A LIKELIHOOD-BASED APPROACH TO DETECTING ABERRANT INDIVIDUALS IN CONFIRMATORY FACTOR ANALYTIC MODELS

John Sideris

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Psychology.

Chapel Hill
2006

Approved by
Advisor: A. T. Panter
Reader: Daniel J. Bauer
Reader: Margaret R. Burchinal
Reader: Robert C. MacCallum
Reader: David Thissen

ABSTRACT
JOHN SIDERIS: A Likelihood-Based Approach to Detecting Aberrant Iindividuals in Confirmatory Factor Analytic Models
(Under the direction of A. T. Panter.)

The study presented here was intended to develop and provide a relatively simple method for detecting aberrant observations in confirmatory factor analysis (CFA). This method exploited a by-product of Full Information Maximum Likelihood (FIML) estimation of these models, the log-likelihood produced for each individual observation. This score, after adjusting for missing data, indexed the degree to which a model fits for a specific individual. A simulation study was run to test this index, labelled adj_lli. Data were simulated under varying levels of covariance structure, proportion of aberrant data, and proportion of missing data. Each cell had 200 samples with n = 200. Additionally, adj_lli was compared to three existing methods: Reise and Widaman's (1999) INDCHI, Yung's (1997) method for detecting outliers in mixture models, and Bollen's A, a general multivariate method (1987). Results indicated that adj_lli was effective in detecting outliers and offered some advantages over three other methods.

ACKNOWLEDGEMENTS

I am grateful to many people for their contributions to this document and to my training. First and foremost is my advisor, Abigail T. Panter. Not only did she provide me with statistical and academic guidance, but her unfailing optimism and belief in me were an inspiration. I am extremely grateful to the readers on my committee. Their contributions to my education go well beyond the work presented in this document. Daniel J. Bauer is a prime example for new faculty, both in the contributions he makes in research and in his care a teacher. Margaret R. Burchinal, as a reader on this project and as my employer, provides me with an outstanding illustration of how to be an applied statistician. Robert C. MacCallum exemplifies both the careful, thoughtful scientist and the patient teacher. David Thissen's clarity of thought and ability to see the meaning in the numbers set standards that should be the goal of any scientist.

Thanks to Kelly Sullivan for her careful read of the final version of this paper. To all of my fellow Zeta members, I am grateful for your friendship and support. I am especially fortunate to have had the friendship of R. J. Wirth, Kristopher Preacher, and Shana Judge. And, most of all, I must recognize the fearsome Michael Edwards, Ariane Kawata, and Viji Sathy. I can not imagine better friends than these.

Finally, thanks go to my parents for their constant and tireless support and belief. Special thanks to Grace and Will for giving me reason enough to get through all of this.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I.

INTRODUCTION

> *"Whoever knows the ways of nature will more easily notice her deviations; and, on the other hand, whoever knows her deviations will more accurately describe her ways."* (Francis Bacon, 1620/1994).

Scientists have long been interested in uncovering and studying unusual observations. Attention to unique data points both may provide insights about attributes of data being analyzed and may suggest new theoretical possibilities (e.g., Behrens, 1997; Billor, Hadi, & Velleman, 2000; Hays, 1994). The detection of aberrant observations remains of great interest in many disciplines. Examples both of analytic techniques and of their application are found in a broad range of fields such as geology (Velasco, Verma, & Guevara, 1999), business (Conklin, 2003), law (Basmann, 2003) and computer science (Hodge & Austin, 2004). In psychology and education, identifying and studying aberrance is often of paramount importance. Clinical psychology, for example, is often concerned with diagnosing and classifying relatively rare individuals with a specific complex of behaviors. Assessment scales often provide cut-off scores, beyond which the individual is considered extreme enough to warrant a special classification (e.g., Lyons & Scotti, 1994; Matthey & Petrovski, 2002; Sheeran & Zimmerman, 2002).

A variety of labels have been applied both to the aberrant data themselves and the methods used to assess them. The text by Barnett and Lewis (1994) is dedicated to assessing, understanding and managing aberrant data and is perhaps the most widely-cited

reference. They prefer the term "outlier," which they defined as "an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data" (p. 20). This statement serves very well as a general definition of aberrant data. The operational definition of what is meant by "inconsistent" will depend on the context in which it is applied. While there are guidelines for classifying an observation as an outlier, ultimately, the decision is left to the individual researcher (Barnet & Lewis, 1994; Chatfield, 2002), based on the specific attributes of the data and the research question.

In addition to theoretical need to identify distinctive observations, there has been considerable research describing the statistical consequences of outliers. This research has largely been concerned with identifying "influential" observations (e.g., Bollen & Jackman, 1985; Fox, 1991). The term "influential" implies more than just the inconsistency indicated by "outlier." Influential data points have a disproportionate effect on the data distribution. In the univariate case, they may simply be extreme cases that skew the data. Their impact on multivariate data may be subtler. Gnanadesikan and Kettenring (1972, p.83) s tate the difficulty colorfully: "a single univariate outlier may be typically thought of as 'the one that sticks out at the end', but no such simple idea suffices in higher dimensions." Numerous techniques have been developed to assess multivariate outliers (for an overview and history, see Barnett & Lewis, 1994; Hawkins, 1980). The consequences of aberrant observations depend too on the analytic context. For example, in regression models, small numbers of cases or even individual cases that are highly discrepant may exert undo influence on parameter estimates (e.g., Belsley, Kuh & Welch, 1980; Bollen, 1989; Bollen & Jackman, 1985; Fox, 1991).

One area of research in which there has been notably little research on aberrant observations is covariance structure analysis (CSA). The lack of attention may stem, in part, from the fact that models are typically based on the covariance or correlation matrix, not on the individual observations. As Bollen (1989, p. 1) points out, this emphasis on covariances requires a "reorientation" for researchers who are accustomed to thinking about individual observations. In a regression equation, for example, the parameters are used to estimate an expected value of some outcome variable given some set of predictor variables. For any individual observation, a predicted value for the outcome variable can be generated. The model itself is evaluated based on the discrepancies or residuals between the expected values and the measured values of the outcome variable over all of the individual observations. A model that minimizes residuals is considered to fit well.

In CSA, however, residuals do not refer to discrepancies at the level of the individual observation. Instead, models are evaluated on the difference between the covariance structure implied by the model and the sample covariances. The difference between the covariance matrix suggested by the model and the observed covariance matrix is the basis of most model fit analyses in CSA (e.g., Amemiya & Anderson, 1990; Browne, 1984; for reviews and discussion see Gerbing & Anderson, 1993; Yuan, 2005).

CSA models assume homogeneity in the population from which the data are sampled (Ansari, Jedidi, & Jagpal, 2000; Bollen, 1989; Yuan, Chan, & Bentler, 2000). In the context of confirmatory factor analysis (CFA) and structural equation models (SEM), if distinct sub-samples are present in a data set or hypothesized in advance, multiple group analyses (e.g., Jöreskog & Sörbom, 1996) manage mixtures well. Models can be developed

and tested to compare the groups on any aspect of the model (e.g., differences between the covariance matrix, the factor loadings, latent means, and so on). However, when the number and composition of distinct samples are unknown model fitting is difficult. The potential for the presence of unexpected and distinctive observations in a study sample has spawned two classes of research. The first, which includes the method proposed in the current paper, focuses on identifying small numbers of aberrant observations (e.g., Riese & Widaman, 1999; Yung, 1997). The second class of research, mixture models, grew out of a need to understand and manage nonnormal data (e.g., Blåfield, 1980). The literature on mixture models has focused on determining whether or not different subsamples exist in a dataset and on estimating the distinct covariance structure of each subset (e.g., Day, 1969; Muthén, 1989; Titterington, Smith & Makov, 1985).

Mixture Models

As the number of aberrant responses increases, the term 'outlier' becomes less apt. Mixture models have been developed to handle the presence of distinct subsets of observations in a sample. When the groups in a data set are unknown, the apparent shape of the distribution is affected (Blåfield, 1980). For example, a sample constituted of two, normally distributed samples whose means are sufficiently separated will appear bimodal. Fitting a model to these data, even if one takes the non-normality into account, will lead one to incorrect theoretical conclusions.

Generally speaking, mixture models have been applied in two contexts. In both cases, researchers are faced with nonnormal data. The first, sometimes referred to as "direct modeling" (Dolan & Van der Maas, 1999) is applied when there is reason to believe that

there are several distinct populations represented in a given sample (e.g., Muthén & Shedden, 1999; Day, 1969). Data mixing from these populations results in the nonnormal shape of the sample distribution. Direct models attempt to uncover the number and members of the separate underlying distributions.

Indirect models (Dolan & van der Maas, 1998) use finite mixture models as a tool to estimate models for data with intractable distributions (e.g., Bauer & Curran, 2003; McLachlan & Peel, 2000). In these cases, the sample represents a single, non-normal population. Fitting a single model to non-normal distributions can be difficult. Instead, separate models are fit to subsets of the data, just as in a direct model. These distinct models are then aggregated into a model of the whole distribution.

There is a large literature exploring methods to determine whether or not different samples exist (e.g., Arminger, Stein & Wittenberg, 1999; Biernacki, Celeux, & Govaert, 1999; Blåfield, 1980; Lo, Mendell, & Rubin, 2001; McLachlan, 1987). Unfortunately, because nonnormal data may look like a mixture of normal distributions, the distinction between indirect and direct models is theoretical, rather than statistical (Bauer & Curran, 2003). Therefore, even if the results of fitting a mixture model suggest multiple groups within the data, the possibility that the data are simply nonnormal remains.

If identified latent classes do indeed exist, their presence can cause a number of statistical problems if they are not identified and modeled. Muthén (1989) provides a variety of cautionary examples where the failure to recognize the presence of heterogeneity will lead to undesirable consequences.

Impact of Outliers on Model Estimation in CSA

The degree to which small numbers of aberrant observations affect model estimation is an open question, and there has been some research thathas begun to address this issue. The effects of outliers on correlation coefficients and covariances are well documented (e.g., Anscome, 1973; Fox, 1991). Bollen (1987) points out the presence of outliers in SEM can lead to "improper solutions," models where parameter estimates are outside of the possible range in the population.

Yuan and Bentler (2001) argue convincingly that aberrant responses can have negative consequences in CFA. They demonstrate analytically and empirically that even a relatively small number of outliers can bias parameter estimates and their associated test statistics. Specifically, the presence of outliers can inflate the noncentrality parameter leading to an exaggeration in the power to reject a model. This inferential decision, of course, can result in the discarding of valid models. The authors provide evidence that these distortions can occur under both maximum likelihood and Browne's (1982, 1984) asymptotically distribution-free procedure.

Detecting Aberrant Observations in CSA

Existing multivariate outlier techniques can be broadly classified into model-free and model-based methods. Model-free methods do not rely on model specification to detect outliers. While such techniques may account for the interdependence among the measured variables in their screening, they do not relate the distinctiveness of a given observation to theoretical model being tested. Model-free methods are typically applied to screen for unusual cases prior to analysis (e.g., Bollen, 1989) and are broadly applicable to virtually

any multivariate data.

In contrast, model-based methods begin with some hypothetical model for the data. Assuming that the model is a good approximation to the process underlying the majority of the data, aberrant observations are those with a distinctive underlying process. Regression diagnostics fit into this class of outlier assessment (for examples, see Belsley et al., 1980; Barnett & Lewis, 1994; Fox, 1991).

The history of both model-free and model-based methods of uncovering aberrant observations in SEM is relatively short. Bollen and Arminger (1991) argue that this short history may be an unfortunate consequence of the nature of the methods themselves. Because the focus is generally on latent variables and the covariances, computing an individual residual is not as straightforward as it is in regression. Once the covariances are computed, "analysts tend to forget about the specific observations that led to them" (Bollen & Arminger, 1991, p. 236).

Model-Free Methods

Bollen (1987) provides a simple, effective model-free method to identify an outlying observation that a significant impact on a factor analytic model. The procedure uses an N (sample size) $\times$ q (number of variables) matrix $\mathbf{Z}$ where each scalar is a deviation score from the mean for that variable. $\mathbf{Z}$ is then used to compute matrix $\mathbf{A}$:

$$\mathbf{A} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' \quad (1)$$

The diagonal of $\mathbf{A}$, $\boldsymbol{a_{ii}}$, contains a measure, ranging from 0 to 1, expressing the distance of each observation from the multivariate mean of the other observations in the

dataset. Further, $\sum_{i=1}^{N} a_{ii} = q$, so the mean of the vector $\boldsymbol{a_{ii}}$ can be computed as q/N. The relative size of each can also be assessed through examination of the univariate distribution of $\boldsymbol{a}_{ii}$.

The Mahalanobis Distance (*MD*) isfrequently used to identify aberrant observations (e.g., Bacon, 1995; Comrey, 1985; Gnanadesikan & Kettenring, 1972; Hardin & Rocke, 2002). *MD* measures the multivariate distance of each point from the centroid and is computed as:

$$D_i = \sqrt{(x_i - \bar{\mathbf{x}})' \mathbf{S}^{-1} (x_i - \bar{\mathbf{x}})} \tag{2}$$

where $\mathbf{S}^{-1}$ is the inverse of the sample covariance matrix and $\bar{\mathbf{x}}$ is the sample mean vector.

Typically, the squared distance (MD^2) is used in assessing the degree to which a given observation can be considered an outlier. The size of the distance with regard to an individual observation can be evaluated either statistically (e.g., Penny, 1996; Rao, 1973, p. 570; Rasmussen, 1988) or graphically (e.g., De Maesschalck, Jouan-Rimbaud & Massart, 2000; Kim, 2000). In the two-variable case, these plots are bivariate plots of the *MD*s for each individual. As one moves to designs with more variables, the construction of graphs obviously becomes increasingly complex and impractical.

Model-Based Methods

Model assessment in CFA and SEM typically has focused on overall model fit and a variety of diagnostic indices have been proposed for this purpose. However, the existing literature on mixture models provides some methods that may be extended to the special case of outlier detection. For example, Blåfield (1980) provided a maximum likelihood

method for simultaneously fitting a CFA model and clustering observations when faced with a mixture of multivariate normal distributions. Yung (1997) offered extensions to this method and suggested that his procedure could be applied to outlier detection in addition to fitting mixture models.

In Yung's (1997) method, model fitting proceeds under the assumption that sample data are drawn from a mixture of populations. For example, a high-performer group and a low-performer group. Distinct CFA models can be fit, even though the population membership of the individual observations is unknown. Fitting the model provides the parameter values for each distinct CFA model as well as the mixing proportions for those models. Yung's method enables the computation of the posterior probability that a given observation belongs to each component sample (e.g., the high-performer group).

Yung (1997) provided examples of his method applied to three different CFA mixture-models: 1) a two-mixture ("high" and "low" groups) unstructured mixed model, 2) a two-mixture mean-shift with different factor matrices, and 3) a two-mixture mean-shift with common factor matrices. All three models were fit under maximum likelihood (ML) estimation. The results of the unstructured model were used to provide starting values for the latter two models. These models were estimated with both expectation-maximization (EM) and approximate scoring (AS) approaches to ML in order to compare these methods. The data analyzed were those used by Sörbom (1974) to illustrate multiple-group factor analysis. The measurement variables were the scores of 145 seventh and eighth grade students on nine achievement tests (Holzinger & Swinford, 1939).

All three models provided similar results. Further, the estimates produced by the

EM and AS methods did not differ appreciably. The mixing proportion for the high group ranged from .08 to .11. Examination of the posterior probabilities classified about 12% of the sample into this group. Unfortunately, because these are empirical, not simulated, data, the true mixing proportions are unknown. This research does provide evidence that model-based clustering is more effective than techniques that cluster in the absence of a model. Further, Yung (1997) notes that the effectiveness of this method increased as the size of the aberrant sample decreased. As such, he argues that this could be applied to outlier detection, but points out that mixture approaches are not widely explored.

