CASE RESIDUALS IN STRUCTURAL EQUATION MODELING

by

John Cardinale

A dissertation submitted to the Graduate Faculty in Educational Psychology
in partial fulfillment of the requirements for the degree of Doctor of Philosophy,
The City University of New York.

2011

This manuscript has been read and accepted for the
Graduate Faculty in Educational Psychology in satisfaction of the
dissertation requirements for the degree of Doctor of Philosophy.

Jay Verkuilen, Ph.D.

Date	Chair of Examining Committee

Mario Kelly, Ed.D.

Date	Executive Officer

David Rindskopf, Ph.D.

Charles Scherbaum, Ph.D.

Louis Primavera, Ph.D

Keith Markus, Ph.D.

Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK

Abstract

CASE RESIDUALS IN STRUCTURAL EQUATION MODELING

by

John Cardinale

Advisor: Professor Jay Verkuilen

From the beginning, lead methodologists in psychometrics and quantitative psychology have been well aware of the problems of fitting structural and confirmatory factor models. The question we approach in our research is how to best detect this misfit and how to identify specific sources of misfit by scrutinizing the data at the case level. Since Anscombe's seminal 1973 paper, detecting problems at the case level in ordinary least-squares regression has become the norm in statistical modeling. In contrast, the usual practice in fitting structural and confirmatory factor models has been to only examine misfit at the variable and sufficient statistic level. This practice ignores a small body of literature that has arisen since the early 1990s about diagnostics of case level and case by variable level misfit. An important paper by Bollen and Arminger (1991) and a follow-up paper by Raykov and Penev (1999), have developed theory behind Individual Case Residuals (ICRs). These papers help lay the ground work for more detailed case and case by variable level diagnostics, without discarding traditional variable oriented procedures. Our goal is to demonstrate uses of multivariate techniques,

such as robust Mahalanobis distances, biplots and cluster analysis to analyze the multivariate dataset of ICRs and thereby detect sources of data problems with respect to a target model. We hope to encourage researchers to make better use of case level diagnostics among the various classes of latent variable models, especially with the advent of multivariate tools in packages such as R and SAS.

Acknowledgements

I would like to thank all of my family and friends for their undying support....

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction to The Problem of Mis-Approximation

Much of the research in quantitative psychology and psychometrics, which dates back over a century, has attempted to understand how well statistical models approximate relationships among observations and constructs; or as with structural equation models (SEMs), the relationship between constructs. These models have been understood to fail in their approximations to varying degrees–a fact which is discussed theoretically in the work of MacCallum and Tucker (1973). Here I attempt to understand this mis-approximation not only through the analysis of indices and sufficient statistics, but also through the examination of the individual-level data, which is often the source of the problem. I accomplish this analysis through semi-exploratory multivariate methods applied to case residuals. I am particularly interested in using robust methods to identify multivariate outliers and anomalous response patterns.

I argue that the researcher can understand mis-approximation by quantifying it as is done with residuals in regression analysis. But how can this quantification be done in the algebraically complex world of CFA models and, especially, in the larger universe of SEMs? If the typically used hypothesis testing methods fall short, then what else is there? One option is the discrepancy matrix. As far back as 1981, Jöreskog suggested that researchers examining the difference between fitted covariances and observed covariances when χ^2 change values are smaller than the degrees of freedom of the model. This research intends to discuss and demonstrate in detail the approach of examining case residuals, in a similar fashion, as a diagnostic for proper SEM and confirmatory factor model specification. As an

introduction to the common practice, I will survey traditional diagnostic procedures beyond hypothesis tests. I do not disparage the use of any of these procedures. Rather, I propose that creating individual case residuals (ICRs) could be used as an adjunct to the current practice. The following is the formal definition of the ICR data matrix, where $\mathbf{P}_f$ is the factor score multiplied by the factor loading matrix, and $\mathbf{X}$ is the matrix of observed indicators. Note well that this metric is only useful for exogenous variables in the SEM framework.

$$\hat{\mathbf{\Delta}} = (\mathbf{I} - \mathbf{P}_f)\mathbf{X} \tag{1}$$

ICRs detect anomalies with respect to the model at hand, and therefore they at least partially answer the aforementioned question: How do we detect SEM mis-approximation. I will discuss the art of SEM approximation throughout this document and will start by emphasizing the prominent issues at hand in the field.

Indeed, it is no surprise that top researchers in the field of model building in SEM and other models with latent variables agree that the process is both an art as well as a science (McDonald, 2010). Perhaps the real issue for researchers is that, in actuality, it is often not apparent what exactly is wrong with the data. Therefore jumping straight to a model-based solution begs the question: What is the appropriate model? Looking at model characteristics in terms of the data matrix is perhaps a sensible approach. The data matrix has rows (cases), columns (variables or items) and cells (which are the intersection of rows and columns). Therefore, data can be viewed in terms of variables, cases, and the interaction of variables and cases (variable by case). As we see later, the ICR is a metric that addresses the

Diagnostic Approaches	
Type	**Model**
Variable (V)	SEM
Case (C)	Regression
V×C	Molenaar (2004)

mis-approximation within each cell of the data matrix. The above scheme summarizes the approach I will discuss here, and does indeed generalize to broader contexts, such as item response theory. These broader contexts, however, will not be covered in this study as they are extensively discussed by researchers, such as Pek and MacCallum (2011). Covered as a side-bar is the notion of the mis-approximation of cases, aptly called person-fit in the item-response theory literature. Millsap (2007) couched the problem of person-fit as the failure to satisfy the property of measurement invariance. Indeed, measurement invariance, unlike factorial invariance, has to do with the individuals being measured, who are highly unlikely with respect to the model at hand. Here, the properties of individuals should have little to do with the characteristics of the persons being measured. In contrast, factorial invariance, which has been studied extensively, is a property that has to do with the invariance of the latent variables. SEM, in contrast, is highly variable oriented as it tests the relationship of manifest to latent variables. In most other areas of statistics, regression has become, as we will see, case oriented because of residual diagnostics. Residuals were closely studied by researchers such as Cook and Weisberg (1982). The psychometric literature, by contrast, focused on very different issues than the literature surrounding regression diagnostics. However, psychometrics refocused, when Molenaar (2004) called for psychologists to carefully examine individuals, while still not ignoring the variable orientation.

Ultimately, to examine mis-approximation in the SEM world, on the level of detail that researchers such as Molenaar call for, requires some discussion of procedure. In positing a procedure, I bear in mind that process of model building and diagnosis, cannot completely be broken into a series of predictable steps. Indeed, this process is marked by branching off onto diagnostic forays, which may lead the researcher back to the first step he took, only to alter parameters slightly. Therefore, I hesitate to posit an "exact order of operations" for fitting factor analytic and structural models, let alone diagnosing them for misfit. I can, with the help of a few demonstrations, posit a check-list, which can, with residuals, and data-level analysis, produce meaningful results. Along the lines of McDonald and Ho's (2002) recommendations, I warn against allowing the process to be driven purely by global fit statistics, but encourage the use of local information. And central to my research is the specific focus on Bollen and Arminger's (1991) observational residuals (or ICRs), which were studied formally by Raykov and Penev (1999).

Below I outline a checklist for SEM approximation and diagnosis, which will be expounded upon through the rest of this document:

- Clean the raw data using Rousseuw's robust algorithms to generate a robust data set.
- Run exploratory factor analysis (EFA) on cleaned data to detect factor structure.
- Fit confirmatory factor analysis (CFA) models, with unweighted-least squares (ULS) or some other robust fitiing method: One to model that represents the

desired structure, and one to the model that "fits" very well, but is not substantively correct.

- Create ICRs and latent individual residuals (LIRs).
- Create Biplot and Dendrogram to study structure of both residuals and latent residuals.
- Run robust algorithms on ICRs to locate cases that are outliers with respect to the matrix of ICRs (which is the discrepancy matrix).
- Examine latent structure by computing and visualizing latent individual residuals (LIRs).
- Attempt to adjust target model by adding terms (covariates, intercepts).
- Fix or free variances and covariances based on modification indices and the discrepancy matrix.

Chapter 2: The Theoretical Underpinnings of SEM

To motivate our discussion of SEM model building as it is often conducted, I now describe a popular SEM. This model, the Latent Structural Relations (LISREL) model, is by no means the only model of this type, but is perhaps the most widely used. This model combines both CFA and path analysis. Both covariance and mean structure can be modeled in the LISREL framework. I will therefore explain all of the parameters and associated statistics in LISREL notation as much as possible. However, I also extend the methodology more broadly to latent profile analysis models, which use continuous indicator variables, but a discrete latent space.

To define the LISREL model, I use the following notion (Jörsekog, et al., 2008). Let:

- $\mathbf{y}$ $p\times1$ vector of observed response, or outcome variables.
- $\mathbf{x}$ $q\times1$ vector of observed predictor, covariates, or input variables.
- $\boldsymbol{\eta}$ $m\times1$ random vector of latent dependent, or exogenous, variables.
- $\boldsymbol{\xi}$ $n\times1$ random vector of latent independent, or exogenous, variables.
- $\boldsymbol{\epsilon}$ $p\times1$ vector of measurement errors in y.
- $\boldsymbol{\delta}$ $q\times1$ vector of measurement error in x.
- $\boldsymbol{\Lambda}_x$ is a $p\times m$ matrix of coefficients of the regression of $\mathbf{x}$ on $\boldsymbol{\xi}$.
- $\mathbf{T}$ an $m\times n$ matrix of coefficients of the $\boldsymbol{\xi}$ variables in the structural relationship.

- **B** an $m \times n$ matrix of coefficient of the $\boldsymbol{\eta}$ variables in the structural relationship. has zeros in the diagonal, and I- is required to be non-singular.

The LISREL model takes the following form:

$$\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\xi} + \boldsymbol{\zeta} \tag{2}$$

$$\mathbf{X} = \boldsymbol{\xi}\boldsymbol{\Lambda}_x + \delta \tag{3}$$

$$\mathbf{y} = \eta\boldsymbol{\Lambda}_y + \epsilon \tag{4}$$

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta} \tag{5}$$

Equation (1) is the structural model comprising latent variables, η, and error, ϵ, as well as coefficients and **B**. Equation (2) is the endogeneous measurement model denoted by observed indicators, **Y**. Lastly, equation (3) is the input, or exogenous variables, denoted by observed indicators **X**.

Specifying the LISREL model requires many decisions. Model parameters must be fixed, freed or constrained to establish a reasonable fit (fixing involves assigning a specific value to a parameter; constraining involve setting a parameter, or group of parameters, equal to other parameters; and freeing involves unconstraining a parameter). Any combination of these techniques are used and decisions are often guided by observation of standard errors (SEs), the change in the χ^2 statistic, and, especially, the change in modification indices (MIs). All of these practices have severe short-comings. First of all, MacCallum's ideas (2004) must be

taken to heart in the SEM fitting process: There is no true and perfectly fitting model. Furthermore, it is possible to fit two or more equivalent models, which may or may not offer meaningful insight despite the fact that the fit indices are favorable.

Known Problems With Social Science Data

The misspecification problems I examine in this research are, in part, traced back to issues involving the sample or realization of the data. Misspecification, or model error as it it called by MacCallum (2001), often arises from systematic error of measurement that is not accounted for by the variables in the model. This error may be of the following types:

- Error amongst the manifest or latent variables, such as non-Gaussianity or Heywood cases (negative variances).
- Omitted variables.
- Case misfit due to aberrant response.
- Variable by Case misfit; for example, NURB, which I will explain in the next section.

Systematic error among persons is often seen in aberrant response patterns. This problem is often referred to as a correlational error because there is non-zero correlation between items on the measurement scale (Viswanathan, 2005). An aberrant response from an six item scale with scores ranging from (1,6) may look like this: (1,6,1,6,1,6). There may be a small number of aberrant response cases or there may be a subset of aberrant responders. Non-English speakers who

misinterpret the questionnaire or instructions from the test administrator may comprise a groups of aberrant responders. Also test administration problems may result in such aberrant patterns if a group of respondents have an administrator who does not give proper instructions for completing the inventory.

A model will suffer from specification error if such aberrant responses are not accounted for. The variability of this response directly affects the unique variances and will ultimately inflate the absolute value of the individual case residuals (ICRs). This is because, as I will demonstrate, the ICRs are differences between each observation and the product of the factor loading and factor score. This product, which can further be decomposed into matrix products of covariance matrix, is bound to be inflated if uniqueness are inflated. Therefore, a researcher can detect the problem cases by examining ICRs, but how do we fix the problem? It is true that a mixture model, may be more appropriate to model such a pattern. A model such as the random intercept model proposed by Maydeu-Olivares and Coffman (2006) can adjust for the systematic variation created by an aberrant response set. It, therefore, may not be necessary to use a mixture model, which is difficult to specify under some conditions. A simple alternative might involve case removal or deletion of clear outliers in a given variable leaving these cases as missing, presuming that these cases can be found.

A second data problem, also representing correlational case by variable (C×V) measurement error, is the single outlier. This error is both person and variable oriented because it is really the result of a small subset of extreme item responses relative to the other responses given. For example, a response to a

unidimensional six item scale, from -3 to 3, may look like: (-2,-1,-2,3,-1,-1). This response directly affects the uniquenesses in the CFA or measurement component of an SEM. Outlying responses such as this will be at least as easy to identify the aberrant response pattern. A situation where an outlier could occur would be an anxiety scale that poses six anxiety provoking situations. The scale may range from least anxious to most anxious. The respondent may favor the middle or lower points on the scale until an item (say item 4 out of 6) regarding public speaking, is presented. If the respondent is calm in most situations except public speaking, the response pattern would resemble the one presented above.

Uniform response bias (URB) leads to variable-oriented measurement error. URB is a type of response set bias that was studied extensively by researchers such as Hui and Triandis (1997), Chan (2001), Bentler and Chan (2003) and Chan and Cheung (2002). URB involves a consistent additive, multiplicative, or ordinal effect constant across a set of individuals (Chan, 2003; Viswanathan, 2005). Uniform bias response style may be extreme, acquiescent or consistently middle of the road. For a more detailed discussion of these three types of response styles and the use of multinomial logistic regression to model them, see Bolt and Johnson (2010).

The more complex case is NURB, where bias exists across cases as well as variables (V×C). A hypothetical example of a data set, which may contain NURB is a survey of American-Iranians (Persians) who are asked their level of agreement on issues of U.S.-Iranian relations. A conservative subpopulation of respondents who came to the U.S. during the 1970's is sampled. This population may have very strong views against the current Iranian government and may uniformly respond

unfavorably to items that support a lenient U.S. policy toward Iran. This bias would create a location shift for this response set, which would have to be corrected for in a factor model, possibly by adding a nuisance factor as described in Maydeu-Olivares' model. Later I will demonstrate a possible solution to this bias using covariates, which creates a multiple-cause multiple indicator (MIMIC) model.

NURB creates a case by variable type of systematic error. Another example of this problem may be seen in a survey of a conservative U.S. community, with a contingent of tea-party members. A scale may feature items both on foreign and domestic policy; the domestic policy including items regarding government spending. The tea-party member may respond particularly unfavorably to items endorse government spending–more so than other conservatives. Therefore, the tea party subgroup would comprise a response set with a high degree of variance across items. NURB, is perhaps the most difficult problem to detect that is discussed in this research. I propose a multi-step exploratory approach if a research has some prior notion of the existence of NURB. Without such a prior notion, however, this problem poses a serious threat to model specification.

Multigroup data structure is an extension of URB where the entire sample is divided in to two or more groups separated by some variable not related to the construct being measured. An example may be seen in Scherbaum's General Self-Efficacy Survey (2006), which is studied in Exhibit 3. The size and direction of ICRs can expose the location shift among subgroups, which is the hallmark of multigroup data.

