UNIVERSITY OF CALIFORNIA

Los Angeles

Case Diagnostics in Categorical Factor Analysis

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Psychology

by

Maxwell Armand Mansolf

2019

ABSTRACT OF THE DISSERTATION

Case Diagnostics in Categorical Factor Analysis

by

Maxwell Armand Mansolf

Doctor of Philosophy in Psychology

University of California, Los Angeles, 2019

Professor Steven Reise, Chair

Case diagnostics in categorical factor analysis include Mahalanobis distance-based statistics, which measure residual and leverage, and adaptations of existing influence diagnostics such as individual contribution to chi-square and generalized Cook's distance which measure each case's influence on statistical results. This dissertation uses two simulation studies to explore issues related to the use of case diagnostics in categorical factor analysis in order to assess the feasibility and utility of an iteratively reweighted least squares estimator for categorical factor analysis and structural equation modeling. In the first simulation, I used large data sets simulated according to a hypothesized model structure to examine the null distributions of Mahalanobis distance-based measures of residual and leverage in categorical factor analysis. Specifically, this study examined the validity of statistical cut-off values derived from continuous distributions in categorical factor analysis and assessed the differences between theoretical and empirical critical values in these models. In most conditions, the distributions of leverage and residual diagnostics in polytomous

data, and of leverage diagnostics in dichotomous data, were similar enough to those in continuous data that existing critical values can safely be used to identify high-leverage cases. In contrast, residual diagnostics in dichotomous data had severely truncated distributions, a result which complicates the choice of critical value for identifying high-residual cases in residual analysis or down-weighting cases in robust estimation. In the second simulation, I examined the relationships between leverage, residual, and influence in categorical and continuous factor analysis and compared those relationships across continuous, polytomous, and dichotomous test conditions. Results were largely consistent between continuous and polytomous data but differed markedly in dichotomous data with high variability across dichotomous test conditions. Together, these findings reveal that, while categorical case diagnostics are well-behaved in polytomous tests under ideal conditions, these diagnostics can behave unpredictably in dichotomous data, and thus caution should be used in interpreting their values directly in dichotomous tests, whether as a means for screening for outliers or for down-weighting cases in robust estimation.

The dissertation of Maxwell Armand Mansolf is approved.

Peter Bentler

Steve Lee

Catherine Sugar

Steven Reise, Committee Chair

University of California, Los Angeles

2019

To Mom & Dad

Thank you

TABLE OF CONTENTS

Vita

Education

2014 (Fall)	M.A., Department of Psychology, UCLA, Quantitative Methods Major *Thesis title*: First and second-order local influence measures in structural equation modeling.
2012 (Spring)	B.S., Department of Psychology, UCLA, Cognitive Science Major Minor in Statistics, Minor in Mathematics, Specialization in Computing

Publications

Ramos, I. F., Guardino, C. M., **Mansolf, M.**, Glynn, L. M., Sandman, C. A., Hobel, C. J., & Schetter, C. D. (2019). Pregnancy anxiety predicts shorter gestation in Latina and non-Latina white women: The role of placental corticotrophin-releasing hormone. *Psychoneuroendocrinology, 99*, 166-173.

Conway, C., **Mansolf, M.**, & Reise, S. P. (2019). The utility of a general factor (p-factor) of psychopathology: Associations with clinical outcomes in 25,000 treatment-seeking adults. *Psychological Assessment. In Press*

Anderson, A. E., **Mansolf, M.**, Reise, S. P., Savitz, A., Salvadore, G., Li, Q., & Bilder, R. M. (2017). Measuring pathology using the PANSS across diagnoses: Inconsistency of the positive symptom domain across schizophrenia, schizoaffective, and bipolar disorder. *Psychiatry Research, 258,* 207-216.

Mansolf, M., & Reise, S.P. (2017). Case diagnostics for factor analysis of ordered categorical data with applications to person-fit measurement. *Structural Equation Modeling: A Multidisciplinary Journal*. doi: http://dx.doi.org/10.1080/10705511.2017.1367926.

Mansolf, M., & Reise, S. P. (2017). When and why the second-order and bifactor models are distinguishable. *Intelligence, 61*, 120-129.

Yang, R., Spirtes, P., Scheines, R., Reise, S. P., & **Mansolf, M.** (2017). Finding pure submodels for improved differentiation of bifactor and second-order models. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(3), 402-413.

Anderson, A. E., Reise, S. P., Marder, S. R., **Mansolf, M.**, Han, C., & Bilder, R. M. (2017). Disparity between General Symptom Relief and Remission Criteria in the Positive and Negative Syndrome Scale (PANSS): A Post-treatment Bifactor Item Response Theory Model. Innovations in clinical neuroscience, 14(11-12), 41.

Enders, C. K., & **Mansolf, M.** (2016). Assessing the fit of structural equation models with multiply imputed data. *Psychological Methods.* http://dx.doi.org/10.1037/met0000102.

Mansolf, M., & Reise, S. P. (2016). Exploratory bifactor analysis: The Schmid-Leiman orthogonalization and Jennrich-Bentler analytic rotations. *Multivariate Behavioral Research, 51*(5), 698-717.

Reise, S. P., Kim, D. S., **Mansolf, M.**, & Widaman, K. F. (2016). Is the bifactor model a better model or is it just better at modeling implausible responses? Application of iteratively reweighted least squares to the Rosenberg Self-Esteem Scale. *Multivariate Behavioral Research, 51*(6), 818-838.

Mansolf, M., & Reise, S.P. (2015). Local minima in exploratory bifactor analysis. *Multivariate Behavioral Research, 50*(6).

Papers in Progress

Mansolf, M., & Enders, C. K. (under revision). A multiple imputation score test for model modification in structural equation models. *Psychological Methods*.

Hobel, C., Guardino, C., **Mansolf, M.**, Dunkel Schetter, C. (in progress). Preterm birth phenotypes: Novel grouping of birth outcomes, their correlates and consequences.

Mansolf, M., Vreeker, A., Reise, S.P., & Bilder, R. (in progress). Extensions of multiple-group item response theory alignment: Application to psychiatric phenotypes in an international genomics consortium.

Conference Presentations

Mansolf, M., Bilder, R., & Reise, S. P. (December 2017). Identifying dimensional phenotypes from clinical assessments. Presentation at the 56th Annual Meeting of the American College of Neuropsychopharmacology, Palm Springs, CA.

Mansolf, M. (May 2017). Harmonizing clinical assessments: Statistical approaches. Presentation at the Whole Genome Sequencing for Personality Disorders working group meeting, Broad Institute, Boston, MA.

Mansolf, M., & Reise, S. P. (2015). Local minima in exploratory bifactor analysis. Presentation at the 55th Annual Meeting of the Society of Multivariate Experimental Psychology, Redondo Beach, CA.

Ren, X.*, Shen, Y.*, **Mansolf, M.**, & Reise, S.P. (2018). Cognitive task harmonization for genetic studies of psychiatric disorders. Presentation at the 27th Annual Psychology Undergraduate Research Conference, Los Angeles, CA.

Ma, J.*, Ng, A.*, **Mansolf, M.**, & Reise, S.P. (2018). Factor analysis of psychiatric symptoms in bipolar I, schizophrenia, and schizoaffective disorder. Presentation at the 27th Annual Psychology Undergraduate Research Conference, Los Angeles, CA.

Li, Z., Hasratian, A., **Mansolf, M.**, Mesri, B., & Craske, M. G. (2016). Using implicit techniques to augment fear extinction as a proxy for anxiety treatment. Abstract accepted by the 96th Annual Convention for the Western Psychological Association, Long Beach, CA.

* indicates undergraduate mentees

Chapter 1 - Introduction

Psychological measures are typically constructed under the assumption that item responses are manifestations of one or more unobserved or "latent" variables representing the construct(s) of interest which are related to the observed item responses through a common cause model (Bollen & Lennox, 1991). When item responses are ordered categorical, the data can be modeled using item response theory (IRT; van der Linden & Hambleton, 1997), also called item factor analysis (IFA; Bock, Gibbons, & Muraki, 1988), and factor analysis (FA; Mislevy, 1986) of polychoric correlation matrices (Muthén, 1984), which can be shown to be formally equivalent (Takane & De Leeuw, 1987).

Typically, model fit is evaluated by using global fit indices (e.g., Hu & Bentler, 1999) to quantify the fit of the model to the entire dataset. However, especially within factor analysis, much less attention has been devoted to the other side of the data matrix, that is, how well the model accounts for an individual's response pattern. Overall model fit does not guarantee that all individuals provide patterns of item response that are consistent with a given model[1]. Psychological theory, and common sense, suggest a multitude of reasons an individual may not respond as predicted by the hypothesized latent structure, including unmodeled multidimensionality (Waller and Reise, 1992), faking (Zickar and Drasgow, 1996; Ferrando & Anguiano-Carrasco, 2013), acquiescence (Curtis, 2004; Reise & Flannery, 1996), sabotage (Ferrando, 2012), and idiosyncratic use of response options (extreme/middling responding; Emons, 2009; Ferrando, 2010), to name just a few. All of these are examples of individuals responding in ways that are not due to the hypothesized latent variable(s). Factor analysis and item response theory models which explicitly incorporate "response styles" such as acquiescence and

[1] In fact, model fit at the sample level in no way guarantees that the model applies to an individual or any subset of individuals (Borsboom, Mellenbergh, & van Heerden, 2003).

extreme/middling responding have recently been developed to account for these phenomena (Falk & Cai, 2016), although such models can only potentially identify patterns of responding which are explicitly accounted for in the model specification. While not all individuals whose response process doesn't match the hypothesized model structure produce model-inconsistent or "aberrant" response patterns, many do; such response patterns are said to exhibit poor "person-fit" and a variety of metrics are available for judging the credibility of a response pattern given a hypothesized model. It is well-known that if a response pattern has poor person-fit, then the precision of the associated latent trait estimate, and the estimate's applicability to the individual, are questionable (Ferrando, 2015)[2].

Also well-known, but less understood in IRT and categorical FA, is that individuals with model-inconsistent response patterns can affect statistical inference by distorting model fit (Reise and Widaman, 1999) and parameter estimation (Pek and MacCallum, 2011). Levine and Drasgow (1983) observed that it is possible for a model to fit a dataset well, even in the presence of individuals whose response patterns cannot be well-explained by the model. Moreover, Reise and Widaman (1999, Table 7) examined the distribution of individual contribution to chi-square (*INDCHI*) in observed and simulated data and found that a small subset of aberrant cases could have a relatively large impact on model fit. Although simulation studies have evaluated the robustness of polychoric correlations to distributional assumptions (Flora & Curran, 2004; Lee & Lam, 1988; Quiroga, 1992; Jin & Yang-Wallentin, 2017), a rigorous study of case diagnostics and influence in categorical SEM has not been conducted.

Iteratively reweighted least squares (IRLS) estimators for SEM with continuous variables are well-known (e.g., Yuan and Bentler, 2000; Yuan and Zhong, 2008), but have not been extended

[2] For this reason such indices have also been referred to as "test score appropriateness" indices.

to the categorical case. This study was designed to investigate potential issues with, and the potential utility of, IRLS for SEM in categorical data, specifically within the context of categorical factor analysis. Specifically, I am interested in whether categorical versions of Mahalanobis distance measures of leverage (d_f) and residual (d_r) can be used to determine case weights in categorical IRLS and in the potential utility of categorical IRLS to appropriately down-weight influential cases. As the number of categories increases, categorical data approach the same quality of information as continuous data, but with fewer categories, categorical responses can differ substantially from continuous responses; thus, the conditions under which categorical data can be treated similarly to continuous data are critical in understanding the potential validity and utility of IRLS in categorical data.

The two primary goals of this research are:

1. To examine the distribution of Mahalanobis distance measures of leverage (d_f) and residual (d_r) in categorical data in order to determine the conditions in which the cutoffs used in continuous data, based on the quantiles of theoretical distributions, will be appropriate for IRLS in categorical data.
2. To characterize the relationship between leverage (d_f), residual (d_r), and influence (generalized Cook's distance, $\Delta\chi^2$) in categorical data under varying test conditions in order to determine the conditions in which leverage and residual function properly as proxies for influence in categorical IRLS, compared to well-known relationships in continuous data (Yuan and Zhong, 2008). Whether these relationships hold in categorical data, as modeled by the polychoric correlation matrix, will determine the potential efficacy of robust procedures which use leverage and residual measures to down-weight aberrant observations during estimation.

These distributional properties and relationships depend, of course, on the properties of the test; with many items and many well-placed thresholds, the differences are likely minimal, but performance is bound to be worse in less ideal conditions. Therefore, it is essential to investigate these issues under varying test conditions. In the following sections, I first review case diagnostics and IRLS in regression and their extensions to structural equation modeling. Then, I will review prior literature on the robustness of categorical factor analysis to violations of distributional assumptions, followed by the goals and research questions for this dissertation. I then discuss the case diagnostics used in this study, followed by two simulation studies examining these diagnostics. A General Discussion assessing the implications of the results of the simulation studies for the development of IRLS for categorical factor analysis and for practical use of categorical factor analysis concludes this dissertation.

1.1. Case diagnostics and IRLS in regression

In linear regression, it is well-known that individual observations that deviate substantially, and in the right ways, from the general trend of the data can distort or invalidate the results of an analysis (e.g., Rousseeuw & van Zomeren, 1990; Wilcox, 2001, pp. 218 –219) and diagnostic measures have been developed to identify such problematic cases (e.g., Belsley, Kuh, & Welsch, 1980; Cohen, Cohen, West, & Aiken, 2003, Chapter 10; Cook & Weisberg, 1982; Rousseeuw & Leroy, 1987). In OLS regression, a case's relationship to the general trend of the data, as represented by the regression line, can be quantified in terms of *residual* and *leverage*. Cases with large residuals lie far from the predicted values based on the regression line (i.e., extreme values of the outcome, conditional on the predictor set), while cases with high leverage have extreme values on the predictors. Additional indices have been developed to specifically measure case influence. In regression: DFBETA (Belsley, Kuh, & Welsch, 1980), quantifies the influence of

cases on individual regression coefficients; Cook's distance (Cook, 1977, 1979) quantifies the influence of cases on the parameter estimates; DFFITS (Belsley, Kuh, & Welsch, 1980) quantifies the influence of cases on predicted values; likelihood distance (Cook and Weisberg, 1982) quantifies the influence of cases on the model (log)likelihood, and many other, often redundant, indices exist (see Belsley, Kuh, & Welsch, 1980; Fox, 1991). The relationships among these diagnostics have been well-known in regression for decades (e.g., Rousseeuw & van Zomeren, 1990); specifically, so-called *bad leverage points*, which have both large residuals and extreme values of the predictors, can have catastrophic effects on a regression analysis, while *good leverage points* generally improve statistical power.

The existence, and potentially disastrous consequences, of such influential cases have motivated the development of robust regression, which in this context refers to estimators that attempt to account for the presence of potentially problematic cases. In linear and logistic regression, one widely used robust procedure is iteratively reweighted least squares (IRLS; Green, 1984; Holland & Welsch, 1977; O'Leary, 1990). In IRLS regression, each case is assigned a case weight during estimation, wherein smaller case weights are assigned to cases that are poorly predicted by the hypothesized model (high residuals), such that cases with small case weights have a reduced impact on estimation. As a result, IRLS estimation yields (1) regression parameters that are less affected by the influence of outliers or unusual observations and (2) case weights that quantify the fit of the model to the individual case that can be used to identify unusual observations. One key advantage of these and similar robust estimators is that they arguably produce parameter estimates that are more replicable across studies – a chronic problem in psychological research in general (Bohannon, 2015; Yuan, Marshall, & Weston, 2002).

1.2. Case diagnostics and IRLS in structural equation modeling

Considering that SEM is a multivariate extension of the regression model, it is no surprise that case diagnostics and influence within SEM have received an increasing amount of attention. It has long been known that outliers can distort assessments of model fit (Bentler, 1989, pp. 117-124; Bollen & Arminger, 1991; Yuan & Zhong, 2008; Zhong & Yuan, 2011), while good leverage observations mainly impact the parameter estimates of a model, and bad leverage observations impact both fit and parameter estimates (Yuan & Zhong, 2008; Zhong & Yuan, 2011). Recently, SEM analogues of regression diagnostics have been studied more rigorously in parallel with the development of IRLS estimators for structural equation models (Yuan & Bentler, 1998, 2000; Yuan & Zhong, 2008). These case diagnostics generalize the concepts of residual and leverage to the structural equation modeling context and downweight cases with high values of these indices.

The SEM case diagnostics used in IRLS in SEM take the form of Mahalanobis distance (M-distance) measures, also known as multivariate Z-scores. Two M-distances, d_c and d_s, are simply multivariate Z-scores using the saturated (d_c) or model-implied (d_s) mean and covariance matrix of all of the observed variables. Two additional M-distances are the factor-score-based M-distance d_f, which uses Bartlett's factor score estimates to calculate an M-distance for factor scores in latent variable models, and the residual-based M-distance d_r, which uses Bartlett's factor score estimates to calculate an M-distance for model residuals. Within the context of confirmatory factor analysis, in which the latent variables are predictors and the observed variables are outcomes, d_f quantifies multivariate leverage and d_r quantifies multivariate residual (Yuan and Zhong, 2008; Zhong and Yuan, 2011; Yuan, Fung, & Reise, 2004).

1.3. Prior research on robustness of categorical factor analysis

The aforementioned SEM case diagnostics and robust estimators assume that the observed data are continuous; however, the vast majority of self-report and clinical diagnostic measures in psychology use ordered categorical measurement. Both asymptotically distribution-free (ADF; Browne, 1984) and maximum likelihood (ML) estimation require the use of a covariance or correlation matrix, rendering both inappropriate for variables measured at an ordinal level. Although ADF estimation can theoretically accommodate distributional violations associated with ordinal item responses, large sample sizes are needed to achieve the desirable asymptotic properties of these estimators (Anderson & Gerbing, 1988). Categorical diagonally weighted least squares (DWLS; Muthén, du Toit, & Spisic, 1997) and categorical unweighted least squares (ULS; Browne, 1974) are widely considered to be ideal estimation approaches for SEM in ordinal item response data (Yang-Wallentin, Jöreskog, & Luo, 2010). Unlike ADF and ML, these approaches use a polychoric correlation matrix that properly accounts for the ordered categorical nature of the indicators by assuming that a latent continuous variable, called a response variable, is discretized according to thresholds to produce the observed ordinal responses. Based on this assumption, a model is estimated that accounts for the correlations among the unobserved response variables, rather than the observed ordinal variables.

As with any statistical model, the accuracy of the results of an analysis of polychoric correlation depends on satisfying the assumption of normality for the underlying response variables. Several articles have investigated the robustness of the polychoric correlation to violations of this distributional assumption (Flora & Curran, 2004; Lee & Lam, 1988; Quiroga, 1992; Jin & Yang-Wallentin, 2017). While Flora and Curran (2004), Lee and Lam (1988), and Quiroga (1992) found that the polychoric correlation estimates based on the normality assumption

were fairly robust against violations of this assumption, a more thorough investigation by Jin and Yang-Wallentin (2017) has cast doubt on these results. These authors studied the misspecification of the underlying distribution in a general sense, including models that assumed non-normal underlying distributions, and found that when the underlying distribution is skewed, assuming an underlying skew-normal distribution during polychoric estimation better recovers the true correlation between the latent response variables compared to the standard normal distribution, albeit with problems estimating the parameters of the underlying distribution. In general, Jin and Yang-Wallentin (2017) found that when the underlying distribution differed substantially from the distribution used to estimate the polychoric correlations, the resulting estimates could be severely biased; specifically, the skew-*t*(4) distribution (Azzalini & Capitanio, 2003) and the pareto distribution (Mardia, 1962) were the most problematic and introduced substantial bias in estimates of the polychoric correlations. While prior studies (Flora & Curran, 2004; Lee & Lam, 1988; Quiroga, 1992) concluded that polychoric correlations were generally robust to discrepancies between the underlying distribution and what was assumed during estimation, the Jin and Yang-Wallentin (2017) study suggests that when the underlying distribution is very heavily kurtotic, as with the skew-*t*(4) distribution and the pareto distribution, the standard normality assumption for polychoric estimation can yield highly biased results. Considering that outliers also contribute to kurtosis, these results suggest that polychoric estimation may not be robust to outliers and potential influential cases. However, to my knowledge, no systematic study of the sensitivity of polychoric estimation, or categorical factor analysis or SEM, to aberrant cases has been performed. Furthermore, Flora and Curran (2004, Table 2) showed that the effects of misspecification of the underlying distribution varied depending on the number of categories and the magnitude of the

polychoric correlations; thus, I expect the potential for case influence to depend on characteristics of the items.