The assessment of latent variable models at the individual level has been called "person-fit" in item response theory (IRT) research (e.g., Levine & Drasgow, 1982, Levine & Rubin, 1979, Meier, 2003; Meier & Sijtsma, 2001). In the context of a well-fitting IRT model, the goal of person-fit analysis is to identify those individuals with highly improbable response patterns. This same notion may be appropriate in CFA contexts as well. Aberrant individuals are not simply those who are distinctive; the person who scores highest on a test is distinctive, but not necessarily aberrant. Rather, person-fit refers to the appropriateness of the model for a specific individual.

Reise and Widaman (1999) proposed a method specifically designed to detect aberrant observations. They suggested that individual contributions to the overall model *misfit* could be assessed. Aberrant observations are those individuals with notably higher contributions than others. In CFA fit under maximum-likelihood estimation, chi-square and other tests of model fit are based on the maximum-likelihood fit-function:

$$F_{ML} = \ln|\Sigma^*| - \ln|S| + tr(S\Sigma^*)^{-1} - p \qquad (3)$$

where S is the observed covariance matrix, Σ^* is the covariance matrix as it is reproduced from the parameter estimates, and p is the number of measured variables in the model. Parameter estimates are found that will produce the smallest possible value for F_{ML} (e.g., Bentler & Bonnett, 1980, Bollen, 1989).

The log-likelihood of this function is:

$$LL = -\frac{[(N)F_{ML}]}{2}, \tag{4}$$

where N is sample size. A model is considered "saturated" when it reproduces S perfectly. Model fit can be evaluated by taking –2 times the difference between the LL for a saturated model and the LL for theoretical model being tested. The statistic proposed by Reise and Widaman (1999) is an attempt to make the same comparison at the level of the individual respondent, rather than at the level of the model as a whole. To build their measure, they begin with a measure of the log-likelihood computed at the individual level (Arbuckle, 1996; Muthén, Kaplan & Hollis, 1987):

$$P_{LL} = -\frac{1}{2}[p\ln(2\pi) + \ln|\Sigma^*| + (x_i - \bar{x})\Sigma^{*-1}(x_i - \bar{x})] \tag{5}$$

Taking the sum of the individual P_{LL} values over the entire sample gives the LL for the overall model. The model LL is routinely computed in software making use of Full-Information Maximum Likelihood (FIML) estimation such as LISREL (Jöreskog & Sörbom, 2004) and Mplus (Muthén & Muthén, 2001). One SEM package (Mx; Neale, Boker, Xie & Maes, 2003) provides the individual P_{LL} values. Reise and Widaman (1999) argue that the scaling of this statistic is not readily interpretable. They suggest using their person-fit index, which they call IND_{CHI}, Their index takes –2 times the difference in the

P_{LL} for the saturated and hypothesized model for each individual:

$$IND_{CHI} = -2(P_{LL(saturated)} - P_{LL(hypothesized)}) \quad (6)$$

IND_{CHI} values can be negative or positive. The upper and lower boundaries vary with the scale of the measurement variables. However, individuals with relatively large negative scores are considered misfitting. The authors compare their new statistic with an IRT person-fit measure, Z_l (Drasgow & Levine, 1986, discussed in greater detail in the following section). They apply both Z_l and IND_{CHI} to the same data and flag respondents indicated as misfitting by each statistic.

In their study, Reise and Widaman (1999) used data from 3,245 respondents to three subscales of Tellegen's (1982) Multidimensional Personality Questionnaire (MPQ). All of the response variables were dichotomous. They used multidimensional scaling to sort items within each scale into "content homogeneous facets." Each facet was treated as an item parcel. The items falling into each parcel were summed together. There were 9, 10, and 10 parcels for the Well-Being, Stress-Reaction, and Traditionalism subscales, respectively. These parcels, in turn, were treated as Likert-type items, each with three- or four-response categories.

Samejima's (1969) graded response model (GRM), an IRT model designed for modeling items with ordered-response categories, was fit to each of the three MPQ subscales. They computed two IRT-based measures of person-fit for each respondent on all three subscales. The first, which they refer to as LL_{IRT}, was simply the log-likelihood for each response pattern (see equation 7). The other was Z_l, the previously mentioned IRT person-fit measure.

$$LL(\hat{\theta}) = \sum_{j=1}^{J}[u_{ij} \times \ln(P_{ij}(\hat{\theta})) + (1 - u_{ij}) \times \ln(Q_{ij}(\hat{\theta}_i))\} \quad (7)$$

Three separate CFA models were fit for each of the three MPQ scales. The authors used Mx (Neale, et al., 2003) to obtain the individual log-likelihood values from each of these models. The first model was a saturated model of the factor structure for the subscale. Second, they fit a one-factor "substantive model" where all factor loadings are estimated. In this model the variance of the latent variable (Φ) was fixed at 1.0. Finally, they fit a null model by setting all factor loadings to zero. Each of these models produced individual *LL* values: LL_{SAT} for the saturated model, LL_{1FAC} for the one factor model, and LL_{NULL} for the null model. Using this notation, equation 6 can be rewritten as:

$$IND_{CHI} = -2(LL_{SAT} - LL_{1FAC}) \quad (8)$$

The first evaluative step the authors took was to correlate the raw log-likelihoods produced in their computations with each other and with the raw log-likelihood from the IRT analyses (LL_{IRT}, LL_{SAT}, LL_{1FAC}, and LL_{NULL}). All four displayed moderate to high correlations with each other and with θ (see Table 1). The correlations with θ were taken as evidence for the need for adjusting LL_{1FAC}, claiming that as underlying latent variable increases so does the log-likelihood. They speculate that extreme LL_{1FAC} values could be reflective of high scores on the measure, not of aberrance. However, the authors may be premature in drawing this conclusion. First, one should note that the pattern of correlations with the latent trait varies across the three scales. All of the LL_{1FAC} -trait correlations for the Stress-Reaction Scale were only moderate and all were negative. Next, even if all the correlations between LL_{1FAC} values and trait scores were strong and positive, this does not

necessarily imply that the most extreme LL_{1FAC} scores are simply the most extreme trait scores. Correlations are indicators of a general linear trend across the data set as a whole and do not address the location of specific data points. A simple comparison of the observations with the most extreme LL_{1FAC} scores and those with the most extreme latent trait scores could have more adequately addressed this concern.

Assuming that the LL_{1FAC} needs to be transformed, however, the more crucial comparisons, is between the actual person-fit indices. The standardized IRT fit-statistic, Z_l, and IND_{CHI} had low correlations with each other and with θ. The low correlations with the latent variable are encouraging. The lack of dependence between the person-fit measures is troubling. If both were supposed to flag observations whose response patterns do not fit with the model, a strong correspondence between the two would support their use.

Looking at the correlations between Z_l and IND_{CHI} (see Table 1) across the three subscales reveals that they were negative and low, -.17, -.18, and -.20 for Well-Being, Stress-Reduction and Traditionalism, respectively. Because high negative values for Z_l indicate poor fit, while high positive IND_{CHI} scores indicate poor fit, the negative relationship is expected. However, the low magnitude of the correlations implies a lack of correspondence between the measures. Further, as seen in Figure 1 (Reise & Widaman, 1999, Figure 9), the variance of IND_{CHI} is dependent on the value of Z_l. Notably, the variance of IND_{CHI} increases as Z_l becomes more negative. As such, among those observations identified as aberrant by Z_l, are both observations as most *and* least aberrant by IND_{CHI}.

This lack of correspondence is further demonstrated by comparing the rank orders

of the 3,245 examinees on both Z_l and IND_{CHI} on the Stress Reaction subscale of the MPQ. The authors then compared the 25 respondents identified by each statistic as the poorest fitting. The worst fitting respondents as indicated by Z_l did tend to have positive values of IND_{CHI}. However, only a small proportion of those 25 identified as aberrant by Z_l had a chance at being classified as aberrant by IND_{CHI} (mean = .282, SD = 1.178); only 16 fell more than a standard deviation above the mean of IND_{CHI} and 9 were more than two standard deviations above the mean. Conversely, all but one of the 25 observations classified as aberrant by IND_{CHI} fell more than a standard deviation below the mean of Z_l (mean = .365, SD = .776), and 14 were two standard deviations below. Finally, the lists have only 7 respondents in common.

The authors conclude that the two methods produce "similar, but certainly not identical judgments regarding fit at the extremes of either index" (Reise & Widaman, 1999, p 18). They go on to note a key difference in the computation of these two statistics that likely accounts for this difference. Z_l is conditional on the examinee's latent trait score, while IND_{CHI} is not. As such, the former, because it is conditional on the latent trait, identifies those whose response patterns are inconsistent with their latent trait scores. The latter identifies respondents whose response patterns deviate from the item means. Given this difference, the lack of correspondence may be more due to a difference in how each defines aberrance rather than an indication than either is misclassifying observations.

A Likelihood-Based Method for Detecting Aberrant Observations in IRT

There is a large IRT literature regarding person-fit (see Meijer & Sijtsma, 2001 for a review). The statistic used as the comparison by Reise and Widaman (1999) above, Z_l[1] (Drasgow, 1982; Drasgow, Levine & Williams, 1985), is particularly relevant here. Person-fit measures in IRT typically identify response score patterns that violate the Guttman (1950) model. If the items on a test or scale are ordered with regard to difficulty, once a person reaches an item that he or she is unable to answer, the remaining items should be beyond his or her

ability[2]. That is, the respondent is expected to fail items that fall above his or her ability and is expected to pass items that fall below it.

There were respondents who violated this expectation. IRT scoring is based on composite probabilities, not sums of individual responses, so minor deviations from the expected response pattern (e.g., a low-ability testee responding correctly to a very difficult item) are not causes for concern. Levine and Rubin (1979) argued that some respondents to a test might be so distinct from other examinees that their score is an inappropriate measure of their ability. They suggested that "appropriateness" could be measured at the individual level. They go on to describe three general classes of appropriateness measures: (1) marginal probability of the response pattern given θ; (2) the ratio of the likelihood of the examinees response pattern with θ held constant over all items or where θ is allowed to

[1] Note that this statistic is sometimes referred to l_z in the literature.

[2] IRT is often applied to educational tests. In this context the meaning of "difficulty" is relatively clear. In other psychological contexts (e.g., self-esteem research) this term may be less appropriate. However, one can assume that in these contexts difficulty reflects a higher level of the trait under examination. Thus, an item of high difficulty on a self-esteem measure is one whose endorsement indicates high self-esteem.

vary over items; and (3) estimated ability variation over the range of items. Most person-fit statistics fall into the first of these categories (Meijer, 2003). The conditional probability of a given response pattern is:

$$L\left(\hat{\theta}\right) = \Pi_{j=1}^{k} p_j\left(\hat{\theta}_j\right)^{u_j}\left[1 - p_j\left(\hat{\theta}_j\right)\right]^{(1-u_j)}, \tag{9}$$

where $L\left(\hat{\theta}\right)$ is the likelihood of the response vector $\boldsymbol{u}_j$ (that is, the observed responses of individual *i* to the 1 through k^{th} items), $p_j\left(\hat{\theta}_j\right)$ is the probability of individual *i* endorsing item *j* and $1 - p_j\left(\hat{\theta}_j\right)$ is the probability of individual *i* not endorsing item *j*. Because the log of the likelihood was computationally simpler, the foundational literature suggested its use (Levine & Rubin, 1979; Levine & Drasgow, 1982). The log-likelihood, *l*, is computed:

$$l = \sum_{j=1}^{k}\left\{X_j \times \ln P_j\left(\hat{\theta}\right) + \left(1 - X_j\right) \times \ln\left[1 - P_j\left(\hat{\theta}\right)\right]\right\} \tag{10}$$

Drasgow (1982) noted that respondents in the extremes of the distribution of θ have response patterns that are unexpected, not because they violate the model, but simply because they are rare. As such, *l* is confounded with trait level, making it less desirable as a measure of person-fit. To correct for this, he and his colleagues (Drasgow et al., 1985; Drasgow & Levine, 1986) proposed a transformation of *l* that is standardized conditionally on θ:

$$Z_l = \frac{E(l)}{\sqrt{\text{var}(l)}} \tag{11}$$

where the expectation of *l* is:

$$E(l) = \sum_{j=1}^{k} \left\{ P_j(\theta) \ln\left[P_j(\theta)\right] + \left[1 - P_j(\theta)\right] \ln\left[1 - P_j(\theta)\right] \right\} \tag{12}$$

and the variance is:

$$\text{var}(l) = \sum P_j(\theta)\left[1 - P_j(\theta)\right] \left[\frac{P_j}{1 - P_j(\theta)} \right]^2 . \tag{13}$$

Large negative values indicate poor person-fit. However, more recent research (Nering, 1995; Reise, 1995) has demonstrated via simulation, that when θ is estimated (as it must be in any practical situation), the distribution is not normal. Specifically, the probabilities in the tails are higher, resulting in greater misclassification of observations as poorly fitting. Snijders (2001) provides methods for deriving an asymptotic sampling distribution for l_z and similar person-fit statistics.

Full-Information Maximum-Likelihood

Full-Information Maximum-Likelihood (FIML) estimation has become increasingly popular in SEM research, due primarily to its capacity to handle missing data (Enders & Bandalos, 2001). FIML provides, as part of its computation, information that could be exploited to make assessment of model-fit at the level of the individual observation. The discrepancy function in FIML is computed simply as the sum of the log-likelihoods from each observation. Thus, an individual with a high value has a disproportionate impact on the discrepancy function

The discrepancy function for FIML is calculated as (Arbuckle, 1996; Finkbeiner, 1979):

$$\log L(\mu, \Sigma) = \sum_{i=1}^{N} \log L_i \tag{14}$$

where μ is a vector of item means and Σ is the estimated variance/covariance matrix. Log L_i is calculated:

$$\log L_i = -\frac{1}{2}[p_i \log(2\pi) + \log|\Sigma_i| + (x_i - \mu)'\Sigma_i^{-1}(x_i - \mu)] \tag{15}$$

where p_i is the number of non-missing responses for observation i, x_i is a vector of non-missing data for observation i, and Σ_i is a the estimated variance-covariance matrix for respondents with the same missing data pattern as observation i. The first half of equation 15 is a constant for all respondents with the same missing data pattern. The variance of the individual log-likelihoods within each pattern arises out of the second half of the equation, which contains deviations between the observed variables and their means. Aberrant response patterns should have higher deviations from the item means. Large negative values imply that the specific response pattern is unlikely, given the model. The presence of these unlikely patterns results in a more extreme model discrepancy function (or log-likelihood function for the restricted model). As this function becomes more negative, the difference between it and the function for unrestricted models becomes greater.

Previous research that used the log L_i as a basis for person-fit computed a measure that was the difference between log L_i for theoretical model and for the saturated model (Reise & Widaman, 1999). However, this step seems unnecessary. The sum produced by Equation 15 grows larger as model fit decreases. Observations who particularly contribute to this sum (i.e., the most extreme observation), should be those who do not conform to the

model. Singling out observations with noticeably high log L_i values should be a useful method for aberrant observation.

The presence of missing data, however, provides a computational confound to using log L_i to identify aberrant data. Looking again at equation 15, note log L_i is weighted by the number of non-missing items (p_i). As such, increases in the amount of missing data should decrease the value of log L_i. To adjust for this weighting, each individual's score was adjusted by dividing it by p_i. This adjusted statistic is referred to as adj_ll_i.