The Problems of Model Building

An appropriate sample size is necessary for some fit statistics in SEM and confirmatory factor models to be useful. An example of such a statistic is the χ^2 statistic, which is an example of how fit indicies can be misleading to the untrained researcher. A small χ^2 will ideally point to a well fitting model. However, the χ^2 test requires reasonable sample sizes to correctly reject a poor fitting model. On the other hand, if the researcher has a very large sample, the test will almost always be rejected. This is also an omnibus test, which says nothing about what is wrong. There are alternative statistics to the χ^2 in the SEM literature. For example, Kaplan (1988) examines the use of the z statistics (or t statistic in LISREL) to evaluate the statistical significance of model parameters. He finds that in maximum likelihood estimation, z is statistically unbiased only when the sample size is optimal. Smaller sample sizes lead to an upward bias in the z statistic.

After fitting a model, the researcher may want to improve fit by using MIs–again, a variable oriented approach. This approach is sensible, but, as with using χ^2 or z, one must proceed with caution. The shortcomings of MIs is largely due to the fact that they are neither independent nor additive. Without these properties, a solution is impossible to find from relying solely on these estimated values. Furthermore, for MIs to be useful at all, the researcher must free them one at a time in between new model fits. It is not uncommon for MIs to fail to help researchers detect incorrect factor structure if similar items are parceled or combined (Bandalos, 2002). These and other shortcomings of diagnosing and fitting SEMs point to the fact that a general and inclusive procedure must be adopted by

researchers. This current research argues that such a procedure can be more effective if diagnostics are also performed at the data level, specifically at the level of individual cases residuals.

One useful strategy mentioned earlier is to examine covariance residuals. The residual covariance matrix, or discrepancy matrix as discussed by McDonald and Ho (2002), is defined as follows:

$$\mathbf{\Sigma}_{\delta} = \mathbf{S} - \hat{\mathbf{\Sigma}} \tag{6}$$

$\hat{\mathbf{\Sigma}}$ is the predicted covariance matrix, which is output by all SEM packages. The descrepancy matrix, which contains zeros along its diagonal, is already part of the common practice of diagnosing and fitting SEMs and CFA.

I propose and additional diagnostic procedure. As mentioned previously, case residual analysis is central to this proposed methodology. ICRs have there own covariance structure, which may be useful in certain circumstances, especially when examining dispersion and independence of error. The following derives the matrix form for the ICR covariance matrix:

$$\begin{aligned}
&\text{cov}(\mathbf{\Delta}_{pq}) = \text{cov}([\mathbf{I} - \xi\mathbf{\Lambda}]\mathbf{X}_{pq}) \\
&= \text{E}([\mathbf{I} - \xi\mathbf{\Lambda}]\mathbf{X}_{pq}[\mathbf{I}' - \xi'\mathbf{\Lambda}']\mathbf{X}'_{pq}) \\
&= \text{E}([\mathbf{X}_{pq} - \xi\mathbf{\Lambda}][\mathbf{X}'_{pq} - \xi'\mathbf{\Lambda}']) \\
&= \text{E}[\mathbf{X}_{pq}\mathbf{X}_{pq}{}' - \mathbf{X}_{pq}\xi'\mathbf{\Lambda}' - \xi\mathbf{\Lambda}\mathbf{X}_{pq} + \xi\mathbf{\Lambda}\xi'\mathbf{\Lambda}'] \\
&= \text{E}[\mathbf{X}_{pq}\mathbf{X}_{pq}{}'] - \text{E}[\xi\mathbf{\Lambda}\xi'\mathbf{\Lambda}'] \\
&= \mathbf{S}\text{–}\Lambda\mathbf{\Phi}\mathbf{\Lambda}'
\end{aligned}$$

In many instances, researchers mention ICRs, but do not develop a scheme of how to use them. For example, McDonald and Bolt (1998) point out that case

residual analysis can be useful for the same reasons it is useful in ordinary least square (OLS) regression. The residuals are functions of the data and therefore help shed light on both model problems and data problems; more specifically *bad data points with respect to the model.* In the next chapter, I will further develop the residualization procedure proposed by Bollen and Arminger in their 1991 paper. First I will discuss ordinary least square (OLS) regression diagnostics, which developed parallel to the development of the SEM.

The Asymmetry Between Usual Model Fitting Practice in SEM and OLS Regression

The above modeling procedures for SEM are very different than the case oriented OLS regression diagnostic procedures. Going back to a seminal paper by Francis Anscombe (1973) researchers began to examine raw data and residuals on the case level. I will go into further detail about Anscombe's idea in Exhibit 1 in Chapter 3. Essentially, Anscombe emphasizes that examining raw data and residuals in conjunction with model specification will allow the researcher to detected masked data anomalies. Problems that arise from not heeding Anscombe's recommendations may be much more subtle than even his initial data set (Figure 1) describes. Consider, for example, a study on the per capita education expenditure of the 50 US states. There were three predictors of achievement in this study, which yielded a seemingly strong linear regression where R^2 was 0.7. Alaska, however, had outliers, and influential ones at that. When Alaska was omitted, the R^2 dropped to 0.6. Anscombe (1973), commented that simply removing Alaska is an inadequate solution, as other outlying states also have to be removed. Reporting the original

regression, while explaining that Alaska's data were influential is a more appropriate approach. The advantage of examining residuals is that the underlying truth of relations among data can be rightly exposed.

As depicted in figure 1, individual case diagnostics became a center piece of statistical research from Anscombe's paper through the early 1980s. Cook and Weisberg (1982) discussed at length the process of detecting violations to the assumptions of linear regression using both so-called model-free and model-based methods. Model-free methods consist of visualizations, such as scatterplots and normal probability plots (q-q plots). Model-based methods employ statistical tests for such statistics as Cook's D, which detects extreme and influential observations. Because residuals are asymptotically normal, they can be tested to a degree as a t-statistic, provided that they are modified by dividing them by a measure of error.

Residuals became even more difficult to study, when researchers, starting in the 1980s, introduced diagnostics in multilevel and longitudinal models. In doing so, they introduced more visualization and model-based tests. These tests and visualizing techniques developed parallel to the discoveries of case residuals in SEM and individual case diagnostics. Both of these areas of research have essentially led to the same place. The difficulty of multilevel and longitudinal models presented here is that residuals are correlated and often heteroscedastic. Correlation matrices are therefore specified to not only contribute to tests of residuals, but also to better specify the model in question. Similarly, multivariate data sets feature high correlation between variables for each case. Therefore researchers can obtain individual latent residuals (LIRs) and study them using methods prevalent in

multilevel and longitudinal regression models. Researchers can apply graphical displays and statistical tests, such as the Kolmogorov-Smirnov test for normality (Sanchez et al., 2009). Fitzmaurice, Laird and Ware (2004) discuss using residual transformations, particularly the Cholesky decomposition to produce residuals, which are uncorrelated and have unit variances. These whitened residuals can be graphed to give better insights to model misspecifications such as missing quadratic terms or violations to Gaussian assumptions. Houseman et al. (2004) recommended the use of cumulative distribution functions of transformed residuals, and q-q plots, to assess normality.

The following relationship holds for the Cholesky decomposition, $\mathbf{U}$, which is an upper-triangular matrix:

$$\mathbf{A} = \mathbf{U}'\mathbf{U} \tag{7}$$

To whiten (decorrelate) this vector, the researcher should pre-multiply the vector of residuals $\boldsymbol{\delta}_i$ by $\mathbf{U}$. This residual vector belongs to the linear model built from a data set with covariance matrix $\mathbf{A}$.

Chapter 3: Case Oriented Tools in Behaviorial Statistical Models

Person Orientation in Psychology

I discussed OLS and then multivariate regression to emphasize that I am taking a variable oriented statistical model and diagnosing misfit by more closely looking at individual observations and residuals as one would in models with only observed variables. This approach essentially reflects Molenaar's approach. This approach also intersects an entirely different set of models from clinical psychology that were discussed by Magnusson and Bergman (1997) years before the Molenaar's work. I note this alternative area of research because it, perhaps the approach of the social science researcher toward closer scrutiny of individual participants. This approach is used in time series and longitudinal analysis. But in the case of non-time varying models, I believe it is also important to supplement variable level analysis with case-level and response set analysis. I will demonstrate later that this approach is especially useful for diagnostics of model mis-specification due to anomalies such as uniform and non-uniform response biases.

Bergman and Magnusson (1997) were particularly critical of SEM because it is only concerned with relationship among constructs, or in the case of the Spearman model, the strength and validity of a single construct. For the research agenda of these psychologists, SEM falls short in that it ignores patterns among individuals. Their procedures were largely case oriented and stem from what these authors call a holistic interactionist approach. Here, cluster analysis is often applied longitudinally to capture changes in patterns over time. In this way the development of individual participant can be better understood. Bergman and Trost (2006)

suggest that the variable oriented approach does have its place and in many cases, the case-oriented and variable-oriented approach can be complimentary. I argue for the third row table 1, that the complimentary use of variable oriented analysis with case oriented procedures is useful, especially in diagnosing fit and model validity.

Case Level Diagnostics

As mentioned earlier, case-level diagnostics have been prominent in the Item-Response Theory (IRT) literature. For a thorough review of various IRT person-fit metrics, see Karabatsos (2003). In the SEM literature, as far back as 1985, Comrey writes about the close examination of individual observations in factor analysis—something that, up to that point, was largely glossed over by researchers. Comrey discussed the use of Mahalanobis-squared distance (MD^2) and the problem of masking associated with them. Masking occurs when outliers cannot detected because they are located directed opposite each other on the hyperplane that contains the data (Atkinson, 2002). Comrey argues that his D_k provides more information about the relationship of the data to the correlation matrix in factor analysis. His statistic is as follows:

$$D_k = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (r_{ij} - Z_{ik}Z_{jk})^2. \tag{8}$$

Here, r are Pearson correlations and Z are standard scores for each person for each item. Individuals with a large D_k deviate from the common response style, while individuals with a small D_k do not deviate from the norm substantially. Comrey applies D_K, and an additional scaled version of this D_k, T_k, to a simulation study. Unfortunately, this study exposes a lack of correspondence between D_k and T_k, a

problem that brings in to question the usefulness of these distance metrics. He then applies both of these distances to a personality inventory of college students to attempt to detect response "faking" and shows that these statistics are somewhat useful, but are by no means definitive in identifying outliers. In fact, they sometimes misidentify cases as errant. Therefore, the researcher should use them primarily to identify cases that require further scrutiny as opposed to identifying outliers. This research will bear out that examining ICRs is a task that can prove to be even more insightful than closely examining the original data as Comrey does.

Case-level Influence Diagnostics

Influence analysis has been a major part of the literature on residuals in regression (Cook and Weisberg, 1982). In its simplest form, the stability of a model is assessed by examining how is model is perturbed by removal of data one point at a time. Also a method of examining stability called forward searching is performed by examining a subset of data and tracking changes in the model when data is added, which helps consider blocks of outliers. The Cook's D statistic is often used in conjunction with removal of data one point at a time to observe changes in parameter estimates, in this case regression coefficients, β_i. A case with a residual leading to the a large change in a parameter may be labeled as an influential observation. D is a distance measure, which is asympotically χ^2 and represents the distance from the centroid of an ellipsoid created by the regression coefficients. Cook's D is of the following quadratic form (similar to all distance measures including Mahalanobis Distances in the multivariate case):

$$D_i = \frac{(\hat{\beta}_{(-i)} - \hat{\beta})' \mathbf{X}' \mathbf{X} (\hat{\beta}_{(-i)} - \hat{\beta})}{ps^2}$$

Here, p is the number of parameters, s is the standard deviation of the data matrix, and $\hat{\beta}_{(-i)}$ is the least square estimate of the regression coefficient vector β with the ith point removed. Effectively, D_i is the distance between estimates $\hat{\beta}_{(-i)}$ and $\hat{\beta}$.

The equivalent of Cook's work among methodologists of case level analysis in psychological models was advanced in earnest in the 1990s. In this instance, studying influence allows the researcher to understand the degree to which aberrant observations perturb the target model. Lange et al. (1976) were among the first researchers to discuss some ways of analyzing individual observations for influence. They suggested computing individual fit statistics for covariance structure analysis (CSA). An example of such a fit statistic is the individual -2 log likelihood. Reise and Widaman (1999), continuing with the Lange et al. approach, computed individual fit statistics by examining the individual likelihood, or P_{ll}:

$$P_{ll} = -\frac{1}{2}[q \ln(2\pi) + \ln |\hat{\mathbf{\Sigma}}|] + (x_i - \overline{x})\hat{\mathbf{\Sigma}}^{-1}(x_i - \overline{x}) \tag{9}$$

Here, $\hat{\mathbf{\Sigma}}$ is the fitted covariance matrix, $\overline{x}$ is the mean vector, q is the number of variables and x_i is the vector of responses. The second part of the equation is the Mahalanobis square distance (seen earlier), which varies across individuals.

Reise and Widaman then subtracted the P_{ll} statistics for the full model from the P_{ll} statistics for the saturated model to yield the individual contribution to the chi-square statistic. These CHI_{ind} statistics revealed that a small percentage of

their data was contributing to model misfit and needed to be examined more closely. For a more recent and detailed discussion of this and other influence metrics, the reader is referred to Pek and MacCallum (2011).

Using latent growth curve models (LGC), Coffman and Millsap (2006) used the P_{ll} statistics to evaluate the fit of individual growth curves. These authors suggested removing curves that do not fit and examining them carefully. They argued that just because global fit statistics indicate lack of fit does not mean the model is not useful. This is especially true for LGC models, where individual growth curves are often of interest.

Variable × Case Methods and Case Residuals: Bollen and Arminger, and Raykov and Penev

As alluded to earlier, Bollen and Arminger (1991) presented the residualization of CFA models as a method of generating a vector of case residuals, which can then be tested for outliers and non-Gaussianity. This methodology has allowed the researcher to apply both statistical tests and visualizations, which originated from OLS regression diagnostic research. As I will show later, it is possible to represent each individual residual as a t statistic, which could identify the residual as representing an outlier. In this research, I attempt to take this methodology further by detecting misspecification using a different approach than examining the residuals as individual statistics. Misspecification occurs when a proposed model does not accurately represent the data collected. ICRs are useful because they reveal anomalies with respect to the model at hand. In this sense, ICRs, represent an alternative to traditional diagnostics such as the χ^2.

Rather that measuring each ICR in terms of a test statistic, I propose examining the matrix of ICRs as a multivariate data set using exploratory multivariate statistical methods specifically the biplot, cluster analysis, and robust Mahalanobis distances. The biplot, is particularly interesting because it uses as an alternative decomposition to Cholesky, the Singular Value Decomposition. I will discuss the biplot in more detail after developing the mathematics behind ICRs.

Many researchers may approach the idea of examining ICRs with concern regarding the fact that they are indeterminate. Indeed, the latent variables on which they are based are not unique (see below). Regression (or Thomson), Bartlett and Anderson-Rubin are just three methods of estimating factor scores, which are quantifications of the latent variables. Each of these methods yield somewhat different scores, an issue that can best be resolved if I understand the nature and definition of latent variables. I follow Bollen's (2004) notion of latent variables serving as realizations of the underlying structure of the target sample. This inclusive definition reflects the latent nature of the error terms in regression analysis as well as factors in CFA.

Raykov and Penev (1999) formally show that the process of residualizing CFA models is equivalent to the process of residualizing SEMs. This is true for both manifest residuals (ICRs) and latent individual residuals (LIRs). Raykov and Penev's proof, discussed below, opens the door to a broad range of diagnostics and applications with ICRs. Indeed, nonlinearity, violations of the normality assumption, and testing equivalent models are all applications that cannot be developed with sufficient statistics. In our research, I are particularly interested in

the following problems, which can be identified through the multivariate analysis of the ICR dataset:

- Different models with equivalent covariance matrices
- Non-linearity
- Multigroup structure
- Response set biases: both uniform and non-uniform
- Individual outlying cases
- Equivalent models

While many of these problems can be partially detected by turning to the original observations without creating residuals, I argue that studying ICRs will add information above and beyond diagnostics of the original data. Anscombe (1973), in discussing residualizing for regression analysis, argues that examining residuals allows for better examination of the data after removing the dominant trend–"minus the regression"–and therefore residual behavior is easiest to see in a 2×2 scatter plot.