1.4. This dissertation: Case diagnostics and influence in categorical factor analysis

Over the last three years, I have been working on extending the case diagnostics in SEM from Yuan and Zhong (2008) and Yuan and Hayashi (2010), specificially d_f, the factor-score-based M-distance, and d_r, the residual-based M-distance, to ordered categorical data (Mansolf & Reise, 2018). As will be discussed in the Method Section, d_f* and d_r* quantify leverage and residual, respectively, with respect to the unobserved latent response variables in the polychoric model based on the estimated thresholds and polychoric correlation matrix. This is done by treating the latent response variables as missing data and integrating the diagnostic functions over the expected conditional distribution of the latent response variables given an observed response pattern. While it would be straightforward to implement an IRLS algorithm for robust estimation in categorical factor analysis using these extensions of leverage and residual diagnostics, two questions remain as to the potential validity and utility of this algorithm. The goal of this dissertation is to address these two questions in order to motivate the development of categorical IRLS.

First, it is not clear whether such an algorithm would achieve the desired goal of using leverage and residual to down-weight potentially influential cases. IRLS in continuous data, as implemented in the literature, uses Huber-type weights, which down-weight cases when values of d_f or d_r exceed some *a priori* critical value based on their theoretical distributions. In categorical data, especially when the number of categories is low, these diagnostics can deviate substantially from their theoretical distributions in continuous data. Should *a priori* critical values and weighting functions from continuous data be applied to categorical data, and does their utility depend on the characteristics of the test? Practically speaking, this is a Type I error issue: in what

conditions do the *a priori* critical values give reasonable Type I error rates? In conditions in which they do not, alternative critical values or weighting schemes will need to be explored.

Second, it is not obvious that leverage (d_f) and residual (d_r) have the same relationships to influence (e.g., on model fit and parameter estimates) in continuous and categorical data. Yuan and Zhong (2008) found that, when no robust procedures are used, good leverage observations (high leverage, low residual) have a small effect on factor variances and covariances[3], while outliers (low leverage, high residual) and especially bad leverage observations (high leverage, high residual) influence model fit as well as parameter estimates. The goal of IRLS estimation is to use leverage and residual to down-weight potentially influential cases, and the success of this goal depends on the existence of these relationships, as d_f and d_r are used as proxies for potential influence in continuous IRLS. Do these relationships hold in categorical data, and how do they depend on the characteristics of the test? If, or when, they hold, the extension of IRLS to categorical data is justified, as d_f* and d_r* would suitably serve their roles as proxies for potential influence. In addition, this would justify use of continuous critical values in categorical data in those conditions; if the d_f* and d_r* in a given instrument never exceed these critical values, and case influence is similarly restricted, then the use of IRLS in such instruments would be not only ineffectual, but pointless in achieving the goal of IRLS. If there are conditions where these relationships are not comparable to those in continuous data, it is important to identify those conditions, as the statistical properties of categorical IRLS may differ in those conditions.

Influence diagnostics, specifically case deletion diagnostics, already exist for SEM and can be adapted without modification to categorical SEM. Measures of residual and leverage in

[3] In this work, all models were identified by standardizing the latent variable, and thus good leverage observations should, based on the findings of Yuan and Zhong (2008), influence factor loading estimates rather than factor variances and covariances.

categorical factor analysis have also been developed and will be introduced in detail in the next section. All of these diagnostics can be computed easily in R.

In this dissertation, I will attempt to answer the two questions outlined above by examining the distributions of, and relationships between, leverage (d_f), residual (d_r), and influence in continuous and categorical factor analysis under a variety of test conditions using simulation studies. The two objectives presented above will help to determine the conditions in which robust estimation for polytomous data is necessary and appropriate and how the characteristics of a test, specifically the number of items, number and placement of item thresholds, and factor loadings, affect these distributions and relationships. Ultimately, I aim to apply this program of research to the development of robust estimators for ordered categorical (dichotomous or polytomous) data, which will greatly expand the applicability of these robust procedures within psychology and the social sciences. This dissertation research will serve to motivate this larger program of research by identifying the conditions in which IRLS in categorical factor analysis would be fruitful. Just as research into regression diagnostics preceded the development of robust estimation in regression, this research precedes the development of robust estimation for categorical factor analysis. Additionally, the study of case diagnostics for categorical data will yield tangible results even without the associated robust estimation. Researchers should be informed of the potential for their analysis results to be distorted by aberrant observations and the statistical properties and limitations of the tools available for identifying such observations.

The remainder of this dissertation will begin with a discussion of case diagnostics in categorical factor analysis. First, diagnostics for leverage and residual (d_f and d_r) in continuous data will be reviewed, followed by the extension of these diagnostics to categorical data using the polychoric model and associated computational issues. Next, case influence diagnostics in SEM

applicable to categorical factor analysis will be reviewed. Next, two simulation studies will be presented. In the first simulation study, I determined empirical *p*-values and critical values for the categorical case diagnostics d_f* and d_r* (Mansolf & Reise, 2018). In the second simulation study, I examined the relationships between these two diagnostics and two case influence diagnostics, $\Delta\chi^2$ and gCD_λ, in categorical factor analysis and compared those relationships to those in continuous factor analysis. This dissertation concludes with a general discussion describing the implications of these results for future applied and methodological work in categorical factor analysis and item response theory.

Chapter 2 – Case Diagnostics

2.1. Residual and Leverage: Four Mahalanobis Distance Measures in Continuous Data

Yuan and Zhong (2008) describe four types of M-distances that can be defined for SEMs: d_c, d_s, d_f, and d_r (see also Yuan, Fung, & Reise, 2004). Although these M-distances tend to be correlated to varying degrees, each has a unique interpretation. In general, an M-distance takes the form

$$d^2 = (\boldsymbol{\alpha}_i - \bar{\boldsymbol{\alpha}})' \boldsymbol{\Psi}^{-1} (\boldsymbol{\alpha}_i - \bar{\boldsymbol{\alpha}})$$

where $(\boldsymbol{\alpha}_i - \bar{\boldsymbol{\alpha}})$ is some vector measuring discrepancy and $\boldsymbol{\Psi}^{-1}$ is the inverse of the covariance matrix of $(\boldsymbol{\alpha}_i - \bar{\boldsymbol{\alpha}})$. M-distances can be interpreted as multivariate Z-scores, where the minimum possible value is zero, the expected value is based on the degrees of freedom of $(\boldsymbol{\alpha}_i - \bar{\boldsymbol{\alpha}})$, and higher values indicate increasing degrees of discrepancy. By condensing the information in the vector $(\boldsymbol{\alpha}_i - \bar{\boldsymbol{\alpha}})$ into a scalar, M-distances allow an investigator to quickly identify highly discrepant cases. One can examine either the squared M-distances d^2 or their square root d to determine case discrepancy; for simplicity, we provide formulas for d^2 only, although our discussion and simulation will focus on d.

The most straightforward M-distance in structural equation modeling, d_c, is calculated using the sample mean and covariance matrix $\bar{\boldsymbol{x}}$ and $\boldsymbol{S}$, and measures the discrepancy between a case and the saturated model in SEM:

$$d_{ci}^2 = (\boldsymbol{x}_i - \bar{\boldsymbol{x}})' \boldsymbol{S}^{-1} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}).$$

This M-distance is very similar to d_s, which simply exchanges the sample mean and covariance matrix with their model-implied counterparts based on an estimated model:

$$d_{si}^2 = \big(\boldsymbol{x}_i - \boldsymbol{\mu}(\boldsymbol{\theta})\big)' \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} \big(\boldsymbol{x}_i - \boldsymbol{\mu}(\boldsymbol{\theta})\big)$$

To the extent that the estimated (i.e., structured) and saturated models (i.e., sample mean and covariance matrix) agree, d_c and d_s will be very highly correlated, and a discrepancy between d_c and d_s likely indicates a major model misspecification. Both d_c and d_s are evaluated on p degrees of freedom for models with p observed variables.

While d_c and d_s are useful in that they quantify a case's deviance from the bulk of the data in multivariate space, they are imperfect as measures of discrepancy between a case and predictions from an estimated model. Yuan and Hayashi (2010) identified that d_c and d_s in factor analysis are functions of both *leverage*, or how extreme a case is in the predictor space, and *outlyingness*, or how well (or poorly) a case is predicted by a model. For a typical structural equation model with a measurement portion consisting of confirmatory factor models and a structural portion containing regression paths among the factors, the two M-distances d_f and d_r quantify leverage and outlyingness, respectively.

The factor-score-based M-distance d_f quantifies how far a case is from the bulk of the data in the factor space, and is given by

$$d_{fi}^2 = (\boldsymbol{f}_i)' \boldsymbol{\Omega}_f{}^{-1} (\boldsymbol{f}_i).$$

where

$$\boldsymbol{f}_i = (\boldsymbol{\Lambda}^T \boldsymbol{\Theta}^{-1} \boldsymbol{\Lambda})^{-1} \boldsymbol{\Lambda}^T \boldsymbol{\Theta}^{-1} \left(\boldsymbol{x}_i - \boldsymbol{\mu}(\boldsymbol{\theta})\right)$$

is Bartlett's factor score estimate for case i and $\boldsymbol{\Omega}_f$ is the covariance matrix of Bartlett's factor score estimates (Lawley and Maxwell, 1971, pp. 106-112; Yuan & Hayashi, 2010). With p observed variables and q latent variables, $\boldsymbol{\Lambda}$ is the p by q matrix of factor loadings in the measurement model and $\boldsymbol{\Theta}$ is the p by p matrix of residual variances and covariances for the observed variables, where latent variables are assumed to have unit variance.

While d_f quantifies leverage in the factor score space, d_r quantifies outlyingness in the residual space. Residuals can be defined using Bartlett's factor score estimates as

$$\boldsymbol{e}_i = \left[\mathbf{I} - \boldsymbol{\Lambda}(\boldsymbol{\Lambda}^T\boldsymbol{\Theta}^{-1}\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}^T\boldsymbol{\Theta}^{-1}\right](\boldsymbol{x}_i - \boldsymbol{\mu}(\boldsymbol{\theta})\big).$$

where $\mathbf{I}$ is the p by p identity matrix. The residual vector $\boldsymbol{e}_i$ is of length p and its elements contain the residuals for the observed variables after controlling for the Bartlett factor score estimates. The covariance matrix of $\boldsymbol{e}_i$ is given by (Bollen and Arminger. 1991, eq. 21)

$$\boldsymbol{\Omega} = \boldsymbol{\Theta} - \boldsymbol{\Lambda}(\boldsymbol{\Lambda}^{\mathrm{T}}\boldsymbol{\Theta}^{-1}\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}^T.$$

However, this covariance matrix is rank-deficient and cannot be inverted to calculate a M-distance directly using $\boldsymbol{e}_i$. Let $\mathbf{A}$ be a p by (p-q) matrix whose columns are orthogonal to $\boldsymbol{\Theta}^{-1}\boldsymbol{\Lambda}$; such a matrix can be defined using the eigenvectors of $\boldsymbol{\Omega}$ corresponding to the ($p - q$) nonzero eigenvalues as columns. Then a residual-based M-distance using $\boldsymbol{e}_i$ can be calculated as (Yuan & Zhong, 2008)

$$d_{ri}^2 = (\mathbf{A}^T\boldsymbol{e}_i)^T(\mathbf{A}^T\boldsymbol{\Omega}\mathbf{A})^{-1}(\mathbf{A}^T\boldsymbol{e}_i).$$

Thus d_r measures the extent to which case i is an outlier. Cases with large d_f or large d_r may be different from those with large d_c or d_s and will also differ depending on the measurement model because different measurement models imply different sets of predictors and thus different factor score estimates and residuals. Yuan and Hayashi (2010) propose using scatterplots of d_f and d_r to identify cases with both high residuals and high leverage, which are most likely to be influential cases.

The Mahalanobis distance measures discussed above are based on normal distribution theory and are typically compared to critical values based on a chi-squared (d^2) or equivalently a chi (d) distribution. Therefore, the validity of these measures must be questioned when data are ordered categorical, as often occurs in the social sciences with self-report and assessment data, as these data often substantially deviate from multivariate normality.

2.2. Mahalanobis Distance Measures in Ordered Categorical Data

Let $\mathbf{Y}_{n,p}$ denote a data matrix of dimension *n* by *p*, where *n* is the sample size and *p* is the number of observed variables, and let $\boldsymbol{y}_i$ denote the *p*-vector of observations for case *i*. For simplicity, we assume that all variables in $\mathbf{Y}_{n,p}$ are measured at the ordinal level and that each ordinal variable y_j can take on values from 1 to $m_j, j = 1, \dots, p$. Extensions to combinations of ordinal and continuous variables are straightforward (Muthén, 1984). The standard statistical models for ordinal data in SEM assume that underlying each ordinal observation $\boldsymbol{y}_i$ is an unobserved, continuous, multivariate normally distributed observation $\boldsymbol{y}_i^*$, called a *response process vector*, assumed to be multivariate normally distributed with components having mean 0 and variance 1. Under the polychoric model, this latent response vector is discretized according to thresholds $\boldsymbol{\tau} = \{\tau_{jl}\}, j = 1, \dots, p,\ l = 1, \dots, m_j - 1$ such that

$$\begin{cases} \boldsymbol{y}_{ij} = 1 & if\ \tau_{j1} < \boldsymbol{y}_{ij}^* \leq\ \tau_{j2} \\ & \vdots \\ \boldsymbol{y}_{ij} = m_j & if\ \tau_{jm_j} < \boldsymbol{y}_{ij}^* \leq\ \tau_{j(m_j+1)} \end{cases}$$

By default, $\tau_{j1} = -\infty$ and $\tau_{j(m_j+1)} = \infty$. When data generated from this model are collected, only the ordered categorical $\boldsymbol{y}_{ij}$ values are observed.

Under these assumptions, the correlations among the $\boldsymbol{y}_i^*$ variables can be estimated using only the observed responses $\boldsymbol{y}_{ij}$ via maximum likelihood using the *p*-way contingency table of the ordinal responses; these correlations are called *polychoric correlations*. For a dataset $\mathbf{Y}_{n,p}$, all *p*(*p*-1)/2 polychoric correlations $\rho_{jk}, j = 2, \dots, p, k = 1, \dots, j-1$ are estimated and are used to construct a *polychoric correlation matrix* $\mathbf{S}^*$ with ones on the diagonal and the polychoric correlation between variables *j* and *k* at the [*j*,*k*] and [*k*,*j*] positions on the off-diagonal. A structural

equation model can then be estimated on the matrix of polychoric correlations, and the resulting model describes the relationships among the unobserved $\boldsymbol{y}_j^*$ variables.

The calculation of M-distances requires a covariance matrix for the observations and implicitly assumes that the variables used are continuous; thus, when using ordinal data, M-distances cannot be calculated. SEM with ordinal data involves modeling the underlying response process, which is assumed to have a normal distribution; therefore, we use the response process $\boldsymbol{y}_i^*$ to calculate M-distances. This is done by estimating the expected M-distance for $\boldsymbol{y}_i^*$ based on the ordinal response vector $\boldsymbol{y}_i$ by integrating over the region of the multivariate normal distribution defined by the ordinal response vector. Thus, for response vector $\boldsymbol{y}_i$, an individual's expected M-distance for $\boldsymbol{y}_i^*$ is given by

$$d^2(\boldsymbol{y}_i^*) = d^* = \int_{m_{y_{i1}-1}}^{m_{y_{i1}}} \int_{m_{y_{i2}-1}}^{m_{y_{i2}}} \dots \int_{m_{y_{ip}-1}}^{m_{y_{ip}}} f(\boldsymbol{y}_i^*) MD(\boldsymbol{y}_i^*) d\boldsymbol{y}^*.$$

where $f(\boldsymbol{y}_i^*)$ is the multivariate normal density function with mean zero and covariance matrix $\mathbf{S}^*$ and $MD(\boldsymbol{y}_i^*)$ is some M-distance measure on $\boldsymbol{y}_i^*$, such as d_c, d_s, d_f, or d_r. Monte Carlo integration, which permits the high-dimensional integration needed to determine M-distances for models with many ordinal items, is used to calculate the integral, although alternative integration techniques (e.g., quadrature, quasi-Monte Carlo) can be considered as well. In Monte Carlo integration, a large sample of observations is drawn from the region of the multivariate normal distribution bounded by the thresholds corresponding to the observed response pattern, as in the integral above. For each sampled observation, the quantity of interest, here the M-distance, is calculated, and the results are averaged across all sampled observation to yield the expected M-distance for the corresponding region.

This procedure yields the ordinal M-distances d_c^*, d_s^*, d_f^*, and d_r^*, each found by integrating the corresponding M-distance measure over the region of the multivariate normal distribution defined by the thresholds that bound the item response. These indices have similar interpretations to the corresponding M-distances in continuous data; d_c^* identifies general multivariate outliers with respect to the saturated (polychoric) correlation matrix, d_s^* does the same with respect to the model-implied (polychoric) correlation matrix, d_f^* is a measure of leverage (extremity in the predictor space), and d_r^* is a measure of residual, or the discrepancy between observed and expected values based on estimated factor scores.

If it is assumed that the observed item responses are generated according to the polychoric model, it is important to examine the extent to which M-distances calculated from ordinal data can identify observations that are aberrant or extreme with respect to the underlying response process variables. Clearly, much information is lost when continuous variables are discretized into categorical variables, and it is unrealistic to expect perfect correspondence between M-distances calculated before and after discretization. However, the extent of this correspondence can inform us on the power of M-distances based on categorical data to identify truly aberrant response patterns. For instance, with a small number of categories, and with thresholds values close to zero, it is unlikely that any categorical response pattern will be highly discrepant from a model, whereas with many varied thresholds, discrepant response patterns will be easier to detect. Indeed, the power to detect person-fit in item response theory is influenced by such factors as test length, the spread of item locations (here, category thresholds) and item discrimination (Ferrando, 2004; Molenaar & Hoijtink, 1990; Reise & Due, 1991). Thus, to the extent that the M-distance d_r^* functions as a person-fit index, the effects of such factors on M-distance recovery must be examined as well.

2.3. Computation Alternatives in M-distance Estimation

The Monte Carlo simulation required to calculate the M-distances in ordered categorical data represents a non-ignorable computation burden. As we will explain, there are several ways to compute these diagnostics, each with distinct advantages and disadvantages.

The estimation of M-distances d_c^*, d_s^*, d_f^*, and d_r^* in ordered categorical data can be viewed as a missing data problem: given observed categorical responses $\boldsymbol{y}$, we use Monte Carlo techniques to integrate the M-distance functions over the expected conditional distribution of the unobserved continuous response vector $\boldsymbol{y}^*$ to obtain the categorical M-distances for $\boldsymbol{y}^*$ (*full MC* approach). An alternative estimation procedure involves estimating the latent response vector $\boldsymbol{y}^*$ for each observation as the mean of the region of the multivariate normal distribution bounded by the thresholds, and then treating the estimate $\hat{\boldsymbol{y}}^*$ as a continuous response vector when calculating M-distances (*latent mean* approach). The latent mean approach has the computational advantage of not requiring M-distances to be calculated for all Monte Carlo draws; however, the expected M-distance will not be the same as the M-distance corresponding to the expected latent response vector, and thus the results of the two approaches can potentially differ. For the simulations below, we used the latent mean approach with separate factor score estimates for each Monte Carlo-sampled latent response vector.

For the residual-based M-distance d_r*, one may also choose to integrate the residual vector with respect to the unconditional distribution of $\boldsymbol{y}^*$ as described above (*unconditional* approach), or to integrate with respect to the conditional distribution of $\boldsymbol{y}^*$ given fixed factor score estimates, where the factor score estimates are EAP estimates from the categorical factor model (*conditional* approach). These approaches differ little in computational burden; while using fixed factor scores reduces the burden of computing these factor scores for each Monte Carlo draw, it also requires

determining factor scores for the categorical factor analysis model using EAP estimation or some other technique, with a computational burden of its own. The parameters of the distribution used to integrate $\boldsymbol{y}^*$ are fairly trivial to compute; for fixed factor score vector $\boldsymbol{f}$, the conditional mean of $\boldsymbol{y}^*$ is given by $\boldsymbol{\Lambda f}'$ and the conditional covariance matrix is given by $\boldsymbol{\Theta}$. Thus, the *conditional* approach requires only marginally less computation than the *unconditional* approach. In total, there are four unique ways to estimate M-distances in ordered categorical data: *full MC* or *latent mean* integration, with the *conditional* or *unconditional* distribution of $\boldsymbol{y}^*$. For this dissertation, the *full MC* with *conditional* $\boldsymbol{y}^*$ will be used.