As outlined above, methods have been suggested for outlier detection in CFA and related CSA models. However, these techniques have failed to be widely implemented outside of the quantitative and statistical literature. There remains a need for a reliable, easy to implement method for assessing person fit with reference to a specific model.

The Current Research

The primary hypothesis under examination is that, in the context of a CFA fit under FIML, highly aberrant respondents will have correspondingly high log-likelihood values (after adjusting for missing data) that can be used to detect them. This possibility has been suggested by others (MacCallum, 2003; Neale et al. 2003), but has not been fully explored. The efficacy of this statistic, adj_ll_i, was tested via a Monte Carlo simulation. Simulation provides a distinct advantage over research using empirical data. Because the aberrant cases are known, the effectiveness of adj_ll_i in finding those cases can be assessed.

Study conditions involved three experimental factors: different sources of aberrance, different levels of missing data, and different proportions of aberrant respondents. Table 2 presents the levels of these conditions.

Experimental Conditions

Missingness

The robustness of adj_ll_i as an aberrant observation detection technique was assessed at three levels of missing data: no missing data, 2% missing, and 10% missing. To meet the assumption that data are Missing Completely at Random (MCAR), data points were randomly deleted from each simulated sample.

Proportion of Aberrant Respondents

The analysis datasets were each composed of a large primary sample and a small aberrant sample. These datasets were fit to the same model used to generate the primary dataset. As such, samples with relatively small numbers of aberrant respondents should still fit the hypothesized model well. As this number increases, however, the discrepancy between the hypothesized model and the total sample should increase. As it does, it may become correspondingly difficult to fit the model and to find the aberrant respondents. Three levels of the proportion of outliers were used in the simulation: no aberrant observations, 2%, and 10% of the total sample.

Source of Aberrance

In his analysis of finite mixtures, Yung (1997) made use of the multiple groups data presented in Holzinger and Swineford (1939). The parameter estimates (using the expectation-maximization algorithm) from this research were used as the basis for the data simulation in the current research. Yung (1997) provided estimates for primary and aberrant observations in two sets of mixture data. These estimates were used in the generating models for the data simulated in this study.

Four sources of aberrant data were explored. The generating models for all conditions are presented in Table 2. The first resulted in aberrant data sets whose covariance structure was the same, but where there is a mean shift in the intercepts of the measurement variables. The second source generated data where the aberrant data had the same means in the measurement variables, but with a difference covariance structure in the latent variables. The third was a combination of the first two, where the aberrant sample is both mean and covariance shifted.

The fourth condition was a variation on a contaminated normal distribution, referred to here as a "halo distribution." The aberrant data was drawn from a population with the same covariance structure and means as the primary dataset. However, the distribution has greater standard deviations on the measurement variables (they were increased by a factor of two) and observations are drawn from the tails ($\pm$ two standard deviations) of the multivariate distribution. Thus, all of the aberrant observations are certain to be significantly distant from the centroid.

CHAPTER II.

METHOD

Procedure

Simulation

Data were simulated for each of the cells described above. There were 200 samples of $n = 200$ within each cell. For each of these samples the primary and aberrant data was generated separately based on the appropriate population parameters and then merged into one set for analysis. Simulation began by specifying the population parameters (see Appendix A). The population covariance matrix (Σ) and mean vector (μ) were computed from these parameters. Observations in each sample were first generated values drawn from a random normal population. The data matrix is then multiplied by Σ (more precisely, the Cholesky decomposition of Σ) and μ is added (see code labeled "Generate Raw Data" on pages 64 and 67 of Appendix B and pages 72 and 74 of Appendix C). Two sets of sample code are provided. Appendix B presents the code for both the simulation and analysis of a mean shift condition, while Appendix C presents a covariance shift condition.

The Halo conditions required additional steps for the aberrant data. First, a large sample ($n = 100{,}000$) was generated. Second, observations who were more than two standard deviations from the centroid, based on Mahalanobis's distance were selected. Finally, random observations from this extreme distribution were selected for inclusion as aberrant observations.

Code for data simulation was written in Gauss version 3.2 (Aptech, 1997). Cai (2005a) has developed analysis software in Gauss that was used to test the models. Although this analysis software is not commercially available, it has been shown to produce results identical to what one would obtain by running the same models in LISREL 8.51 (Jöreskog & Sörbom, 2003) with the same data (Cai, 2005b). Making use of this software allowed the data to be simulated and analyzed in one program, greatly simplifying and speeding up the process. After data are simulated and analyzed, the individual log-likelihood (ll_i) values are output. For each observation, the number of non-missing items is counted. ll_i is divided by this count to produce *adj_lli*.

Computation of Other Aberrance Statistics

Both adj_ll_i and IND_{CHI} (Reise & Widaman, 1999) could easily be computed within the simulation process. The individual log-likelihoods based both on the fitted model (ll_i) and on the saturated model (ll_{sat}) can be requested from the software. IND_{CHI} is computed for each observation by taking the difference between these two values.

Bollen's *A* required the matrix arithmetic described in Equation 1. Unfortunately, because these operations will not allow for missing data, all observations with missing data were dropped. *A* was computed for the subset of the observations with complete data.

Yung (2006) provided computer code for implementing his method via SAS "proc IML" (SAS Institute, 2004). This code requires the user to input starting values for model parameters for both mixture components expected in the data. In these data, the two components are the primary and aberrant samples. Further, the expected mixing proportions

are input into the code. The generating values for the simulation were used as starting values for the parameter estimates. Because the proportion of aberrant and primary respondents were known, these proportions were entered as the mixing proportions.

CHAPTER III

RESULTS

Results indicated that adj_ll_i was able to identify aberrant observations and that its effectiveness may have been affected by the presence of missing data and by the number of aberrant observations. As shown in Table 3, the average adj_ll_i for the aberrant sample was higher than the average for the primary sample across all but the covariance shift only conditions. In general, these differences were at least a standard deviation. The pattern was reversed in three of the four covariance shift only conditions, with the exception of the 2% missing, 10% aberrance cell. Notably, the size of the difference for this cell was comparable to the differences produced by the other sources of aberrance. The three cells where the primary group had higher scores showed primary vs. aberrance differences that were very small.

A regression model was fit that included the four way interaction of group, missingness, proportion of aberrance, and source of aberrance, as well as all lower order interactions. Given the sample size (N = 640,000), it wasn't surprising that all main and interaction effects were statistically significant. However, the model provided a parameter estimate for the group (aberrant vs. primary) effect, controlling for the impact of the experimental conditions. The main effect for group was clear; the aberrant observations had an average adj_ll_i that was .94 points higher than the primary observations.

The interaction effects in the model were all significant. Again, given the large sample size, this does not necessarily imply practically significant results. Figure 2 provides a graphical description of these effects.

The source of aberrance that led to the greatest differences between aberrant and primary groups was the contaminated normal, followed by the combined mean and covariance shift. The smallest differences were seen in the covariance shift only condition.

The effect of missingness on the difference score was trivial across all levels of the other experimental factors. While there did appear to be a general trend for the proportion of aberrance, the effect was tempered by the source of aberrance. Interestingly, the size of the interaction mirrored the main effect for the source of aberrance. The impact of the proportion of aberrance was greatest for the contaminated normal condition. The difference between the primary and aberrant groups was dramatically smaller when 10% of the observations were aberrant than when 2% were. The direction of this effect was the same for the combined mean and covariance shift condition, but size of the effect was notably smaller. The effect is further reduced in the mean shift only condition. Finally, in the covariance shift only condition, the impact of the proportion of aberrance is negligible.

Accuracy of classification

While the specific value of adj_ll_i is of interest, of greater importance is its ability to correctly classify observations as either aberrant or not. Because true group (aberrant or primary) membership is known, it is possible to determine the proportion of observations accurately classified as aberrant (true positive fraction, TPF) and the proportion of primary observations misclassified as aberrant (false positive fraction, FPF). Classification of a given

observation would require the specification of a "cut point;" a value for adj_ll_i beyond which all observations are classified as aberrant. Setting a low value for the cut point will have the desirable effect of increasing the number of observations correctly classified as aberrant (TPF), but also increases the number of primary observations *mis*classified as aberrant (FPF). Conversely, decreasing the FPF will also decrease the TPF. A Receiver Operating Characteristic (ROC) curve plots TPFs as a function of FPFs and so is an efficient method of displaying the cost (i.e., increase in FPF) of higher TPF rates. If group membership (primary vs. aberrant) had no relationship with adj_ll_i, classification would be arbitrary. In this case, the ROC curve would be a diagonal line bisecting the graph. As the strength of the relationship between adj_ll_i and true group membership increases, so does the area of the graph captured by the ROC curve. The area under the curve (AUC) provides a numerical assessment of the efficacy of the classification of observations. Its value ranges from .5 (indicating that observations are essentially classified at random) to 1.0 (indicating perfect classification).

These TPF and FPF values can be easily computed by fitting a logistic regression model with group membership as the dependent variable and one of the above aberrance statistics as the predictor variable. The SAS (SAS Institute, 2004) system procedure "proc logistic" provides both the TPF, the FPF and the AUC. Graphs of the ROC curves from all samples in each cell are presented in Figures 3, 5, 7, and 9 for adj_ll_i, IND_{CHI}, A, and Yung's mixture method respectively. Note that the flat, diagonal line on each graph is provided for reference and is not one of the sample curves.

Given the variability of the curves displayed within some of these cells, the

computation of a single curve for each was desirable. To generate this curve all of the samples in a given cell were combined and a logistic model was fit. These ROC graphs are presented in Figures 4, 6, 8 and 10, again for adj_ll_i, IND_{CHI}, A, and Yung's mixture method, respectively. Means of the AUC values by study condition and aberrance statistic are presented in Figures 11 through 13.

Efficacy of adj_ll_i

Focusing first on the efficacy of adj_ll_i, Figures 3 and 4 imply that under some conditions this technique is extremely effective, while under others its utility is questionable. The AUC indexes how well adj_ll_i classifies aberrant and primary observations where .5 indicates the poorest possible performance and 1.0 indicates perfect performance. A regression model was fit predicting the AUC from all three study conditions. Note that the level of observation is "sample" rather than the individual observations modeled earlier. The model is effective in predicting the AUC, $F(14, 3184) = 529.12$, $p < .05$; $R^2 = .70$. Tests of the individual effects in the model are presented in Table 4. All main effects and interactions were significant predictors of the AUC. Close examination of Figure 3 reveals the pattern of results. Generally speaking, increases in the proportion of aberrant responses decrease the difference between the scores between the primary and aberrant conditions. There were two exceptions to this pattern. At 2% missing, both the covariance shift and mean-plus-covariance shift conditions show increases in the difference score as the percentage of aberrant observations increases. Finally, the percentage of aberrant observations had essentially no impact on the difference score when 15% of the data were missing from the covariance shift condition.

Proportion of aberrant observations, source of aberrance, and their interaction, however, are all significant predictors of the effectiveness of adj_ll_i as measured by the AUC. Figure 11 presents the mean of the AUC for each condition. Above and beyond the interaction between source and proportion of aberrant observation, there is evidence to support the inference of an independent effect for the source of aberrance. This pattern is reminiscent of the pattern seen earlier in model of adj_ll_i (see Figure 2). The predictive power of adj_ll_i was strongest in the contaminated normal conditions. The two means shift conditions followed, with the combined mean and covariance shift condition outperforming the mean shift only condition. The performance of adj_ll_i was poorest in the covariance shift only condition.

It can not be argued that there is an effect for the proportion of aberrance on the AUC that is independent of the effect of source of aberrance. In the contaminated normal condition, there is no effect for the proportion of aberrant respondents. However, increasing the proportion of aberrant responses significantly decreased the AUC in both mean shift conditions. This decrease was the same for both of these conditions. The effect is similar in the covariance shift only condition, although it is less pronounced.

Close examination of the graphs presented in Figure 3 confirms the findings for the statistical tests of the AUC. Recall that each of these graphs contains 200 lines, one for each simulated sample. Across all four of the contaminated normal conditions, there is a strong tendency for the curves toward the outside of the graph. The technique was very effective on almost all samples simulated as contaminated normal distributions.

The graphs for the mean and covariance conditions are also very encouraging with regard to the efficacy of adj_ll_i. The densest portions of the graphs, however, are lower than

in the contaminated normal conditions indicating somewhat decreased efficacy. One unanticipated effect that begins to be apparent here is the increased variance of areas under the curves in the 2% aberrant conditions.

The densest portion of the graph is again lower in the mean shift only condition. Even so, the bulk of the lines are well above the portion of the graph indicating arbitrary classification. The increased variance for the 2% aberrant conditions noted in previous conditions is somewhat more apparent here.

Finally, the graphs confirm the finding that adj_ll_i performs the worst in the covariance only conditions. Not only is the densest portion of the graph the lowest of all conditions, but a sizable number of the lines actually fall below the diagonal reference line. In these cases, the true positive rate (TPF) did not consistently increase with the (FPF). As such, lowering the cut point increased the number of primary observations classified as aberrant, but there was no corresponding benefit of increased accurate classification of aberrant observations.

As for the previous two sources of aberrance, there is clear evidence of increased variance in the curves for the 2% aberrance conditions compared to the 10% aberrant conditions.

Comparison of adj_ll_i With Other Indices

Scores on the three additional indices described earlier were estimated for the simulated data used to evaluate adj_ll_i. ROC curves were drawn and areas under the curve (AUC) were estimated using "proc logistic" in the same way as for adj_ll_i.

Yung's mixture method presented problems for analysis. First, about five percent of

the logistic models used to draw the ROC curves could not be fit. As such, any statistical analyses intended to compare these four indices would have non-randomly missing data. Further, examination of the ROC graphs that could be drawn (Figures 9 and 10) make it clear that this method performed poorly under all conditions. For these reasons, a decision was made to exclude this method from formal statistical comparison.

A mixed model regression was used to compare the three indices on the AUC. Within each experimental condition, the same simulated samples were used to compute the AUC for each index. The repeated use of the same samples created non-independence between observations in the model. Mixed models are effective in managing this non-independence while estimating the effects of interest. In addition to the comparison of the three indices, the four way interaction of index, missingness, proportion of aberrance, and source of aberrance is tested along with all lower order interactions. The tests of the experimental conditions are presented in Table 5. The model provides evidence that these three indices do perform differently. The analysis supported the conclusion of a main effect for index. Estimation of the least-squared means of the AUC for each index yielded values of .84, .83, and .70 for adj_ll_i, IND_{CHI}, and Bollen's A respectively.

Exploration of the interaction effects is done by examination of the AUC graphs (Figures 12, 13, and 14) and the ROC graphs (Figures 3 through 8). Looking first at Figures 12 and 14, the similarities between adj_ll_i and A are remarkable. For both statistics, there is little, if any, effect for missingness. The most notable effect for missingness is for the covariance shift conditions with regard to the change from 2% to 10% aberrant. There is very little difference between the 2% and 10% samples for the 15% missing condition, while there

is a small but evident effect for the change in the 2% missing condition. Other than this, though, the graphs are quite similar.

The graph for IND_{CHI}, however, is very different. The highest AUCs were found for the contaminated normal samples and the lowest for the covariance shifted data, just as was seen for the other indices. The order of the mean shifted samples is reversed for IND_{CHI}, however. The interaction of source with proportion of aberrant observations is very different from adj_ll_i and A. Rather than remaining flat in the contaminated normal sample, the AUC decreases as the proportion increased.

Perhaps the most notable difference is in the mean shift only condition. Unlike any other condition, the AUC increased as the proportion of aberrant responses increased.