Stout's Contribution

Stout (1987, 2002) was the first to develop the process of creating residuals in item-response theory (IRT) models. IRT models are, of course, factor models with dichotomous indicators. In formulating his theories, he emphasized the importance of removing the dominant trend from the items, which is not unlike

removing the regression when creating OLS residuals. It is no coincidence that Anscombe spoke similarly of this process in his 1973 paper.

Stout's models were variable oriented, and he was therefore faced with strict IRT assumptions, such as local independence, which he had to relax for the purpose showing unidimensionality in confirmatory factor analysis. Stout addressed this issue (1987, 2002) as he discussed essential unidimensionality and relates it to factor analysis. Essential unidimensionality is a weaker version of strict unidimensionality and assumes the following relationship holds:

$$\frac{1}{N(N-1)} \sum_{1 \leq i \neq j \leq N} |Cov(U_i, U_j|\theta)| \approx 0. \tag{10}$$

The above statement of unidimensionality is conditional only on the latent trait. Therefore, local independence does not necessarily have to hold, but the above relationship does. In this way items will load on the dominant traits in the factor analytic sense. Essential unidimensionality is relevant to our research because many attempts at measurement in SEM revolve around proving that constructs are unitary. Indeed, the goal of much of confirmatory factor analysis is to test the unidimensionality of the target data; this translating to a Spearman factor model, or congeneric test model. In other words, researchers want all items in a scale to load on a single factor as in the Spearman model.

Stout argued that this unidimensionality is unrealistic, though often reasonable enough. Often models have to be modified by imposing a multigroup structure or by adding a covariate as in the multiple-cause multiple indicator (MIMIC) model. I argue that in diagnosing violations to essential unidimensionality, it behooves the research to examine residuals, which can point

out anomalies not readily scene by examining fit indicies, sufficient statistics or even discrepancies between the covariance matrices.

A Closer Look at ICRs

Researchers and statisticians see the general linear model (GLM), that is the OLS regression model, as the basis for applied statistical methods. To develop the algebraic principles behind ICRs, I present the multivariate regression model in matrix form:

$$\mathbf{Y}_{ij} = \mathbf{X}_{imm} + \mathbf{E}_{ij} \tag{11}$$

Here, $\mathbf{Y}_{ij}$ is a matrix of observed manifest response variables, $\mathbf{B}_m$ is the matrix of regression coefficients, $\mathbf{E}_{ij}$ is the residual matrix,and $\mathbf{X}_{im}$ is a matrix of predictors. In the exogeneous measurement model of SEM, which is similar to a regression model, $\mathbf{X}_{pq}$ are the manifest factor indicators:

$$\mathbf{X}_{pq} = \xi_{pm}\mathbf{\Lambda}_{mp} + \Delta_{pq} \tag{12}$$

The above regression model can be used as a prediction model, as is the case with GLM. The GLM uses a matrix $\mathbf{H}$ that is in quadratic form, that is $\mathbf{ABA}'$. This so-call "hat" matrix is:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \tag{13}$$

It is the Mahalanobis distance among $\mathbf{X}$ that scales the response variables $\mathbf{Y}$, to give the predicted variables in col($\mathbf{X}$) as seen in matrix form:

$$\mathbf{X}\hat{\mathbf{B}} = \mathbf{X}(\mathbf{X}'\mathbf{X}\mathbf{X}')^{-\mathbf{1}}\mathbf{X}'\mathbf{Y} \tag{14}$$

and in turn,

$$\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y} \tag{15}$$

Residuals in OLS regression can be computed in terms of the hat matrix:

$$\mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y} = \mathbf{\acute{E}}$$

Note that $\mathbf{H}$ is idempotent so $\mathbf{H^2}$=$\mathbf{H}$, $(\mathbf{I} - \mathbf{H})^{\mathbf{2}}$=$\mathbf{H}$ and $(\mathbf{I} - \mathbf{H})\mathbf{H}$=0.

In SEM the factor score pre-multiplied by the matrix of factor loadings serves as the projection matrix, $\mathbf{P}$, serves a similar function as the $\mathbf{H}$ matrix in OLS regression:

$$\hat{\mathbf{X}}_{pq} = \mathbf{\Lambda}_{qm}(\mathbf{\Lambda}'_{mq}\mathbf{\Theta}^{-1}_{qq}\mathbf{\Lambda}_{qm})^{-1}\mathbf{\Lambda}'_{mq}\mathbf{\Theta}^{-1}_{qq}\mathbf{X}_{pq} \tag{17}$$

Below is the form of the Bartlett projection matrix (hence the subscript b):

$$\mathbf{P}_b = \mathbf{\Lambda}_{qm}(\mathbf{\Lambda}'_{mq}\mathbf{\Theta}^{-1}_{qq}\mathbf{\Lambda}_{qm})^{-1}\mathbf{\Lambda}'_{mq}\mathbf{\Theta}^{-1}_{qq} \tag{18}$$

Here we see the computation of Bartlett ICRs in matrix form; the ICR matrix being the $\hat{\mathbf{\Delta}}$ matrix. The Bartlett factor scores are preferred by many researchers, including McDonald (2011):

$$\hat{\mathbf{\Delta}} = (\mathbf{I} - \mathbf{P}_b)\mathbf{X} \tag{19}$$

Expanding $\mathbf{P}_b$ gives:

$$\hat{\mathbf{X}}_{pq} = \mathbf{\Lambda}_{qm}(\mathbf{\Lambda}'_{mq}\mathbf{\Sigma}^{-1}_{qq}\mathbf{\Lambda}'_{mq})^{-1}\mathbf{\Sigma}^{-1}_{qq}\mathbf{X}_{pq} \tag{20}$$

ICRs can also be computed using the Thomson (Regression) factor score (denoted with the subscript r with the projection matrix). The Thomson scores are Bayes estimates if the prior is normal (McDonald, in press):

$$\mathbf{P}_r = \mathbf{\Lambda}_{qm}(\mathbf{\Lambda}'_{mq}\mathbf{\Sigma}^{-1}_{qq}\mathbf{\Lambda}_{qm})^{-1}\mathbf{\Lambda}'_{mq}\mathbf{\Sigma}^{-1}_{qq} \tag{21}$$

$$\hat{\mathbf{\Delta}} = (\mathbf{I} - \mathbf{P}_r)\boldsymbol{X} \tag{22}$$

Lastly, we can form ICRs from the Anderson-Rubin factor scores, which are used by LISREL (Joreskog, et al., 2006). Here we see the form of the factor score. $\mathbf{Z}_{pq}$ is a vector of standardized observations. $\mathbf{V}$ is a matrix of eigenvectors and $\mathbf{G}$ is a matrix of eigenvalues (DiStefano, et al., 2009).

$$\mathbf{P}_a = \mathbf{Z}_{pq}\mathbf{\Theta}^{-1}_{qq}\mathbf{\Lambda}_{qp}\mathbf{V}_{pq}\mathbf{G}_{qq}\mathbf{V}'_{pq} \tag{23}$$

And the form of the ICRs matrix:

$$\hat{\mathbf{\Delta}} = (\mathbf{I} - \mathbf{P}_a)\boldsymbol{X} \tag{24}$$

As I mentioned earlier, the meaning of the ICR, can be looked upon it as a distance, which can be standardized to form a t statistic. Bollen and Arminger (1991) posit the following nave standardization of the ICRs, which uses the square root of the uniqueness matrix in the denominator. Raykov and Penev call this

standardization the "extended" ICR. While a t test can be used with this metric, it is by definition post-hoc and is extremely liberal and unreliable (Longford, 2001).

$$\frac{\delta_b}{\sqrt{\theta}_{jj}} \tag{25}$$

The above standardization involves dividing the Bartlett ICR by the uniqueness matrix. A better standardization is computed by dividing the Regression (Thomson) ICRs by the uniqueness matrix:

$$\frac{\delta_r}{\sqrt{\theta}_{jj}} \tag{26}$$

Finally dividing the Bartlett ICRs by the covariance matrix of residuals (solved for in chapter 1) yields scores that are $N(0,1)$ according to Bollen and Arminger's simulations (1991). This is perhaps the best option for standardized residuals:

$$\frac{\delta_b}{\sqrt{[Var(\delta_b)]}_{qq}} \tag{27}$$

The relationships discussed above holds only for the Confirmatory factor models, which is the measurement component of the LISREL model. Raykov and Penev (1999) extended the definition of Bartlett and Regression ICRs to all SEMs that have fewer variables than observations (i.e. for $p<q$).

The extension and proof of existence of $\mathbf{\Delta}_{pq}$ for SEM in general is facilitated by the following lemma from Raykov and Penev (1999):

The column space of matrix $\mathbf{C}$ *and the matrix product,* $\mathbf{CK}$*, are identical iff* $\mathbf{K}$ *is nonsingular.*

The following is the equation for the structural component of an SEM, which includes the term for the path coefficient from the exogenous to endogenous latent variables. Raykov and Penev defined the equation below:

$$\boldsymbol{\eta} = \boldsymbol{\eta} + \boldsymbol{\zeta} \tag{28}$$

Then they defined the equation for the endogenous variables as seen below:

$$\mathbf{y} = \boldsymbol{\Lambda}\eta + \epsilon \tag{29}$$

Finally, Raykov and Penev combine the above two equations and, in doing so, come to a new form for a factor loading matrix as seen below:

$$\mathbf{y} = \boldsymbol{\Lambda}(\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\zeta} + \boldsymbol{\epsilon} \tag{30}$$

They define this factor loading matrix as follows:

$$\mathbf{A} = \boldsymbol{\Lambda}(\mathbf{I} - \mathbf{B})^{-\mathbf{1}} \tag{31}$$

By lemma 1, the null spaces of $\mathbf{A}$ and $\boldsymbol{\Lambda}$ are equivalent:

$$Null(\mathbf{A}) = Null(\boldsymbol{\Lambda}) \tag{32}$$

And therefore, the column spaces of $\mathbf{A}$ and $\boldsymbol{\Lambda}$ are equal:

$$R(\mathbf{A}) = R(\boldsymbol{\Lambda}) \tag{33}$$

By using the Moore-Penrose inverse, Raykov and Penev prove that the following holds completing the extension of ICRs to all SEMs:

$$\hat{\mathbf{\Delta}}_{pq} = (\mathbf{I} - \mathbf{A}_{pq}(\mathbf{A}_{pq}{}'\mathbf{\Theta}_{pq}{}^{-1}\mathbf{A}_{qp})^{-}\mathbf{A}_{pq}\mathbf{\Theta}^{-1}\mathbf{X}_{pq} \tag{34}$$

This extention of ICRs goes beyond LISREL type SEMs. Lazarfeld and Henry (1968) state, the following factor structure holds, linking the LPA to the CFA.

$$\mathbf{A_l} = \sqrt{\pi_k}(\mu_k(k) - E(x_i)) \tag{35}$$

In appendix 4, I prove the following theorem from the above relationship:

A matrix of ICRs, Δ of the same form as (32) can also be computed for Latent Profile Analysis (LPA) models. Therefore, both the LPA and CFA are equivalent models, in the sense that they share the same null space, and ICRs cannot distinguish between them.

The drawback of computing ICRs is that $\mathbf{\Delta}$ is rank deficienct. This can be shown by examining the theorem of rank products (Kwak, J.H. & Hong, S., 2004).

$$Rank(\mathbf{\Delta}) \leq \min(Rank(\mathbf{A}), Rank(\mathbf{\Theta}^{-1}), Rank(\mathbf{X})) \tag{36}$$

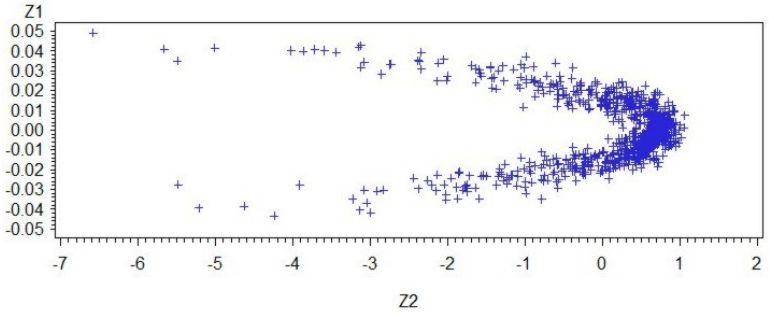
$\mathbf{\Delta}$ cannot be full row rank because it is necessary that $q<p$ for ICRs to be computed. In other words, there must be fewer variables than observations, or fewer columns than rows. Therefore

$$rowrank(\mathbf{\Delta}) < min(rowrank(\mathbf{A}), rowrank(\mathbf{\Theta}), rowrank(\mathbf{X})) \tag{37}$$

Latent Residuals

As mentioned previously, much of the research up to this point has pointed to the use of residuals in SEMs and factor analysis in detecting non-linear trends. A

Figure 1: The Curvilinear Relationship between LIRs



well formulated example of this application is presented by Raykov and Penev (2002). Here I are presented with a two factor model, with a causal path connecting factors. The model fit well, however, the path coefficient, **B** did not bring the causal relationship between latent variable into question. However, Raykov and Penev designed the simulation so that there was a parabolic relationship between factors. They were only able to see this relationship when making a scatterplot of the LIRs. A modified version of their simulation is presented here. The following is the equation for computing LIRs, which in this case uses the Bartlett factor score estimator:

$$\boldsymbol{\zeta} = (\mathbf{I} - \mathbf{B})\boldsymbol{\eta} = (\mathbf{I} - \mathbf{B})(\boldsymbol{\Lambda}'\boldsymbol{\Theta}^{-1}\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}'\boldsymbol{\Theta}^{-1}\mathbf{y} \tag{38}$$

where $\boldsymbol{\zeta}$ is the LIR and is the latent path coefficient. The LIRs from our simulation (z_1 and z_2) are plotted against each other in the scatter plot in figure 1.

This scatterplot of LIRs displays model misspecification problems as the LIRs are scattered in a parabolic fashion. This suggests that specifying a curvilinear model is, in fact, correct; and furthermore that there is indeed a causal relationship between latent variables. LIRs will be used again in the next chapter in the context

of Anscombe's (1973) displays of residuals. The key to LIRs is that in they are on a lower dimension than the raw data, which means that errors within the data become dissattenuated and, therefore, more readily visualized. Indeed Steinley and McDonald (2007) demonstrate that the natural structure of data is more readily seen in latent space. They do this by demonstrating the equivalence between latent profile and continuous confirmatory factor analysis models at the level of sufficient statistics but showing that the distinction is clear at the latent level.

Robust Statistics

To advance this process of examining residuals on the case level, a very important multivariate method I suggest in the checklist from Chapter 1 is screening by applying hypothesis tests with the Mahalanobis distance (MD) metric. MDs are geometrically the distance from an observation to the center of the ellipsoid formed by a metric of a covariance matrix. This matrix may be the sample, estimated or residual covariance matrix. MDs are asymptotically χ^2 distributed, and therefore the χ^2 test is often applied. As I explained, unfortunately, MDs are not immune to masking (Stevens, 2002). No real outlier detection method is completely immune to the problem of masking. However, some algorithms can be applied to minimize the chance of missing an outlier. As described in Appendix 2, I favor the use robust statistical methods as developed by Rousseuw (1981), to refine the covariance matrix of the sample, $\mathbf{S}$. This robust process is an initial screening for outliers. These are not automatically removed from the analysis but should not affect the process of model diagnosis.