2.4. Measures of Influence in Structural Equation Modeling

Pek and MacCallum (2011) discuss several case deletion diagnostics that generalize regression-based influence measures to SEM. While some of the measures are directly drawn from regression, others are unique to SEM. Importantly, these measures generalize directly to categorical factor analysis, as they all simply involve estimating a model with and without a case included in the sample and examining the effect of case deletion on model statistics. These measures will be used in this dissertation to quantify case influence.

A direct carry-over from regression is the likelihood distance (Cook, 1977, 1986; Cook & Weisberg, 1982):

$$LD_i = 2\left[L\left(\widehat{\boldsymbol{\theta}}\right) - L\left(\widehat{\boldsymbol{\theta}}_{(i)}\right)\right]$$

where $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}_{(i)}$ denote the k by 1 vectors of estimated model parameters for the same model based on the original sample ($\widehat{\boldsymbol{\theta}}$) and the sample with the *i*'th case deleted ($\widehat{\boldsymbol{\theta}}_{(i)}$), $i = 1, \ldots, N$. While this likelihood distance generalizes to the polychoric model, it does not relate to impact on model fit directly, as the model is generally evaluated using estimators other than direct maximum likelihood. Typically, the polychoric model is estimated using a three-stage procedure

(Lee, Poon, & Bentler, 1990; Muthén, 1984), in which thresholds are estimated based on univariate marginal proportions, then polychoric correlations are estimated based on bivariate marginal proportions, and finally the model is estimated using some variant of an ADF/GLS estimator (GLS; Browne, 1984; DWLS; Muthén, du Toit, & Spisic, 1997; ULS; Browne, 1974). These estimators attempt to find model parameters that minimize the following fit functions. Let $\widehat{\boldsymbol{\theta}}$ again be the vector of model parameters, and for simplicity assume no mean or threshold structure is imposed, reducing the problem to estimating a covariance structure only. Let $\boldsymbol{\sigma}(\widehat{\boldsymbol{\theta}})$ = vech($\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})$) denote the vector containing the elements in the lower triangle of the polychoric correlation matrix (not including the diagonal, which is constrained to 1 for polychoric correlation matrices) implied by the model parameters $\widehat{\boldsymbol{\theta}}$, and similarly let $\boldsymbol{s}$=vech(**S**) denote the vector containing the elements in the lower triangle of the sample polychoric correlation matrix. Lastly, let **W** denote the asymptotic covariance matrix of the estimates of the polychoric correlations in **s**, and let $\mathbf{I}_q$ denote the q by q identity matrix, where for p observed variables, $q = p(p\text{-}1)/2$. Then the fit functions for GLS, DWLS, and ULS are given by

$$F_{GLS} = \left(\boldsymbol{s} - \boldsymbol{\sigma}(\widehat{\boldsymbol{\theta}})\right)' \mathbf{W}^{-1} \left(\boldsymbol{s} - \boldsymbol{\sigma}(\widehat{\boldsymbol{\theta}})\right)$$

$$F_{DWLS} = \left(\boldsymbol{s} - \boldsymbol{\sigma}(\widehat{\boldsymbol{\theta}})\right)' \left(\text{diag}(\mathbf{W})\right)^{-1} \left(\boldsymbol{s} - \boldsymbol{\sigma}(\widehat{\boldsymbol{\theta}})\right)$$

$$F_{ULS} = \left(\boldsymbol{s} - \boldsymbol{\sigma}(\widehat{\boldsymbol{\theta}})\right)' \mathbf{I}_q \left(\boldsymbol{s} - \boldsymbol{\sigma}(\widehat{\boldsymbol{\theta}})\right)$$

Each of these functions attempts to minimize the (weighted) discrepancy between the observed and model-implied polychoric correlation matrix; in brief, F_{GLS} uses the inverse of the full asymptotic covariance matrix, F_{DWLS} uses only the diagonal of the full asymptotic covariance matrix, and F_{ULS} weights all polychoric correlations equally during estimation. In this dissertation, F_{DWLS} was used for estimation.

Once the model is estimated, model fit is evaluated using the test statistic

$$T = (N - 1) * \hat{F}.$$

where $\hat{F}$ is the fit function used to estimate the model evaluated at the parameter estimates $\widehat{\boldsymbol{\theta}}$ which optimize the fit function. Of these fit functions, only F_{GLS} approximates a chi-square distribution with (q - k) degrees of freedom, but only under asymptotic sample sizes (Flora & Curran, 2004; see also Browne, 1982, 1984). However, using any of these estimators, a "chi-square" distance[4] $\Delta\chi^2$ can be calculated as

$$\Delta\chi 2 = T_{(i)}^2 - T^2$$

where T and $T_{(i)}$ denote test statistics for the same model based on the original sample (T) and the sample with the i'th case deleted ($T_{(i)}$) respectively, i = 1, …, N. Note that here, no correction (e.g., Satorra and Bentler, 2001, 2010) is made to these test statistics.

Generalized Cook's distance (*gCD*; Cook, 1977, 1986) can be written for categorical factor analysis as follows:

$$gCD_i = \left(\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_{(i)}\right)' \widehat{VAR}\left(\widehat{\boldsymbol{\theta}}_{(i)}\right)^{-1} \left(\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_{(i)}\right)$$

where $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}_{(i)}$ are defined as above and $\widehat{VAR}\left(\widehat{\boldsymbol{\theta}}_{(i)}\right)$ is the estimated asymptotic variance-covariance matrix of the parameter estimates obtained in the reduced sample (Pek and MacCallum, 2011). Note that gCD_i quantifies the total change in all parameter estimates; in this study, I will consider only generalized Cook's distance for factor loadings, denoted gCD_λ.

[4] In this study, the resulting distance is not distributed as χ^2 nor as a difference of χ^2 variates; nevertheless, this notation is used to agree with prior literature (Pek and MacCallum, 2011; Pastore & Altoe, 2018).

Chapter 3 – Simulation Studies

3.1. Common Simulation Details

To achieve the two research goals outlined above, two simulation studies were conducted. These simulation studies involved (1) simulating data from a continuous factor analysis model, (2) calculating case diagnostics (residual d_r, leverage d_f, influence $\Delta\chi^2$, gCD_λ) based on the continuous factor model (influence diagnostics in second simulation only), (3) categorizing the continuous data according to pre-specified thresholds to generate ordered categorical data consistent with the tetrachoric/polychoric model, and (4) calculating case diagnostics (residual d_r*, leverage d_f*, influence $\Delta\chi^2$, gCD_λ) based on the categorical factor model for cases in the simulated data set (influence diagnostics in second simulation only). The simulation conditions corresponded to characteristics of the simulated test and were as follows: number of items (p = 5, 20), factor loadings in the data-generating model (λ = .3, .7), and number and placement of thresholds, described below.

There were three conditions for number and placement of thresholds: two for dichotomous data ("narrow" and "wide" threshold conditions), and one for polytomous data. In the "narrow" dichotomous data (one threshold) condition, the position of item thresholds varied across items, with the first item's threshold at -0.5, the last (5^{th} or 20^{th}) item's threshold at 0.5, and thresholds for intermediate items (2^{nd} to 4^{th} or 2^{nd} to 19^{th}) positioned in regular increments between -0.5 and 0.5 (e.g., thresholds of -0.5 for Item 1, -0.25 for Item 2, 0 for Item 3, 0.25 for Item 4, 0.5 for Item 5 for a five-item test). In the "wide" dichotomous data condition, thresholds were evenly spaced across items from -1.5 to 1.5 in a manner identical to the "narrow" dichotomous data condition. These dichotomous data conditions were intended to simulate a typical dichotomous test in which item threshold parameters, which are related to item difficulty parameters in item response theory

and determine the proportion of correct responses to each item, vary across items from "easy" items (negative threshold values) to "difficult" items (positive threshold values), In the polytomous condition, all items had the same four thresholds, set at τ = -1.5, -0.5, 0.5, and 1.5, resulting in five item categories; this condition mimics the "symmetric" threshold condition in Rhemtulla, Brosseau-Liard, and Savalei (2012).

These conditions resulted in a 2 (number of items) by 2 (factor loadings) by 3 (number and placement of thresholds) design for a total of 12 conditions. These conditions were used in both simulation studies.

3.2. Study 1 – Critical Values for d_f* and d_r* and Their Relationship to Influence

3.2.1. Method.

In the first simulation, I estimated empirical *p*-values for leverage (continuous d_f; categorical d_f*) and residual (continuous d_r; categorical d_r*) and compared the empirical critical values of these indices to the theoretical continuous critical values. The purpose of this simulation was to characterize how these *p*-values and critical values change depending on the properties of the test.

The simulation study proceeded as follows for each condition. First, a model-implied covariance matrix was generated according to the specified factor model. For simplicity, all factor loadings were equal and all observed variables were specified to have zero mean and unit variance in the data-generating model. I then simulated a large sample (n = 100,000) of multivariate normal cases from this population model. Note that, because of this large sample size, only a single simulated data set is needed to characterize empirical *p*-values and critical values. The data-generating model was estimated using normal-theory maximum likelihood estimation, and the *p*-value was extracted. The data were then categorized according to the threshold parameters in that

condition. In dichotomous conditions, categorization was done by recoding all values less than the threshold parameter for each item to 1 and all values greater than the threshold parameter to 2; see the previous section for a detailed description of threshold parameter specification in dichotomous tests. In the polytomous conditions, categorization was done by recoding all values less than the lowest threshold (τ = -1.5) to 1, all values between the two lowest thresholds (τ = -1.5 to τ = -0.5) to 2, and so on, with values greater than the highest threshold (τ = 1.5) recoded to 5. After categorization, the data-generating model was estimated on the categorical data using diagonally weighted least squares (DWLS) estimation with polychoric correlations, and the *p*-value was calculated.

Two checks were performed to ensure the integrity of the simulated data. First, in order to ensure that all threshold values are estimated, datasets must have contained at least one response in each response category for all items, ensuring that the same number of threshold parameters would be estimated in all conditions. Second, in order to ensure that the simulated data suitably represents the hypothesized model structure, datasets must have had *p*-values of at least .5 in the estimated continuous and categorical models. If a dataset failed to meet these requirements, another continuous dataset was generated with the same conditions and the categorization and estimation were repeated until these two conditions were satisfied. This quality check was used to ensure that the simulated data conformed to the data-generating model, which was important because only one large data set was used in each condition.

Once the data were simulated, d_f and d_r were calculated for models estimated on continuous data and d_f* and d_r* were calculated for models estimated in categorical data. Categorical case diagnostics d_f*and d_r* were calculated using 100,000 Monte Carlo draws. Empirical critical values (95%, 99%) for continuous and categorical leverage and residual were calculated, along with

empirical p-values based on the 95% and 99% critical values of the appropriate chi distribution for each index. Degrees of freedom for d_f and d_f* were 1 and degrees of freedom for d_r and d_r* were (p – 1) (Yuan & Bentler, 1998, 2000). These p-values and critical values were examined to answer the first research question regarding the utility of the theoretical continuous cutoffs in categorical data and were compared across conditions to determine the effect of test characteristic on these distributional properties.

3.2.2. Results and discussion.

Figures 1 and 2 contain histograms of d_f in continuous (Figure 1, top panels), polytomous (Figure 1, remaining panels), and dichotomous data (Figure 2), while Figures 3 and 4 contain the corresponding histograms for d_r and d_r*. Unlike d_f in continuous data, the polytomous and dichotomous distributions of d_f* vary by test condition (Figure 1). While a smaller number of items (p = 5) leads to a multimodal distribution due to the limited number of possible response patterns, the distribution of d_f* in longer polytomous tests (p = 20) is nearly identical in shape to the distribution of d_f in continuous tests. The distributions of d_f* in dichotomous data (Figure 2) are similar to those in continuous data for longer tests (p = 20), but for shorter tests these distributions are highly multimodal, reflecting the limited number of possible latent trait values that can be calculated based on a short, dichotomous test. In Figure 3, as in Figure 1, the distribution of d_r* in polytomous tests is similar to that of d_r in continuous data; while the 5-item polytomous tests contain the same multimodality in d_r* as observed in Figure 1, the distribution of d_r* in 20-item polytomous tests is nearly identical in shape to that of d_r in continuous tests. The distributions of d_r* in dichotomous tests (Figure 4), however, differ markedly from the distribution of d_r in continuous and polytomous data, most notably in the highly restricted range of d_r* across all conditions.

Table 1 contains empirical *p*-values for categorical d_f* and d_r*. Empirical *p*-values for continuous d_f and d_r were within reasonable ranges for $\alpha = .05$ (d_f: *MEAN* = .0499, *MIN* = .0491, *MAX* = .0508. d_r: *MEAN* = .0499, *MIN* = .0489, *MAX* = .0510) and $\alpha = .01$ (d_f: *MEAN* = .0100, *MIN* = .00959, *MAX* = .0104. d_r: *MEAN* = .00995, *MIN* = .00970, *MAX* = .0103), did not appear to depend on the test conditions, and will not be discussed further as they approximate their theoretical expectations.

In polytomous data, d_f* exhibited positively biased *p*-values, with bias decreasing as the number of items increased. These *p*-values were more biased when factor loadings were low (λ= .3); in contrast, when factor loadings were high ($\lambda = .7$) and the number of items was high ($p = 20$), empirical *p*-values were close to the theoretical continuous *p*-values ($p = .056$ for $\alpha = .05$; $p = .012$ for $\alpha = .01$). More extreme values of d_f* result from low factor loadings because when factor loadings are low, there is necessarily more unique item variance in item responses; in polychoric and tetrachoric models, the sum of the variance explained by the factor and by the item uniqueness must equal one, and an increase in one entails a decrease in the other. In the presence of high unique item variance, a higher factor score is needed to yield response patterns that are uniformly in the highest or lowest response category than when unique item variance is low or, equivalently, factor loadings are high. Thus, when attempting to predict factor scores from item responses, low factor loadings lead to more extreme (read: further from the mean) factor score estimates for responses with uniformly extreme responses than high factor loadings. This phenomenon has also been observed in IRT (Embretson & Reise, 2000, p. 170). The biases in *p*-values for d_f* due to factor loading magnitude are also evident in Figures 1 and 2, with low factor loadings ($\lambda = .3$) leading to a larger proportion of cases falling past the 95% and 99% theoretical critical values compared to continuous d_f, especially in the 5-item test condition.

As in the polytomous conditions, empirical *p*-values for d_f* in the dichotomous conditions were greater when factor loadings were low ($\lambda = .3$) than when factor loadings were high ($\lambda = .7$). When $-0.5 \leq \tau \leq 0.5$, *p*-values for d_f* were, in general, positively biased in dichotomous data when factor loadings were low and negatively biased when factor loadings were high. All empirical *p*-values were zero for conditions with $-0.5 \leq \tau \leq 0.5$, $\lambda = .7$ at $\alpha = .01$, and at $\alpha = .05$ for $0.5 \leq \tau \leq 0.5$, $\lambda = .7$, indicating that no simulated cases yielded d_f* past the theoretical 95% critical value in these conditions. When $-1.5 \leq \tau \leq 1.5$, *p*-values were all positively biased except for conditions with $-1.5 \leq \tau \leq 1.5$, $p = 5$, $\lambda = .7$ at $\alpha = .01$ (empirical *p*-value of zero) and $-1.5 \leq \tau \leq 1.5$, $p = 20$, $\lambda = .7$ at $\alpha = .05$ (empirical *p*-value of .037), with decreasing bias with increased test length.

Empirical *p*-values for d_r* were zero for all dichotomous conditions with narrow threshold values ($-0.5 \leq \tau \leq 0.5$). If values of d_r* are interpreted as measures of outlying-ness with respect to the latent response variables, and 95% or 99% theoretical critical values are treated as "objective" standards for identifying outlying-ness, these results indicate that it is nearly impossible to be an "outlier" in dichotomous data under these conditions; for conditions with empirical *p*-values estimated at zero, no cases out of samples of 100,000 could be categorized by these standards as "outliers". A small number of cases passed the 95% critical values of d_r* in dichotomous data when $-1.5 \leq \tau \leq 1.5$, but no cases passed the 99% critical value. For polytomous tests, empirical *p*-values for d_r* were negatively biased (~0.25 for $\alpha = .05$; ~.003 for $\alpha = .01$); this bias did not appear to depend on test conditions. In these cases, it is possible to be considered an outlier, but it is more difficult to be considered so than if the underlying response variables were observed directly.

This right-side truncation of the distribution of d_r* is also evident in Figure 4. The left-side truncation in Figure 4 is due to the fact that, in calculating d_r* in dichotomous data, d_r is averaged over truncated distributions bounded on one side, in all dimensions, by negative or positive

infinity; in other words, it is possible that any of the underlying response variables was extremely large or small, depending on whether the item was "correct" (score of 2) or "incorrect" (score of 1). Averaging over this possibility makes it essentially impossible to be abnormally "well-fitting" case (low d_r*) in dichotomous data, just as it is very difficult to be an "outlier" (high d_r*) according to the metrics considered here.

Table 2 contains empirical critical values for d_f* and d_r*. These critical values can be illuminating in demonstrating the magnitude of the discrepancy between empirical and theoretical critical values. To put these differences on a meaningful metric, I calculated the proportion of the theoretical χ distribution which falls below the corresponding empirical critical value for d_f* and d_r* (Q in Table 2). These Q values can be interpreted as the percentile of the distribution of continuous d_f and d_r that would be considered as "high-leverage" or "outliers", respectively, if the empirical critical values for categorical d_f* and d_r* were treated as the standards for identifying cases as such.

For most conditions, the critical values and Q values are directly related to the empirical p-values in Table 1, where higher p-values correspond to higher empirical critical values and higher Q values, and thus much of the information in Table 2 is redundant with Table 1. However, Q values for d_r* for the large set of conditions for which empirical p-values were close to or exactly zero provide additional information on the severity of truncation in the distribution of d_r when only categorical manifestations of continuous variables are observed (Figures 3 and 4). Specifically, for dichotomous items with $-0.5 \leq \tau \leq 0.5$, Q values for d_r* in Table 2 range from .65 to .75 for α = .05 and .70 to .84 for α = .01, indicating that observations which would be considered anomalous according to the empirical 95% critical value for d_r* would score at the 65-91th percentile when judged according to the theoretical critical value. For $-1.5 \leq \tau \leq 1.5$, dichotomous items again

exhibited low Q values (.74 to .81 for $\alpha = .05$; .86 to .91 for $\alpha = .01$), albeit not as low as when $-0.5 \leq \tau \leq 0.5$, Polytomous items demonstrated reasonable Q values close to the corresponding χ quantiles.

The low Q values observed for d_r* bear on the decision of which critical value to use (theoretical or empirical) for identifying cases as "high-leverage" or "outliers" in assessing the magnitude of these diagnostics for exploratory purposes or for setting criteria for down-weighting in a robust estimation procedure for categorical factor analysis. First, quantitative and applied researchers should note that, if critical values with nominal Type I error rates are to be based on the distribution of the categorical case diagnostic (d_f* or d_r*), this distribution needs to be simulated to determine this critical value empirically based on the test conditions, especially if the data are dichotomous. In contrast, if critical values are to be based on the theoretical distribution of the continuous case diagnostic (d_f or d_r), these critical values will not yield nominal Type I error rates in all test conditions; in polytomous data, the differences in Type I error rates are minimal, but these differences are substantial in dichotomous data. In the dichotomous case with narrow threshold values, using the theoretical cutoffs as criteria for "outlying-ness" would render it impossible for a response pattern to be considered an "extreme" based on d_r* unless the α level was set higher than .1. Depending on the researcher's perspective on definitions of response aberrance and the reliability of these diagnostics to quantify response aberrance, this may lead to the conclusion that such dichotomous tests have no utility in identifying outlying response patterns or that robust estimation would be fruitless in such test conditions. This issue will be revisited in General Discussion at the end of this dissertation.

3.3. Study 2 – Relationships Between Leverage, Residual, and Influence

3.3.1. Method.

In the second simulation, I examined the relationships between leverage (d_f*), residual (d_r*), and influence ($\Delta\chi^2$, gCD_λ) in categorical data in order to compare these relationships between categorical and continuous data and, where these relationships differed, to characterize them in categorical data. Most importantly, the goal of this simulation is to characterize how these relationships change depending on the properties of the test. Data simulation, categorization, and quality checks were identical to the first simulation but with a smaller sample size (n = 2,500) per condition.

Once the data were simulated, d_f, d_r, $\Delta\chi^2$, and generalized Cook's distance for factor loadings (gCD_λ) were calculated for the models estimated in categorical and continuous data separately. As in the first simulation, categorical case diagnostics d_f*and d_r* were calculated using 100,000 Monte Carlo draws. The reduced sample size in Study 2 was due to the increased computational burden of calculating the influence diagnostics, which requires re-estimating the model n times in each condition.