Close examination of the graphs of the ROC curves (Figures 3 through 8) sheds further light on the differences and similarities between the three indices. Again, the patterns for adj_ll_i and for Bollen's A are strikingly similar. In each experimental cell, the densest portion of the graph for A is the same as what was seen for adj_ll_i. The increased variance in 2% aberrant cells is just as apparent for this index as it was for adj_ll_i.

The graphs for IND_{CHI} are quite distinctive. The first feature that reveals itself is the relatively large number of lines that drop below the .5 reference line. This seems to be particularly problematic at low false positive fractions. Thus, in order to correctly classify large proportions of the aberrant observations, a relatively large number of non-aberrant observations will be misclassified as aberrant.

While all of the indices had their poorest performance in the covariance shift only conditions, IND_{CHI} demonstrated essentially chance level performance in those conditions. IND_{CHI} underperforms the other two indices in every experimental condition.

CHAPTER IV

DISCUSSION

A method to identify aberrant or outlying observations in CFA models was presented. This method, labeled adj_ll_i, was evaluated under a variety of conditions (different sources of aberrance, different levels of missing data, and different proportions of aberrant respondents) via a Monte Carlo simulation. It was also compared to three other methods for identifying these distinctive observations: Reise and Widaman's (1999) IND_{CHI}, Bollen's A (1987), and Yung's (1997) method for evaluating mixtures. The primary hypothesis under examination was that, in the context of a CFA fit under FIML, highly aberrant respondents will have correspondingly high adj_ll_i and that these values could be used to detect those aberrant respondents.

Receiver Operating Characteristic (ROC) curves and the areas under those curves (AUC) provided the method for assessing the efficacy of adj_ll_i. The curves themselves provide a graphical description of the balance between correctly classifying aberrant observations and misclassifying primary observations as aberrant. The AUC quantifies the ROC and thus was used as the dependent variable in statistical assessment of adj_ll_i and of the factors that may influence it.

The proposed technique was effective in identifying aberrant observations. Its effectiveness, however, changed as a function of study condition. Increasing the proportion of aberrant observations from two to 10% tended to flatten the ROC curves for three of the four sources of aberrance (mean shift only, covariance shift only, and the combination of

these two), implying that *adj_ll*$_i$ became less effective with that increase. The contaminated normal was unexpectedly resilient to increases in the proportion of aberrant observations.

The ROC curves showed more variability in the 2% aberrant conditions than in the 10% conditions. The cause of this increased variability is unclear, although it may be related to the distinctiveness of the aberrant sample. There was the greatest amount of variance in the ROC curves in the covariance only conditions and the least amount in the contaminated normal. Not coincidently, these were also the sources of aberrance where *adj_ll*$_i$ performed the poorest and the best, respectively.

In addition to manipulating the amount of aberrant data, the amount of missing data was varied so that half the samples had 2% missingness and the other half had 15% missingness. This manipulation had very little impact on the results in any condition. *Adj_ll*$_i$ was about as effective in identifying aberrant observations regardless of the amount of missing data.

Perhaps the most important study factor was the source of aberrance. *Adj_ll*$_i$ performed best in contaminated mixture condition. Three explanations for the size of this effect are apparent. First, the simulated observations were drawn from particularly extreme portions of the populations. As such, the effectiveness of *adj_ll*$_i$ in these conditions may be an artifact of the simulation method. Second, because the aberrant observations do not cluster, they do not pull the overall model in a consistent direction. In all of the other source conditions, the aberrant observations are made to cluster together and have a common impact on the centroid of the dataset, moving it closer to the aberrant observations. This may also explain why increasing the number of aberrant observations

from two to 10% had the smallest impact on adj_ll_i in the contaminated normal conditions. Finally, again because the aberrant observations in this condition do not cluster together, they are not only distinct from the primary population, but are also more likely to be distinct from each other.

Adj_ll_i performed poorly in the covariance shift only condition. This poor performance is highlighted by the fact that the primary group had slightly higher adj_ll_i in three of the four cells under this source of aberrance. This is a particularly disappointing result. It was hoped that this method would be sensitive to theoretically distinct outliers. While mean shifts can be theoretically important, the covariance shift gets to the heart of factor analysis. The weak showing of this technique in the covariance shift calls into question its effectiveness as a model based method.

The mean shift only condition provided results that were notably superior to what was seen in the covariance shift only condition. The combined covariance and mean shift condition showed even stronger effects. Much of the variance in the individual log-likelihoods comes from the difference between the means of the measurement variables and the means estimated by the model. As such, the effectiveness in the mean only condition is not surprising. Combining this with the effect of the change in the factor structure increased the effect even more.

Adj_ll_i compared favorably to the other methods under examination. The mixture method suggested by Yung (1997) performed particularly poorly. This should have been anticipated. While Yung (1997) suggested that his method might be adaptable for outlier detection, he also noted theory for doing so had not yet been developed (1997, 2006). His

method is effective for dealing with mixture models, but performs best when the specifics of the distinct models and approximate mixing proportions are known. While these may be reasonable assumptions when fitting mixture models, having this information when searching for smaller numbers of aberrant observations is unlikely.

Adj_ll$_i$ outperformed Reise and Widaman's (1999) IND_{CHI} across all conditions. Both techniques performed best on the contaminated normal samples. However, IND_{CHI} was not distinguishable from chance in the covariance shift only conditions. Similar to *adj_ll*$_i$, IND_{CHI} did better on the mean shifted samples, with higher performance for the combined mean and covariance shift conditions. Note, though, that these are relative statements; even when IND_{CHI} was performing at its best, it was not performing particularly well.

Finally, Bollen's A had nearly identical performance to *adj_ll*$_i$. Not only were both techniques affected in similar ways by the experimental factors in the study, but the proportion of observations that were correctly classified were essentially the same. Given complete data, I would expect these techniques to identify the same observations as aberrant. The effectiveness of Bollen's A is tempered by the fact that it can only be applied on complete data. Despite the apparent effectiveness of A, its inability to manage missing data severely limits its widespread use. *Adj_ll*$_i$ does not share this limitation.

As it was originally conceived, one of the strength's of *adj_ll*$_i$ is that it is model based. As such, it is not just that aberrant observations are distinctive, but that a hypothetical model is a particularly poor fit for those observations. While no model is a perfect fit for any specific observation, *adj_ll*$_i$ was intended to identify those for whom the

model is particularly poor relative to the other observations in the sample. The similarity of the results for adj_ll_i and Bollen's A makes the assertion that this method is model based questionable. However, the models tested here were specified to fit the primary data as they were simulated. Future research on adj_ll_i will need to include experimental factors that manipulate model misspecification.

Another component of this method's strength is that the information required to implement it should be readily available. The usefulness of any technique is limited by its accessibility and ease of use. The method proposed here exploits information computed as a matter of course in models using FIML estimation. In fact, the individual log-likelihood values can already be requested in two widely available software packages, Mx (Neale et al., 2003) and Mplus (Muthen, 2005). Even when using software that does not produce the individual log-likelihoods, computation of the statistic would be relatively simple from the estimated covariance matrix and population means. It should be noted, however, that computing the adj_ll_i will require extra effort beyond what is produced by most software. This necessity will limit the widespread implementation of this method.

The results of the analyses reveal not only some of the strengths and weaknesses of adj_ll_i, but also expose some limitations of the research itself and point to some future directions. Addressing these issues in future research will aid in the understanding of utility of this approach.

First, the research was limited to one type of model, a confirmatory factor analysis. While this is a common analytic framework (a Social Sciences Citation Index (SSCI, Thompson Corporation, 2006) search found over 350 publications using or studying CFA

in the year 2005) it is a constrained example of possible CSA models. Beyond this, the model studied here is a relatively simple CFA. More complex models, both factor analytic and models with structural components, should shed light on the conditions under which adj_ll_i performs well.

The specification of the models provides another source of criticism. Because the hypothesized model was also the generating model for the primary subsample, it was perfectly specified for that subsample. This is a particularly artificial situation; in real-world research situations models are at least slightly misspecified. This decision was made to simplify the research situation and provide a clearer assessment of the efficacy of adj_ll_i. However, in the future consideration should be given to more realistic research conditions.

Beyond the conditions where adj_ll_i can be applied, concern about the utility of the method may still exist. Because the data here were simulated, assessment and comparison of true and false positives was possible. In real research situations, though, more guidance will be needed in making use of adj_ll_i to identify aberrant observations. This guidance may take the form of finding cut-points for dividing the sample into aberrant or not-aberrant observations. More research into the distribution of the statistic will be required to make general statements about determining the cut-point in a given sample.

Perhaps the greatest flaw in planning of this research involves the definition of aberrance. It was hoped that this technique would be a general approach for detecting poorly fitting observations without reference to the cause of that aberrance. In retrospect, however, this is likely a critical consideration. Adj_ll_i was best able to identify those observations whose scores were distant from the multivariate center of the observation.

Attention to the situations which might lead to such outlying observations will be needed before this technique becomes useful in applied research. Future research on this topic should begin with reflection on why aberrant data might exist. From this an inference about how that aberrance will express itself in the data can be made. Finally, based on that inference a technique can be developed to target those data. This is a strategy that has been followed in the IRT literature successfully (see Meijer, 1996, for a review). The results of this research strongly imply that adj_ll_i is an effective method for identifying multivariate outliers. It was, for example, very effective in the halo conditions. Future research on this technique should begin by suggesting research situations which would lead to small numbers of cases with extreme values on the measurement variables. The simulations could then be designed to mimic these situations. This approach would not only shed light on the circumstances where adj_ll_i is likely to be most effective, but could set the stage for a series of CSA person-fit statistics, each tailored to the research situations where one is likely to find specific forms of person-misfit.

Appendix A
Parameters of Generating Models

Factor Loadings
Primary and Aberrant

	K1	K2	K3
x1	3.97	--	--
x2	1.93	--	--
x3	1.24	--	--
x4	--	8.75	--
x5	--	3.57	--
x6	--	3.73	--
x7	--	--	3.77
x8	--	--	0.96
x9	--	--	2.17

Inter-Factor Covariances
Primary

	K1	K2	K3
K1	1	.68	.91
K2	.68	1	.41
K3	.91	.41	1

Inter-Factor Covariances
Aberrant

	K1	K2	K3
K1	4.96	.95	3.17
K2	.95	1.47	.23
K3	3.17	.23	2.14

Theta-Delta (diagonal elements)
Both Primary and Aberrant
26, 15, 6, 51, 7, 12, 26, 11, 12,

Means of the Measurement Variables:
Primary: 29, 25, 14, 44, 19, 28, 103, 6, 9
Aberrant: 31, 24, 16, 51, 22, 33, 110, 17, 15

Appendix B
Gauss Code for Simulating and Analyzing Data
Mean Shift Only, 2% Missing, 2% Aberrant

Note: The "include#" statements call the analysis software developed by Cai (2005a)

```
#include "C:\Gauss3\SEM\LISREL1.txt";
#include "C:\Gauss3\MI\MI.txt";


totn = 200;
for r (1,totn,1);
print "RUN" r;
/* Cell 2, Mean Shift, 2% Missing, 2% Aberrant*/
/* Setup a cfa model for the primary group*/
_LA = {
   . 0 0,
   . 0 0,
   . 0 0,
   0 . 0,
   0 . 0,
   0 . 0,
   0 0 .,
   0 0 .,
   0 0 .};

_BE = {
   0 0 0,
   0 0 0,
   0 0 0};

_PS = {
   1 . .,
   . 1 .,
   . . 1};

_PH = {
   .,
   .,
   .,
   .,
   .,
   .,
   .,
   .,
   .};
_PH = dg(_PH);

_AL = {0,
   0,
   0};
```

```
_KA = {.,
        .,
        .,
        .,
        .,
        .,
        .,
        .,
        .};

LA = {
   1 0 0,
   2 0 0,
   3 0 0,
   0 4 0,
   0 5 0,
   0 6 0,
   0 0 7,
   0 0 8,
   0 0 9};

BE = {
   0 0 0,
   0 0 0,
   0 0 0};

PS = {
   1 10 11,
   10 1 12,
   11 12 1};

PH = {
   13,
   14,
   15,
   16,
   17,
   18,
   19,
   20,
   21};
PH = dg(PH);

AL = {
   0,
   0,
   0};

 KA = {22,
        23,
        24,
        25,
        26,
```

```
        27,
        28,
        29,
        30};

theta = {3.97,
          1.93,
          1.24,
          8.75,
          3.57,
          3.73,
          3.77,
          0.96,
          2.17,
          0.68,
          0.91,
          0.41,
          26, 15, 6, 51, 7, 12, 26, 11, 12,
         29,
         25,
         14,
         44,
         19,
         28,
        103,
          6,
          9} ;

{Lambda, Beta, Psi, Phi, Alpha, Kappa} =
SetModel(theta,LA,BE,PS,PH,AL,KA,_LA,_BE,_PS,_PH,_AL,_KA);

p = rows(Lambda);
q = cols(Lambda);
A = inv(eye(q)-Beta);
mu1 = Lambda*A*Alpha+Kappa;
Sigma1 = Lambda*A*Psi*A'*Lambda'+Phi;

/* Generate Raw Data */
n1 = 196;
Y1 = rndn(n1,p)*chol(Sigma1)+mu1';
print "PSI Ab" psi;

/*   Y1;*/

/* Setup a cfa model for the aberrant group*/
_LA = {
   . 0 0,
   . 0 0,
   . 0 0,
   0 . 0,
   0 . 0,
   0 . 0,
   0 0 .,
```

```
   0 0 .,
   0 0 .};

_BE = {
   0 0 0,
   0 0 0,
   0 0 0};

_PS = {
   1 . .,
   . 1 .,
   . . 1};

_PH = {
   .,
   .,
   .,
   .,
   .,
   .,
   .,
   .,
   .};
_PH = dg(_PH);

_AL = {0,
   0,
   0};

_KA = {.,
          .,
          .,
          .,
          .,
          .,
          .,
          .};

LA = {
   1 0 0,
   2 0 0,
   3 0 0,
   0 4 0,
   0 5 0,
   0 6 0,
   0 0 7,
   0 0 8,
   0 0 9};

BE = {
   0 0 0,
   0 0 0,
```

```
   0 0 0};

PS = {
   1 10 11,
   10 1 12,
   11 12 1};

PH = {
   13,
   14,
   15,
   16,
   17,
   18,
   19,
   20,
   21};
PH = dg(PH);

AL = {
   0,
   0,
   0};

 KA = {22,
         23,
         24,
         25,
         26,
         27,
         28,
         29,
         30};

theta = {3.97,
           1.93,
           1.24,
           8.75,
           3.57,
           3.73,
           3.77,
           0.96,
           2.17,
           0.68,
           0.91,
           0.41,
           26, 15, 6, 51, 7, 12, 26, 11, 12,
         31,
         24,
         16,
         51,
         22,
         33,
```

```
          110,
           17,
           15} ;

{Lambda, Beta, Psi, Phi, Alpha, Kappa} =
SetModel(theta,LA,BE,PS,PH,AL,KA,_LA,_BE,_PS,_PH,_AL,_KA);

p = rows(Lambda);
q = cols(Lambda);
A = inv(eye(q)-Beta);
mu1 = Lambda*A*Alpha+Kappa;
Sigma1 = Lambda*A*Psi*A'*Lambda'+Phi;

/*Generate Raw Data*/
n2 = 4;
Y2 = rndn(n2,p)*chol(Sigma1)+mu1';
print "PSI Ab" psi;

/*Combine Primary and Aberrant Data*/
n = n1+n2;
Y = Y1|Y2;

S = vcx(Y);
m = meanc(Y);

{start,F0,ets}=SEMfit(S,m,theta);
/* generate missing */
for i (1,196,1);
  isMis =  rndu(1,1) le .02;
  j = trunc(rndu(1,1)*p)+1;
  if isMis;
    Y[i,j] = nill;
  endif;
endfor;
for i (197,199,1);
  isMis =  rndu(1,1) le .02;
  j = trunc(rndu(1,1)*p)+1;
  if isMis;
    Y[i,j] = nill;
  endif;
endfor;
for i (200,200,1);
j=trunc(rndu(1,1)*p)+1;
endfor;
/* EM to start */
{startmu,startSigma} = GenStartVals(Y);
{muHat,SigmaHat,Yc,EC,parmMatrix} = EMmvn(Y,startmu,startSigma,1e4,1e-10);

{thetahat,LLfit,LLsat,chisq,ets,ret}= FIMLfit(Y,start,&dFIMLlogLKHD,muHat,SigmaHat);

indLLfit = -FIMLlogLKHDi(thetahat,Y);
indLLsat = -FIMLlogLKHDsati(muHat|vech(SigmaHat),Y);
indchi =   indLLfit-indLLsat;
```

```
subj = zeros(n, 1);
          for b (1, rows(subj), 1);
              for c (1, cols(subj), 1);
                  subj[b,c] = b*c;
              endfor;
          endfor;

logs1=  (subj|indllfit|indllsat|indchi);
logs1a = (reshape(logs1,4,n))';

logs1=  (subj|indllfit|indllsat|indchi);
yt=y';
ylong=(reshape(yt,n*p,1));
ylogs1=(logs1|ylong);
logs1b = (reshape(ylogs1,13,n))';

output file="C:\Gauss3\Project3\Cell2A.txt" on;
print logs1b;
output off;

endfor;
```