Recently, robust methods of fitting structural and factor analysis models has

appeared in the literature. Much of these robust model fitting ideas have come from Hubert and Rousseuw's (1997) research, which exposed many of the shortcomings of maximum likelihood estimators, and offers robust solutions. For a description of these model—fitting methods, see Zhong and Yuan (2011).

Search methods for outliers have been prominent in the statistical literature since the seminal book by Cook and Weisberg (1982). The robust methods, in fact, often compliment the forward search method of Atkinson et al. (2002). Here of target data set is selected with Rousseuw's Minimum Volume Ellipsoid method. A process of adding observations to that target data set to observe a change in the Cook's D metric ensues. This search process is designed to refine the target data set to the point that the prominent outliers are removed. Rousseuw argues that the forward search is not necessary given the robustness of the MVE and Minimum Covariance Determinant (MCD) algorithms (1987). The robustness may falsely identify outliers, but will also precisely identify the problem outliers. The data can further be refined in the model-building process.

Visualizations

Like the scatterplot, the multivariate techniques I propose using are largely exploratory as well as visual. In terms of the style guidelines put forth by the American Psychological Association, this visual approach is not only acceptable, but encouraged by researchers, such as Wilkinson (2003). I will cover exploratory techniques briefly here, but will summarize them in more detail in Appendix 2. The first is the aforementioned principal component biplot, which is a visualization that collapses multidimensional data to two dimensions. This visualization may reveal

non-normality, or even influence among sets of variables (Gabriel, 1971). Outliers are identified by examining the extreme cases on the biplot display. I propose using this graphic as a first step in a diagnostic procedure for SEMs in general.

To look at possible anomalies more carefully, cluster analysis tools are useful exploratory techniques. Cluster analysis helps the researcher understand the underlying data, particularly the structure of the ICR data set. In clustering, groups of data are isolated into subsets by a recursive algorithm. This method is unsupervised and therefore there is no target variables and no underlying model. Rather, there is a top-down separation (partitioning methods) or a bottom-up joining (agglomerative methods). Both of these methods are in a class called hierarchical cluster analysis (HCA). There is also the k-means clustering method, where the number of clusters are specified a priori (Lattin, 2002). This research only makes used of agglomerative HCA, largely because it yields a useful display called the dendrogram, or tree diagram. Here, multigroup structure across cases, or non-independence across variables, can be visualized. Furthermore, clustering patterns can be analyzed to identify specific anomalous observations.

Chapter 4: Exhibits

To demonstrate the usefulness of individual case residuals and latent individual residuals, I present here four exhibits where ICRs and LIRs detect problems with respect to a given target model. These exhibits display how residuals go beyond simple examination of raw data. They uncover structures that in some cases can be only visualized on the latent level. Three multivariate tools play a central role to these exhibits: Robust Mahalanobis Distances (RMDs), the Biplot and Hierarichal Cluster Analysis using Ward's method. As discussed earlier, in the first case, I use specific hypothesis testing for identifying outliers that are measured by robust statistics.

Exhibit 1: Latent Anscombe Cases

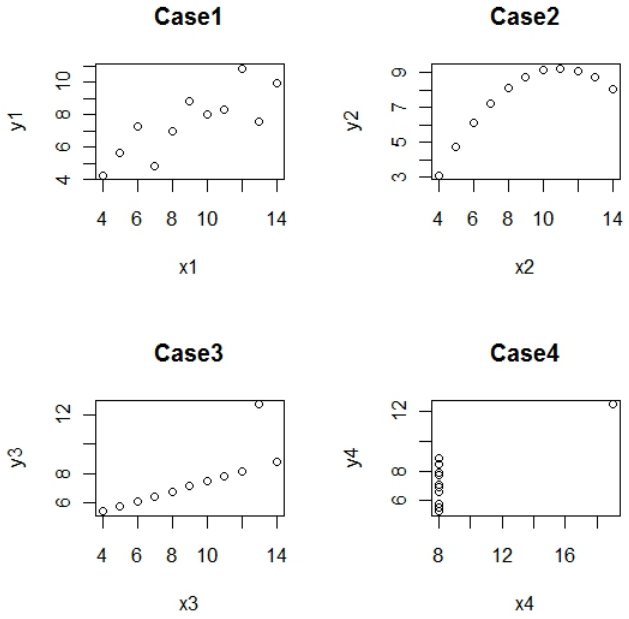
Anscombe (1973) vividly illustrated two essential characteristics of data analysis and model diagnostics. The first is the "hidden" structural relationships of model outcomes that can most clearly be seen by graphing residuals. The second is the dramatic consequences of failing to detect influential observations. Anscombe created four markedly different artificial datasets, yielded the same sufficient statistics and thus the same linear model estimates. The following matrix scatterplot of the fitted response variables against the residuals exhibits how these bivariate data truly differ (see figure 2).

In the first case, I have a scatter of points that is explained well by the line:

$$y = 3 + .5x \tag{39}$$

Few anomalies are noticed in this first graph. The second plot displays a curvilinear

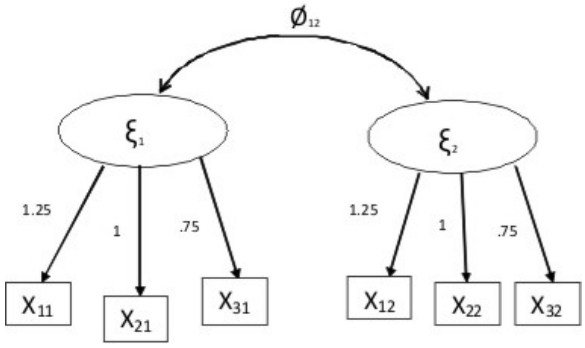
Figure 2: The Original Anscombe Quartet



trend which is not at all represented by this line. The third is a case where a single outlier diverts the regression to the above linear model, although this model is inappropriate. This model would not at all fit if that single outlier were removed; however, another straight line would. The same is true for the fourth display. Without the outlier, a regression could not be fit, as the slope of the left-most vertical stack of points cannot be estimated.

I tried to demonstrate Anscombe's phenomenon with four latent variable models depicted in figure 3.

Figure 3: The Path Model for the Anscombe Simulation



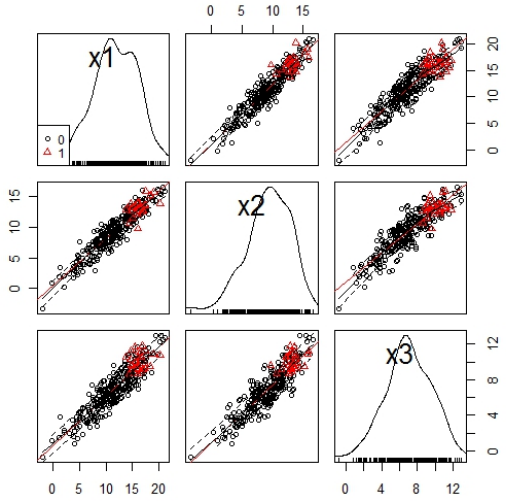
Specifically, I used Anscombe's original four bivariate data sets of 11 cases and extended them to four data sets with six variables of 300 cases each. Graham-Schmidt orthonormalization facilitated my extention of Anscombe's variable, and spline functions allowed us to extend the number of cases. As a result, the following covariance matrix is identical (save rounding error) for all four of these data sets.

$$\begin{bmatrix} 17.3 & & & & & \\ 12.9 & 11.1 & & & & \\ 9.5 & 7.5 & 6.5 & & & \\ 7.8 & 6.3 & 4.5 & 8.8 & & \\ 6.2 & 5.1 & 3.6 & 6.4 & 6.2 & \\ 4.8 & 3.9 & 2.8 & 4.8 & 3.9 & 3.9 \end{bmatrix}$$

The following matrix scatterplot displays bivariate relationships between each of the three pairs of variables for case 3. The other three cases yield similar results.

The above relationships are not as clear-cut as those seen in Anscombe's demonstration because the correlations between observed indicator variables are attenuated; however, I hypothesize that the following may be revealed when examining models specified to fit these data (assuming the same model is fit for all four sets):

Figure 4: Observations for Anscombe's Case 3



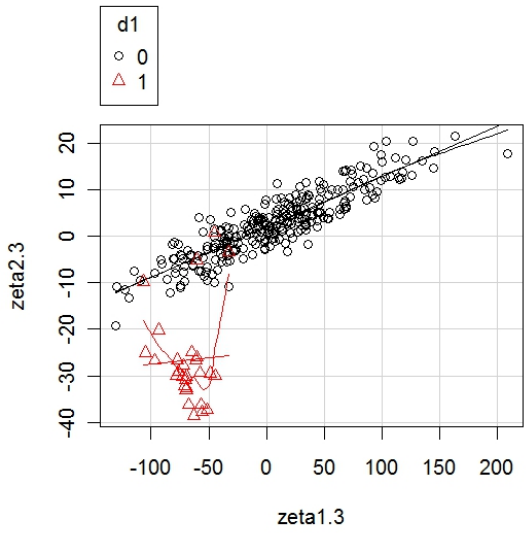
- A typical linear relationship without a large degree of influence from any one observation.
- A curvilinear relationship between each pair of variables or that same relationship between latent variables, which were realized from these observations.
- A series of influential points changing the slope of the overall linear relationship.
- A series of points, that when isolated, form a similar scatter to that seen on Anscombe's fourth graph.

I fit four two-factor models, one to each artificial data set. Three variables load on each factor and the correlation between factors is free. Freeing the correlation would prevent me from identify these models; however I set one factor loading equal to one one each latent variable. The correlation between the two

Table 1: Simple Statistics of an Anscombe Exhibit Model

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Z1	300	11.49879	4.10451	3450	-1.80574	21.13312
Z2	300	9.08303	3.30557	2725	-2.25163	17.46116
Z3	300	6.98688	2.64027	2096	-0.67855	12.56121
Z4	300	9.42287	3.00772	2827	-11.77275	13.54583
Z5	300	7.63665	2.49505	2291	-10.30331	11.82309
Z6	300	5.77323	1.86568	1732	-5.42270	9.15907

Figure 5: The Latent Residuals of the Third Anscombe Case



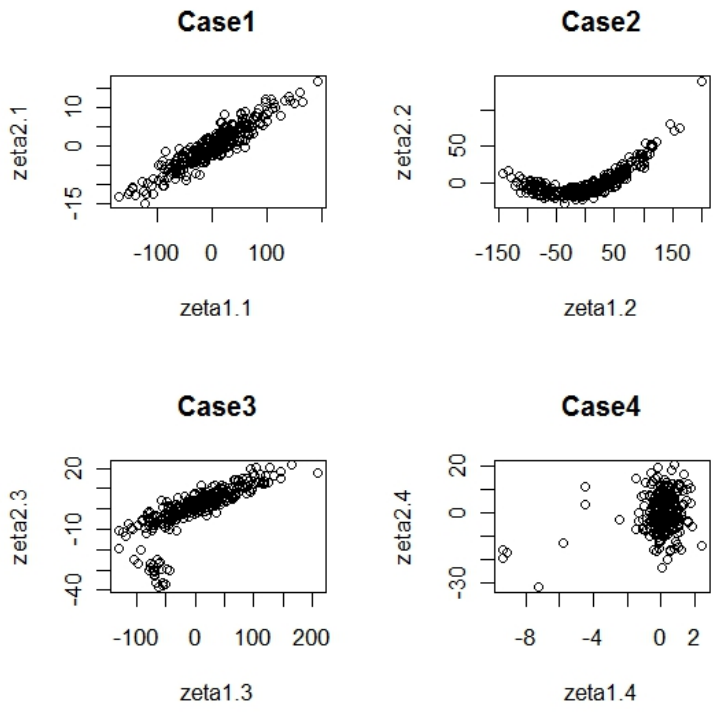
latent variables, allowed me to compute LIRs, which could potential show problems inherent in the data.

All four of these models share identical fit indices, and the same sufficient statistics as seen in table 1.

To more carefully examine the underlying relationships in these data, and the consequences of fitting these models, I plotted the LIRs for one case.

Our predictions about the latent variable relationships as a consequence of anomalies in our data sets, proved to be accurate. The dissattentuation due to the latent variables shows the true anomalies in our simulated data set. In the raw data,

Figure 6: The Four Latent Individual Residuals

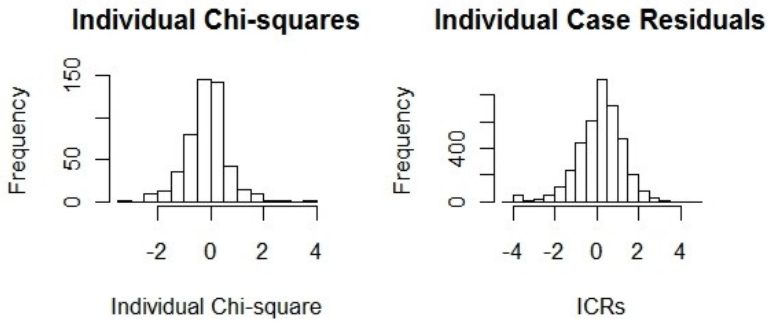


seen above, the signatures of Anscombe's data are certainly present, but it is not entirely obvious that these small trends are influential. However, when I plot the LIRs against each other, I see that Anscombe's relationships are very much present:

Exhibit 2: ICRs versus Raw Data Maximum Likelihood Using a Simulated Data Set

As discussed in Chapter 2, Raykov and Penev (1999) argue that ICRs are more appropriate diagnostic tools than what they refer to as Raw Data Maximum Likelihood (RDML) diagnostics. As mentioned previously, these metrics are discussed in detail in Pek and MacCallum (2011). They denote each cases' contribution to the log-likelihood or χ^2 (Reise & Widaman 1999; Coffman & Millsap, 2006). I attempted to show in this demonstration why ICRs have an

Figure 7: The Distributions of Case-Level Metrics



advantage; specifically in that ICRs exhibit mis-approximation for each item per each case. It is the response patterns that create outliers, and therefore, examining response patterns help determine if a given observation is errant. The display in figure 5 shows histograms of the empirical distributions of ICRs and RDML metrics. Clearly both are symmetric, but ICRs offer a closer look at the data.

RDML metrics, which may or may not be based on maximum likelihood estimation are still useful. They answer Molenaar's (2004) concerns about the variable orientation of psychometrics and the concerns of Bergman and Magnusson (1997). The individual χ^2 distribution is symmetric, and is therefore useful in identifying the direction of misfit. ICRs, however, have the advantage of showing distances of predicted values to the actual observations, which allow them to function similarly to regression residuals. As a consequence, ICRs are sensitive to inconsistencies in response patterns where RDML residuals are not. As a demonstration I simulated 500 cases of 8 items per case from the following model:

$$\mathbf{X}_{500,8} = \mathbf{\Lambda}_{1,8}\xi_{500,1} + \mathbf{\Delta}_{500,8} \tag{40}$$

Table 2: Comparison of Mahalanobis Distances of ICRs to IND_{chi}

item	MD	Robust MD	Item	IND_{chi}
500	2.72	7.04	107	3.59
499	2.72	7.04	418	2.59
498	2.72	7.04	101	2.52
497	2.72	7.04	277	2.42
496	2.72	7.04	56	2.07
495	2.72	7.04	356	1.97
494	2.72	7.04	171	1.94
493	2.72	7.04	116	1.85
492	2.72	7.04	230	1.80
491	2.72	7.04	73	1.77

where $\mathbf{X}$ is the matrix of observations, ξ is the vector of factor scores, $\mathbf{\Lambda}$ is the vector of factor loadings, and $\mathbf{\Delta}$ is the error matrix (also called disturbance terms). I contaminated the last 10 cases with one extreme response in each case. One such response pattern was as follows on a continuous scale from -3 to 3:

(3,3,3,3,3,3,3,-3)

I computed the matrix of ICRs, which the $\mathbf{\Delta}$ matrix and applied the Minimum Covariance Determinant (MCD) algorithm to identify outliers according to the robust Mahalanobis distance (RMD) criterion. The 10 contaminated cases were immediately detected. However, the CHI_{ind} did not identify these cases. The table below shows that the RMD, was markedly more effective in identifying the ten pre-constructed outliers. This is largely due to the robustness of the MCD algorithm which readily identified the cases that had higher variability. Higher variability meant larger determinants, which were not easily minimized in the MCD algorithm, hence bringing these cases outside of the range of tolerance.