I then constructed diagnostic plots with d_r* on the x-axis and d_f* on the y-axis separately for categorical and continuous data in each condition in order to assess the relationship between leverage, residual, and influence in categorical data. The size of the points in these plots was scaled according to each point's influence as measured by each of the influence diagnostics ($\Delta\chi^2$, gCD_λ), with separate plots for each influence diagnostic.

Lastly, I estimated linear regression models to assess these results quantitatively, with d_f*, d_r*, and their interaction as independent variables and each influence diagnostic as a separate dependent variable, comparing the resulting regression coefficients across conditions. To make the

resulting models comparable across conditions, the following standardizations were performed. The categorical M-distances d_f* and d_r* were transformed to approximate a normal distribution using the *L* transformation defined in Canal (2005, p. 806). The influence diagnostics were standardized using a truncated mean and standard deviation, which were calculated using the set of all values of the corresponding influence diagnostic calculated in all conditions. The truncated mean and standard deviation were calculated by first removing any observations outside the range (Q_1 – 1.5*IQR, Q_3 + 1.5*IQR) for the corresponding influence diagnostic, where Q_1 is the first quartile, Q_3 is the third quartile, and IQR = Q_3 – Q_1, and then calculating the mean and standard deviation of the resulting set of influence diagnostics. This was done to remove the influence of severe outliers on the standardization. After these transformations, regression models were estimated using the *L*-transformed d_f, d_r, d_f*, and d_r* values (depending on whether the data were categorical or continuous) and their interaction as predictors and the standardized influence diagnostic as the outcome. In these models, severe outliers were included during estimation to aid in identifying conditions in which the relationship between leverage, residual, and influence would potentially be very strong.

3.3.2. Results.

The regression models predicting influence from leverage, residual, and their interaction are efficient summaries of the relationships between these quantities in continuous and categorical data. Parameter estimates for these models are presented in Table 3 ($|\Delta\chi^2|$) and Table 4 (gCD_λ). While the plots of leverage, residual, and influence (Figures 5 to 15) contain much richer knowledge about these relationships, there are too many of such plots to be displayed efficiently in this dissertation. In this section, I will discuss each influence measure by first describing the relationships revealed in the corresponding regression models and then referring to only a small

subset of the resulting plots: one continuous example per index, to illustrate the relationships between leverage, residual, and each influence diagnostic in the continuous case; and at least one categorical example per index, to illustrate these relationships in the categorical case and to draw attention to specific interesting cases revealed by the set of regression models. See Supplemental Materials for the complete set of plots. Based on the discrepancies between theoretical and empirical critical values in categorical data (Study 1), regression models in categorical data were estimated separately with both indices centered at their theoretical critical values and with both indices centered at the empirical critical values, allowing the first-order effects of leverage and residual for theoretical and empirical critical values to be compared.

3.3.2.1. Influence on model fit.

Inspection of graphs revealed that the magnitude of $\Delta\chi^2$ values, rather than their numerical values, depends on residual and leverage. Thus, the models presented here use the standardized (as described above) absolute value of $\Delta\chi^2$, denoted $|\Delta\chi^2|$, as the dependent variable. Models with $\Delta\chi^2$ as the dependent variable can be found in Supplemental Materials.

Table 3 contains regression coefficients for linear models predicting $|\Delta\chi^2|$ from leverage, residual, and their interaction. In continuous data, $|\Delta\chi^2|$ is primarily a function of residual d_r, with higher d_r values corresponding to higher influence on model fit and little effect of leverage d_f and little interaction. In short tests ($p = 5$), this effect is small (standardized b[5] $\approx .3$), while in long tests ($p = 20$), the effect is large (standardized $b \approx 1$). The relationship between leverage and $|\Delta\chi^2|$ in continuous data does depend on the number of items in the test, with more items yielding increased effects. Figure 5 displays an example of this effect for a 20-item test with $\lambda = .3$. In polytomous

[5] Because each variable was not standardized according to its own distribution, the symbol β was not used.

data, the effects of residual and leverage on $|\Delta\chi^2|$ are nearly identical to those in continuous data; Figure 6 illustrates these relationships in a 20-item polytomous test with $\lambda = .3$.

In dichotomous data, the relationships between leverage, residual, and $|\Delta\chi^2|$ vary widely across conditions. As in continuous and polytomous data, residual d_r* has a strong effect on influence on model fit in dichotomous data; however, unlike in continuous and polytomous data, this effect is still strong (standardized $b \approx 1.25$) in 5-item dichotomous tests when $-1.5 \leq \tau \leq 1.5$. Figure 7 illustrates this condition with $\lambda = .7$. The effect of residual d_r* on $|\Delta\chi^2|$ in long tests was also stronger in dichotomous than in polytomous data, with stronger effects for lower factor loadings. Figure 8 illustrates these effects for a 20-item dichotomous test with $-0.5 \leq \tau \leq 0.5$, $\lambda = .3$. When thresholds are modest ($-0.5 \leq \tau \leq 0.5$), small negative effects of d_f* were observed on $|\Delta\chi^2|$ in some dichotomous conditions, indicating that influence on model fit decreased with increasing leverage in these conditions (Figure 8). Lastly, there were small interaction effects (standardized $b \approx .3$) for dichotomous tests with high factor loadings. These effects were generally negative, indicating that increased d_f* decreased the influence of d_r* on $|\Delta\chi^2|$; Figure 9 demonstrates this interaction in a 20-item dichotomous test with $-1.5 \leq \tau \leq 1.5$, $\lambda = .7$. One exception arose in 5-item dichotomous tests with $-1.5 \leq \tau \leq 1.5$ and $\lambda = .7$, which had a positive interaction effect (Figure 7).

Lastly, across conditions, the first-order effects of leverage d_f* on $|\Delta\chi^2|$ tended to be stronger at the theoretical critical value than at the empirical critical value, while the first-order effects of residual d_r* on $|\Delta\chi^2|$ were generally unaffected by the critical value used for centering. These differences appeared when the empirical critical value for d_r* differed from the theoretical critical value, indicating that the effect of leverage on $|\Delta\chi^2|$ is stronger at the more extreme

theoretical critical value for d_r* than at the more modest empirical critical value. This difference is observed, in general, for the gCD_λ as well.

3.3.2.2. Influence on factor loading estimates.

Table 4 contains regression coefficients for linear models predicting gCD_λ from leverage, residual, and their interaction. In continuous data, gCD_λ is most strongly predicted by d_f, followed by d_r and their interaction. All coefficients were positive, indicating that influence on factor loadings increases high leverage and/or residual with a small superadditive interaction. These relationships increased in magnitude with increased factor loadings and increased numbers of items, the latter of which is due to the larger number of parameters used in the calculation of gCD_λ in long tests. Figure 10 illustrates these effects for a 20-item test with $\lambda = .7$. In polytomous tests, the same effects were observed, albeit smaller in magnitude; Figure 11 illustrates these effects for a 20-item polytomous tests with $\lambda = .7$.

In dichotomous tests, the relationship between leverage, residual, and influence on factor loading estimates varied substantially with all independent variables. In tests with wide-ranging thresholds ($-1.5 \leq \tau \leq 1.5$), these effects were similar to those in polytomous and continuous data, albeith with larger effects of residual d_r* than leverage d_f*; in fact, for tests with a small number of items there was almost no effect of leverage d_f* and no interaction. Figures 12 and 13 show these effects in 5-item and 20-item dichotomous tests, respectively, with $-1.5 \leq \tau \leq 1.5$ and $\lambda = .7$. The relationships between leverage, residual, and influence for 5-item dichotomous tests with narrow thresholds ($-0.5 \leq \tau \leq 0.5$) were similar to those for 5-item dichotomous tests with wide-ranging thresholds ($-1.5 \leq \tau \leq 1.5$); however, the relationships for 20-item dichotomous tests with narrow thresholds ($-0.5 \leq \tau \leq 0.5$) were unique to those conditions. In 20-item dichotomous tests with low ($\lambda = .3$) factor loadings, the effect of d_r* and the interaction effect were negative,

indicating reduced influence on factor loading estimates with increased residual, an effect that increased with increasing leverage. Figure 14 illustrates these effects; note the curvilinear relationship between d_f* and d_r*, which results in the highest-leverage points (which are the most influential) having the lowest residuals. In contrast, all coefficients were large and positive in 5-item dichotomous tests with $-0.5 \leq \tau \leq 0.5$ and $\lambda = .7$, with the effect of residual being the largest. Figure 15 illustrates these effects; note the restricted range of both d_f* and d_r*, which makes these relatively small effects seem larger when quantified as regression coefficients.

3.3.3. Discussion.

These results indicate that across a variety of test conditions, the relationships between leverage, residual, and influence in factor analysis differ between continuous and categorical data and that these relationships depend heavily on test conditions. In continuous data, these relationships are generally stable except for the variability in the sign of $\Delta\chi^2$; high-residual cases influence model fit, while cases with high leverage have the most influence on factor loading estimates, with high-residual cases exerting some influence as well. Leverage and residual have similar relationships to influence in polytomous tests as in continuous tests according to the diagnostics considered here, although the magnitude of these relationships tends to be slightly lower in polytomous data. In dichotomous data, especially when thresholds are far from the mean, the relationships between leverage, residual, and influence can differ dramatically, both from continuous/polytomous tests or from dichotomous tests with other test conditions. In real data, where items vary in their threshold parameters and factor loadings and when these values are sometimes unpredictable *a priori*, these effects are likely to be unpredictable but strong, an unfriendly combination for practitioners.

These findings have implications for detection of aberrant response patterns; in categorical factor analysis with dichotomous indicators, cases with high leverage, residual, or both can have a wide variety of sometimes unpredictable effects on model fit and parameter estimates. If a researcher's goal is to remove cases with influence on particular model quantities in categorical factor analysis, he or she should use a targeted approach based on the case diagnostic specifically relevant to that influence goal, with the understanding that he or she is intentionally manipulating the data to achieve their desired statistical results, a practice which is generally frowned upon. If researchers instead wish to remove or down-weight cases with high residual or leverage, for example in a robust procedure, they should be aware that this approach may not yield the same effects in all test conditions and may not solve a given statistical problem, such as distorted parameter estimates or aberrant model fit results. Alternatives and recommendations are given in the General Discussion below.

Chapter 4 – General Discussion

This dissertation examined the distributions of, and relationships between, case diagnostics in categorical factor analysis under a variety of test conditions to assess the potential utility and behavior of a case-robust categorical iteratively reweighted least squares (IRLS) estimator. Three types of case diagnostics were investigated: leverage diagnostics (d_f in continuous data, d_f* in categorical data), which quantify a case's potential for influencing parameter estimates; residual diagnostics (d_r in continuous data, d_r* in categorical data), which quantify the difference between the observed and model-predicted response patterns; and leverage diagnostics which quantify a case's influence on model fit ($\Delta\chi^2$) and factor loading estimates (gCD_λ). Case diagnostics quantifying leverage and residual in categorical factor analysis are relatively new, and although it would be fairly easy to implement categorical IRLS by simply substituting categorical case diagnostics (Mansolf & Reise, 2018) and weighting operations (Asparouhov, 2005) into the IRLS estimation functions for continuous data (Yuan & Bentler, 2000), such an approach is ill-advised without first investigating the distributions of the categorical case diagnostics d_f* and d_r* and their relationships to case influence. This dissertation represents that investigation.

One concern with using categorical case diagnostics d_f* and d_r* in robust estimation is determining how they might be used to down-weight "extreme" or "outlying" cases in an IRLS estimator. In continuous data, the statistics d_f and d_r have known distributions under standard assumptions (multivariate normality, properly specified model), and thus one can simply determine the theoretical critical values of d_f and d_r and down-weight cases that fall beyond those critical values. The first goal of this dissertation was to determine whether these same critical

values could be used to down-weight cases according to a given Type I error rate, and if not, how the critical values of d_f* and d_r* differ based on test characteristics.

To this end, I examined empirical *p*-values and critical values of the leverage index d_f and the residual index d_r, as well as their categorical counterparts d_f* and d_r*, in the first simulation study. While the empirical *p*-values and critical values of the continuous indices d_f and d_r conformed well to their theoretical properties, the empirical *p*-values and critical values of categorical d_f* and d_r* did not correspond to those of their continuous counterparts. The distribution of the leverage diagnostic d_f* differed considerably depending on test conditions and only approximated the distribution of d_f under a small number of combinations of test conditions. The empirical *p*-values and critical values of d_f* were, however, similar to those of d_f, and based on these results it is reasonable to judge values of d_f* by using the distribution of d_f as a reference distribution. In long polytomous tests d_r* approaches the same distributional properties as d_r, but in dichotomous tests and short polytomous tests very few simulated cases had d_r* values past the 95% theoretical critical value. In other words, in these conditions, it is very difficult to be an "outlier" according to these metrics. Empirical critical values for d_r* were much lower than the theoretical values, corresponding to roughly the 65-90th percentile of the continuous distribution of d_r, and the left-hand side of the distribution of d_r* was similarly truncated, with very few cases having "low" residuals according to the distribution of d_r. If one defines an "outlier" as a high-residual case relative to the distributions of latent response variables, one would conclude that it is very difficult to be an outlier in dichotomous data. Alternatively, if one defines an "outlier" as a high-residual case relative to the distribution of other cases, or relative to the potential for cases to be outliers, one would conclude instead that critical values based on continuous data are not useful for identifying outliers in dichotomous data.

These simulation results suggest potential difficulties in developing a robust estimator for categorical factor analysis which down-weights potentially influential cases analogously to existing IRLS estimators in continuous structural equation models (Yuan & Bentler, 1998, 2000). Specifically, such estimators generally rely on a "cut point" in the distribution of measures of discrepancy from the estimated model, which is often the critical value (95%, 99%, or other) of d_f or d_r. In categorical data, these cut-points depend on the properties of the test, raising the question of how to determine a measure of discrepancy to use in down-weighting cases.

For instance, consider a situation in which one wishes to construct a categorical IRLS estimator which down-weights cases according to the residual index d_r*. One option is to use the theoretical cut-points based on the known critical values of continuous d_r, which would result in a mismatch between the theoretical and actual percentage of cases that would be down-weighted. The biggest potential discrepancy is in dichotomous data; in dichotomous test conditions considered here, nearly no cases would have been down-weighted by an IRLS estimator using d_r* as the index for down-weighting and using the theoretical critical values of d_r as the criteria for down-weighting (Table 1). Such an approach would render a categorical IRLS estimator essentially useless in dichotomous data under the test conditions considered here.

Alternatively, one could use empirical critical values to determine the cut-points for identifying outliers. At the implementation level, this would require a real-data-based simulation study to determine the empirical critical values of d_r* given the test properties of the data set of interest. To be most precise, such a simulation would be required at each iteration of the IRLS algorithm because item parameter estimates, which influence the critical values of d_r*, change at each step of IRLS estimation. Another complication in implementing categorical IRLS is that, as shown in the second simulation, the relationships between leverage, residual, and influence vary

according to the test properties and the influence diagnostic used. In general, cases with high values of d_f* and/or d_r* tended to be influential, but sometimes the reverse was true; see Figure 14, where cases with the lowest residuals were the most influential on factor loading estimates. Therefore, an IRLS estimator applied to dichotomous data which uses empirical critical values may not down-weight the most influential cases as intended.

The truncated distribution of d_r* reflects similar distributional issues in item response theory, where researchers have used person-fit indices, corresponding roughly to the residual indices d_r and d_r*, to identify cases with response patterns that deviate from their expected values given an IRT model. Findings that the ostensibly standardized log-likelihood l_z did not follow a normal distribution (Drasgow, Levine, & Williams, 1985; van Krimpen-Stoop & Meijer, 1999), prompted the development of increasingly well-standardized versions of the index (Snijders, 2001; Sinharay, 2016). Similar efforts to standardize d_r* would require considerable computational labor considering the distributional (truncated distributions) and computational (high-dimensional integration) idiosyncrasies of the polychoric model. Such standardization may not be necessary in a pure measurement context considering the high correlation of d_r* with l_z (Mansolf & Reise, 2018), as researchers who need a standardized index could simply use l_z. However, the calculation of l_z only involves a measurement model whereas d_r* can be calculated using a full structural equation model, and there remains a place in the literature for well-standardized indices which incorporate both measurement and structural portions of an SEM.

A second concern with using categorical case diagnostics d_f* and d_r* in an iteratively reweighted estimator is that the effect of cases with high values of d_f* and d_r* on statistical results (model fit, parameter estimates) need to be understood in order to predict the effects of applying such an estimator. Because the distributions of d_f* and d_r* are truncated relative to continuous d_f*

and d_r*, and because of the non-independence of these diagnostics in categorical data, it is not obvious that a case with high d_f* and high d_r* would have the same effect on model results (worsened model fit, biased parameter estimates, "bad" leverage point) in categorical factor analysis as a case with high d_f and high d_r in continuous factor analysis. In short, if a categorical estimator is to be used to down-weight cases with high values of d_f* and/or d_r*, it is important to understand the potential effects of down-weighting those cases.

To this end, in the second simulation study I examined the relationships between leverage, residual, and influence to determine the effect of down-weighting cases with high d_f* and/or d_r* and compared these relationships to those in continuous data. In dichotomous tests, the bivariate relationship between d_f* and d_r* was distorted relative to continuous and polytomous data, with a very narrow spread of d_r* in those tests; in continuous data, d_f and d_r are independent by construction (Yuan & Hayashi, 2010). More problematically, the characteristics of highly influential response patterns differed between test conditions in dichotomous data. Influence on model fit generally increased with residual across test conditions in dichotomous data, although strong effects of leverage were observed in some dichotomous conditions, sometimes in the absence of effects of residual. In addition, the influence of high-residual cases on model fit in dichotomous data was generally higher than that of high-leverage cases, and in one case (20-item test, widely-spaced thresholds, high factor loadings; see Figure 13) low-leverage cases had the most influence on factor loading estimates. These relationships deviate markedly from those presented in Yuan and Zhong (2008) and those in the corresponding continuous conditions. From these results, we can conclude that relationships between leverage, residual, and influence, as operationalized here, deviate from conventional wisdom and past research when calculated in dichotomous data, and that these deviations depend, sometimes heavily, on test conditions.

Therefore, a case-robust IRLS estimator using d_f* and/or d_r* would have unpredictable behavior in dichotomous data.

In contrast, the findings of Yuan and Zhong (2008) generally held in continuous and polytomous tests, and the relationships between leverage, residual, and influence were roughly comparable between continuous tests and the polytomous test conditions studied in this work. Cases with high leverage had the strongest effect on factor loading estimates in continuous and polytomous data, followed by residual and with a small positive interaction, results which are consistent with Yuan and Zhong (2008). Additionally, empirical critical values of d_f* and d_r* in polytomous tests were close to the theoretical critical values of d_f and d_r. Unlike in dichotomous data, a categorical IRLS estimator would likely have similar statistical properties to continuous IRLS when applied to polytomous data.

These results should be taken with caution, however, because only a single polytomous item type (5 categories, symmetric and evenly spaced thresholds) was considered here. It is likely that case diagnostics for other polytomous item types would behave differently; for example, a test consisting of three-category items with thresholds of (1.25, 1.75) would likely behave nearly identically to a dichotomous test with a single threshold of 1.5; such test properties are rare, but do arise in clinical psychology when assessing psychiatric symptoms, for example in the Structured Clinical Interview for DSM-5 (SCID; First, 2014) which rates symptoms as “not present”, “unsure or equivocal”, or “present”. Additional research is needed to study the behavior of case diagnostics in irregular polytomous test conditions, and a well-behaved IRLS estimator which generalizes to dichotomous data would be necessary to accommodate such irregular polytomous tests.

The findings on case influence in this work add to the IRT literature assessing the effects of misfitting response patterns on model fit and parameter estimates. Consistent with the results presented here, other researchers have reported that contaminating data with careless or misfitting response patterns can lead to worse model fit (Hoijtink, 1987; Phillips, 1986), biased item parameter estimates (Clark, Gironda, & Young, 2003; Oshima, 1994; Wise, Kingsbury, Thomason, & Kong, 2004; van Barneveld, 2007) and biased latent trait estimates (De Alaya, Plake, & Imparta, 2001; Meijer & Sijtsma, 2001; Nering & Meijer, 1998). In a recent study using 40-item tests with relatively wide location parameters (*MEAN* = -0.11, *SD* = .90; Patton, Cheng, Hong, & Diao, 2019), an iterative procedure similar to categorical IRLS was used to remove misfitting response patterns from estimation, resulting in substantially reduced bias in item discrimination and location parameters in the two-parameter logistic IRT model. Although Patton et al. did not use a formalized IRLS estimator, which down-weights cases at each step of estimation, but simply iteratively removed cases with l_z values below a critical value from fully estimated models, these results illustrate the promise of more formalized estimation procedures which down-weight aberrant cases in categorical measurement models.