Appendix C
Gauss Code for Simulating and Analyzing Data
Covariance shift Only, 2% Missing, 2% Aberrant

```
#include "C:\Gauss3\SEM\LISREL1.txt";
#include "C:\Gauss3\MI\MI.txt";

totn = 200;
for r (1,totn,1);
print "RUN" r;
/* Cell 8, Covariance shift, 2% Missing, 2% Aberrant*/
/* Setup a cfa model for the primary group*/
_LA = {
   . 0 0,
   . 0 0,
   . 0 0,
   0 . 0,
   0 . 0,
   0 . 0,
   0 0 .,
   0 0 .,
   0 0 .};

_BE = {
   0 0 0,
   0 0 0,
   0 0 0};

_PS = {
   1 . .,
   . 1 .,
   . . 1};

_PH = {
   .,
   .,
   .,
   .,
   .,
   .,
   .,
   .,
   .};
_PH = dg(_PH);

_AL = {0,
   0,
   0};

_KA = {.,
        .,
```

```
          .,
          .,
          .,
          .,
          .,
          .,
          .};

LA = {
   1 0 0,
   2 0 0,
   3 0 0,
   0 4 0,
   0 5 0,
   0 6 0,
   0 0 7,
   0 0 8,
   0 0 9};

BE = {
   0 0 0,
   0 0 0,
   0 0 0};

PS = {
   1 10 11,
   10 1 12,
   11 12 1};

PH = {
   13,
   14,
   15,
   16,
   17,
   18,
   19,
   20,
   21};
PH = dg(PH);

AL = {
   0,
   0,
   0};

 KA = {22,
         23,
         24,
         25,
         26,
         27,
         28,
```

```
         29,
         30};

theta = {3.97,
          1.93,
          1.24,
          8.75,
          3.57,
          3.73,
          3.77,
          0.96,
          2.17,
          0.68,
          0.91,
          0.41,
          26, 15, 6, 51, 7, 12, 26, 11, 12,
         29,
         25,
         14,
         44,
         19,
         28,
        103,
          6,
          9} ;

{Lambda, Beta, Psi, Phi, Alpha, Kappa} =
SetModel(theta,LA,BE,PS,PH,AL,KA,_LA,_BE,_PS,_PH,_AL,_KA);

p = rows(Lambda);
q = cols(Lambda);
A = inv(eye(q)-Beta);
mu1 = Lambda*A*Alpha+Kappa;
Sigma1 = Lambda*A*Psi*A'*Lambda'+Phi;

/*Generate Raw Data*/
n1 = 196;
Y1 = rndn(n1,p)*chol(Sigma1)+mu1';
print "PSI Ab" psi;

/*   Y1;*/

/* Setup a cfa model for the aberrant group*/
_LA = {
   . 0 0,
   . 0 0,
   . 0 0,
   0 . 0,
   0 . 0,
   0 . 0,
   0 0 .,
   0 0 .,
   0 0 .};
```

```
_BE = {
   0 0 0,
   0 0 0,
   0 0 0};

_PS = {
   . . .,
   . . .,
   . . .};

_PH = {
   .,
   .,
   .,
   .,
   .,
   .,
   .,
   .,
   .};
_PH = dg(_PH);

_AL = {0,
   0,
   0};

_KA = {.,
         .,
         .,
         .,
         .,
         .,
         .,
         .};

LA = {
   1 0 0,
   2 0 0,
   3 0 0,
   0 4 0,
   0 5 0,
   0 6 0,
   0 0 7,
   0 0 8,
   0 0 9};

BE = {
   0 0 0,
   0 0 0,
   0 0 0};
```

```
PS = {
   10 11 12,
   11 13 14,
   12 14 15};

PH = {
   16,
   17,
   18,
   19,
   20,
   21,
   22,
   23,
   24};
PH = dg(PH);

AL = {
   0,
   0,
   0};

 KA = {25,
         26,
         27,
         28,
         29,
         30,
         31,
         32,
         33};

theta = {3.97,
           1.93,
           1.24,
           8.75,
           3.57,
           3.73,
           3.77,
           0.96,
           2.17,
           4.96,
           0.95,
           3.17,
           1.47,
           0.23,
           2.14,
           26, 15, 6, 51, 7, 12, 26, 11, 12,
         29,
         25,
         14,
         44,
         19,
```

```
          28,
         103,
           6,
           9} ;

{Lambda, Beta, Psi, Phi, Alpha, Kappa} =
SetModel(theta,LA,BE,PS,PH,AL,KA,_LA,_BE,_PS,_PH,_AL,_KA);

p = rows(Lambda);
q = cols(Lambda);
A = inv(eye(q)-Beta);
mu1 = Lambda*A*Alpha+Kappa;
Sigma1 = Lambda*A*Psi*A'*Lambda'+Phi;

/* Generate Raw Data*/
n2 = 4;
Y2 = rndn(n2,p)*chol(Sigma1)+mu1';
print "PSI Ab" psi;

/*Combine Primary and Aberrant Data*/
n = n1+n2;
Y = Y1|Y2;

S = vcx(Y);
m = meanc(Y);

{start,F0,ets}=SEMfit(S,m,theta);
/* generate missing */
for i (1,196,1);
  isMis =  rndu(1,1) le .02;
  j = trunc(rndu(1,1)*p)+1;
  if isMis;
    Y[i,j] = nill;
  endif;
endfor;
for i (197,199,1);
  isMis =  rndu(1,1) le .02;
  j = trunc(rndu(1,1)*p)+1;
  if isMis;
    Y[i,j] = nill;
  endif;
endfor;
for i (200,200,1);
j=trunc(rndu(1,1)*p)+1;
endfor;
/* EM to start */
{startmu,startSigma} = GenStartVals(Y);
{muHat,SigmaHat,Yc,EC,parmMatrix} = EMmvn(Y,startmu,startSigma,1e4,1e-10);

{thetahat,LLfit,LLsat,chisq,ets,ret}= FIMLfit(Y,start,&dFIMLlogLKHD,muHat,SigmaHat);

indLLfit = -FIMLlogLKHDi(thetahat,Y);
indLLsat = -FIMLlogLKHDsati(muHat|vech(SigmaHat),Y);
```

```
indchi =   indLLfit-indLLsat;

subj = zeros(n, 1);
          for b (1, rows(subj), 1);
             for c (1, cols(subj), 1);
                subj[b,c] = b*c;
             endfor;
          endfor;

logs1=  (subj|indllfit|indllsat|indchi);
logs1a = (reshape(logs1,4,n))';

logs1=  (subj|indllfit|indllsat|indchi);
yt=y';
ylong=(reshape(yt,n*p,1));
ylogs1=(logs1|ylong);
logs1b = (reshape(ylogs1,13,n))';

output file="C:\Gauss3\Project3\Cell8A.txt" on;
print logs1b;
output off;

endfor;
```

Table 1:

Correlations Among Person-Fit Indices (Reise & Widaman, 1999)

	1	2	3	4	5	6	7
				Well-Being Scale			
1. $\hat{\theta}$	--						
2. Z_l	.17	--					
3. LL_{IRT}	.74	.66	--				
4. LL_{SAT}	.75	.70	.86	--			
5. LL_{BAS}	.77	.24	.41	.69	--		
6. LL_{1FAC}	.75	.72	.87	.98	.70	--	
7. IND_{CHI}	-.07	-.17	-.11	.00	.11	-.18	--
				Stress Reaction Scale			
1. $\hat{\theta}$	--						
2. Z_l	-.04	--					
3. LL_{IRT}	-.36	.64	--				
4. LL_{SAT}	-.35	.81	.78	--			
5. LL_{BAS}	-.45	.22	-.27	.19	--		
6. LL_{1FAC}	-.35	.84	.79	.96	.21	--	
7. IND_{CHI}	.00	-.18	-.11	.07	-.06	-.19	--
				Traditionalism Scale			
1. $\hat{\theta}$	--						
2. Z_l	.16	--					
3. LL_{IRT}	.67	.72	--				
4. LL_{SAT}	.58	.80	.81	--			
5. LL_{BAS}	.67	.42	.40	.71	--		
6. LL_{1FAC}	.58	.83	.83	.97	.73	--	
7. IND_{CHI}	-.05	-.20	.83	.04	-.21	-.19	--

Table 2:

Study Design: Covariance Structures, Proportions of Aberrant Observations and Missing Data

	Percentage of Observations with Missing Data					
	2%			*15%*		
	Percentage Aberrant Observations			*Percentage Aberrant Observations*		
	0%	2%	10%	0%	2%	10%
Mean Shift						
Cov. Shift						
Mean + Cov.						
Halo						

Table 3:

Means and Standard Deviations of adj_ll_i by Condition and Sample (Primary vs. Aberrant)

		Percentage of Observations with Missing Data					
		2%			*15%*		
		Percentage Aberrant Observations			*Percentage Aberrant Observations*		
		0%	2%	10%	0%	2%	10%
Mean Shift	P	6.20 (.48)	6.19 (.48)	6.27 (.48)	6.23 (.51)	6.21 (.47)	6.23 (.45)
	A		7.08 (.75)	6.74 (.58)		7.07 (.71)	6.74 (.57)
Cov. Shift	P	--	6.20 (.47)	6.22 (.45)	--	6.21 (.47)	6.21 (.47)
	A		6.08 (.45)	6.73 (.58)		6.10 (.44)	6.09 (.44)
Mean Cov.	P	--	6.20 (.46)	6.02 (.52)	--	6.21 (.47)	6.23 (.46)
	A		6.97 (.66)	6.95 (.81)		6.94 (.66)	6.20 (.55)
Halo	P	--	6.22 (.44)	6.55 (.96)	--	6.22 (.44)	6.33 (.39)
	A		10.20 (2.63)	8.99 (1.67)		10.77 (2.58)	8.99 (1.72)

Table 4:

Tests of Model Effects, AUC for adj_ll_i

Source	DF	Type III SS	Mean Square	F Value	*p*
Missingness	1, 3184	.06	.06	9.39	<.05
Aberrance	1, 3184	1.87	1.87	278.55	<.05
Missingness*Aberrance	1, 3184	.24	.23	35.61	<.05
Source	3, 3184	40.21	13.40	2000.49	<.05
Missingness*Source	3, 3184	1.69	.56	83.93	<.05
Aberrance*Source	3, 3184	4.10	1.37	203.76	<.05
Missingness*Aberrance*Source	3, 3184	1.61	.80	119.96	<.05

Table 5:

Tests of Model Effects, AUC

Effect	Num DF	Den DF	F Value	Pr > F
Index	2	6616	1301.65	<.05
Missingness	1	3172	12.67	<.05
Index*Missingness	2	6616	1.15	.32
Aberrance	1	2980	299.79	<.05
Index*Aberrance	2	6616	62.41	<.05
Missingness*Aberrance	1	3175	54.12	<.05
Index*Missingness*Aberrance	2	6616	.01	.99
Source	3	2982	2627.91	<.05
Index*Source	6	6615	55.83	<.05
Missingness*Source	3	3175	96.62	<.05
Index*Missingness*Source	6	6616	2.31	.03
Aberrance*Source	3	2957	138.25	<.05
Index*Aberrance*Source	6	6616	70.82	<.05
Missingness*Aberrance*Source	3	3175	105.65	<.05
Index*Missingness*Aberrance*Source	6	6616	11.48	<.05

Figure 1:

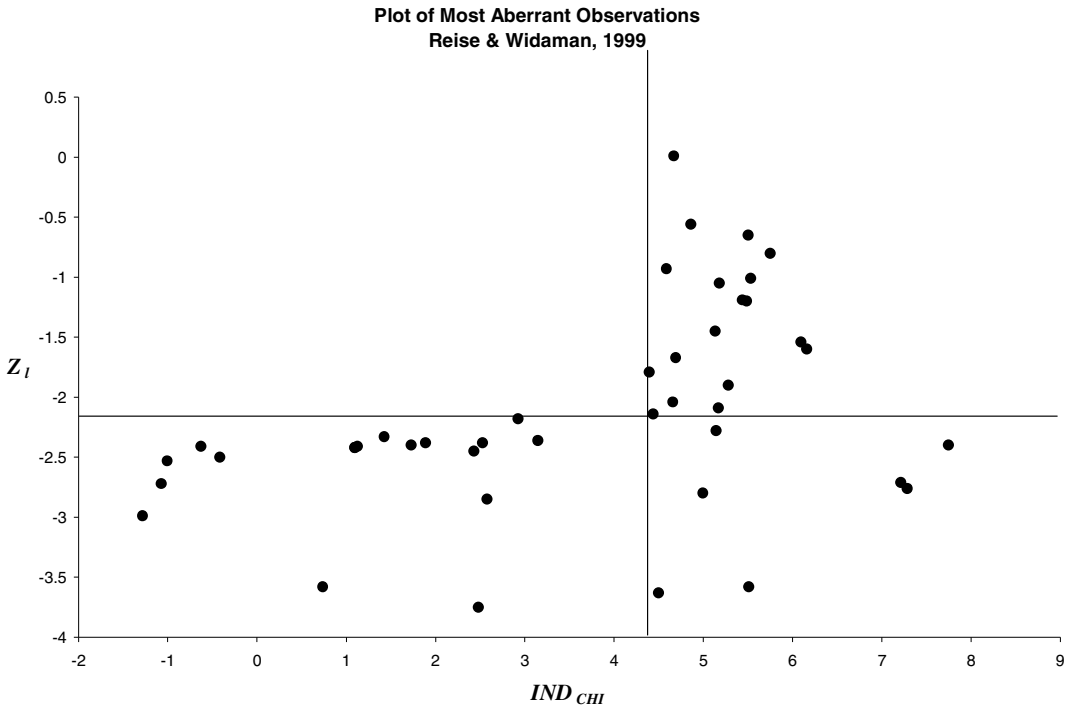

Plot of Most Aberrant Observations
Reise & Widaman, 1999
0.5
0
-0.5
-1
-1.5
-2
-2.5
-3
-3.5
-4
Z_l
-2
-1
0
1
2
3
4
5
6
7
8
9
IND_{CHI}