Table 3: Patterns of ICRs for the top 20 Robust Mahalanobis Distances

item	ICR1	ICR2	ICR3	ICR4	ICR5	ICR6	ICR7	ICR8	RMD
491-500	1.2	0.8	0.5	0.4	0.6	0.5	0.2	-3.8	7.5
276	0.0	-0.4	0.9	-0.3	0.6	0.8	1.0	3.1	5.5
393	-1.1	-0.1	0.7	-1.5	1.5	1.2	0.6	0.6	5.1
182	-0.4	0.0	1.6	0.6	-0.7	-0.3	1.1	1.1	5.1
202	-1.0	0.3	-0.3	2.0	0.1	0.1	0.9	1.6	4.9
111	-0.3	0.8	-0.1	1.0	2.7	-0.9	-1.3	-1.4	4.8
72	-1.6	-0.4	0.5	1.2	1.2	1.2	-0.5	0.3	4.7
376	1.1	0.0	1.5	0.2	0.2	0.2	-0.6	1.1	4.6
243	0.6	0.4	0.0	-0.4	1.5	-0.2	0.1	0.8	4.6
67	-0.8	-1.8	1.3	1.2	0.2	-0.5	1.8	-0.9	4.6

Table shows the patterns of the 20 ICRs, which were identified as outliers for the artificial data set. It is clear that there is heterogeneity across response–something that cannot be revealed by CHI_{ind}.

To further support my case of the inconsistency of CHI_{ind} as compared to ICRs in detecting outliers, I formed a scatter plot of the two and find that the correlation between them is only 0.18, which indicates that the two metrics do not identify the same problems. See figure 6 for the relationship between case-level metrics.

The major drawback of the CHI_{ind} is that it confounds the response vector with cases. Molenaar (2004), acknowledging this fact, warns against intrapersonal response metrics. I believe that the CHI_{ind} is valuable in that it provides a direction to the misfit. However, I strongly favor ICRs if the researcher's intention is to screen for outliers.

Figure 8: The Relationship Between Case-level Metrics

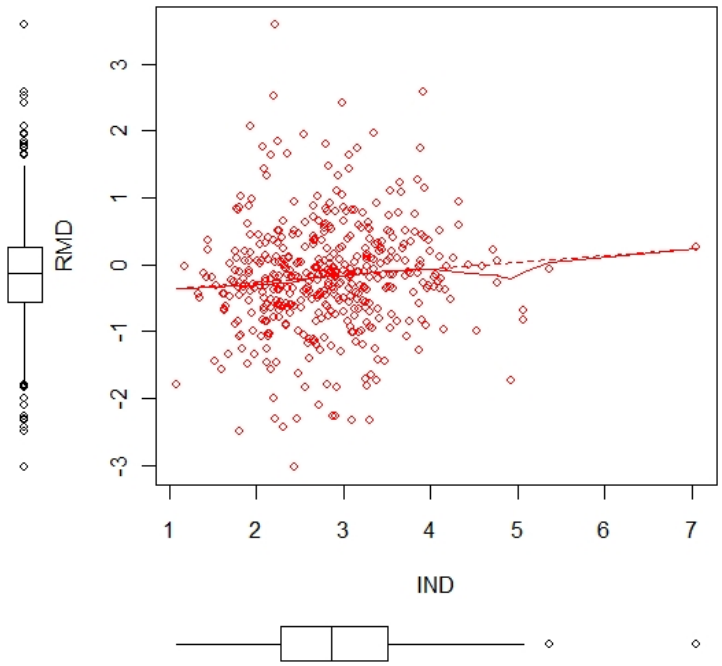


Exhibit 3: Building Models and Diagnosing them: The GSE Scale

In the interest of following the model building and diagnostic checklist, which I proposed in Chapter 1, I worked with data collected by Scherbaum (2006) from undergraduates at an urban college in the northeastern US. Specifically, I examined responses to eight items from Chen's General Self-Efficacy Scale (GSE) (2001), all of which are related stimuli. The eight items are listed below:

1. I will be able to achieve most of the goals that I have set for myself.

2. When facing difficult tasks, I am certain that I will accomplish them.

3. In general, I think that I can obtain outcomes that are important to me.

4. I believe I can succeed at most any endeavor to which I set my mind.

5. I will be able to successfully overcome many challenges.

6. I am confident that I can perform effectively on many different tasks.

7. Compared to other people, I can do most tasks very well.

8. Even when things are tough, I can perform quite well.

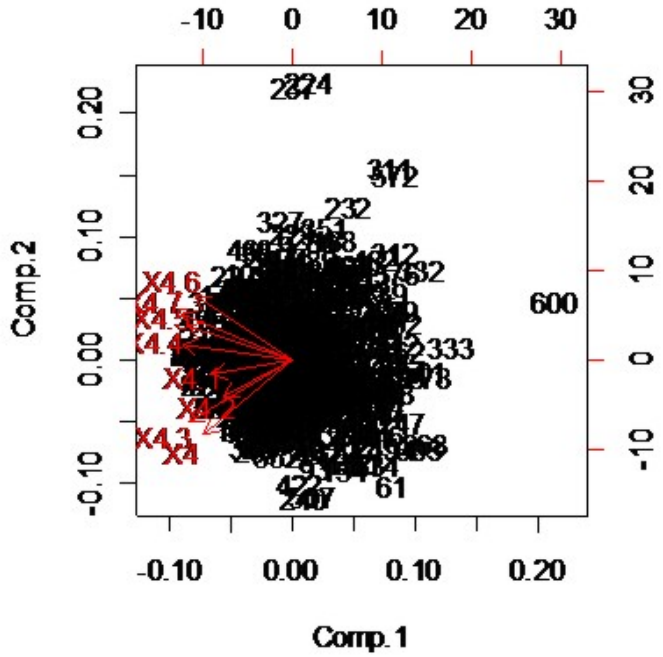
According to Chen (2001), there is no reason to assume multidimensionality with this scale. The empirical sample comprised 606 cases: 375 females and 230 males. To begin, I focused only on cleaning the raw data in the interest of fitting the target models to a robust data set. I screened for outliers that based on Robust Mahalanobis Distances with respect to the sample covariance matrix. These procedures, discussed in detail in Appendix 2, would ultimately allow me to fit a model as if there were no extreme responses. The following is the sample covariance matrix for the GSE data:

$$\begin{bmatrix} 0.64 & & & & & & \\ 0.25 & 0.51 & & & & & \\ 0.21 & 0.17 & 0.45 & & & & \\ 0.30 & 0.21 & 0.25 & 0.70 & & & \\ 0.23 & 0.22 & 0.18 & 0.33 & 0.71 & & \\ 0.22 & 0.21 & 0.19 & 0.28 & 0.35 & 0.59 & \\ 0.19 & 0.20 & 0.15 & 0.21 & 0.27 & 0.35 & 0.64 \end{bmatrix}$$

The biplot in figure 7 shows that there were indeed outlying responses, so running the minimum covariance determinant (MCD) algorithm and minimum volume ellipsoid (MVE) algorithms were necessary.

The following table displays the top ten outlying responses with respect to the centroid of the GSE data matrix. These responses are from a group who tended to use the extremes of the scale–a response style that may have been attributable to

Figure 9: Biplot of GSE Data



misinterpretation or language comprehension difficulties. More data would need to be collected to account for these responses.

My goal was to fit a model to the responses that were not identified as outliers, which are displayed in the table above. I am by no means attempting to ignore the fact that these unusual response patterns exist. For the purpose of this demonstration, however, I intended to fit a model to responses that are more typical, and in doing so, displayed how a model can fail to properly account for response sets and possible biases. The following shows the covariance scatter matrix,

Table 4: Outliers with Response to the Sample of GSE Data

ID	MD	RMD	Response Patterns							
408	8.08	10.34	4	3	4	4	9	4	3	4
238	7.34	9.89	1	4	5	1	5	5	5	5
557	7.66	9.66	3	4	3	5	4	4	3	3
577	7.21	8.89	3	4	4	1	1	4	5	1
333	7.78	8.70	4	4	1	1	1	1	4	4
483	7.26	8.30	1	4	1	4	4	3	2	4
312	6.61	8.32	1	4	4	1	1	4	4	4
225	6.15	8.03	2	3	3	1	5	5	5	5
458	6.08	7.93	1	3	4	3	5	3	1	3
63	6.09	7.83	1	4	5	4	5	3	4	5

which is the refined covariance matrix after the above outliers were removed.

$$\begin{bmatrix} 0.463 & & & & & & & \\ 0.194 & 0.399 & & & & & & \\ 0.157 & 0.120 & 0.358 & & & & & \\ 0.217 & 0.149 & 0.195 & 0.499 & & & & \\ 0.199 & 0.199 & 0.117 & 0.233 & 0.467 & & & \\ 0.209 & 0.175 & 0.130 & 0.249 & 0.291 & 0.496 & & \\ 0.170 & 0.191 & 0.127 & 0.195 & 0.243 & 0.316 & 0.571 & \\ 0.211 & 0.227 & 0.163 & 0.230 & 0.279 & 0.302 & 0.354 & 0.605 \end{bmatrix}$$

This refined covariance matrix is clearly different than the original matrix. The variances and covariances are substantially smaller, which reflects the removal of influential response patterns.

Turning now to exploratory factor analysis using this refined data I used PROMAX oblique rotation. I fit three models as displayed in the table below:

It may be tempting to judge the structure of the GSE data to have a three dimensions given the above metrics. However, the 3 factor EFA model contained a Heywood case. The unique variance for item4 was -78.040 rendering this solution impossible. With the intention of preserving the unidimensional structure of the GSE scale, I decided attempt to fit a congeneric test model, which I would then

Table 5: ICR Patterns

								MD	RMD	Sex
-0.4	-1.1	-0.4	-0.4	-2.3	-2.3	-2.1	-2.1	4.3	5.1	M
0.7	0.0	-0.3	0.7	0.8	-0.2	-1.0	1.0	3.9	5.0	M
0.0	-0.8	0.9	0.0	-0.8	0.2	-0.7	-0.6	4.3	5.0	F
-0.2	0.1	0.8	-1.2	0.0	-1.0	-0.8	-0.8	4.0	4.8	M
0.3	0.5	0.1	0.3	0.6	0.6	0.7	0.8	4.1	4.8	F

alter to accommodate the extra dimensions induced by the data. I also decided to use ULS in the process of fitting a factor model, especially with ordinal data, which I were treating as continuous for the purposes of this demonstration. I decided not to use ML estimation at least until I obtained a model that fit. Indeed, the fit with ULS appeared to be better based on a few metrics. The GFI for ML was .93 and for ULS was .98. The χ^2 was 157.5 for ML and the χ^2 could not be computed for ULS.

I knowingly left two cases with the missing data code "9" for one item in each case. These cases should be detected if the Minimum Covariance Determinant (MCD) algorithm were sensitive. Unfortunately, the MCD, which is the less robust of the two methods and less sensitive to outliers which as masked. Therefore, MCD failed to detect any "bad" points. This is disappointing because sensitivity analysis summarized in Appendix 2 reveals that MCD is more appropriate for samples below 800. I turned to the Minimum Volume Ellipsoid (MVE) method, which is very robust (as described in Appendix 2), but has no efficiency unless a large subset of data is chosen to start the minimization algorithm. In SAS IML I opted to start MVE with h=.75n, or quantile=.75. This means that the initial subset comprised 75 percent of the data, allowing MVE to function more efficiently than at the default setting, which was h=0.5. 56 points out of the 606 were identified as outliers. The top five ICRs are shown in table 5 below. The missing data codes were

identified by MVE. The other 54 cases seemed to represent atypical responses with respect to the data matrix and possibly group difference that would have to be modeled for using covariate or a random intercept. However, I recommend that a working model with the robust data should be fit to obtain parameter estimate and prediction. I therefore fitted a model on the robust subset of the GSE data using ULS and obtained a χ^2 of 128.2.

Examining discrepancy matrices reveal problems with regards to the one factor solution, versus the two factor solution. The discrepancy matrix for he one factor solution is displayed below:

$$\begin{bmatrix}
0 & & & & & & & \\
0.11 & 0 & & & & & & \\
\mathbf{0.19} & 0.06 & 0 & & & & & \\
\mathbf{0.24} & -0.17 & 0.17 & 0 & & & & \\
-0.01 & 0.02 & -0.01 & \mathbf{0.19} & 0 & & & \\
0 & -0.01 & -0.07 & 0.13 & 0.1 & 0 & & \\
0.07 & 0.01 & 0.05 & -0.14 & -0.07 & -0.09 & 0 & \\
0 & -0.03 & -0.03 & -0.02 & -0.04 & -0.05 & 0.13 & 0
\end{bmatrix}$$

Here, item 1 and item 2, item 1 and item 3, and item 5 and item 4 have relatively large residual covariances. These covariances are evidence that the one factor solution is inadequate. Here, I display the two factor solution:

$$\begin{bmatrix}
0 & & & & & & & \\
0.13 & 0 & & & & & & \\
\mathbf{0.18} & 0.03 & 0 & & & & & \\
\mathbf{0.23} & -0.18 & 0.17 & 0 & & & & \\
-0.02 & 0.02 & 0 & \mathbf{0.18} & 0 & & & \\
0 & -0.01 & -0.06 & 0.12 & 0.1 & 0 & & \\
0.07 & 0.01 & 0.06 & -0.15 & -0.07 & -0.09 & 0 & \\
0 & -0.03 & -0.02 & -0.03 & -0.04 & -0.04 & 0.13 & 0
\end{bmatrix}$$

The covariance residuals are somewhat smaller than with the one factor solution. Still there are some large covariances indicating that I needed to search for

Table 6: ICR response patterns for GSE data

								Sex	MD	RMD
4	4	3	4	5	4	4	3	*M*	4.32	4.93
5	5	5	5	5	5	5	5	*M*	4.36	4.67
5	4	4	5	4	4	4	3	*F*	4.21	4.67
4	4	4	4	5	4	4	4	*M*	4.03	4.63
5	4	5	4	4	4	3	4	*F*	4.11	4.60

a better solution.

At this point I computed ICRs to detect misfit with respect to this target model. We again used MVE on the data set of ICRs and detected the following top 5 outlying cases:

Now 23 out of 526 cases were identified as outliers with respect to the model, and I examined closely for response patterns. At this point we can detect uniform response bias and non-uniform response bias. It is interesting to note that 16 out of 23 outliers were female responses. That is 70 percent compared, the sample proportion of females of 0.62. This evidence is hardly definitive, but it at least steers me toward examining an approach involving a multigroup model. I compared the mean vectors for males and female responses below. It appears that females are different in their response style compared to males.

Table 7: Mean Vectors of Responses

Male	4.2	4.0	4.2	4.2	4.1	4.1	4.0	3.9
Female	4.1	3.8	4.2	4.1	3.9	3.9	3.8	3.7

Table 8: Mean Vectors of ICRs

Male	0.05	-0.04	0.001	0.03	-0.002	-0.002	-0.03	-0.04
Female	-0.07	0.06	-0.07	-0.05	-0.001	0.01	0.05	0.06

In terms of raw effect size the difference between males and females was 0.96–a large effect considering the GSE scale is five points. According to the Hotellings T^2 test, there was no difference in the mean structure between males and females. The $T^2 = 1.0919$, $df_1 = 8$, $df_2 = 515$, and p-value = 0.3671. The same was true between the male and female ICR data structure. $T^2 = 0.8294$, $df_1 = 8$, $df_2 = 515$, and p-value = 0.5769. The ICRs clear accentuate the sameness of the mean responses of males and females.

