates since the number of brand observations per individual is a function of $\text{deff}$ itself (see equation 2). However, the above manipulation resulted in unequal sample sizes in the three configurations as in (a) $n=1,912$, in (b) $n=1,195$ and in (c) $n=717$.

Notably, it could be argued that the three models are not comparable because the variability from one data subset to another dictates that the parameter estimates will vary according to what data are observed as well as the sample sizes. Yield estimates from a particular model depend on the parameter estimates, and so the yield estimates themselves are variable. However, because we are using real data, we cannot specify a priori parameter values for these models to produce the desired $\rho$ (as one would normally do in a Monte Carlo simulation). As an alternative solution, we present the analysis of 6 models whereby, following the three configurations proposed above, we compare the $SEs$ of three models in (a), (b) and (c) obtained from normal asymptotic theory (maximum likelihood estimation) with their respective models calculated with the “sandwich” estimator (pseudo maximum likelihood estimation as portrayed in Pfeffermann, Skinner, Holmes, Goldstein, & Rasbash, 1998) to produce robust estimates using LISREL 8 (Jöreskog & Sörbom, 2006). In summary, we compare the $SEs$ of:

- model (a) with 239 respondents with 8 brand observations each, obtained from normal asymptotic theory with model with (a’) having 239 respondents with 8 brand observations each, calculated with the sandwich estimator;
- model (b) having 239 respondents with 5 brand observations each obtained from normal asymptotic theory with model (b’) having 239 respondents with 5 brand observations each, calculated with the sandwich estimator;
- model (c) having 239 respondents with 3 brand observations each obtained from normal asymptotic theory with model (c’) having 239 respondents with 3 brand observations each, calculated with the sandwich estimator.

Since the $SEs$ of the models (’) are “robust”, we take the $SE$ estimates of these as baselines to be compared with the estimates (and consequent misspecifications) of the models calculated with normal asymptotic theory. The percentage $\Delta$ in $SEs$ between the two models represents the $SE$ misspecifications given by the SRS design.

5. Findings

Table 1 shows $\rho$, $\text{deff}$ and $SEs$ of the six models for the $\lambda$ as well as the $\Phi_{\text{inveq,quan}}$ parameter estimates. $\rho_s$, $\text{deff}_s$ and $SE_s$ refer to both models (a and a’) with 8 repeated measures, $\rho_s$, $\text{deff}_s$ and $SE_s$ to models (b and b’) with 5 repeated measures and $\rho_s$, $\text{deff}_s$ and $SE_s$ to models (c and c’).
with 3 repeated measures per respondent. Bias % measures the bias of the asymptotic (ML) estimation vs. the sandwich estimation and is calculated as follows: \[
\frac{SE\text{asymptotic (ML)} - SE\text{sandwich est.}}{SE\text{asymptotic (ML)}}
\]

### Table 1

<table>
<thead>
<tr>
<th>( \lambda )</th>
<th>( \rho_5 )</th>
<th>( \rho_8 )</th>
<th>( \phi )</th>
</tr>
</thead>
<tbody>
<tr>
<td>qual 02</td>
<td>0.22</td>
<td>0.25</td>
<td>0.016</td>
</tr>
<tr>
<td>qual 03</td>
<td>0.24</td>
<td>0.27</td>
<td>0.027</td>
</tr>
<tr>
<td>qual 04</td>
<td>0.29</td>
<td>0.35</td>
<td>0.041</td>
</tr>
<tr>
<td>Invest 01</td>
<td>0.20</td>
<td>0.27</td>
<td>0.049</td>
</tr>
<tr>
<td>Invest 02</td>
<td>0.26</td>
<td>0.31</td>
<td>0.050</td>
</tr>
</tbody>
</table>

The findings suggest a relationship between the \( \text{deff} \) and the \( SEs \) bias. If, for instance, we observe the \( \text{deff} \) coefficients in all the \( \lambda \) estimations we notice that all the parameter estimates \( \text{deff}_8 > \text{deff}_5 > \text{deff}_4 \) and respectively \( \text{bias}_8 > \text{bias}_5 > \text{bias}_3 \) suggesting the bigger the \( \text{deff} \) the bigger the bias. It is also interesting to note that relative to \( \Phi \text{ invested qual} \) \( \text{bias}_8 > \text{bias}_5 > \text{bias}_3 \). However, it is somewhat surprising to see that not much difference in \( \rho \) relative to the number of observations per individual. In fact, relative to the \( \lambda \) parameters, \( \rho_3 \sim \rho_5 \sim \rho_8 \).

### 6. Conclusion

We conclude this study observing that taxonomic outcomes based on relationships between \( \rho \), \( \text{deff} \) and the \( SEs \) misspecifications cannot be rigorously drawn. Admittedly the biggest limitation of this research is that the three case studies examined therein must be interpreted independently since, in order to manipulate \( \rho \) and \( \text{deff} \), three different data subsets were used. However, considering the methodological constraints in employing real data and the impossibility to link the three independent analyses, we still provided some “case study” evidence that might suggest the existence of such a relationship. Using real data, our study confirms the findings of Julian (2001), Shackman (2005) and Maas & Hox (1998), that is: (a) when \( \rho > .05 \) the multilevel structure of the data cannot be disregarded in estimating a proper \( SE \), and (b) as previously suggested there would seem to be a relationship between the \( \text{deff} \) and \( SE \) estimates misspecifications.
References