These results, and those referenced above, underscore the need to develop case-robust estimators for categorical factor analysis and structural equation modeling. At the moment, the only available option for practitioners interested in mitigating the effects of case influence in these models is to use case deletion diagnostics to assess the influence of each case on model results of interest. Unlike in regression, where some of these diagnostics can be determined analytically, in categorical factor analysis their calculation requires re-estimating the model of interest a number of times equal to the number of unique response patterns, each time removing one such unique

response pattern from the analysis. This presents a considerable computational burden for large data sets.

In addition, SEM packages vary in their implementation of these diagnostics. For example, the R package *influence.sem* (Pastore & Altoe, 2018) can calculate $\Delta\chi^2$, generalized Cook's distance, and likelihood distance, but cannot calculate *COVRATIO* or decompose generalized Cook's distance by parameter type. The R package *semdiag* (Yuan & Zhang, 2012) interfaces with the program EQS (Bentler & Wu, 2005) to calculate leverage and residual values d_f and d_r in continuous data only. Mplus (Muthén & Muthén, 1988-2017) allows the user to save Cook's distance, Mahalanobis distance d_c (for continuous observed variables only) and likelihood distance, which can be used to calculate $\Delta\chi^2$, but cannot perform case-robust estimation. If the model of interest is estimated using a software package which does not permit the straightforward calculation of the case diagnostic(s) of interest, some programming will be required to implement their calculation, presenting an additional burden to practitioners. Proliferation of software for calculating case diagnostics in SEM, many of which (importantly, case deletion diagnostics) generalize straightforwardly to categorical factor analysis, would aid researchers in understanding and mitigating the effects of influential cases until a reliable robust estimator is developed.

Like person-fit indices in item response theory, the indices evaluated herein can, in principle, be useful for practitioners interested in evaluating the validity of individual response patterns or for detecting aberrant response processes such as cheating or guessing. While this goal remains a primary motivator for the development of person-fit indices, attempts to apply these indices in practice and understand their relationships with psychological variables have yielded mixed results (Reise & Waller, 1993; Reise & Flannery, 1996. For instance, Birenbaum (1986) studied the relationships between person-fit values and scores on an anxiety test, a lie scale, and a

general cognitive ability scale and found that, contrary to expectation, person-fit values were most strongly correlated with general ability ($r \approx .50$), rather than with anxiety ($r \approx .14$) or the lie scale ($r \approx .10$). In contrast, Schmitt, Chan, Sacco, McFarland, and Jennings (1999) found that, in personality tests, test-taking motivation and conscientiousness correlated .26 and .34, respectively with the person-fit index l_{zm}, with lower correlations for cognitive tests. The person-fit index d_r* evaluated in this dissertation can be used for similar research where models are estimated and evaluated using categorical factor analysis rather than item response theory; however, such research must be undertaken with a clear understanding of what person-fit indices are supposed to measure from a psychological standpoint. Tellegen (1988) outlined a multitude of possible explanations for intraindividual inconsistency in item responses, and fruitful use of these indices in a pure measurement context requires the researcher to first understand the meaning of the indices with respect to the instrument-population combination under investigation.

The indices d_f* and d_r*, and the studies contained herein, augment this literature in two important ways. First, person misfit is operationalized here (d_r*) in terms of geometric distance from expectation, rather than by a low conditional probability of endorsement given the estimated latent trait value(s) (l_z and related indices). The simulations presented herein take advantage of the geometric interpretation of person-fit by enabling the comparison of the influence of misfitting cases to the influence of cases with high residuals in regression and structural equation modeling. This comparison adds an additional, highly practical utility to person-fit indices which had previously only been explored in principle: person-fit indices are diagnostic of a case's influence on model results in a manner that, in some conditions, is predictable from their geometric properties. Thus, researchers should be interested in cases which deviate from model results not only when testing hypotheses about the relationship between intra-individual response

inconsistency and other behaviors of interest (cheating, guessing, etc.) but as a diagnostic tool to ensure the validity of all inferences drawn from the models.

Second, these indices permit the evaluation of person-fit within the context of a full structural equation model, rather than simply a measurement model. A potentially fruitful future direction for person-fit research involves determining not only which individuals have aberrant response patterns, but also which individuals have latent trait values which do not predict external correlates well; in short, studying cases with high d_r* calculated from a full structural equation model. For instance, an individual with aberrant responses due to "sleeping" (responding poorly to the first few items on an ability test, but improving in performance over the course of the test) could be identified by including their previous and subsequent test scores in a single model; this individual would have high residual d_r* in a full structural equation model not only because their responses are inconsistent within a specific test, but because those responses are inconsistent with their behavior in other contexts. Likewise, the estimated latent trait values of cheaters, as predicted from responses to the exams on which they cheated, would not be expected to relate to other measures of studiousness (attendance, class participation, etc.) in the same way as with non-cheaters. Combining residual analysis of test scores (person-fit) with residual analysis of other variables in a full structural model may lead to higher power to detect specific behaviors. This approach can be useful for overcoming the well-documented low power of person-fit indices to identify responses generated according to a particular deviant response model (Meijer & Sijtsma, 2001; Meijer, 1996; Karabatsos, 2003).

As with all simulation research, the results presented here generalize best to the test conditions used in the simulations. Other test conditions may yield different results; for instance, a very long test with a wider spread of item response thresholds may yield distributions of d_f* and

d_r* which more closely approximate the theoretical distributions, or a test with threshold values far from the mean may yield empirical critical values that coincide better with the theoretical critical values based on continuous data. However, note that with long tests it becomes very computationally difficult to simulate from the small tail of the multivariate normal distribution bounded by the highest and/or lowest threshold values for each item. In previous versions of these simulations with threshold parameters of ±2, the Gibbs sampler often broke down when attempting to calculate d_r* and d_f* for extreme response patterns in 20-item tests. Thus, it may be difficult to calculate these diagnostics when response thresholds are extreme and/or there are many items, exactly those conditions in which outliers and high-leverage cases are expected to arise.

Furthermore, only a single polytomous item type, with four evenly spaced threshold parameters between -1.5 and 1.5, was examined here; skewed or asymmetrical threshold parameters may yield different results, as observed in Rhemtulla, Brosseau-Liard and Savalei (2012). More threshold parameters would also serve to "de-coarsify" the distributions of d_f* and d_r*, although items with more than five categories can usually be safely treated as continuous (Rhemtulla, Brosseau-Liard, & Savalei, 2012). All simulated data in this study had an underlying normal distribution in the latent response variables underlying each item; considering that skew and kurtosis of this distribution can affect model estimation (e.g., Roscino & Pollice, 2006), results may differ if the underlying normal distribution assumption is violated, as is likely to happen in real data.

In summary, three general observations about case diagnostics in categorical factor analysis can be made based on this research. First, leverage d_f* in these models is bounded, but can still take on large values. Second, residual d_r* in these models can seldom take on large (or, in the dichotomous case, small) values, even when considered across a variety of test conditions.

Lastly, the relationships between leverage, residual, and influence, while largely similar in continuous and polytomous data (in the polytomous test conditions considered here), can vary substantially in dichotomous data depending on the test conditions. While these results suggest that IRLS estimation in polytomous factor analysis will have similar statistical properties to IRLS estimation in continuous factor analysis, these findings complicate the extension of IRLS estimation to dichotomous factor models in two ways. First, because there are large and meaningful differences between the distributions of d_f* and d_r* in dichotomous data and the distributions of continuous d_f and d_r, it is difficult to determine criteria for down-weighting in IRLS. Second, because the relationships between leverage, residual, and influence vary depending on test conditions, it is not clear whether a hypothetical dichotomous IRLS estimator with a stable down-weighting rule would successfully reduce the impact of potentially influential cases comparably to a continuous IRLS estimator. Regardless, this work has revealed that there is considerable potential for case influence in categorical factor analysis, and that researchers should be taking steps to mitigate the effects of influential cases, whether by examining case deletion diagnostics or by developing estimators robust to these effects.

Tables

Simulation Conditions				d_f*		d_r*	
τ	m	p	λ	α = .05	α = .01	α = .05	α = .01
0.5	2	5	.3	.117	.085	.000	.000
0.5	2	5	.7	.000	.000	.000	.000
0.5	2	20	.3	.076	.013	.000	.000
0.5	2	20	.7	.067	.000	.000	.000
1.5	2	5	.3	.232	.020	.005	.000
1.5	2	5	.7	.072	.000	.004	.000
1.5	2	20	.3	.063	.011	.000	.000
1.5	2	20	.7	.037	.011	.001	.000
1.5	5	5	.3	.204	.100	.028	.003
1.5	5	5	.7	.084	.019	.028	.004
1.5	5	20	.3	.104	.031	.026	.002
1.5	5	20	.7	.056	.012	.026	.003

Table 1. Empirical *p*-values for categorical *M*-distances. m = number of item categories, p = number of items, λ = population factor loading. τ denotes the maximum threshold value for dichotomous and polytomous tests, and -τ denotes the minimum threshold value for dichotomous and polytomous tests. Cells are coded based on bias in *p*-value, where darker shades indicate increased bias (white indicates no bias), red indicates negative bias (darkest shade for zero) and green indicates positive bias (darkest shade for .1 for α = .05 or .02 for α = .01).

Simulation Conditions				d_f 95% CV			d_f 99% CV			d_r 95% CV			d_r 99% CV		
τ	m	p	λ	***The.***	*Emp.*	*Q*	***The.***	*Emp.*	*Q*	***The.***	*Emp.*	*Q*	***The.***	*Emp.*	*Q*
0.5	2	5	0.3	**1.96**	2.75	0.99	**2.58**	3.03	1.00	**3.08**	2.26	0.72	**3.64**	2.36	0.77
0.5	2	5	0.7	**1.96**	1.62	0.89	**2.58**	1.62	0.89	**3.08**	2.32	0.75	**3.64**	2.56	0.84
0.5	2	20	0.3	**1.96**	2.14	0.97	**2.58**	2.83	1.00	**5.49**	4.56	0.65	**6.02**	4.66	0.70
0.5	2	20	0.7	**1.96**	2.09	0.96	**2.58**	2.12	0.97	**5.49**	4.61	0.68	**6.02**	4.74	0.74
1.5	2	5	0.3	**1.96**	2.27	0.98	**2.58**	3.35	1.00	**3.08**	2.48	0.81	**3.64**	2.86	0.91
1.5	2	5	0.7	**1.96**	2.09	0.96	**2.58**	2.10	0.96	**3.08**	2.31	0.74	**3.64**	2.76	0.89
1.5	2	20	0.3	**1.96**	2.09	0.96	**2.58**	2.59	0.99	**5.49**	4.82	0.77	**6.02**	5.05	0.86
1.5	2	20	0.7	**1.96**	1.92	0.94	**2.58**	2.64	0.99	**5.49**	4.83	0.78	**6.02**	5.11	0.87
1.5	5	5	0.3	**1.96**	3.16	1.00	**2.58**	3.89	1.00	**3.08**	3.01	0.94	**3.64**	3.31	0.97
1.5	5	5	0.7	**1.96**	2.06	0.96	**2.58**	2.78	0.99	**3.08**	2.85	0.91	**3.64**	3.31	0.97
1.5	5	20	0.3	**1.96**	2.35	0.98	**2.58**	3.13	1.00	**5.49**	5.32	0.92	**6.02**	5.71	0.97
1.5	5	20	0.7	**1.96**	1.99	0.95	**2.58**	2.67	0.99	**5.49**	5.27	0.91	**6.02**	5.72	0.97

Table 2. Empirical critical values for categorical *M*-distances. m = number of item categories, p = number of items, λ = population factor loading, *The.* = theoretical critical value, *Emp* = empirical critical value, Q = quantile of χ distribution corresponding to empirical critical value. τ denotes the maximum threshold value for dichotomous and polytomous tests, and $-\tau$ denotes the minimum threshold value for dichotomous and polytomous tests. Columns of empirical critical values are color-coded based on bias in those values, where darker shades indicate increased bias (white indicates no bias), red indicates negative bias (darkest shade for one) and green indicates positive bias (darkest shade for ten). Columns of quantiles are color-coded based on bias in those values, where darker shades indicate increased bias (white indicates no bias), red indicates negative bias (darkest shade for .5) and green indicates positive bias (darkest shade for one).

Simulation Conditions				Continuous			Categorical					
				Theoretical CV Centered			Theoretical CV Centered			Empirical CV Centered		
τ	m	p	λ	d_f	d_r	d_f*d_r	d_f	d_r	d_f*d_r	d_f	d_r	d_f*d_r
0.5	2	5	0.3	-0.05	0.29	-0.02	-0.32	0.45	-0.16	-0.15	0.29	-0.16
0.5	2	5	0.7	0.04	0.26	0.01	***-0.85***	-0.24	-0.40	-0.47	-0.07	-0.40
0.5	2	20	0.3	-0.02	***1.09***	-0.03	0.11	***2.48***	0.09	-0.01	***2.50***	0.09
0.5	2	20	0.7	0.30	***1.33***	0.13	***-1.08***	***1.89***	-0.27	*-0.76*	***1.87***	-0.27
1.5	2	5	0.3	0.00	0.21	0.00	0.37	***1.18***	0.20	0.22	***1.24***	0.20
1.5	2	5	0.7	0.01	0.30	0.01	0.41	***1.46***	0.30	0.10	***1.50***	0.30
1.5	2	20	0.3	0.22	***1.27***	0.07	0.19	***3.06***	0.11	0.09	***3.07***	0.11
1.5	2	20	0.7	-0.18	***0.94***	-0.10	-0.50	***2.61***	-0.22	-0.30	***2.62***	-0.22
1.5	5	5	0.3	0.02	0.37	0.01	-0.02	0.33	-0.01	-0.02	0.32	-0.01
1.5	5	5	0.7	0.00	0.36	0.00	-0.03	0.14	-0.01	-0.03	0.14	-0.01
1.5	5	20	0.3	-0.24	***1.16***	-0.11	-0.16	***1.14***	-0.06	-0.15	***1.11***	-0.06
1.5	5	20	0.7	-0.14	***1.12***	-0.04	-0.04	*0.76*	0.03	-0.05	*0.76*	0.03

Table 3. Regression coefficients predicting $|\Delta\chi^2|$ from d_f, d_r, and their interaction. m = number of item categories, p = number of items, λ = population factor loading. τ denotes the maximum threshold value for dichotomous and polytomous tests, and -τ denotes the minimum threshold value for dichotomous and polytomous tests. Regression coefficients with magnitude less than .2 are presented in gray, coefficients with magnitude greater than .5 are presented in italics, and coefficients with magnitude greater than .8 are presented in boldface.

Simulation Conditions				Continuous Theoretical CV Centered			Categorical Theoretical CV Centered			Categorical Empirical CV Centered		
τ	m	p	λ	d_f	d_r	d_f*d_r	d_f	d_r	d_f*d_r	d_f	d_r	d_f*d_r
0.5	2	5	0.3	***0.98***	***0.83***	0.22	***1.18***	***1.59***	0.24	***0.93***	***1.83***	0.24
0.5	2	5	0.7	***1.33***	***0.94***	0.34	***2.60***	***4.34***	***1.12***	***1.52***	***3.88***	***1.12***
0.5	2	20	0.3	***4.50***	***2.18***	***0.82***	0.15	***-1.34***	***-1.30***	***1.76***	***-1.56***	***-1.30***
0.5	2	20	0.7	***4.76***	***2.41***	***0.98***	***7.80***	***10.87***	***3.62***	***3.46***	***11.23***	***3.62***
1.5	2	5	0.3	***0.95***	*0.76*	0.20	0.06	***2.19***	-0.27	0.27	***2.10***	-0.27
1.5	2	5	0.7	***1.39***	***0.92***	0.35	0.17	***4.56***	-0.07	0.24	***4.55***	-0.07
1.5	2	20	0.3	***4.46***	***1.97***	*0.74*	***4.64***	***5.07***	***1.22***	***3.54***	***5.18***	***1.22***
1.5	2	20	0.7	***4.40***	***1.91***	*0.73*	***1.78***	***5.88***	0.30	***1.52***	***5.88***	0.30
1.5	5	5	0.3	***1.03***	***0.88***	0.25	*0.63*	*0.71*	0.14	*0.61*	***0.91***	0.14
1.5	5	5	0.7	***1.39***	***0.97***	0.36	0.45	*0.78*	0.07	0.43	*0.79*	0.07
1.5	5	20	0.3	***4.09***	***1.80***	*0.63*	***3.67***	***1.93***	*0.68*	***3.47***	***2.22***	*0.68*
1.5	5	20	0.7	***5.11***	***2.55***	***1.06***	***2.90***	***2.59***	*0.69*	***2.70***	***2.62***	*0.69*

Table 4. Regression coefficients predicting gCD_λ from d_f, d_r, and their interaction. m = number of item categories, p = number of items, λ = population factor loading. τ denotes the maximum threshold value for dichotomous and polytomous tests, and -τ denotes the minimum threshold value for dichotomous and polytomous tests. Regression coefficients with magnitude less than .2 are presented in gray, coefficients with magnitude greater than .5 are presented in italics, and coefficients with magnitude greater than .8 are presented in boldface.

Figures

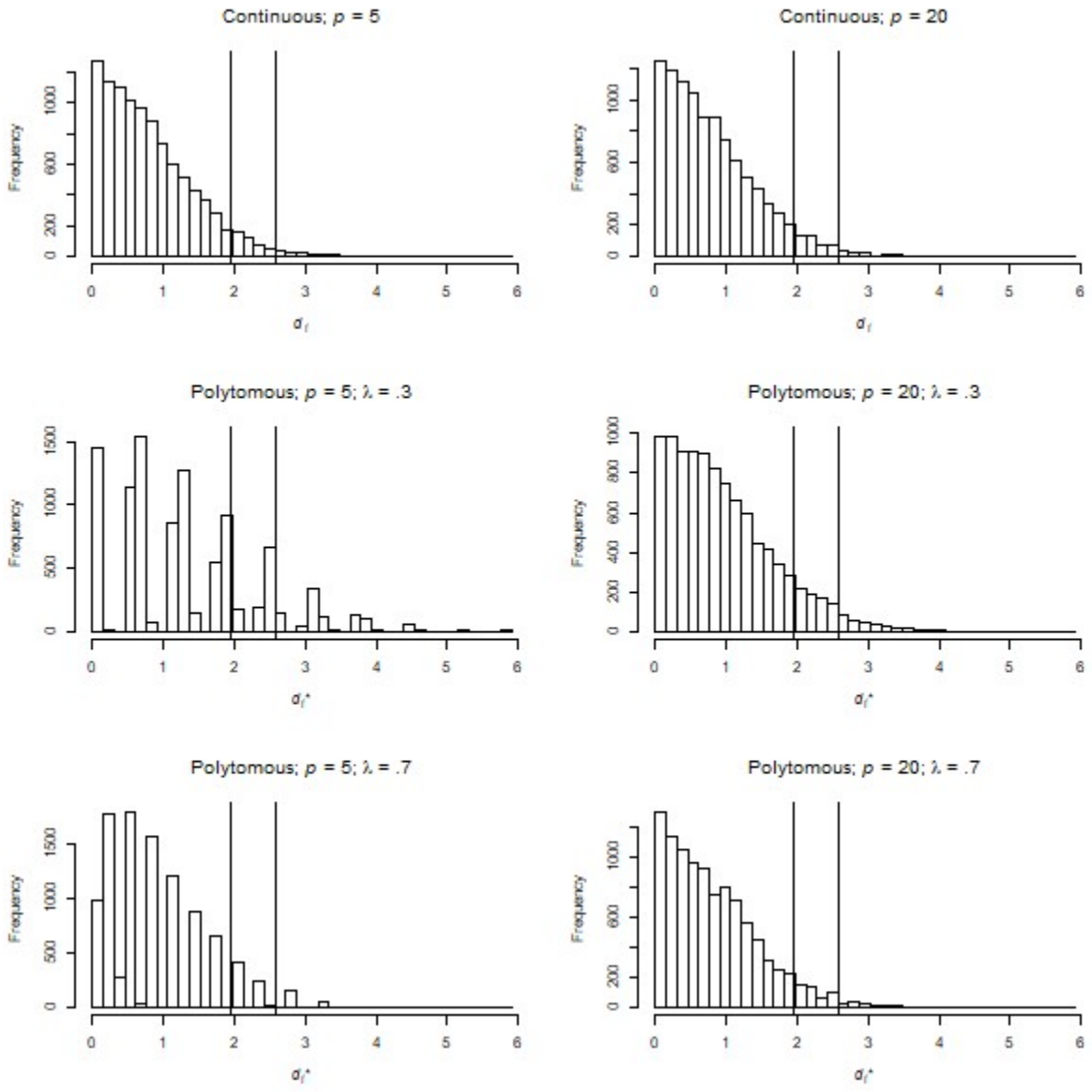

Figure 1. Histograms of d_f in continuous tests and d_f* in polytomous tests. p = number of items, λ = population factor loading. τ denotes the maximum threshold value, and -τ denotes the minimum threshold value. Vertical lines correspond to the 95% and 99% critical values of d_f based on a χ distribution with 1 degree of freedom.