Figure 2:

Difference (Aberrant – Primary) adj_ll_i Scores by Experimental Condition

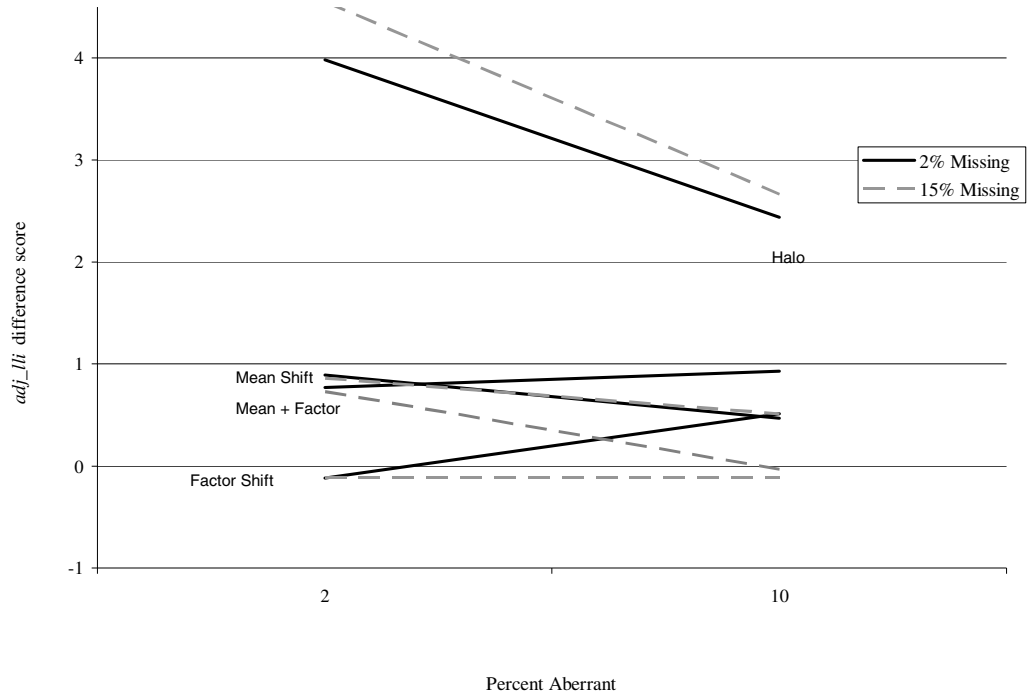


Figure 3:

ROC Curves for adj_ll_i

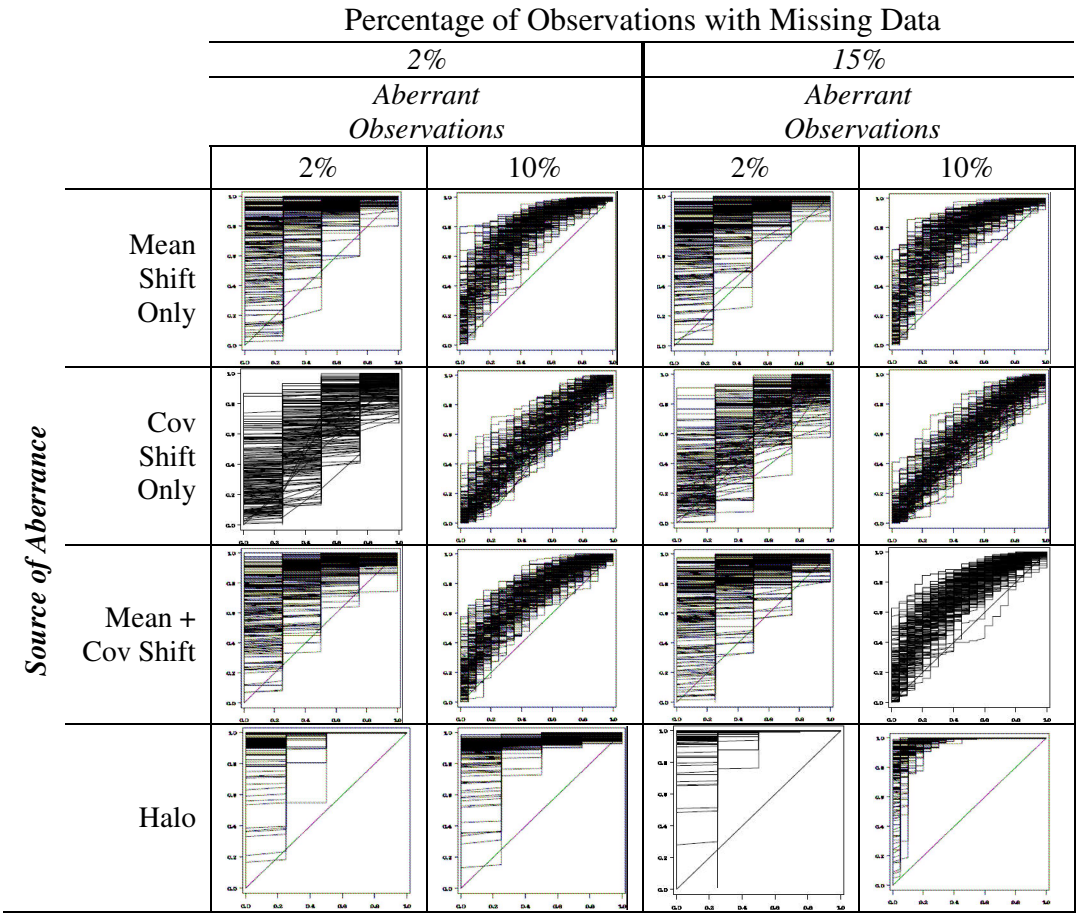


Figure 4:

Collapsed ROC Curves for adj_ll_i

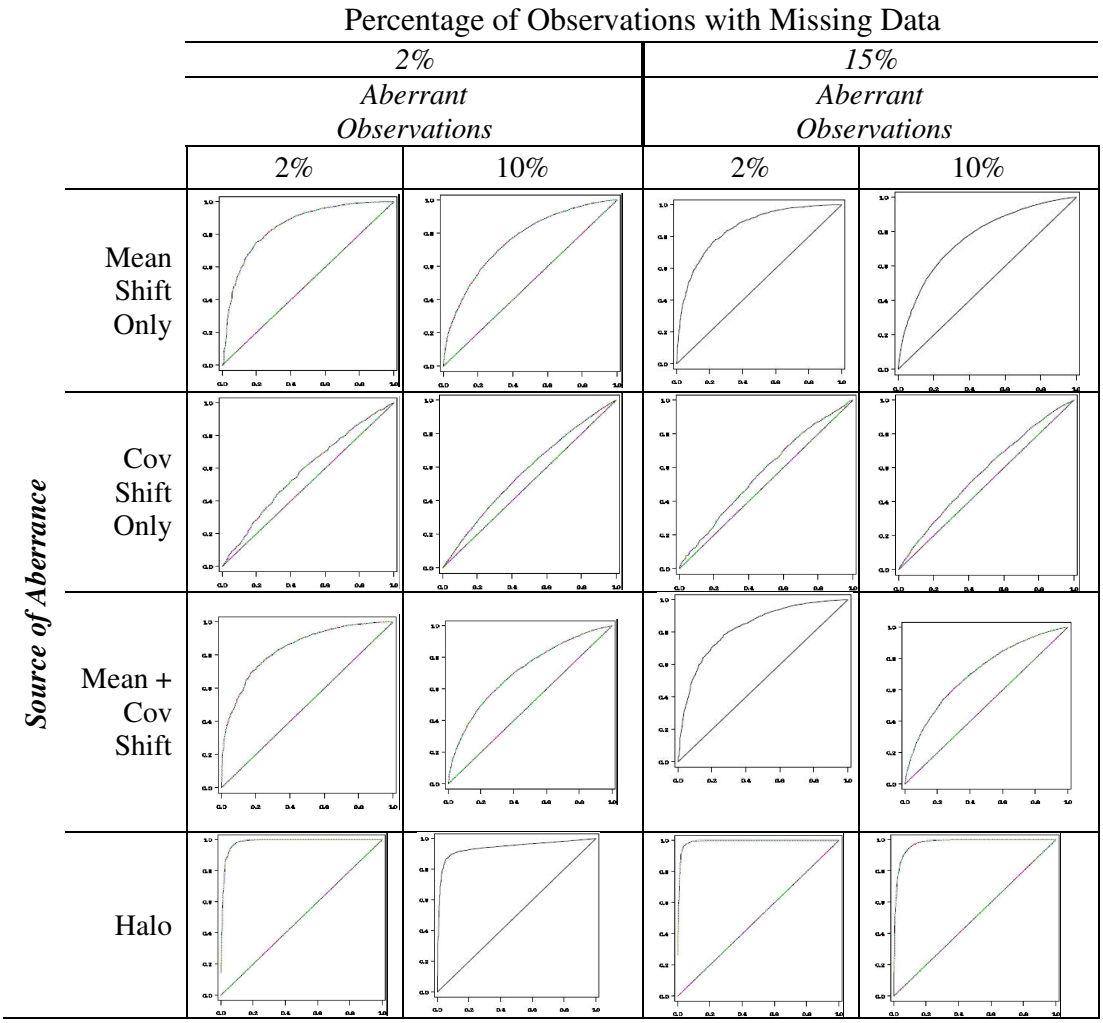


Figure 5:

ROC Curves for IND_{CHI}

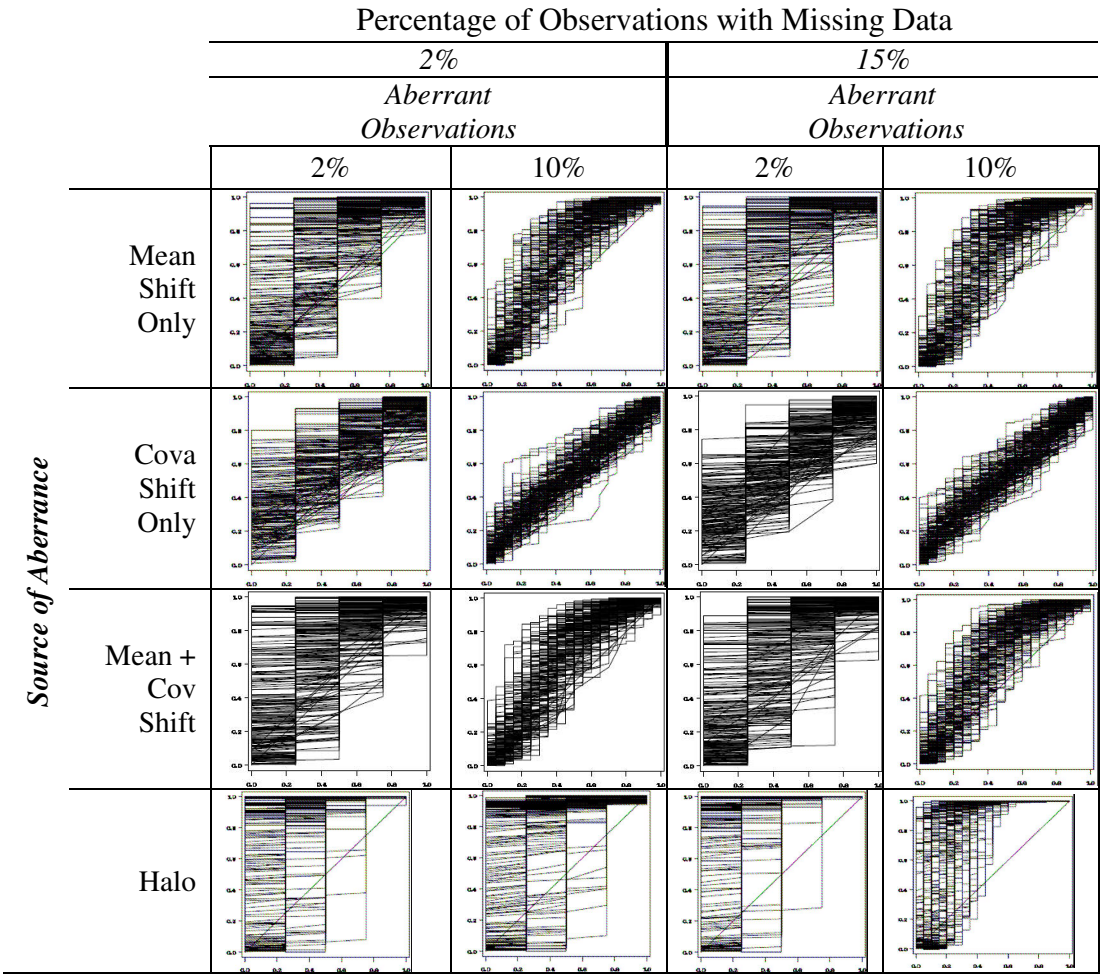


Figure 6:

Collapsed ROC Curves for IND_{CHI}

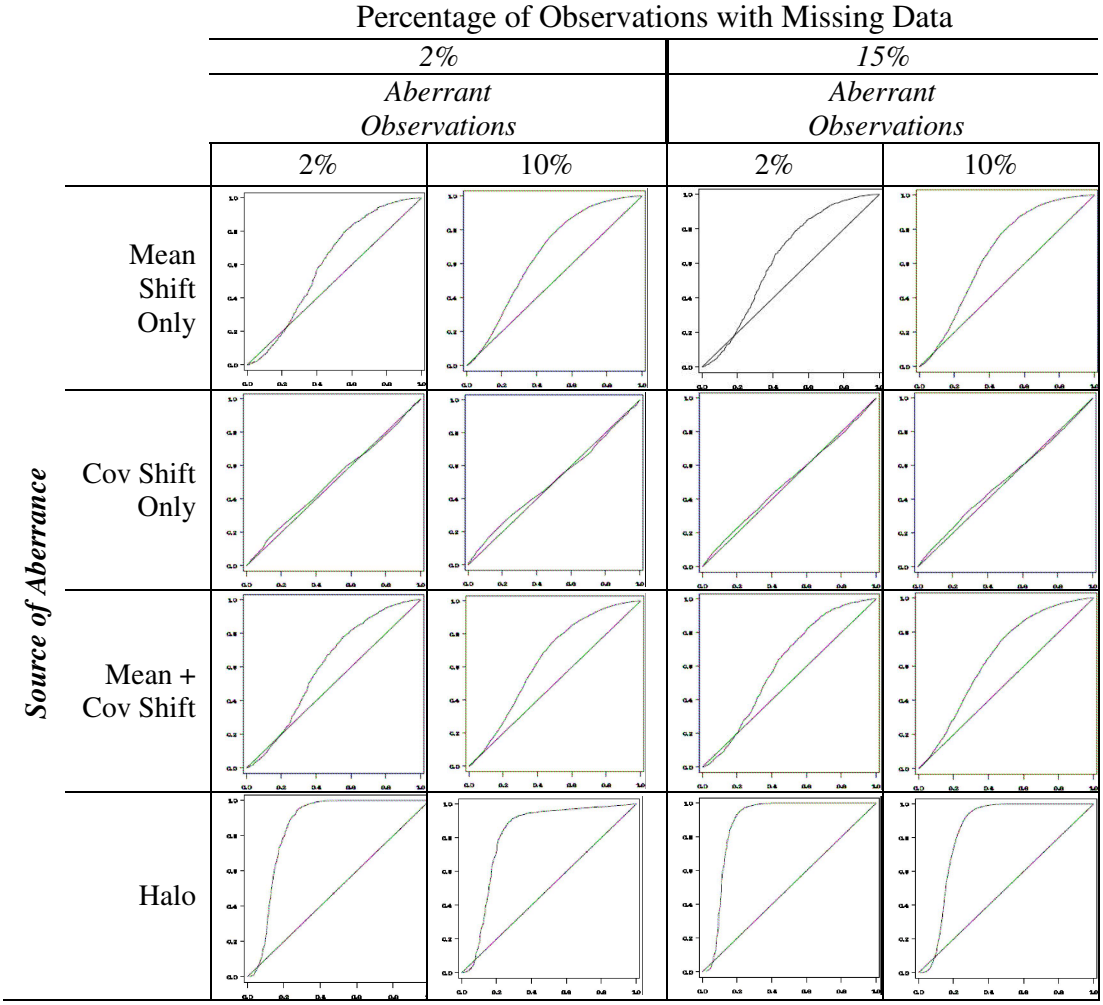


Figure 7:

ROC Curves for Bollen's *A*

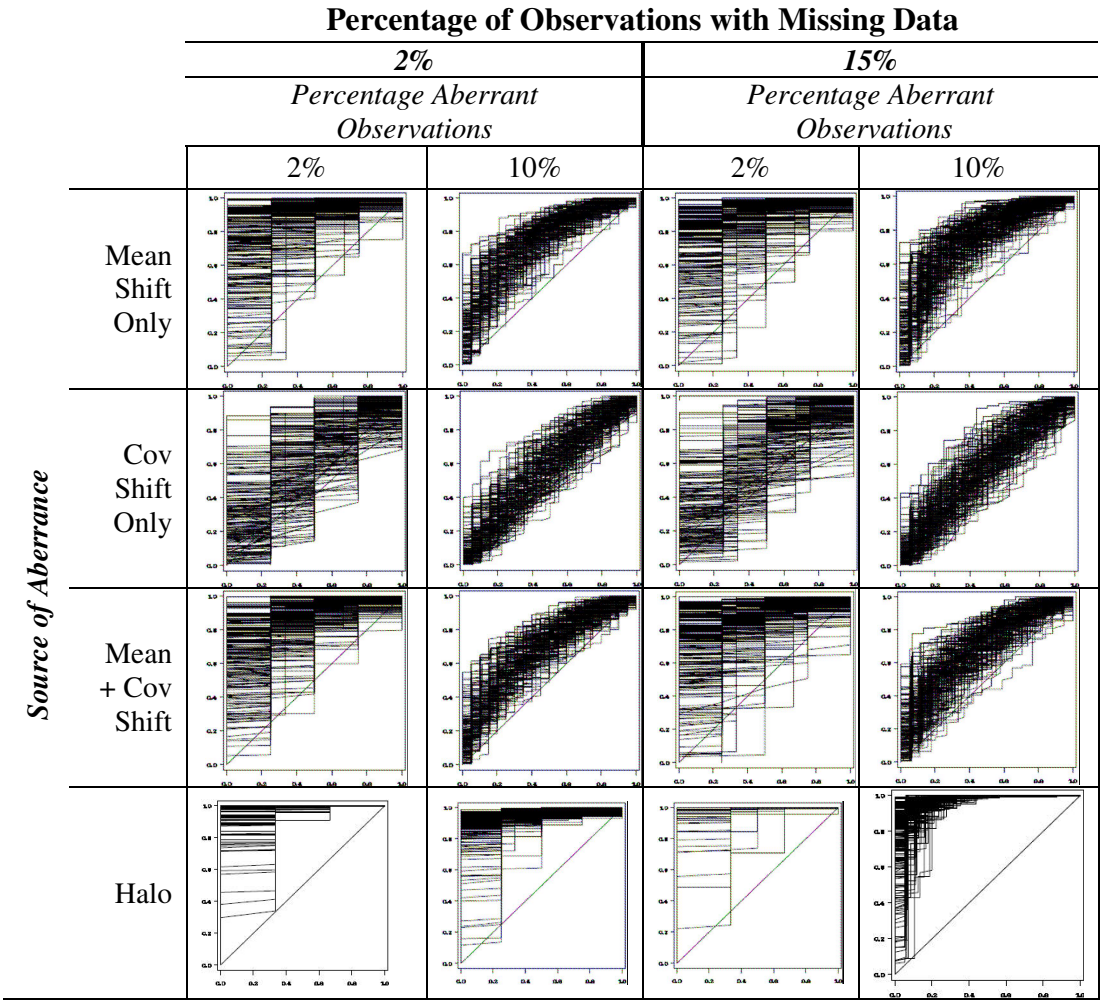


Figure 8:

ROC Curves for Bollen's *A*

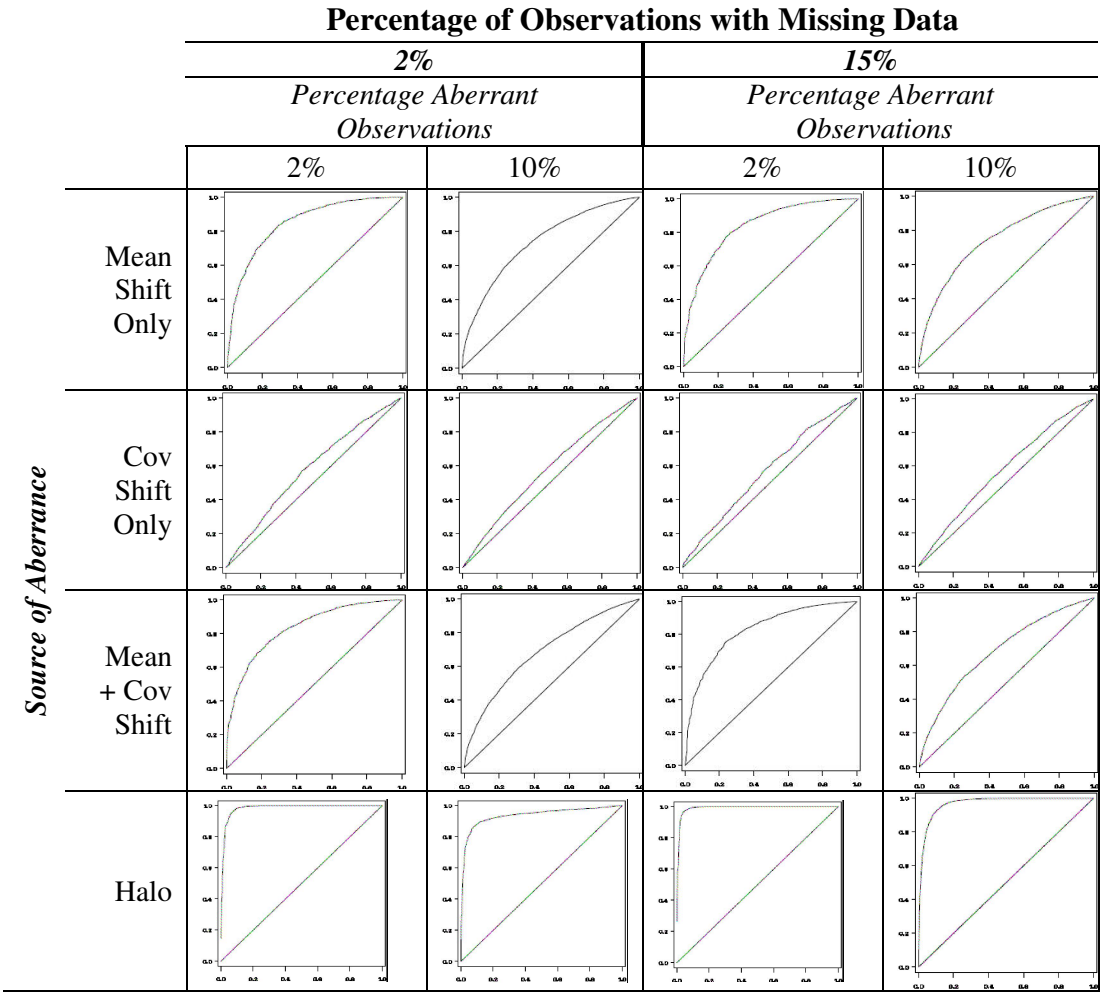


Figure 9:

ROC Curves for Yung's Mixture Method

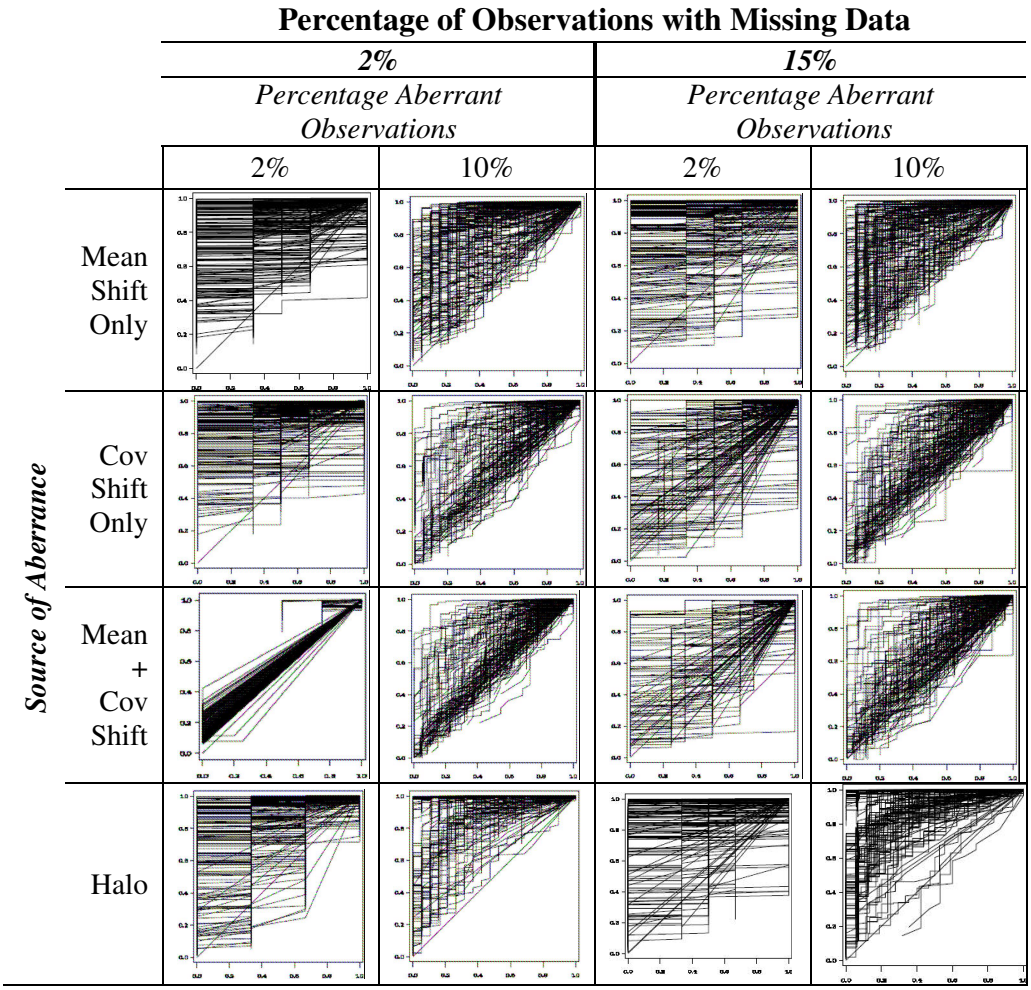


Figure 10:

Collapsed ROC Curves for Yung's Mixture Method

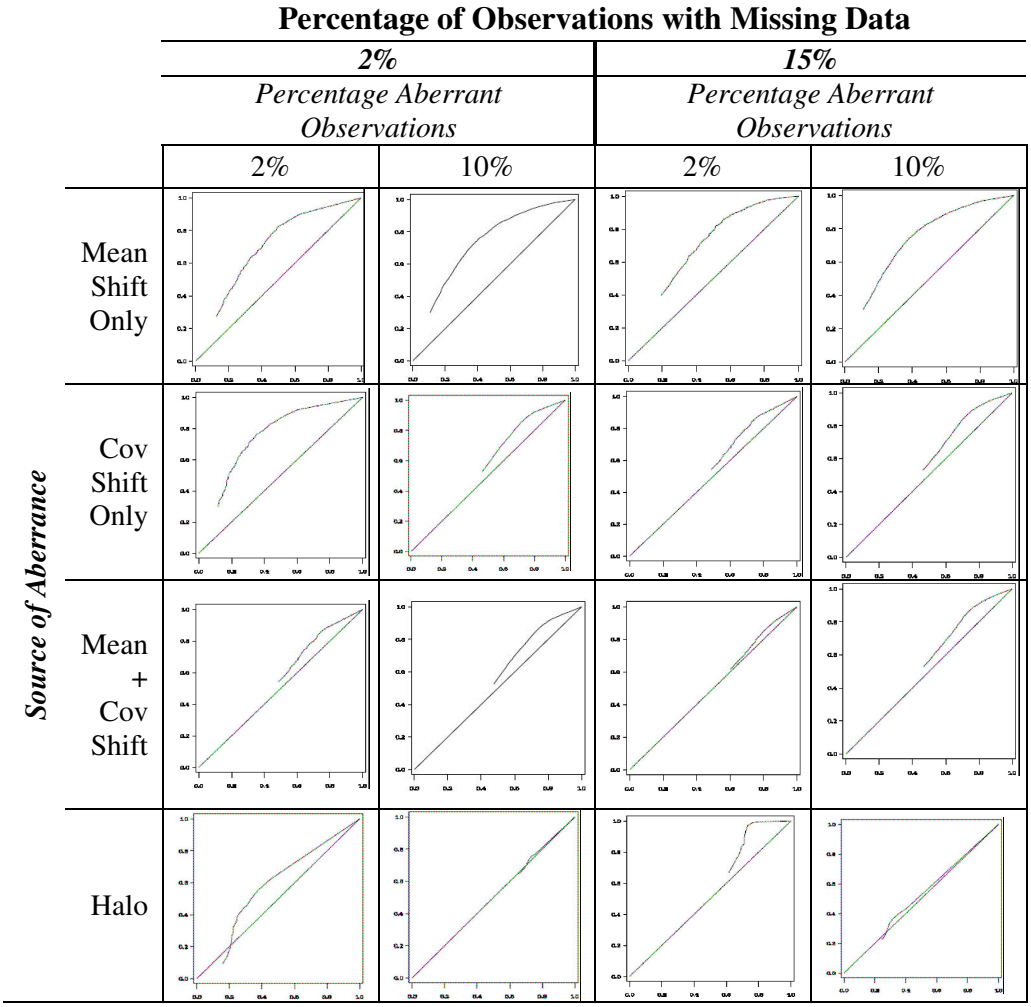


Figure 11:

Area Under the ROC Curve Based on adj_ll_i

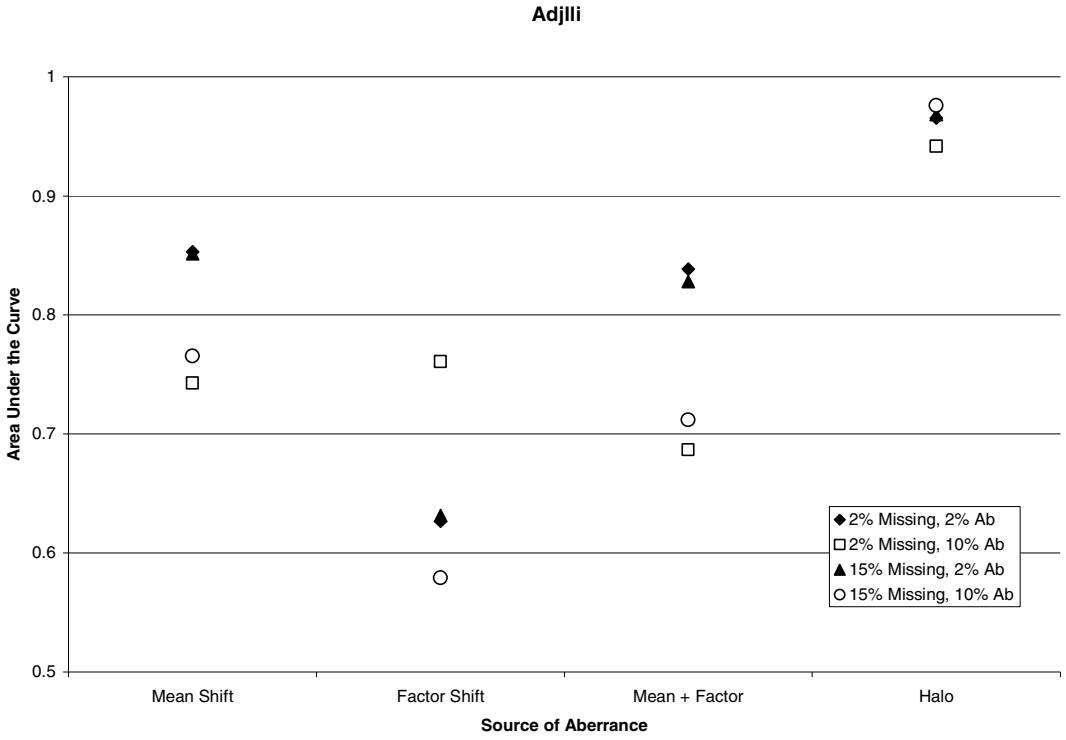


Figure 12:

Area Under the ROC Curve Based on IND_{CHI}

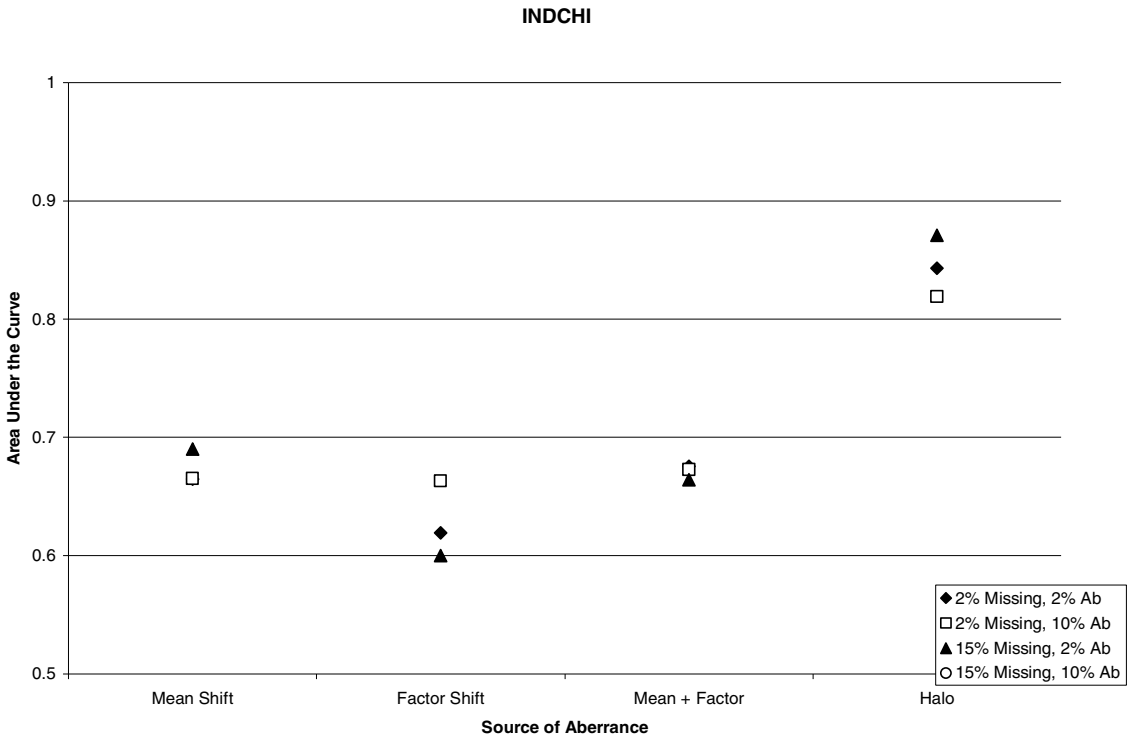
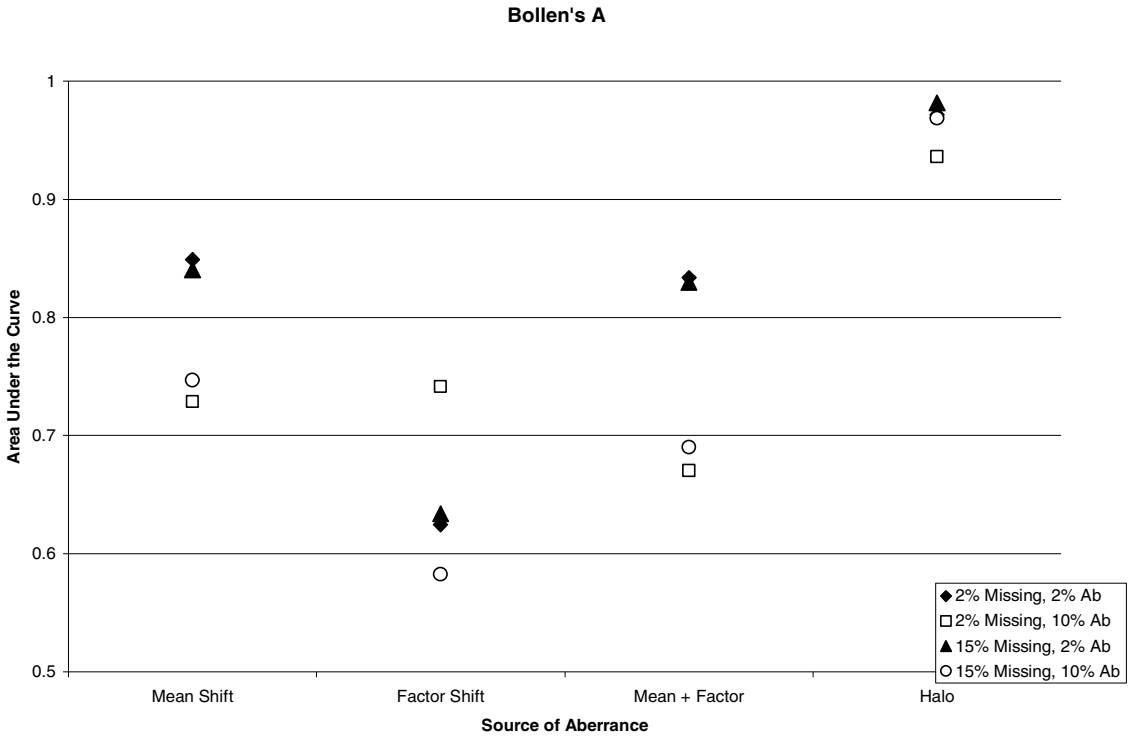


Figure 13:

Area Under the ROC Curve Based on Bollen's *A*



REFERENCES

Amemiya, Y., & Anderson, T. W. (1990). Asymptotic chi-square tests for a large class of factor analysis models. *Annals of Statistic*s, *18*, 1453–1463.

Anscome, F. J. (1973). Graphs and statistical analysis. *American Statistician, 27,* 17-22.

Ansari, A., Jedidi, K., & Jagpal, S. (2000) A hierarchical Bayesian methodology for treating heterogeneity in structural equation models. *Marketing Science, 19*, 328-247.

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumacker (Eds.), Advanced structural equation modeling: Issues and techniques (pp. 243-277). Mahwah, NJ: Lawrence Erlbaum.

Arminger, G., Stein, P., & Wittenberg, J (1999). Mixtures of conditional mean- and covariance-structure models. *Psychometrika*, *64* (4), 475-494.

Bacon, F. (1620/1994). *Novum organum ; with other parts of the great instauration.* Chicago: Open Court.

Barnett, V., & Lewis, T. (1994). *Outliers in Statistical Data*. New York: Wiley.

Bauer, D. J., & Curran P. J. (2003). Distributional assumptions of growth mixture models: Implications for overextraction of latent trajectory classes. *Psychological Methods, 8*(3), 338–363.

Basmann, R. L., (2003). Statistical outlier analysis in litigation support: the case of Paul F. Engler and Cactus Feeders, Inc., v. Oprah Winfrey et al. *Journal of Econometrics*, *113* (1), 159-200 .

Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. *Psychological Methods, 2*, 131-160.

Belsley, D. A., Kuh, E., & Welsch, R. E., (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: Wiley.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness-of-fit in the analysis of covariance structures. *Psychological Bulletin, 88*, 588-600.

Biernacki, C., Celeux, G., & Govaert, G. (1999). An improvement of the NEC criterion for assessing the number of clusters in a mixture model. *Pattern Recognition Letters, 20*, 267-272.

Billor, N., Hadi, A. S., & Velleman, P. F. (2000). BACON: blocked adaptive computationally efficient outlier nominators. *Computational Statistics & Data Analysis, 34*(3), 279-298.

Blåfield, E. (1980). Clustering of observations from finite mixtures with structural information. *Jyvaskyla Studies in Computer Science, Economics, and Statistics, 2.*

Bollen, K.A. & Jackman, R. (1985). Regression diagnostics: An expository treatment of outliers and influential cases. *Sociological Methods and Research* 13, 510-42.

Bollen, K.A. (1987). Outliers and improper solutions: A confirmatory factor analysis example. *Sociological Methods and Research, 15*, 375-384.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York: Wiley.

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 37,* 62-83.

Cai, L. (2005a). Structural Equation Modeling software.

Cai, L. (2005b). Personal communication.

Chatfield, C. (2002). Confessions of a pragmatic statistician. *The Statistician, 51*(1), 1-20.

Comrey, A. L. (1985). A method for removing outliers to improve factor analytic results. *Multivariate Behavioral Research, 20*, 273-281.

Day, N. E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika*, *56*, 463-474.

De Maesschalck, R., Jouan-Rimbaud, D., & Massart, D. L. (2000). The Mahalanobis distance *Chemometrics and Intelligent Laboratory Systems, 50, 1-18*

Dolan, C. V., and van der Maas, H. J. L. (1998). Fitting multivariate normal mixtures subject to structural equation modelling. *Psychometrika, 63*, 227-253.

Drasgow, F., & Levine, M.V., Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology. 38*, 67-86.

Drasgow, F., & Levine, M. V. (1986). Optimal detection of certain forms of inappropriate test scores. *Applied Psychological Measurement, 10*, 59-67.

Enders, C. K. & Bandalos, D. L. (2001). The relative performance of Full Information Maximum Likelihood Estimation for missing data in structural equation models. *Structural Equation Modeling, 8*(3), 430-457.

Fox, J. (1991). *Regression Diagnostics*. Thousand Oaks, CA: Sage.

Gerbing, D. W., & Anderson, J. C. (1993). Monte Carlo evaluations of goodness-of-fit indices for structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 40–65). Newbury Park, CA: Sage.

Gnanadesikan, R., & Kettenring, J. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics 28*, 81–124.

Guttman, L. (1950). The principal components of scale analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and Prediction*, pp. 312-361. New York: Wiley

Hardin, J., & Rocke, D. (2002). The distribution of robust distances. http://www.cipic.ucdavis.edu/~dmrocke/preprints.html.

Hawkins, D. M. (1980). *Identification of Outliers.* Chapman & Hall: London.

Hays, W. H., (1994). *Statistics*, (5th ed.), Fort Worth, TX: Harcourt.

Hodge, V. H., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review, 22*, 85-126.

Holzinger, K., & Swineford, F. (1939). A study in factor analysis: The stability of a bifactor solution. *Supplementary Educational Monograph No. 48.* University of Chicago Press: Chicago.

Jöreskog, K. & Sörbom, D. (1996). *LISREL 8: Users's reference guide*. Lincolnwood, IL: Scientific Software.

Kim, M. G. (2000). *Communications in statistics - theory and methods, 29,* 1511-1526

Levine, M. V., & Drasgow, F. (1982). Appropriateness measurement: Review, critique and validating studies. *British Journal of Mathematical and Statistical Psychology, 35,* 42-52.

Levine, M. V. & Rubin, D. F. (1979). Measuring the appropriateness of multiple choice test scores. *Journal of Educational Statistics*, 4, 269-290.

Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika 88*(3), 767-778

Lyons, J. A., & Scotti, J. R. (1994). Comparability of two administration formats of the Keane Posttraumatic Stress Disorder Scale. *Psychological Assessment, 6*, 209-211.

MacCallum, R. A. (2003). Personal communication.

Matthey, S., & Petrovski, P. (2002) The Children's Depression Inventory: Error in cutoff scores for screening purposes. *Psychological Assessment, 14*, 146-149.

McLachlan, G. J. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Applied Statistics, 36*, 318-324.

McLachlan, J., & Peel, D. (2000). *Finite mixture models*. Wiley: New York.

Meijer, R. R. (1996). Person-fit research: An introduction. *Applied Measurement in Education, 9*, 3-8.

Muthén, B. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54*, 557–585.

Muthén, L. K., & Muthén, B. O. (2001). *Mplus user's guide.* Los Angeles: Muthén & Muthén.

Muthén, B., & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics, 55,* 463-469.

Neale M. C., Boker S. M., Xie G., & Maes, H. H. (2003). *Mx: Statistical Modeling.* Richmond, VA: Virginia Commonwealth University.

Rao, C. R. (1973). *Linear Statistical Inference and Its Application, (2nd ed.)*. New York: John Wiley.

Rasmussen, J. L. (1988). Evaluating outlier identification tests: Mahalanobis *D* Squared and Comrey's *D*. *Multivariate Behavioral Research, 23*, 189-202.

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.

Samejima, F. (1969). Estimation of ability using a response pattern of graded scores. *Psychometrika Monograph,* No. 17.

Sheeran, T, & Zimmerman, M. (2002) Screening for posttraumatic stress disorder in a general psychiatric outpatient setting. *Journal of Consulting & Clinical Psychology, 70*, 961-966.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structures between groups. *British Journal of Mathematical and Statistical Psychology, 27,* 229-239.

Tellegen, A. (1982). *Brief manual of the Multidimensional Personality Questionnaire.* Unpublished Manuscript.

Titterington, D. M., Smith, A. F. M., & Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*.

Velasco, F., Verma, S. P., & Guevara, M. (2000). Comparison of the performance of fourteen statistical tests for detection of outlying values in geochemical reference material databases. *32, Mathematical Geology,.*

Yuan, K. H. (2005). Fit indices versus test statistics. *Multivariate Behavioral Research*, *40*, 115-148.

Yuan, K. H. & Bentler, P. M. (2001). Effect of outliers on estimators and tests in covariance structure analysis. *British Journal of Mathematical and Statistical Psychology, 54*, 161-175.

Yuan, K. H., Chan, W., & Bentler, P. M. (2000). Robust transformation with applications to structural equation modeling. *British Journal of Mathematical and Statistical Psychology, 53, 31-50.*

Yung, Y. F. (1997). Finite mixtures in confirmatory factor-analytic models. *Psychometrika, 62*, 297-330.

Yung, Y. F. (2006). Personal communication.