I observed the difference between genders, however, when I compared males versus female covariance matrices. The female covariance matrix is displayed below:

$$\begin{bmatrix} 0.59 & & & & & & & \\ 0.28 & 0.42 & & & & & & \\ 0.25 & 0.16 & 0.28 & & & & & \\ 0.26 & 0.15 & 0.17 & 0.34 & & & & \\ 0.26 & 0.21 & 0.12 & 0.2 & 0.35 & & & \\ 0.32 & 0.25 & 0.16 & 0.25 & 0.35 & 0.64 & & \\ 0.2 & 0.16 & 0.14 & 0.17 & 0.22 & 0.34 & 0.39 & \\ 0.34 & 0.27 & 0.2 & 0.22 & 0.32 & 0.41 & 0.35 & 0.69 \end{bmatrix}$$

The male covariance matrix is displayed below:

$$\begin{bmatrix} 0.31 & & & & & & & \\ 0.19 & 0.29 & & & & & & \\ 0.11 & 0.1 & 0.25 & & & & & \\ 0.18 & 0.12 & 0.17 & 0.47 & & & & \\ 0.16 & 0.19 & 0.11 & 0.28 & 0.39 & & & \\ 0.18 & 0.19 & 0.18 & 0.3 & 0.3 & 0.48 & & \\ 0.09 & 0.13 & 0.06 & 0.13 & 0.17 & 0.24 & 0.26 & \\ 0.18 & 0.24 & 0.18 & 0.28 & 0.26 & 0.38 & 0.26 & 0.61 \end{bmatrix}$$

A test with Wilks' lambda showed that there was a significant difference between these covariance matrices ($p<0.05$). Therefore I had strong evidence that a multigroup model was warranted.

I fit the multigroup model using modification indices to guide our fixing of variances and free of covariances. I executed the following modifications:

- For the females, set the error variances of item 1 item 6 item 8 be fixed
- For the females, set the covariance of item 1 and item 6 free
- For the females, set the covariance of item 1 and item 7 free
- For the males set the covariance of item 1 and item 2 free
- For the males set the covariance of item 1 and item 2 free
- For the males set the variance of item 6 fixed

This procedure yielded an the RMSEA to 0.13 and a GFI of 0.99. Furthermore, examining residuals added to my confidence that the multigroup structure was appropriate. In fact, it is possible that the factorial invariance explains the weak unidimensionality evidenced by the two highly correlated factors in the EFA that I conducted at the start of the analysis.

I decided to run a factor mixture model to see if the gender predicted group membership. Indeed gender was a useful variable in predicted class in the mixture model. It is clear from these discrepancy matrices that the multigroup model adds to reducing deviance from observed values. The following are the discrepancy matrices for males and females, which feature very small values (<0.1). For males:

$$\begin{bmatrix} 0.03 & & & & & & & \\ 0.05 & 0.03 & & & & & & \\ 0.03 & 0.04 & 0.01 & & & & & \\ 0.05 & -0.01 & 0.05 & -0.02 & & & & \\ -0.01 & 0.01 & -0.03 & 0.01 & -0.01 & & & \\ -0.03 & -0.02 & -0.04 & -0.01 & 0.03 & 0 & & \\ -0.02 & -0.04 & -0.01 & -0.01 & -0.01 & 0.02 & 0.01 & \\ -0.03 & 0 & 0 & -0.03 & 0 & -0.02 & 0.03 & 0.02 \end{bmatrix}$$

For Females:

$$\begin{bmatrix} -0.04 & & & & & & & \\ -0.02 & -0.05 & & & & & & \\ -0.02 & 0.01 & -0.01 & & & & & \\ 0.03 & -0.05 & 0.04 & 0.02 & & & & \\ 0 & 0.02 & -0.02 & 0.02 & 0.01 & & & \\ 0 & -0.01 & 0.02 & 0 & -0.01 & 0 & & \\ 0.02 & 0.01 & -0.04 & -0.04 & -0.02 & 0.02 & -0.02 & \\ -0.01 & 0.03 & 0.02 & -0.03 & 0 & 0 & 0.04 & -0.02 \end{bmatrix}$$

The MVE analysis of the residuals to the multigroup model reduce the number of outliers with respect to the model to 15. The biplot in figure 8 shows the reduced amount of outliers as compared to figure 7:

I found it unfortunate to discover poor results in terms of change in the χ^2. The multigroup χ^2 is 233, which is considerably larger than the that of 128 for the one factor CFA. This poor result could be explained by the factor that the χ^2 is estimated in ULS under the assumption of multivariate normality, which is not met with these data. After all, these data are ordinal and most likely negatively skewed. At this point, further model fitting with modification indices would have to be used to refine the multigroup model as per the covariance structure of the single group model. Also an ordinal model is warranted using polychoric correlations. Jöreskog and Sörbom (1993) outline this process in detail, and we therefore will not attempt to duplicate their procedure here. We do recommend that researchers keep in mind the short-comings of modification indices discussed earlier (Bandalos, 2002).

Figure 10: Biplot of ICRs from Multigroup Model

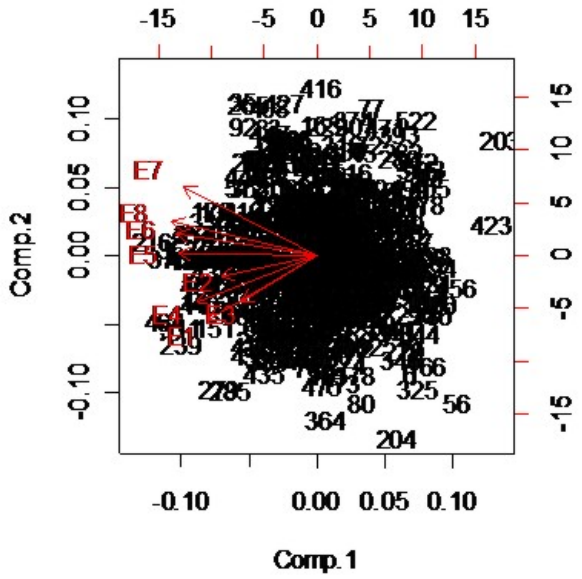


Exhibit 4: The Risk Perception Scale

My next data set was gathered from three different populations in Los Angeles by UCLA in 1997. Possible multigroup structure is again present in these data. However, here we intend to eventually side-step the multigroup model building process by using the 4 covariates in a multiple indicator multiple cause (MIMIC) style model, where a male-female dummy variable can be used as a covariate along with other covariates including age and ethnicity. This is of course a pre-conceived idea about how these data should be modeled, and should therefore not be trusted until the proper exploration, with and without ICRs is completed. The covariance matrix for the Perceived Risk dataset is as follows:

Table 9: Correlations for the Risk Perception Scale

GOLD	MED20	MED80	SILV	STK20	STK80	APPL	BIKE	BRCA	CAR	DIAB	DOC	FLU	GREY	HRT	NUC	PLANE	POOL	SWAT	XRAY
584.06																			
120.55	544.01																		
66.12	160.08	566.93																	
179.02	71.90	-58.54	533.25																
126.23	100.33	38.19	94.43	507.38															
57.28	93.62	248.12	-12.88	208.89	650.89														
74.81	-13.97	-42.89	103.16	86.98	48.81	430.48													
101.17	112.93	34.60	80.92	28.23	52.05	123.28	499.68												
38.77	48.90	50.82	76.02	61.82	87.18	199.44	95.73	764.17											
71.89	39.24	-7.98	104.20	127.54	89.38	174.09	164.12	90.22	600.24										
33.72	58.21	12.63	63.92	88.48	84.55	169.95	67.80	558.44	67.91	713.62									
85.61	57.87	0.46	95.33	103.10	61.16	133.12	103.21	178.36	113.17	133.09	485.20								
65.08	70.53	-50.23	65.43	57.72	-24.81	99.96	95.63	182.64	77.92	202.34	105.49	828.78							
59.50	53.62	-14.14	157.02	138.93	14.29	172.21	57.96	399.96	73.04	417.07	175.29	129.39	678.94						
34.94	46.20	80.80	88.05	95.07	133.72	188.94	90.76	623.10	97.80	561.84	156.75	157.52	183.66	706.90					
-20.98	0.19	61.69	-15.98	83.48	119.39	77.71	16.21	93.51	99.74	111.49	16.78	71.94	115.27	238.92	504.72				
67.05	36.94	1.83	91.92	143.86	145.58	177.44	97.87	177.34	221.07	160.57	85.73	56.80	159.82	109.94	46.82	713.62			
55.34	39.80	-46.99	98.58	108.47	56.57	167.08	128.31	96.60	161.44	109.20	144.25	193.13	175.23	130.62	60.16	133.09	485.20		
61.81	80.21	99.79	25.15	58.96	131.16	39.67	39.28	30.75	61.00	45.16	24.71	99.96	95.63	182.64	77.92	202.34	105.49	501.21	
5. 46	52.18	67.07	-8.58	78.15	85.52	69.74	68.53	180.53	62.47	184.18	60.17	172.21	57.96	399.96	73.04	417.07	175.29	62.95	754.31

The responses, yielding the above covariance matrix are subjective risk assignments. The items of this perceived risk inventory are listed in Appendix 3. They are abbreviated here. The first 6 items were financially related and the final 16 were health/lifestyle related. We excluded two health items, risk of ovarian cancer, and risk of prostate cancer because they were gender specific. Respondents were asked to rate each item on a scale of 1 to 100 based of risk (with 100 being the highest). An important paper written about these data was by Carlstrom, et al. (2000) which did not feature latent variable modeling, but rather regression techniques and probabilistic modeling. Skewness and kurtosis are inevitably going to be different for each of the 20 items that we are retaining in the scale. This problem is visualized in figure, an empirical graph ICRs across items with respect to a two factor model.

Part of this study was the World View Category classifications (WVCAT) for each participant which we used as a covariate in our confirmatory factor modeling. The world view categories were:

- Unclassifiable
- Individualist
- Hierarchicalist
- Egalitarian

The two factor one factor for finance and one for health. Observing the covariance matrix reveals possible structure among the health items. I therefore

Figure 11: The Items Across the Risk Perception Scale

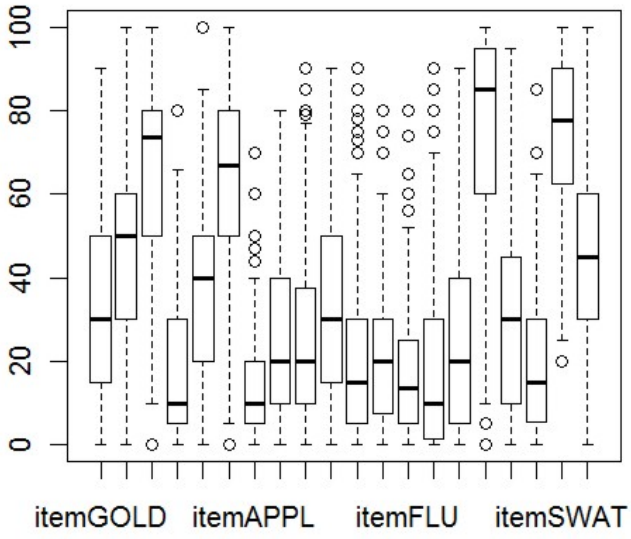


Table 10: EFA Diagnostics for the Risk Perception

Factor	χ^2	CFI	RMSEA
1	683.4	0.743	0.114
2	243.2	0.924	0.068
3	127.8	0.968	0.049

used exploratory factor analysis on the health items alone, to see how the one, two, and three factor solutions fit. Table 7 shows the results.

Considering the lack of fit of the three factor structure for the health related items, I decided to fit a four factor confirmatory factor analysis with the following structure:

Table 11: EFA Factor Loadings

item	1	2	3
APPL	0.145	0.556	0.025
BIKE	-0.017	0.461	-0.071
BRCA	0.891	0.007	-0.001
CAR	-0.036	0.502	0.125
DIAB	0.884	-0.030	0.037
DOC	0.176	0.390	-0.029
FLU	0.240	0.287	-0.033
GREY	0.594	0.110	-0.018
HRT	0.830	0.003	0.007
NUC	-0.019	-0.006	0.735
PLAN	0.034	0.446	0.319
POOL	-0.044	0.650	0.005
SWAT	-0.028	-0.031	0.479
XRAY	0.195	0.037	0.345

The factor correlation matrix for the above three factor Geomin EFA is as follows:

$$\begin{bmatrix} 1.00 & & \\ 0.38 & 1.00 & \\ 0.10 & 0.20 & 1.00 \end{bmatrix}$$

The uniqueness matrix from the above 3 factor CFA are displayed in the following table:

Next I fit a four factor structure to the health-related items which is shown in the table below:

As with the GSE data, there was a larger female group of respondents of 384, versus a male group of respondents of 226. There was also a great deal of missing data. List-wise deletion followed by Rousseuw's MCD algorithm reduced the data set down to 304 from 610. The obvious solution to a missing data problem such as this one is to employ full information maximum likelihood. In the interest of rigor, and out of our belief that ULS would perform better in data such as these, we fit

Table 12: EFA Unique Variances

APPL	1
BIKE	14.205
BRCA	15.996
CAR	24.222
DIAB	17.283
DOC	11.458
FLU	21.334
GREY	16.083
HRT	16.761
NUC	9.201
PLANE	15.125
POOL	14.180
SWAT	10.471
XRAY	15.238

Table 13: Factor Structure for the Risk Perception Model

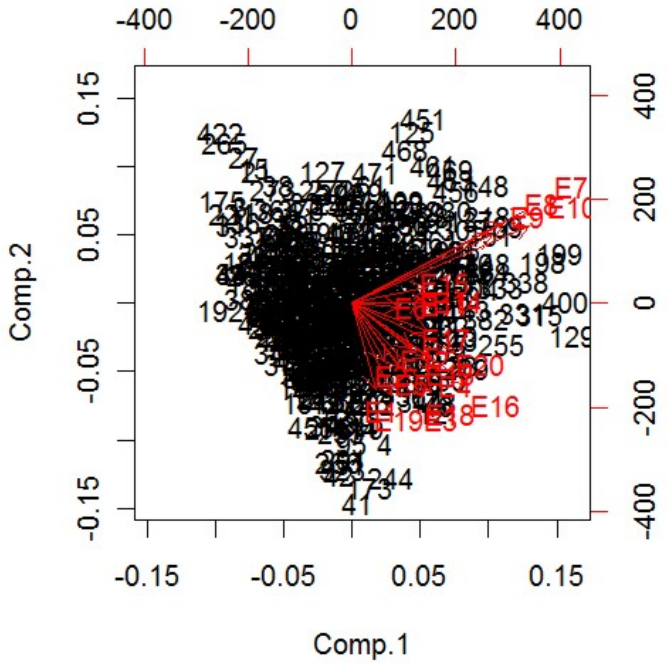
Factor1	Factor2	Factor3	Factor4
hline MED80	BRCA	APPL	NUKE
MED20	GREY	BIKE	SWAT
STK80	HRT	CAR	XRAY
STK20	DIAB	DOC	
GOLD		FLU	
SILV		PLAN	

two four-factor models, one by FIML and one by ULS with list-wise deletion. First of all, the RMSEA for FIML was 0.07 versus an RMSEA of 0.064 for ULS. Also the residual variance for FIML ranged from 171 to 616, where the residual variances for ULS ranged from 157 to 560.

The better fit of the the list-wise deleted ULS data set was overshadowed by the fact that between robust cleaning with the MCD algorithm and the deletion, I was only left with 304 out of the original 610 cases. Furthermore, only 100 male cases that had full responses remained. Examining outliers amongst the ICRs of these 304 data revealed that female respondents tended to be outlying. This female variance is unlikely to be due to true relationships, and more likely to point to my inadequately addressing the missing data problem. Perhaps the group differences pointed to non-uniform response bias and the need to control for gender in the analysis.