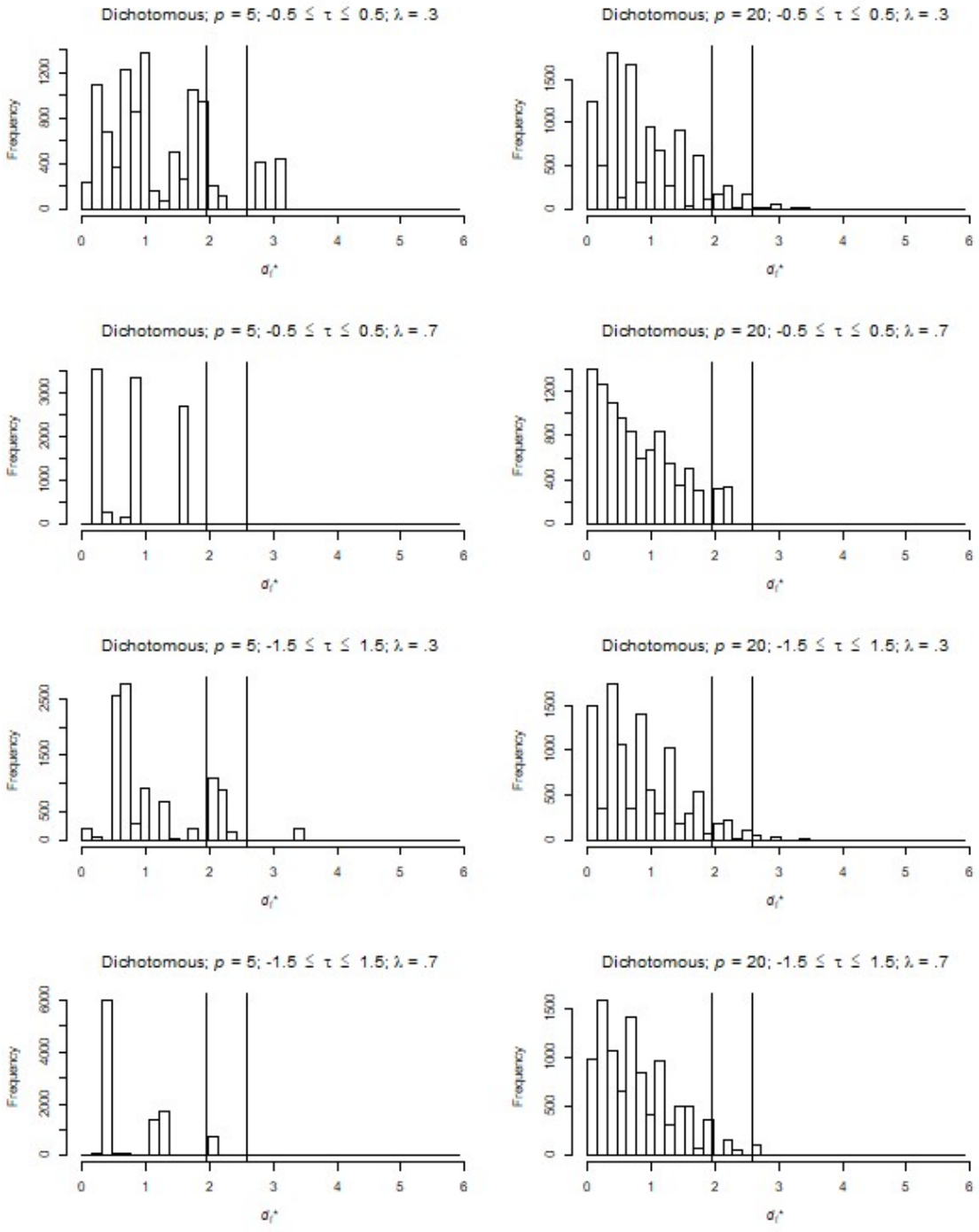

Figure 2. Histograms of d_f^* in dichotomous tests. p = number of items, λ = population factor loading. τ denotes the maximum threshold value, and $-\tau$ denotes the minimum threshold value. Vertical lines correspond to the 95% and 99% critical values of d_f based on a χ distribution with 1 degree of freedom.

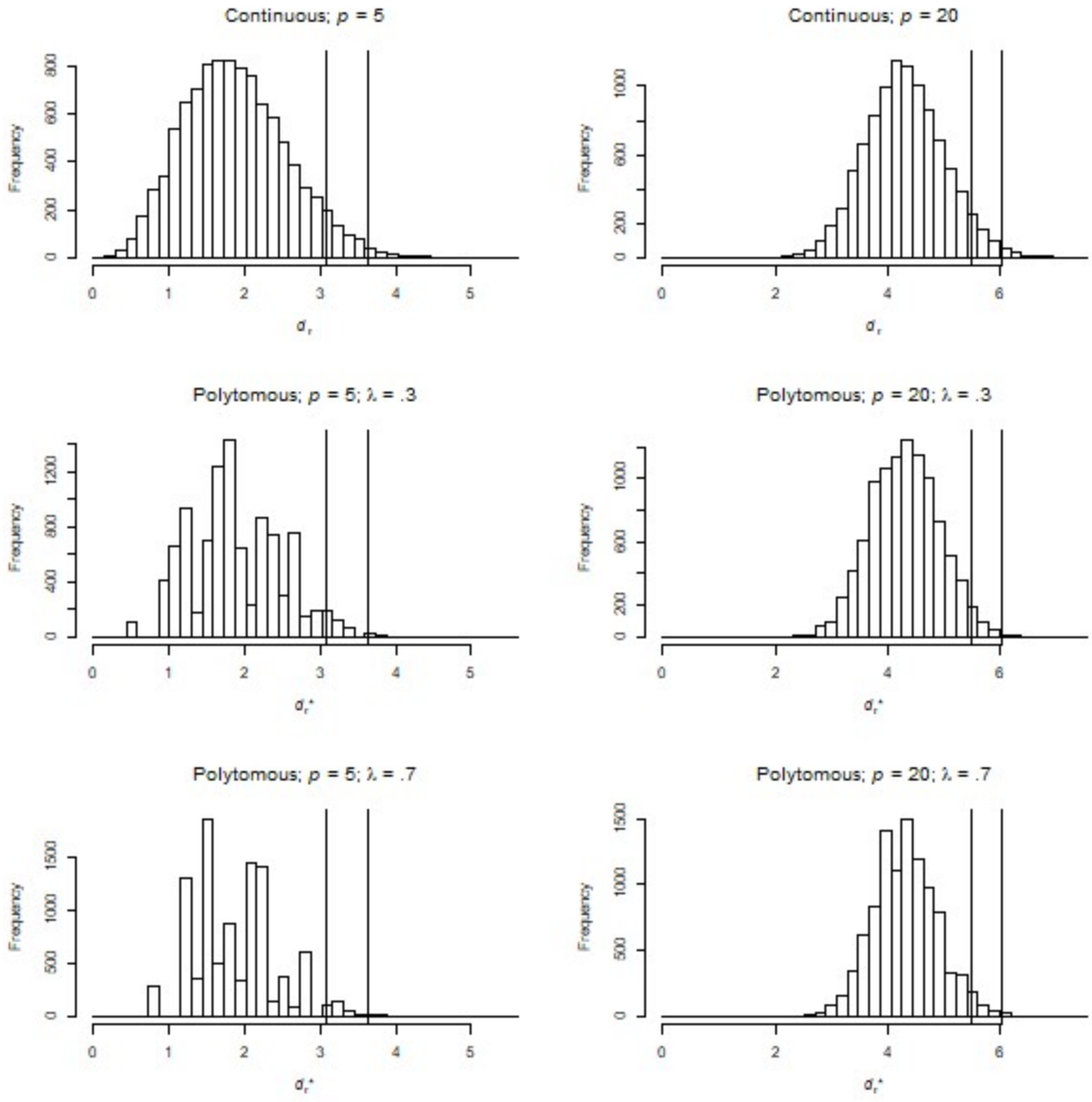

Figure 3. Histograms of d_r in continuous tests and d_r* in polytomous tests. p = number of items, λ = population factor loading. τ denotes the maximum threshold value, and -τ denotes the minimum threshold value. Vertical lines correspond to the 95% and 99% critical values of d_r based on a χ distribution with 4 (p = 5) or 19 (p = 20) degree of freedom.

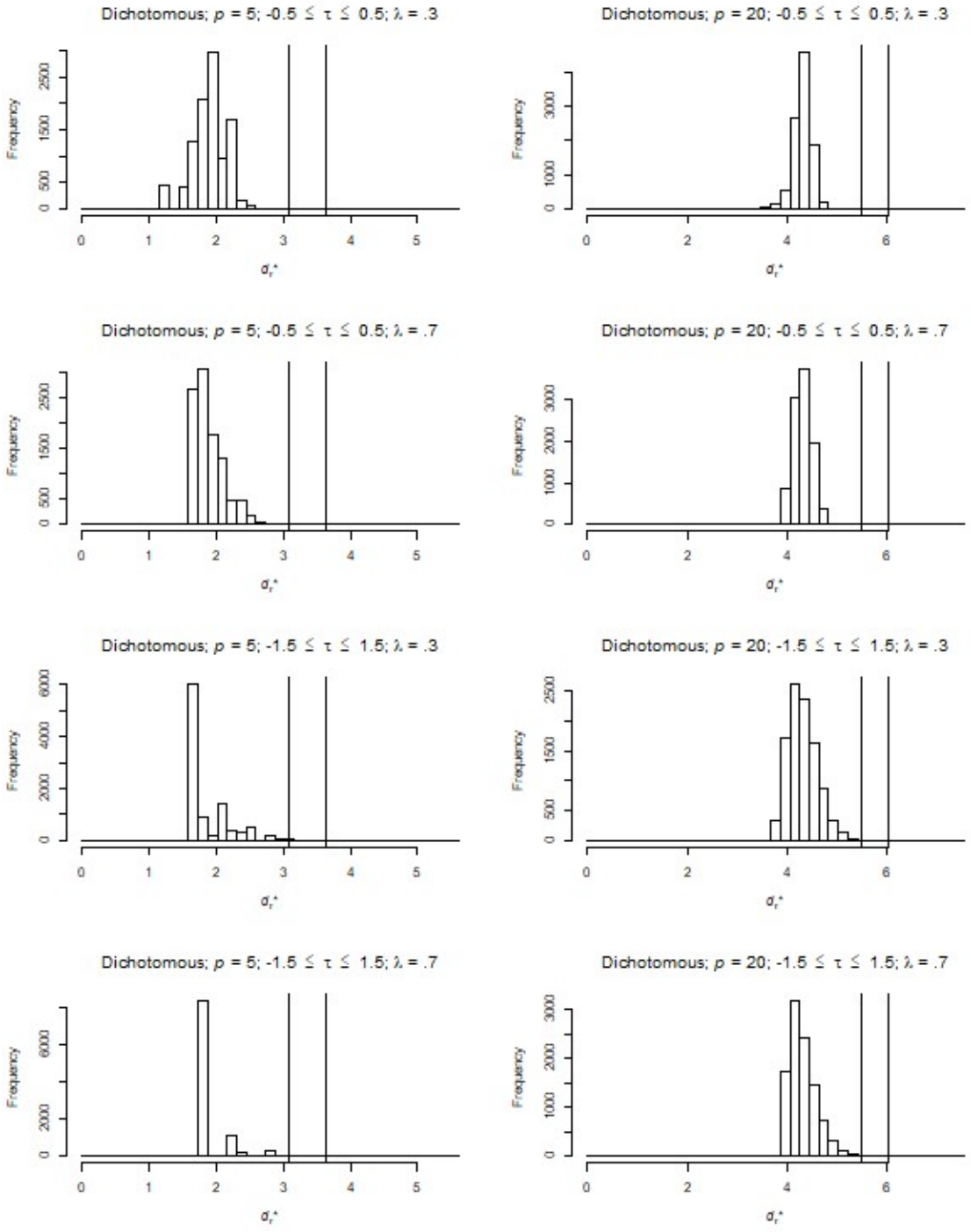

Figure 4. Histograms of d_r* in dichotomous tests. p = number of items, λ = population factor loading. τ denotes the maximum threshold value, and -τ denotes the minimum threshold value. Vertical lines correspond to the 95% and 99% critical values of d_r based on a χ distribution with 4 (p = 5) or 19 (p = 20) degree of freedom.

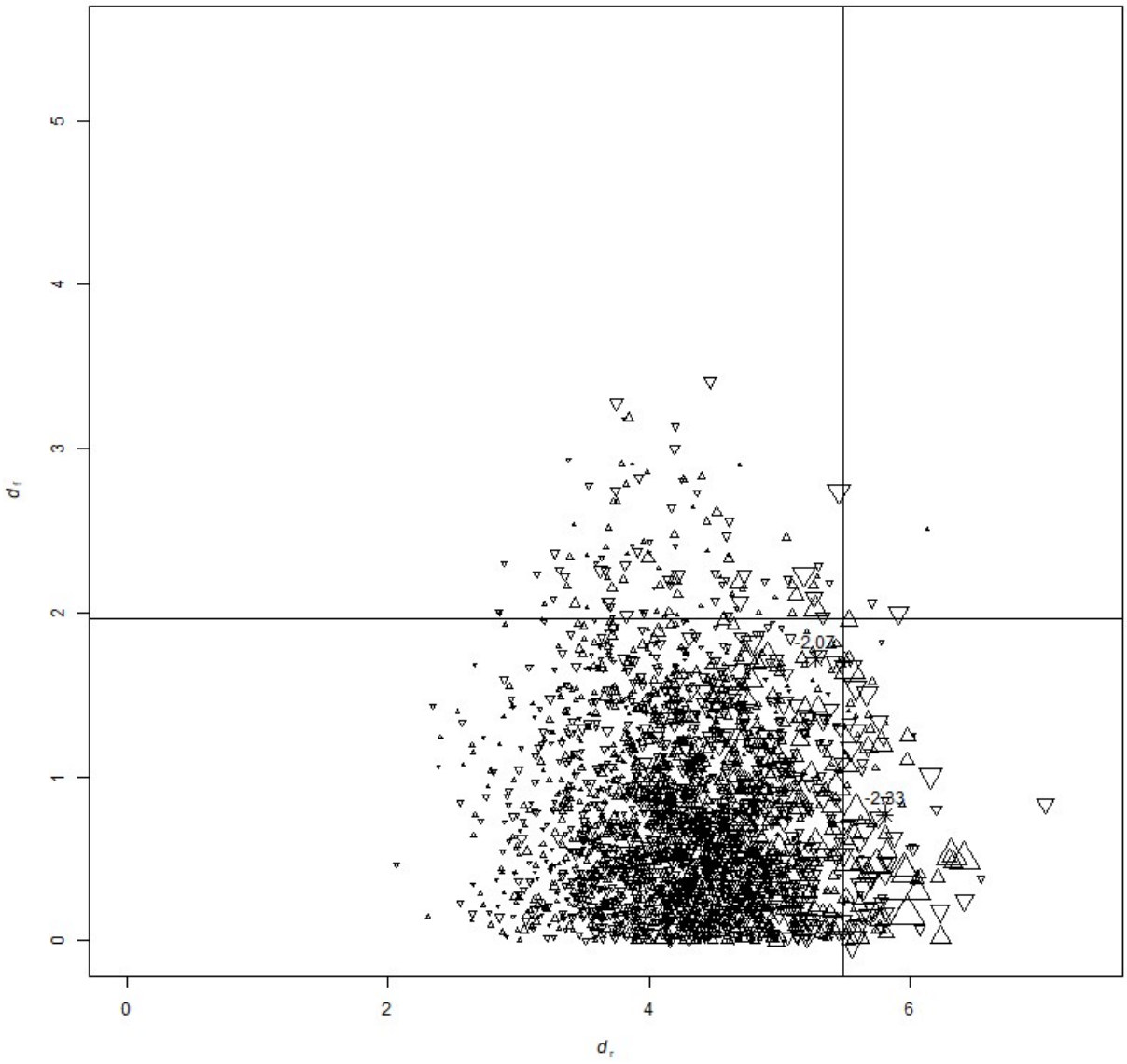

Figure 5. Scatterplot of d_f and d_r in a continuous 20-item test with $\lambda = .3$. Positive values of $\Delta\chi^2$ are marked with upward-pointing triangles, while negative values of $\Delta\chi^2$ are marked with downward-pointing triangles. The size of the triangles is scaled to the absolute value of $\Delta\chi^2$, where the largest size is given by the largest absolute value of $\Delta\chi^2$ across all conditions, determined after removing the ten highest and lowest values of $\Delta\chi^2$; see Supplemental Materials for the complete set of plots.

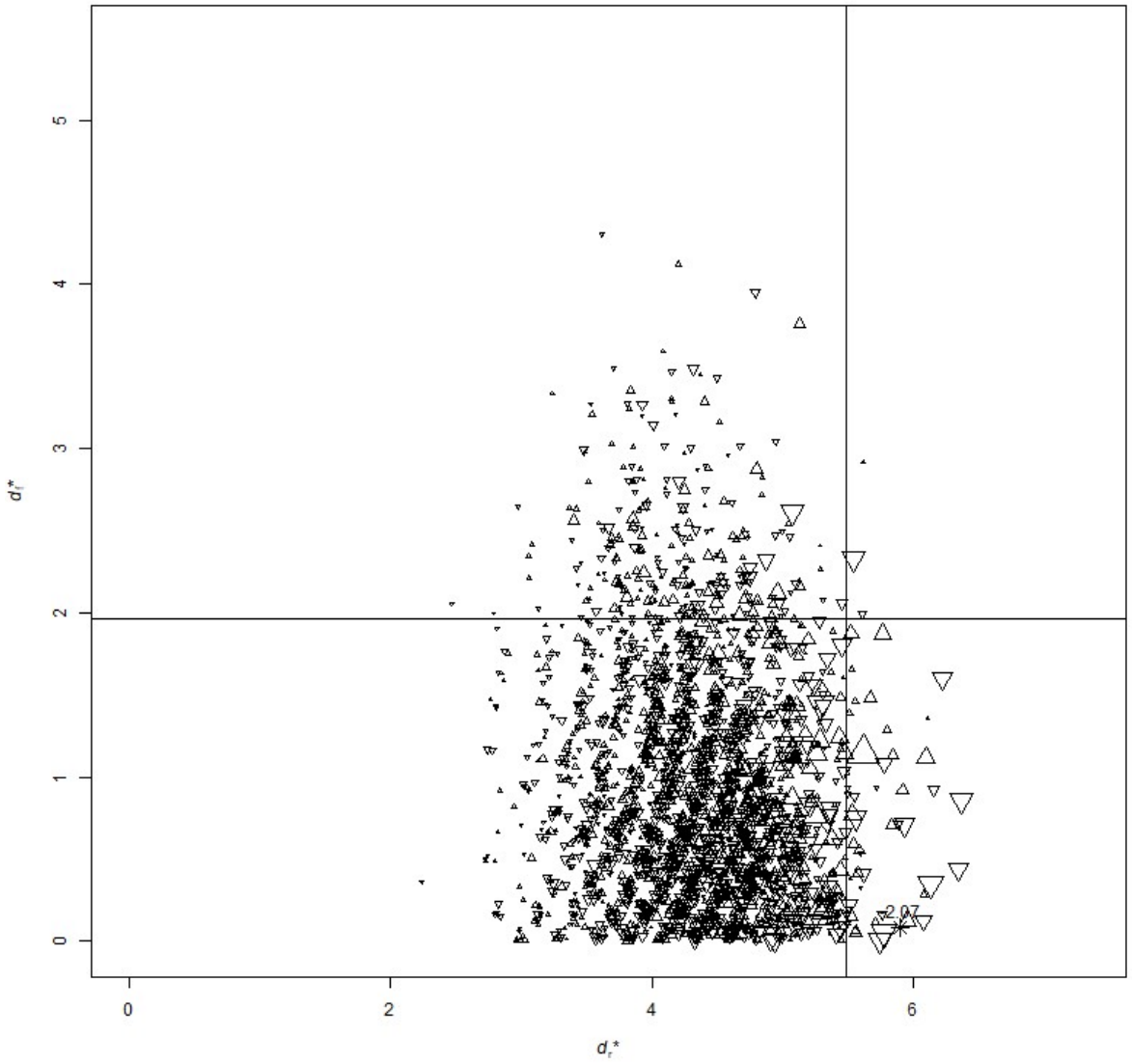

Figure 6. Scatterplot of d_f* and d_r* in a 20-item polytomous test with $\lambda = .3$. Positive values of $\Delta\chi^2$ are marked with upward-pointing triangles, while negative values of $\Delta\chi^2$ are marked with downward-pointing triangles. The size of the triangles is scaled to the absolute value of $\Delta\chi^2$, where the largest size is given by the largest absolute value of $\Delta\chi^2$ across all conditions, determined after removing the ten highest and lowest values of $\Delta\chi^2$; see Supplemental Materials for the complete set of plots. Values of $\Delta\chi^2$ which rank among the ten highest or lowest across all conditions are denoted by asterisks.

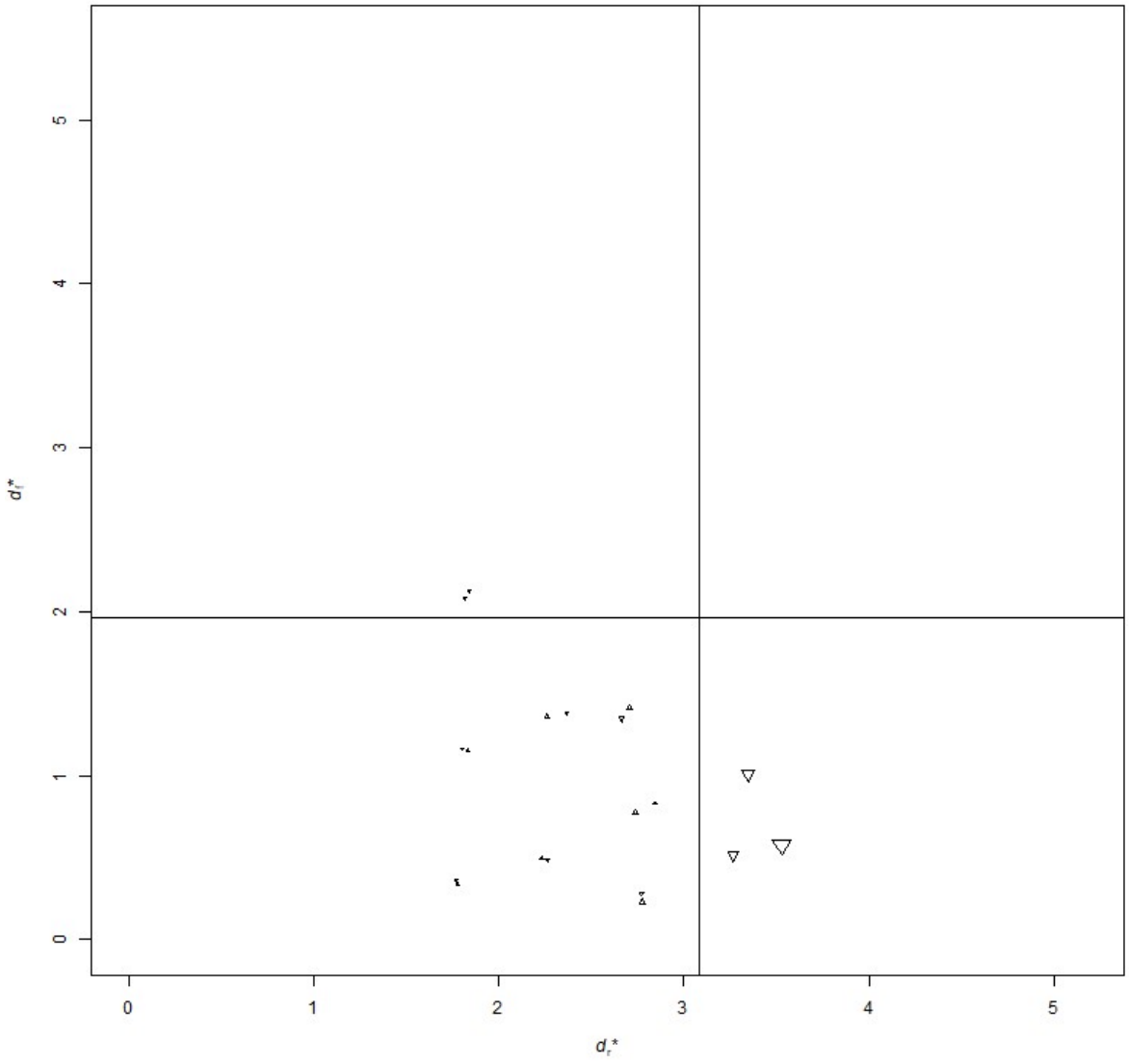

Figure 7. Scatterplot of d_f* and d_r* in a 5-item dichotomous test with $-1.5 \leq \tau \leq 1.5$ and $\lambda = .7$. Positive values of $\Delta\chi^2$ are marked with upward-pointing triangles, while negative values of $\Delta\chi^2$ are marked with downward-pointing triangles. The size of the triangles is scaled to the absolute value of $\Delta\chi^2$, where the largest size is given by the largest absolute value of $\Delta\chi^2$ across all conditions, determined after removing the ten highest and lowest values of $\Delta\chi^2$; see Supplemental Materials for the complete set of plots.

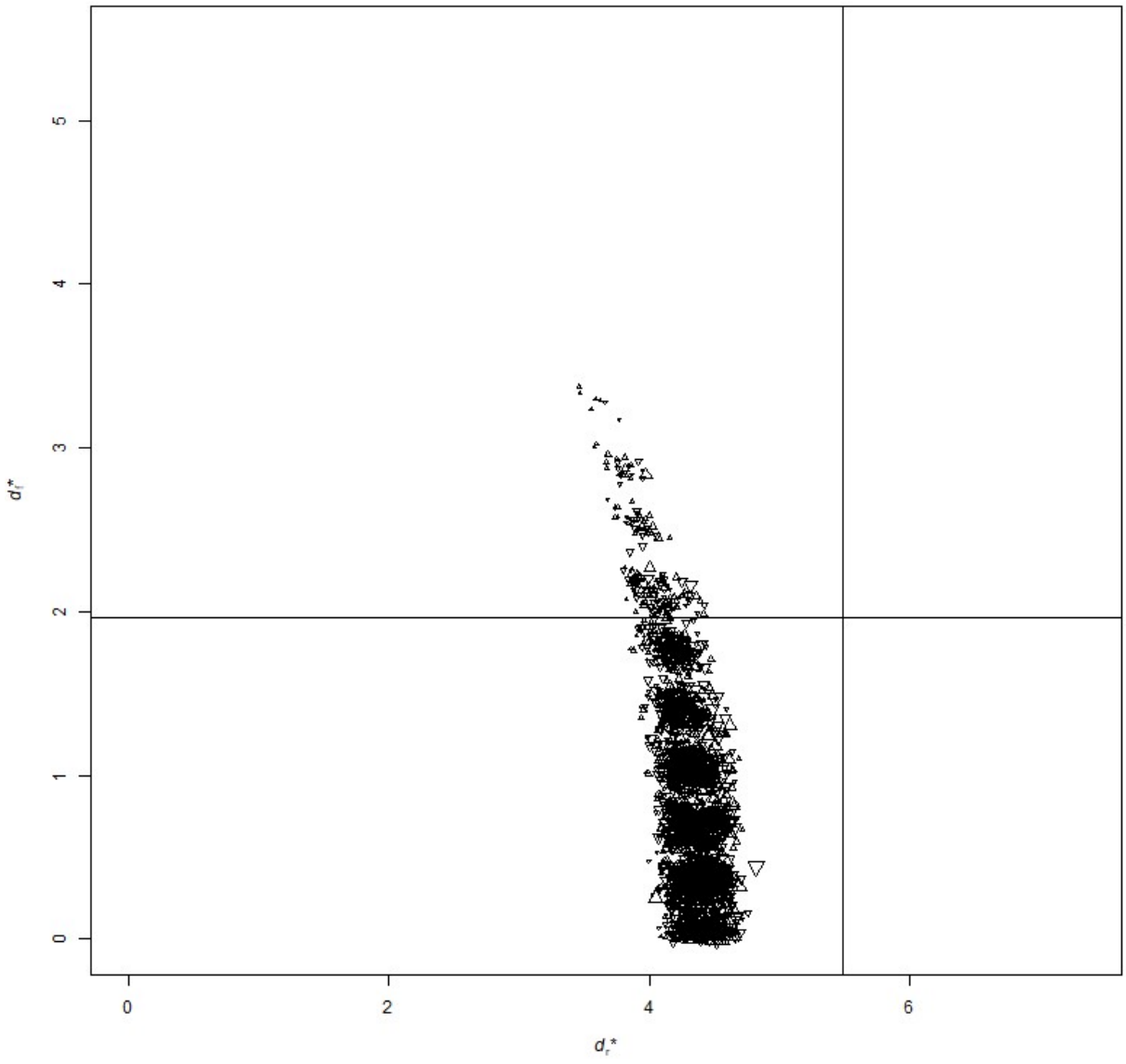

Figure 8. Scatterplot of d_f* and d_r* in a 20-item dichotomous test with $-0.5 \leq \tau \leq 0.5$, and $\lambda = .3$. Positive values of $\Delta\chi^2$ are marked with upward-facing triangles, while negative values of $\Delta\chi^2$ are marked with downward-facing triangles. The size of the triangles is scaled to the absolute value of $\Delta\chi^2$, where the largest size is given by the largest absolute value of $\Delta\chi^2$ across all conditions, determined after removing the ten highest and lowest values of $\Delta\chi^2$; see Supplemental Materials for the complete set of plots.

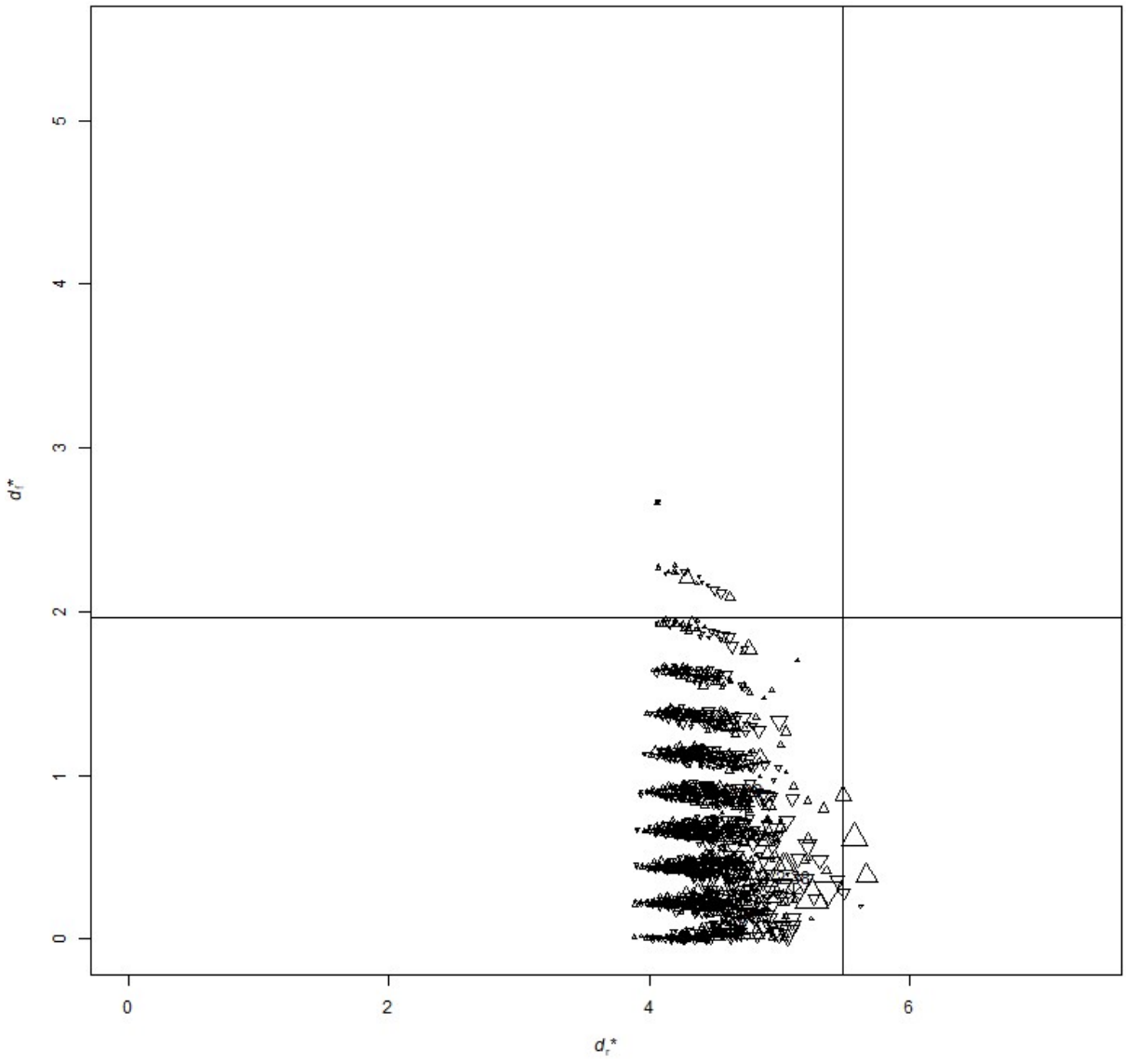

Figure 9. Scatterplot of d_f* and d_r* in a 20-item dichotomous test with -1.5 ≤ τ ≤ 1.5, and λ = .7. Positive values of $\Delta\chi^2$ are marked with upward-facing triangles, while negative values of $\Delta\chi^2$ are marked with downward-facing triangles. The size of the triangles is scaled to the absolute value of $\Delta\chi^2$, where the largest size is given by the largest absolute value of $\Delta\chi^2$ across all conditions, determined after removing the ten highest and lowest values of $\Delta\chi^2$; see Supplemental Materials for the complete set of plots.

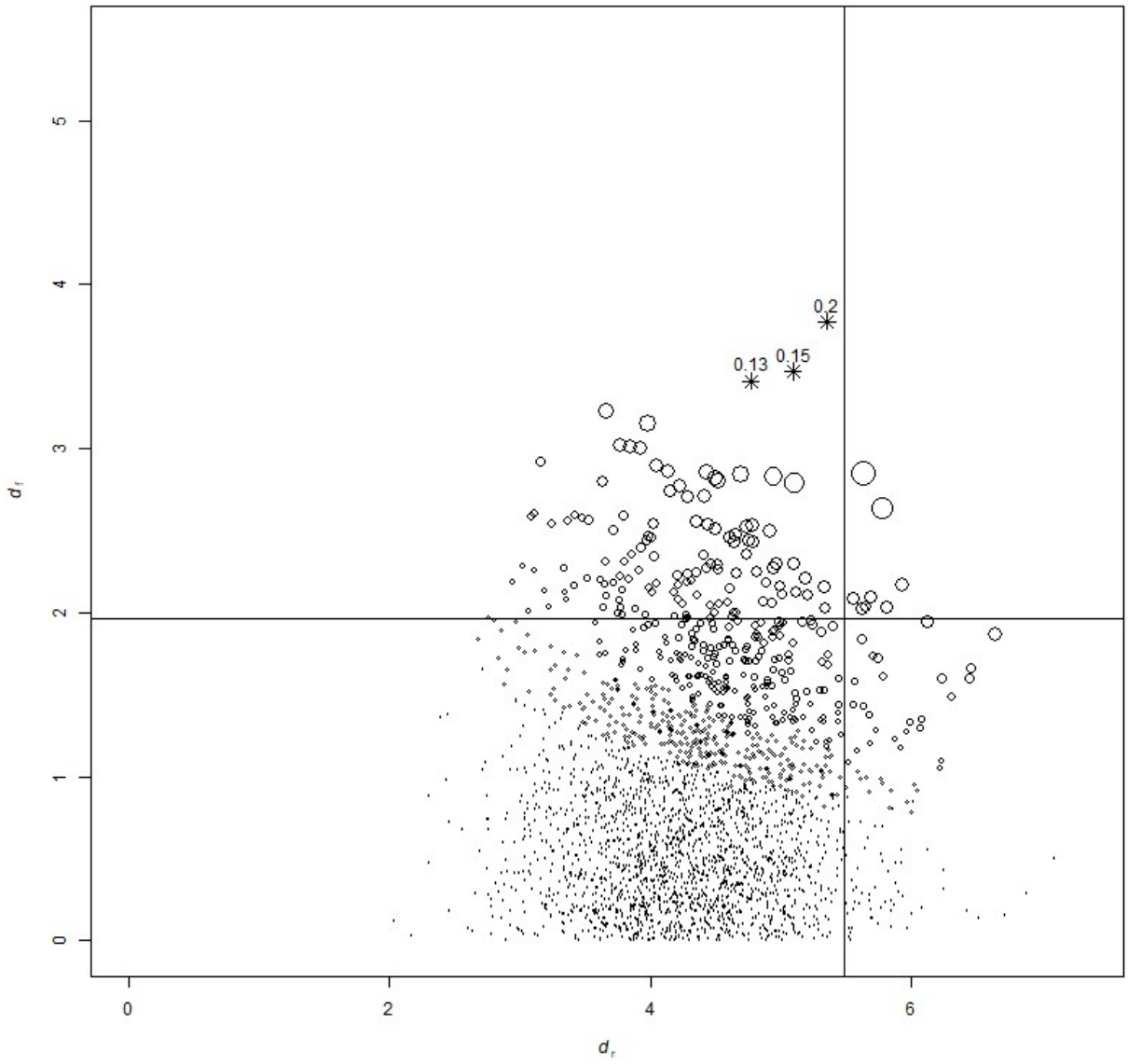

Figure 10. Scatterplot of d_f and d_r in a continuous 20-item test with $\lambda = .7$. The size of the circles is scaled to gCD_λ, where the largest size is given by the largest value of gCD_λ across all conditions, determined after removing the ten highest values of gCD_λ; see Supplemental Materials for the complete set of plots. Values of gCD_λ which rank among the ten highest or lowest across all conditions are denoted by asterisks.

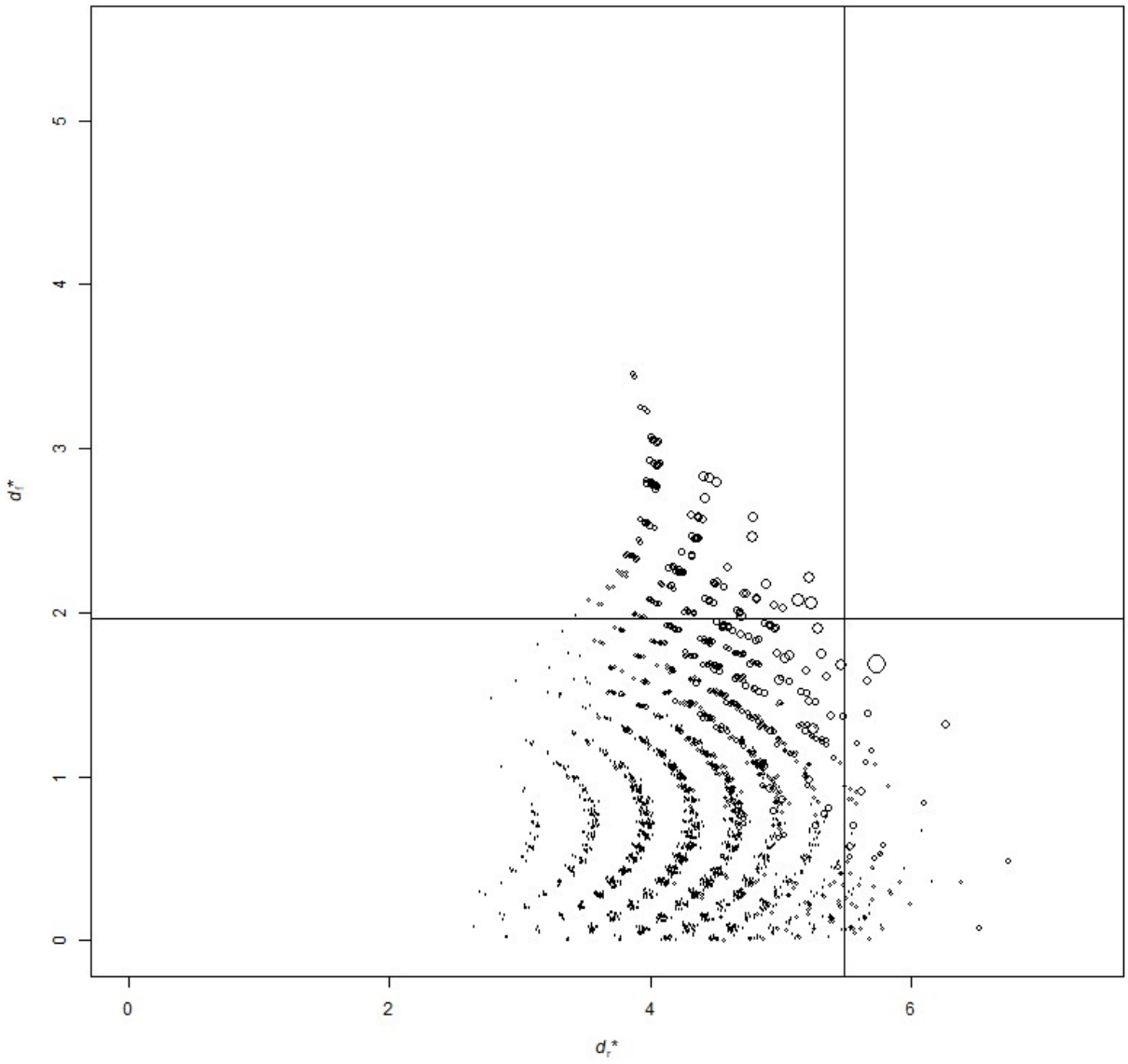

Figure 11. Scatterplot of d_f and d_r in a 20-item polytomous test with $\lambda = .7$. The size of the circles is scaled to gCD_λ, where the largest size is given by the largest value of gCD_λ across all conditions, determined after removing the ten highest values of gCD_λ; see Supplemental Materials for the complete set of plots.