Therefore I turned to LISREL 8.8 and used the Markov-chain monte carlo imputation algorithm developed by Jrekog and Srbom (1993). Here we found 68 missing data patterns, which were corrected by imputation based on likely responses for a particular response. This allowed us to recover all 610 cases and still use the preferred ULS fit function. I next cleaned the imputed data set using the robust MCD algorithm, which is appropriate for a data set of this size (Hubert, et al., 2008) and lost 138 data points. I then fit the 4 factor ULS model and achieved a satisfactory GFI of 93.3. The computation and analysis of the ICRs from this model revealed some interesting findings, not seen when we lost so much data due to list-wise deletion. Now 64 percent of the 130 MCD outliers were female; therefore

Figure 12: The Biplot of ICRs from the Risk Perception CFA Model



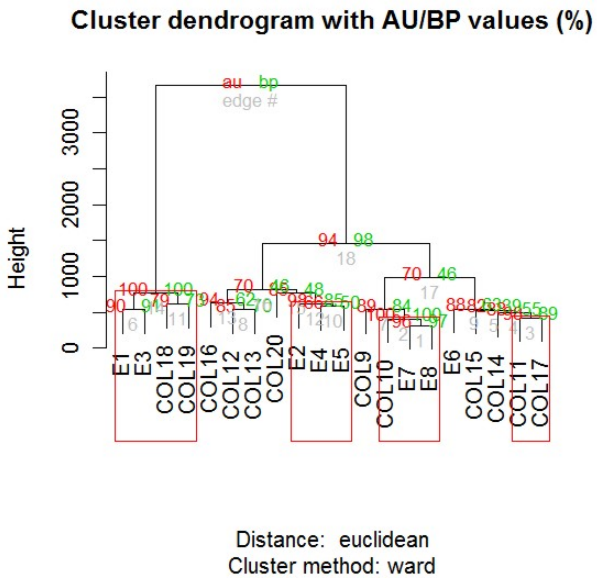
the bias toward female aberrant response was eliminated by the imputation.

Unfortunately, specification issues in this model are readily apparent in the biplot of ICRs seen in figure . All twenty principal components point in the same direction, indicating a lack of independence among errors.

Further, the dendrogram of a cluster analysis of these ICRs reveals they are conflated in the agglomerative joining process. We used the common Ward's joining method for this cluster analysis, as this method was favored by Maydeu-Olivaries in his 1996 paper due to its reliable results in his simulations. Each ICR should be at the root of the joining process; due to lack of independence, this branching pattern is not observed (see figure 10).

At this point, I realized that it was time to fit the full model with covariates. One more use of ICRs that we wanted to attempt is to cluster the ICRs by a

Figure 13: Dendrogram for Risk Perception Data



covariate both across variables and across cases. PROC CLUSTER and PROC VARCLUS in SAS 9.2 offer us this capability. Now, it is true that at the outset of our analysis, we performed exploratory factor analysis and discovered a substantively satisfactory factor structure. Now that we have fit the model, we can examine the residual structure across the 20 items to test the validity of the model specifications, which is indeed tenuous. As the table below exhibits, there is a missing factor arising from the structure amongst the first six financial items.

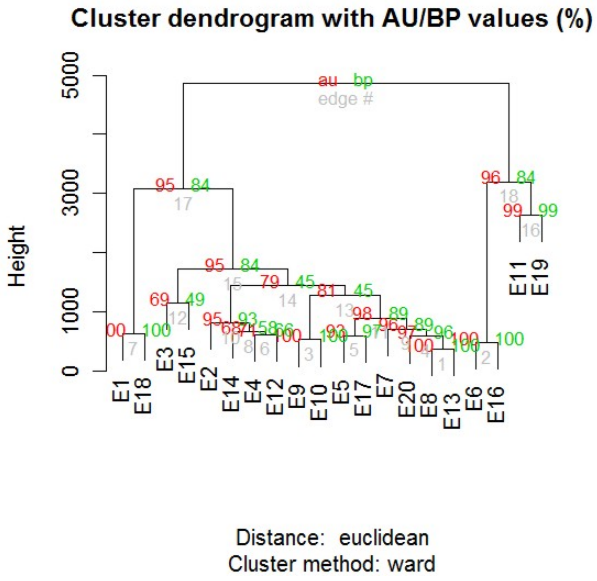
This structure is not readily seen by examining the raw data, because an oblique principal components analysis reveals a four factor structure. Looking at ICRs helps us identify structure in the residue of the fit model that can guide in better specification. And here we increase the GFI of the ULS CFA to 96.5 from 93.3 by adding a fifth factor.

The advantage of the five factor solution is further evidenced by examining the two dendrograms from Ward's Hierarchical Cluster analysis performed on the 20

Table 14: Standardized Scoring Coefficients from Variable Clustering

Cluster	1	2	3	4	5
E1	0	0	0.425	0	0
E2	0	0	0	0	0.418
E3	0	0	0.505	0	0
E4	0	0	0.396	0	0
E5	0	0	0	0	0.526
E6	0	0	0	0	0.462
E7	0.305	0	0	0	0
E8	0.298	0	0	0	0
E9	0.255	0	0	0	0
E10	0.291	0	0	0	0
E11	0	0.270	0	0	0
E12	0	0.219	0	0	0
E13	0	0.238	0	0	0
E14	0	0.209	0	0	0
E15	0	0.171	0	0	0
E16	0	0.240	0	0	0
E17	0	0.276	0	0	0
E18	0	0	0	0.543	0
E19	0	0	0	0.412	0
E20	0	0	0	0.462	0

Figure 14: Dendrogram for the Five Factor Solution of the Risk Perception Scale



ICRs (last 2 figures). The first dendrogram displayed is for the four factor solution and the second is for the five factor solution. It is apparent that the ICRs (denoted by E) are independent in the second dendrogram (See Figure 11).

Discussion and Future Directions

Perhaps the major question left by the above research presentation is: does the added value to creating individual and/or latent residuals for diagnosing misfit of structural or confirmatory factor models, outweigh the extra effort? I believe the answer to this question is yes given the advent of new multivariate diagnostic tools in packages such as SAS, R, Mplus and LISREL. Furthermore, in the near future, more ready-to-use diagnostics will be available, which will make the creation of these residuals much more feasible. Currently, LISREL 8.8 produces Andersen-Rubin case residuals and factor scores (Jöreskog & Sörbom, 2008).

Given the development of data-level diagnostics in latent variable models, I believe it is safe to conjecture that the person-orientation in psychology over the past 15 years has impetus to migrate from parameter space to the space of the observations, which are ultimately realized by latent variables. Also, judging individual case misfit with respect to the researchers' target model is valuable in guiding decision making in fitting models, especially when it comes to response biases and multigroup structure. Furthermore, understanding each individual response pattern beyond the individual contribution to the $log - likelihood$ adds to understanding the response patterns of subpopulations, which can in turn provide reason to create a multigroup model, extra factors or add covariates.

Now that I have put the procedure into practice, I now attempt to expound upon the SEM/CFA model building and diagnosis checklist, which I first proposed in the introductory chapter:

- Clean the raw data using Rousseuw's robust algorithms to generate a robust

data set. Identified outlying cases should not be discarded completely, but examined for anomalies that could explain response style, biases or subpopulations.

- Run exploratory factor analysis (EFA) on cleaned data to detect factor structure. The EFA should be semi-exploratory as there should be a preconceived notion as to the number of emerging factors.
- Fit confirmatory factor analysis (CFA) models, with unweighted-least squares (ULS): One to model that represents the desired structure, and one to the model that "fits" very well, but is not substantively correct. At this point the models should be informed by the initial EFA procedure.
- Create ICRs and latent individual residuals (LIRs).
- Create a scatterplot, biplot and dendrogram to study structure of residuals on both the manifest and latent level.
- Run robust algorithms on ICRs to locate cases that are outliers with respect to the matrix of ICRs (which is the discrepancy matrix). These outliers will give clues to specification problems.
- Attempt to adjust target model by adding terms (covariates, intercepts).
- If a multigroup model is warrented, fix or free variances and covariances based on modification indices and the discrepancy matrix.

Future research into SEM and CFA residuals can extend to latent profile analysis, latent class analysis, latent growth curve models (building upon the

research of Coffman & Millsap (2006)) and adjusting to ordinal and binary data. Such ordinal and binary person oriented SEM and CFA bridges the gap between SEM and item-response theory. It would also be interesting to research the application of ICRs in detecting sources of improper solutions such as, large SEs, negative unique variances (Heywood cases), or underidentification exhibited by the information matrix.

Perhaps the important lesson from this area of research speaks to the model building and diagnostic process, which thus far, does not match the rigorous case oriented practice in OLS regression. Anscombe (1973) argued for statisticians to look look at raw data and residuals. Today we are at the point where the same can be done in latent variable modeling. And with the development of better software options, in tools such as SAS IML, MPlus or R's MIX package, such diagnostic capability is practically feasible.

References and Appendices

Appendix 1: Notation Bible

$\mathbf{y}$ p×1 vector of observed response, or outcome variables.

$\mathbf{x}$ q×1 vector of observed predictor, covariates, or input variables.

$\boldsymbol{\eta}$ m× random vector of latent dependent, or exogenous, variables.

$\boldsymbol{\xi}$ n×1 random vector of latent independent, or exogenous, variables.

$\boldsymbol{\epsilon}$ p×1 vector of measurement errors in y.

$\boldsymbol{\delta}$ q×1 vector of measurement error in x.

$\mathbf{e}$ q×1 vector of individual case residuals.

$\boldsymbol{\Lambda}_y$ is a p×m matrix of coefficients of the regression of $\mathbf{y}$ on η.

$\boldsymbol{\Lambda}_x$ is a p×m matrix of coefficients of the regression of $\mathbf{x}$ on $\boldsymbol{\xi}$.

$\mathbf{T}$ an m×n matrix of coefficients of the $\boldsymbol{\xi}$ variables in the structural relationship.

$\mathbf{B}$ an m×n matrix of coefficient of the $\boldsymbol{\eta}$ variables in the structural relationship.$\mathbf{B}$ has zeros in the diagonal, and I-$\mathbf{B}$ is required to be non-singular.

$\boldsymbol{\zeta}$ an m×1 vector of equation errors (random disturbances) in the structural relationship between $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$.

$\text{cov}(\boldsymbol{\xi}) = \boldsymbol{\Phi}(n \times n)$.

$\text{cov}(\boldsymbol{\epsilon}) = \boldsymbol{\theta}$

$(p \times p)$.

$\text{cov}(\boldsymbol{\zeta}) = \boldsymbol{\Psi}(m \times m)$.

$\text{cov}(\boldsymbol{\delta}) = \boldsymbol{\theta}$

δ

$(q \times q)$.

π_k the probability of membership to class k for latent class and latent profile models.

Appendix 2: Glossary of Multivariate Techniques

subsectionThe Biplot

The biplot can serve as a convenient first step in understanding problems in the ICR multivariate data set. The SVD is used here to generate a the popular Principal Components Biplot; the principal components being the result of projecting the original set of variables onto a lower dimensional vector space.

Cluster Analysis

Cluster analysis provides an opportunity to detect outliers among the ICR data, as this methodology groups and classifies similar observations, while relying on a set of explanatory variables (Lattin et al., 2002). The many types of clustering algorithms all depend on a specified linkage rule, which is applied to a distance metric. This metric may be the Euclidean distance (single linkage, average linkage),

it may rely on the modality of the data (density linkage), or may rely on differences between error sums-of-squares (Ward's method). We use Ward's method because it is a popular method, which was used by Maydeu-Olivares in his 1996 on prison furlows. Maydeu-Olivares favors Ward's method because his simulations showed accurate joining patterns, which were based on common variances among observations.

Robust Statistics

The advantage of robust statistics and algorithms of robust statistics, such as the MVE and MDC, is their high breakdown points (Rousseuw, 1997). The breakdown point is the fraction of contaminated observations than can cause an estimate to become unbounded or arbitrarily distant from the center of a dataset (or from the centroid with a covariance matrix). We have found in our research that MCD has a high breakdown point for the type of residual datasets we are studying. We use the MDC, by implementing the robust regression algorithm used in the Interactive Matrix Language (IML) of SAS 9.2, accurately implements the fast MDC procedure. This fast version of MDC, as described in Rousseuw and Van Driessen (1999), uses a selective iteration algorithm where . This algorithm allows the procedure to converge more rapidly than Rousseuw's original procedure.

The two procedures discussed here involve finding minimum distances in multivariate vector space. If we look carefully at the general distance metric, we see that we can derive many types of distances provided that we specificy at matrix as the middle term. For example, we can take the following distance with respect to the sample covariance matrix, S:

$$MD_S = \sqrt{(x_i - \bar{x})\mathbf{S}^{-1}(x_i - \bar{x})} \tag{41}$$

The above distance metric is the Mahalanobis distance (MD), which was mentioned earlier. This distance may be taken with respect to the sample, residual, or estimated covariance matrices. Here we will focus on MDs with respect to S. The MD as defined geometrically, is the distance of a given observation to the centroid a the ellipsoid formed by matrix S (Fischer, 2004). The critical values of the χ^2 distribution, with degrees of freedom equal to the number of parameters in the model can point out extreme observations and outliers. Unfortunately, if the matrix S is contaminated, this simple statistical test will not perform well due to masking.

Rousseuw (1997) uses robust techniques to define the minimimun volume ellipsoid (MVE) and the minimum determinant covariance (MDC) matrix to further refine the specific covariance matrix. Rousseuw (1997) defines his MVE method in developing an algorithm as marked by determining the ellipsoid with the smallest volume that contains h observations, where h=n/2. The distance from the centroid of covariance matrix C and the ellipsoid is defined by the Robust Mahalanobis Distance (RMD).

Risk Perception Stimuli

Financial Activities.

1. Invest 80 % of savings in new medical research firm.
2. Invest 20 % of savings in new medical research firm.
3. Invest 80 % of savings in blue-chip stock.

4. Invest 1 oz. of gold: now worth about $ 383.

5. Invest 1 oz. of silver: now worth about $ 5.

Health Activities

1. Work as a family physician in rural areas.

2. Work as a member of SWAT police team.

3. Ride bicycle 1 mile each day in an urban area.

4. Receive annual preventative flue vaccination.

5. Drive automobile 10 miles each day in an urban area.

6. Swim in indoor public pool each weekend.

7. Live near nuclear power station.

8. Receive diagnostic X-rays every 6 months.

9. Fly on commercial airplanes every month.

10. Use household appliances.

11. Tested for gene that predisposes to heart disease.

12. Tested for gene that predisposes to diabetes.

13. Tested for gene that predisposes to breast cancer.

14. Tested for gene that predisposes to premature grey hair.

Appendix 4: Proofs

Now we formally prove a characteristic of $\mathbf{P}_b$ and $\mathbf{P}_r$. A projection matrix is idemopotent by definition. Idempotency means that $\boldsymbol{A}=\boldsymbol{A}^2=\boldsymbol{A}\boldsymbol{A}$. For the "hat" matrix, be show $\boldsymbol{H}^2 = \boldsymbol{H}$ below:

$$\mathbf{H}^2 = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \tag{42}$$

Theorem 1: The Bartlett and Thomson (regression) scores are idempotent.

Proof:

$$\mathbf{P}_b{}^2 = \mathbf{\Lambda}_{qm}(\mathbf{\Lambda}'_{mq}\mathbf{\Theta}^{-1}_{qq}\mathbf{\Lambda}_{qm})^{-1}(\mathbf{\Lambda}'_{mq}\mathbf{\Theta}^{-1}_{qq}\mathbf{\Lambda}_{qm})(\mathbf{\Lambda}'_{mq}\mathbf{\Theta}^{-1}_{qq}\mathbf{\Lambda}_{qm})^{-1}\mathbf{\Lambda}'_{mq}\mathbf{\Theta}^{-1}_{qq} = \mathbf{P}_b \tag{43}$$

$$\mathbf{P}_r^2 = \mathbf{\Lambda}_{qm}(\mathbf{\Lambda}'_{mq}\mathbf{\Sigma}^{-1}_{qq}\mathbf{\Lambda}_{qm})^{-1}(\mathbf{\Lambda}'_{mq}\mathbf{\Sigma}^{-1}_{qq}\mathbf{\Lambda}_{qm})(\mathbf{\Lambda}'_{mq}\mathbf{\Sigma}^{-1}_{qq}\mathbf{\Lambda}_{qm})^{-1}\mathbf{\Lambda}'_{mq}\mathbf{\Sigma}^{-1}_{qq} = \mathbf{P}_r \tag{44}$$

Theorem 3: A matrix of ICRs, Δ of the same form as (32) can also be computed for LPA models.

Proof:

$$\mathbf{A_l} = \sqrt{\pi_k}(\mu_k(k) - E(x_i)) \tag{45}$$

Steinley and McDonald (2008) show the following relationship holds:

$$\mathbf{A_l} = \mathbf{\Lambda}\mathbf{\Phi}_n^{-\frac{1}{2}} \tag{46}$$

Therefore by lemma 1 ,

$$Null(\mathbf{A_1}) = Null(\mathbf{\Lambda}) \tag{47}$$

and, therefore, the general equation for matrix $\hat{\mathbf{\Delta}}_{pq}$ holds for LPA models as well.

Bibliography

[1] Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, 27(1), 17-21.

[2] Atkinson, A. C. (1986). Masking Unmasked. *Biometrika*, 73(3), 53 3-541.

[3] Atkinson, A. C., Riani, M., & Cerioli, A. (2004). *Exploring Multivariate Data with the Forward Search*. New York: Springer.

[4] Atkinson, A. C. &. Riani, M. (2001). Regression Diagnostics for Binomial Data from the Forward Search. *Journal of the Royal Statistical Society. Series D,* 50(1), 63-78.

[5] Bandalos, D. L. (2002). The effects of item parcelling on goodness-of-fit and parameter estimate bias in structural equation modeling. *Structural Equation Modeling*, 9(1), 78-102.

[6] Bergman, L. R., & Magnusson, D. (1997). A person-oriented approach in research on developmental psychopathology. *Development and psychopathology*, 9(02), 291-319.

[7] Bergman, L. R., & Trost, K. (2006). The person-oriented versus the variable-oriented approach: are they complementary, opposites, or exploring different worlds? *Merrill-Palmer Quarterly*, 52(3), 601-632.

[8] Bollen, K.A. and D.J. Bauer. 2004. Automating the Selection of Model-Implied Instrumental Variables. *Sociological Methods and Research* 32, 425-52.

[9] Bollen, K. A. (1987). Outliers and Improper Solutions: A Confirmatory Factor Analysis Example. *Sociological Methods Research*, 15, 375-386.

[10] Bollen, K. A., Kirby, J.B., Curran, P.J., Paxton, P.M., & Chen, F. (2007). Latent Variable Models Under Misspecification. *Sociological Methods Research*, 36(1), 48-86.

[11] Bollen, K. A. &. Arminger, G. (1991). Observational Residuals in Factor Analysis and Structural Equation Models. *Sociological Methods*, 21, 235-262.

[12] Carlstrom, L. K., Woodward, J. A., & Palmer, C. G. S. (2000). Evaluating the simplified conjoint expected risk model: comparing the use of objective and subjective information. *Risk Analysis*, 20(3), 385-392.

[13] Caroll, J. D., Green, P.E. & Chaturvedi, A. (1997). Mathematical Tools for Applied Multivariate Analysis, Revised Edition. San Diego: Academic Press.

[14] Chan, W. (2003). Analyzing ipsative data in psychological research. *Behaviormetrika*, 30(1), 99-121.

[15] Chan, W., & Bentler, P. M. (1993). The covariance structure analysis of ipsative data. *Sociological Methods* & research, 22(2), 214.

[16] Chen, G., Gully, S.M. & Eden, D. (2001). Validation of a new General Self-Efficacy Scale. *Organizational Research Methods*, 4, 62.

[17] Cheng, Y., & Yuan, K. H. The Impact of Fallible Item Parameter Estimates on Latent Trait Recovery. *Psychometrika*, 1-12.

[18] Cheung, M. W. L., & Chan, W. (2002). Reducing uniform response bias with ipsative measurement in multiple-group confirmatory factor analysis. *Structural Equation Modeling*.

[19] Coffman, D. L. & Millsap, R.E. (2006). Evaluating Latent Growth Curves Using Individual Fit Statistics. *Structural Equation Modeling*, 13(1), 1-27.

[20] Comrey, A. L. (1985). A Method For Removing Outliers to Improve Factor Analysis Results. *Multivariate Behavioral Research*, 20, 273-281.

[21] Cook, R. D., & Weisberg, S. (1982). *Residuals and Influence in Regression*. New York: Chapman and Hall.

[22] Cudeck, R., & MacCallum, R. (2007). *Factor analysis at 100: historical developments and future directions*, Lawrence Erlbaum.

[23] DiStefano, C., Zhu, M., & Mindrila, D. Understanding and Using Factor Scores: Considerations for the Applied Researcher. *Practical Research Evaluation.* 14(20), 1-11.

[24] Drineas, P., Frieze, A., Kannan, R., Vempala, S., & Vinay, V. (2004). Clustering large graphs via the singular value decomposition. *Machine Learning*, 56(1), 9-33.

[25] Ferrando, P., & Lorenzo-Seva, U. (2009). Acquiescence as a source of bias and model and person misfit: A theoretical and empirical analysis. *British Journal of Mathematical and Statistical Psychology*,(in press, pre-print available online).

[26] Fitzmaurice, G. M., Laird, N.M., and Ware, J.H. (2004). *Applied Longitudinal Analysis*. New York: Wiley.

[27] Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58(3), 453.

[28] Gower, J. C., & Hand, D. J. (1996). *Biplots*: Chapman & Hall/CRC.

[29] Hausman, J. A. (1978). Specification Tests in Econometrics. *Econometrica*, 48(6), 1251-1271.

[30] Houseman, E. A., Ryan, L. M., & Coull, B. A. (2004). Cholesky residuals for assessing normal errors in a linear model with correlated outcomes. *Journal of the American Statistical Association*, 99(466), 383-394.

[31] Hubert, M. Rousseeuw, P.J (1997) Journal of Statistical Planning and Inference, 57, 153-163

[32] Hubert, M., & Rousseeuw, P. J. (2005). ROBPCA: a new approach to robust principal component analysis. *Technometrics*, 64-79.

[33] Hui, C. H., & Triandis, H. C. (1985). The instability of response sets. *Public Opinion Quarterly*, 49(2), 253.

[34] Johnson, T. R., & Bolt, D. M. On the use of factor-analytic multinomial logit item response models to account for individual differences in response style. *Journal of Educational and Behavioral Statistics*, 35(1), 92.

[35] Jöreskog, K., Sörbom, D., & Inc, S. (1996). *LISREL 8: User's reference guide: Scientific Software*.

[36] Jöreskog, K., Sörbom, D., Magidson, J., & Cooley, W. (1979). *Advances in factor analysis and structural equation models* Abt Books Cambridge, MA.

[37] Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language: Scientific Software.*

[38] Jöreskog, K. G., & Sörbom, D. (1993). Testing structural equation models. *Testing structural equation models*, 294, 316.

[39] Kaplan, D. (1988). The impact of specification error on the estimation, testing, and improvement of structural equation models. *Multivariate Behavioral Research*, 23(1), 69-86.

[40] Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. Applied Measurement In Education 16, 277-298.

[41] Kline, R. B. *Principles and practice of structural equation modeling* The Guilford Press.

[42] Lange, K., Westlake, J., & Spence, M. (1976). Extensions to pedigree analysis III. Variance components by the scoring method. *Annals of Human Genetics*, 39(4), 485-491.

[43] Lange, N. a. R., L. (1989). Assessing Normality in Random Effects Models. *The Annals of Statistics*, 17(2), 624-642.

[44] Langford, I. H. a. L., T. (1998). Outliers in Multilevel Data. *Journal of the Royal Statistical Society. Series A*, 161(2), 121-160.

[45] Lattin, J., Carroll, J., & Green, P. (2003). *Analyzing Multivariate Data* Thomson Brooks/Cole.

[46] Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis* Houghton Mifflin Co.

[47] Ling, R. F. (1973). A probability theory of cluster analysis. *Journal of the American Statistical Association*, 68(341), 159-164.

[48] Longford, N.T. (2001). Simulation-Based Diagnostics in Random-Coefficient Models *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 164(2), 259-273. Published by: Blackwell Publishing

[49] MacCallum, R. C. (2003). Working with imperfect models. *Multivariate Behavioral Research*, 38(1), 113-139.

[50] MacCallum, R. C., Browne, M.W., and Cai, L. (Ed.). (2007). *Factor Analysis Models as Approximations: Some History and Some Implications.* Mahwah: Laurence Erlbaum.

[51] MacCallum, R. C. & Tucker, L.R. (1991). Representing Sources of Error in the Common Factor Model: Implications for Theory and Practice. *Psychological Bulletin*, 109, 502-511.

[52] Mavridis, D., & Moustaki, I. (2009). The forward search algorithm for detecting aberrant response patterns in factor analysis for binary data. *Journal of computational and graphical statistics*, 18(4), 1016-1034.

[53] Maydeu-Olivares, A. (1996). Classification of prison inmates based on hierarchical cluster analysis. *Psicothema*, 8(3), 709-715.

[54] Maydeu-Olivares, A. a. C., D.L. (2006). *Random Intercept Item Factor Analysis*. Psychological Methods, 11(4), 344-362.

[55] McDonald, R. P. (2010). Structural Models and the Art of Approximation. *Perspectives on Psychological Science*, 5(6), 675.

[56] McDonald, R.P (2011). Measuring Latent Quantities. *Psychometrika, online now.*

[57] McDonald, R. P., & Bolt, D. M. (1998). The Determinacy of Variables in Structural Equation Models. *Multivariate Behavioral Research*, 33(3), 385-401.

[58] McDonald, R. P., & Ho, M. H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, 7(1), 64-82.

[59] McLeod, L. D. (2000). Minutes of the Psychometric Society business meeting July 8, 2000. Psychometrika, 65(4), 561-562.

[60] Molenaar, P. (2004). A Manifesto on Psychology as Idiographic Science: Bringing the Person Back Into Scientific Psychology, This Time Forever. *Measurement: Interdisciplinary research and perspectives*.

[61] Pek, J.

[62] Raykov, T. (2005). Residuals in Structural Equation, Factor Analysis, and Path Analysis Model. In B. a. H. Everitt, D. (Ed.), *Encyclopedia of Statistics in Behavioral Science*. Hoboken, N.J.: John Wiley and Sons.

[63] Millsap, R.E. (2007). Invariance in measurement and prediction revisited. *Psychometrika*. 72(4), 461-473.

[64] Pek, J. & MacCallum, R. C.(2011). Sensitivity Analysis in Structural Equation Models: Cases and Their Influence, [hMultivariate Behavioral Research, 46: 2, 202 228

[65] Raykov, T., & Marcoulides, G. A. (2000). *A first course in structural equation modeling*. Mahwah, N.J.: Lawrence Erlbaum Associates.

[66] Raykov, T., & Marcoulides, G. A. (2008). *Introduction to applied multivariate analysis*. New York: Routledge.

[67] Raykov, T.,& Penev, S. (2001). The problem of equivalent structural equation models: An individual residual perspective. *New developments and techniques in structural equation modeling*, 297-321.

[68] Raykov, T., & Penev, S. (2001). The problem of equivalent structural equation models: An individual residual perspective. *New developments and techniques in structural equation modeling*, 297-321.

[69] Raykov, T., & Penev, S. (2001). The problem of equivalent structural models: An individual residual perspective. *New developments and techniques in structural equation modeling*, 297-321.

[70] Raykov, T., & Penev, S. (2002). Exploring Structural Equation Model Misspecifications via Latent Individual Residuals. *Latent variable and latent structure models*, 121.

[71] Raykov, T. a. P., S. (1999). On Structural Equation Model Equivalence. *Multivariate Behavioral Research*, 2(34), 199-244.

[72] Raykov, T. & Penev, S. (Ed.). (2001). The Problem of Equivalent Structural Equation Models: An Individual Residual Perspective. Mahwah, N.J.: Lawrence Erlbaum.

[73] Raykov, T. & Penev, S. (Ed.). (2002). Exploring structural equation model misspecifications via latent individual residuals (Vol. Latent Variables and Latent Structure Models). Mawah, N.J.: Lawrence Earlbaum.

[74] Reise, S. P., & Widaman, K. F. (1999). Assessing the fit of measurement models at the individual level: A comparison of item response theory and covariance structure approaches. *Psychological Methods*, 4, 3-21.

[75] Rousseeuw, P. J. (1997). Introduction to positive-breakdown methods. *Handbook of statistics*, 15, 101-121.

[76] Rousseeuw, P. J., Leroy, A. M., & Wiley, J. (1987). Robust regression and outlier detection (Vol. 3): Wiley Online Library.

[77] Rousseeuw, P. J., & Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3), 212-223.

[78] Roussos, L., Stout, W., & Marden, J. (1998). Using new proximity measures with hierarchical cluster analysis to detect multidimensionality. *Journal of Educational Measurement*, 35(1), 1-30.

[79] Salvador, S., & Chan, P. (2004). Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms.

[80] Sanchez, B. N., Houseman, E.A. and Ryan, L.M. (2009). Residual-Based Diagnostics for Structural Equation Models. *Biometrics*, 65, 104-115.

[81] Scherbaum, C. A., Cohen-Charash, Y., and Kern, M.J. (2006). Measuring General Self-Efficacy: A Comparison of Three Measures Using Item Response Theory.*Educational and Psychological Measurement*, 66.

[82] Searle, S. R. (1982). *Matrix Algebra Useful for Statistics* (Wiley Series in Probability and Statistics).

[83] Shy-Modjeska, J. S., Riviere, J. E., & Rawlings, J. O. (1984). Application of biplot methods to the multivariate analysis of toxicological and pharmacokinetic data. *Toxicology and applied pharmacology*, 72(1), 91-101.

[84] Steinley, D., & McDonald, R. (2007). Examining factor score distributions to determine the nature of latent spaces. *Multivariate Behavioral Research*, 42(1), 133-156.

[85] Stevens, J. (2002). *Applied Multivariate Statistics for the Social Sciences*, Lawrence Erlbaum.

[86] Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, 52(4), 589-617.

[87] Stout, W. (2002). Psychometrics: From practice to theory and back. *Psychometrika*, 67(4), 485-518.

[88] Sugawara, H. M. a. M., R.C. (1993). Effect of Estimation Method on Incremental Fit Indexes for Covariance Structure Models. *Applied Psychological Measurement*, 17, 365-377.

[89] Viswanathan, M. (2005). *Measurement error and research design*: Sage Publications, Inc.

[90] Waller, N. G., Meehl, P.E., Yonce, L.G & Grove, W.M. (Ed.). (2006). *A Paul Meehl reader: Essays on the practice of scientific psychology*, Lawrence Erlbaum.

[91] Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American psychologist*, 54, 594-604.

[92] Zijlstra, W., Van Der Ark, L., & Sijtsma, K. (2007). Outlier detection in test and questionnaire data. *Multivariate Behavioral Research*, 42(3), 531-555.

[93] Zhong, X. Yuan, K.H. (2011). Bias and Efficiency in Structural Equation Modelling: Maximum Likelihood Versus Robust Method. *Multivariate Behavior Research*, 46(2), 229-265.