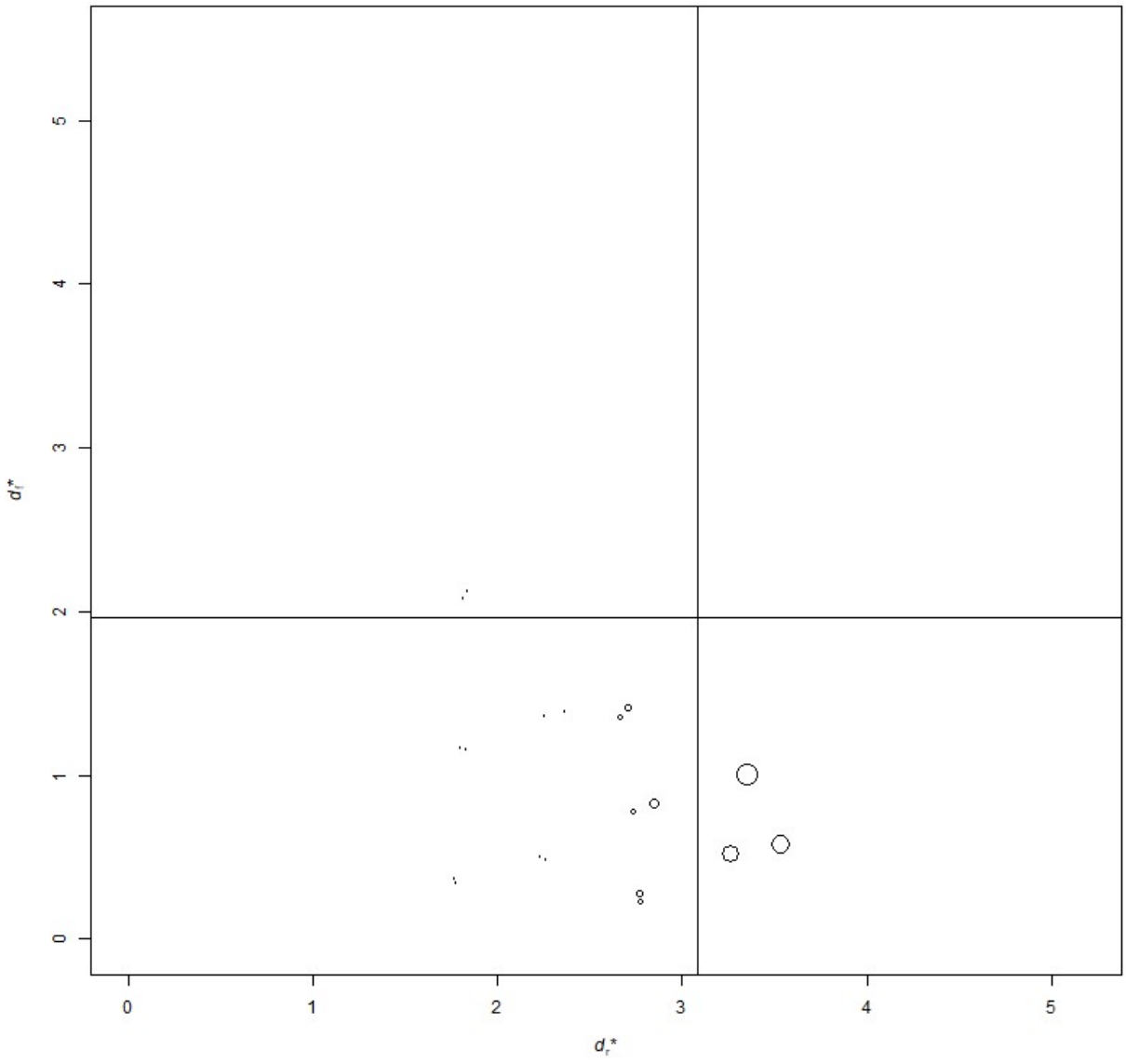

Figure 12. Scatterplot of d_f and d_r in a 5-item dichotomous test with $-1.5 \leq \tau \leq 1.5$ and $\lambda = .7$. The size of the circles is scaled to gCD_λ, where the largest size is given by the largest value of gCD_λ across all conditions, determined after removing the ten highest values of gCD_λ; see Supplemental Materials for the complete set of plots.

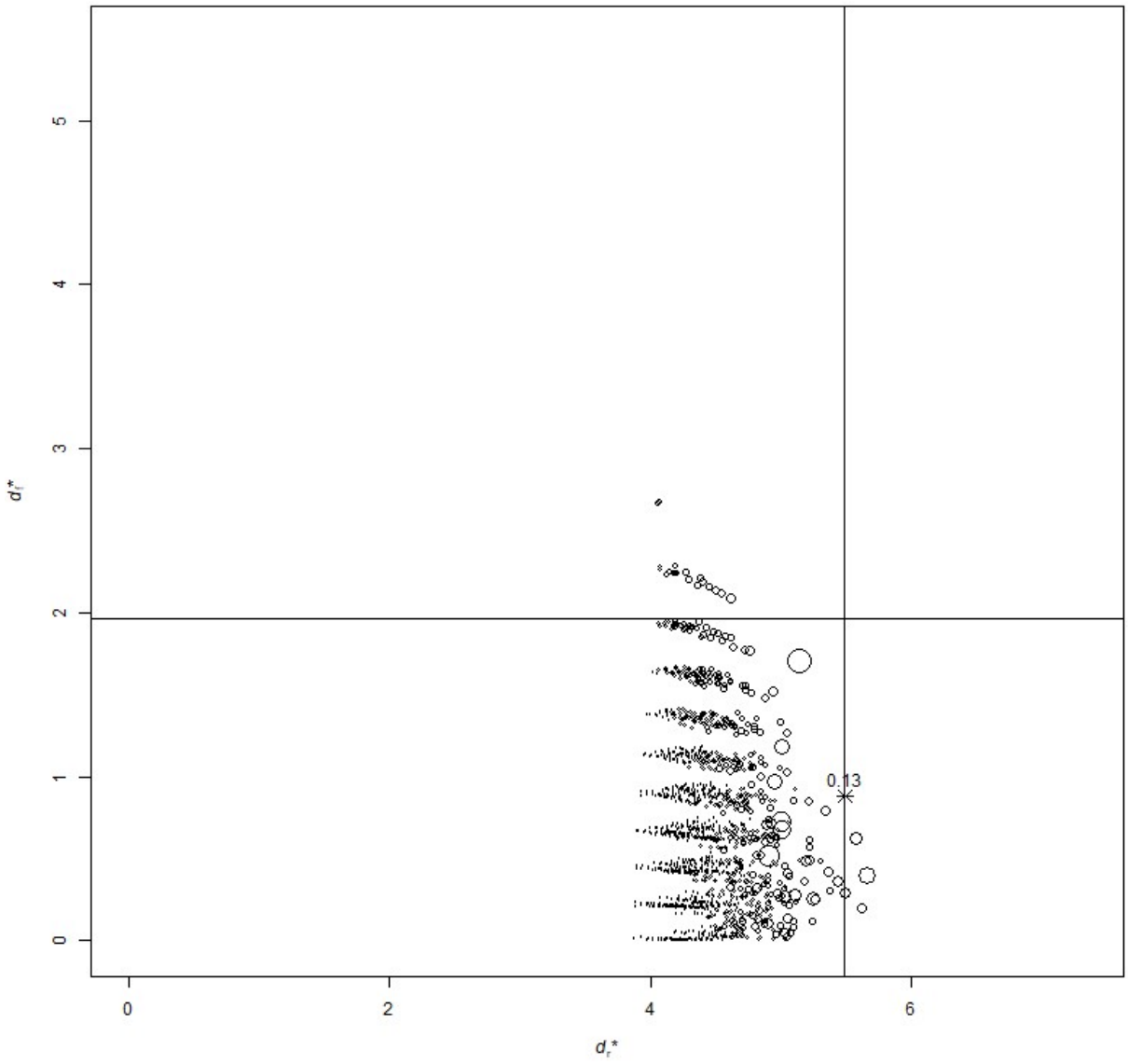

Figure 13. Scatterplot of d_f and d_r in a 20-item dichotomous test with $-1.5 \leq \tau \leq 1.5$ and $\lambda = .7$. The size of the circles is scaled to gCD_λ, where the largest size is given by the largest value of gCD_λ across all conditions, determined after removing the ten highest values of gCD_λ; see Supplemental Materials for the complete set of plots. Values of gCD_λ which rank among the ten highest or lowest across all conditions are denoted by asterisks.

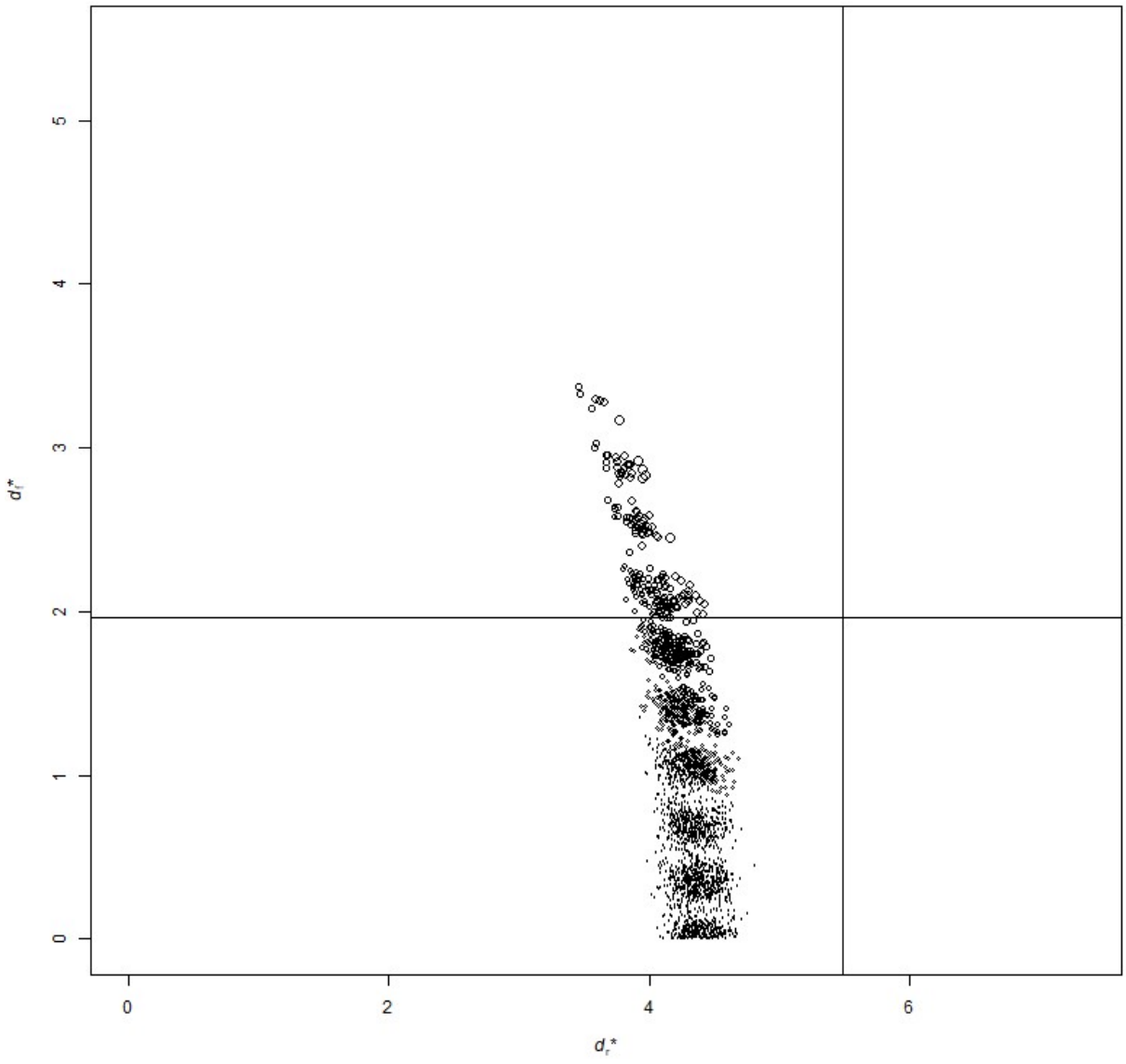

Figure 14. Scatterplot of d_f and d_r in a 20-item dichotomous test with $-0.5 \leq \tau \leq 0.5$ and $\lambda = .3$. The size of the circles is scaled to gCD_λ, where the largest size is given by the largest value of gCD_λ across all conditions, determined after removing the ten highest values of gCD_λ; see Supplemental Materials for the complete set of plots.

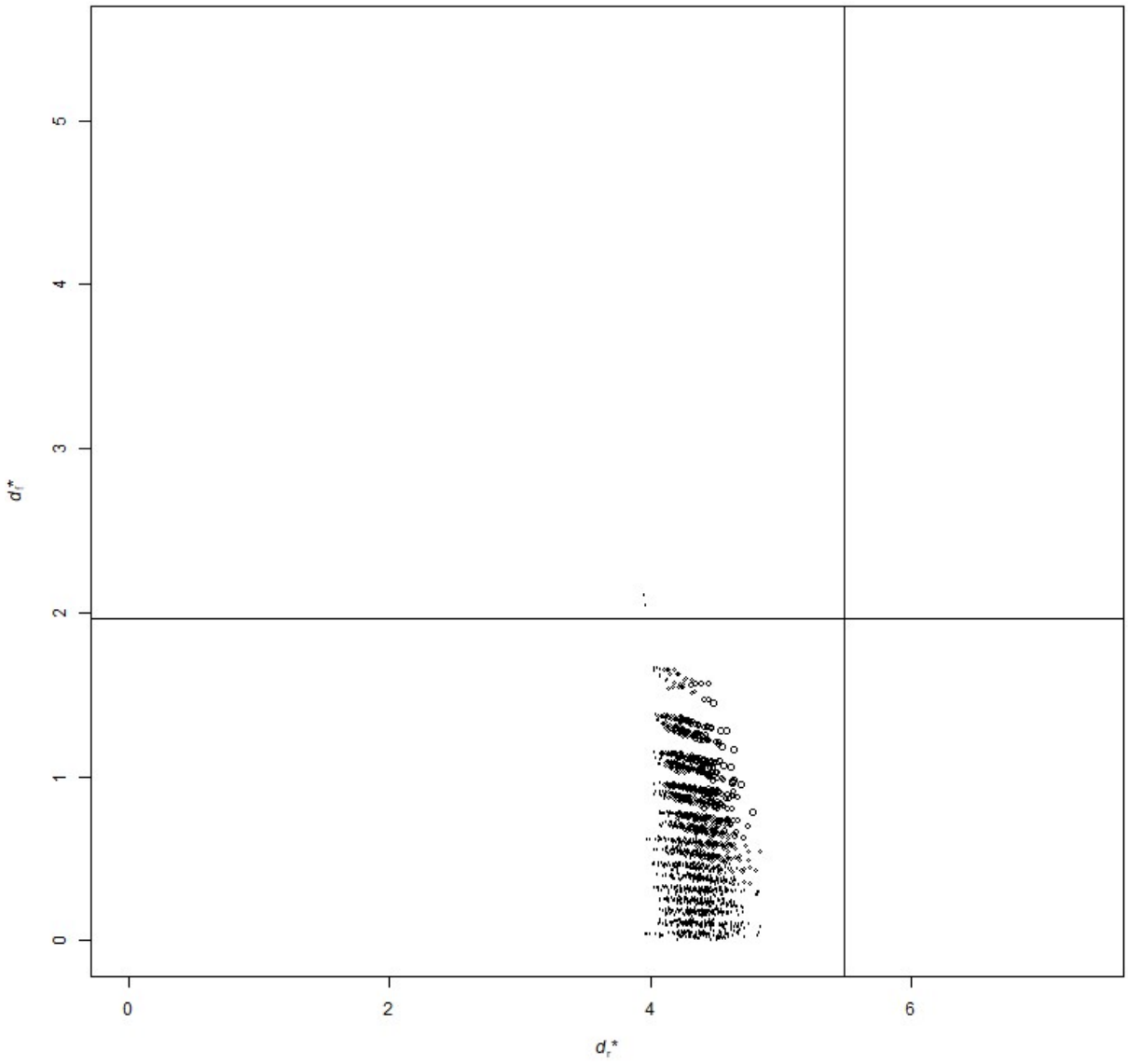

Figure 15. Scatterplot of d_f and d_r in a 20-item dichotomous test with $-0.5 \leq \tau \leq 0.5$ and $\lambda = .7$. The size of the circles is scaled to gCD_λ, where the largest size is given by the largest value of gCD_λ across all conditions, determined after removing the ten highest values of gCD_λ; see Supplemental Materials for the complete set of plots.

References

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103*(3), 411.

Asparouhov, T. (2005). Sampling weights in latent variable modeling. *Structural Equation Modeling, 12*(3), 411-434.

Azzalini, A., & Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 65*, 367–389.

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. New York, NY: Wiley.

Bentler, P. M. (1989). *EQS structural equations program manual.* Los Angeles, CA: BMDP Statistical Software.

Bentler, P. M., & Wu, E. J. (2005). EQS 6.1 for Windows. Encino, CA: Multivariate Software INC.

Birenbaum, M. (1986). Effect of dissimulation motivation and anxiety on response pattern appropriateness measures. *Applied Psychological Measurement, 10,* 167–174.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement*, *12*(3), 261-280.

Bohannon, J. (2015). Many psychology papers fail replication test. *Science, 349*(6251), 910-911.

Bollen, K. A., & Arminger, G. (1991). Observational residuals in factor analysis and structural equation models. *Sociological Methodology, 21*, 235–262. doi:10.2307/270937

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, *110*(2), 305-314.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110*(2), 203.

Browne, M. W. (1974). Generalized least squares estimators in the analysis of covariance structures. *South African Statistical Journal, 8*(1), 1-24.

Browne, M. W. (1982) Covariance structures. In: Hawkins, D. M. *Topics in applied multivariate analysis*. Cambridge University Press. pp. 72–141.

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, *37*(1), 62-83.

Canal, L. (2005). A normal approximation for the chi-square distribution. *Computational Statistics & Data Analysis, 48*(4), 803-808.

Clark, M. E., Gironda, R. J., & Young, R. W. (2003). Detection of back random responding: Effectiveness of MMPI-2 and personality assessment inventory validity indices. *Psychological Assessment, 15*, 223–234.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*. Mahwah, NJ: Erlbaum.

Cook, R. D. (1977). Detection of influential observations in linear regression. *Technometrics*, *19*(1), 15-18.

Cook, R. D. (1979). Influential observations in linear regression. *Journal of the American Statistical Association*, *74*(365), 169-174.

Cook, R. D. (1986). Assessment of local influence. *Journal of the Royal Statistical Society B, 48*(2), 133–169.

Cook, D. R., & Weisberg, S. (1982). *Residuals and influence in regression*. New York, NY: Chapman & Hall.

Curtis, D. D. (2004). Person misfit in attitude surveys: Influences, impacts and implications. *International Education Journal, 5*, 2125-2144.

De Ayala, R. J., Plake, B., & Impara, J. C (2001). The effect of omitted responses on the accuracy of ability estimation in item response theory. *Journal of Educational Measurement, 38*, 213–234.

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*(1), 67-86.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory*. Psychology Press.

Emons, W. H. (2009). Detection and diagnosis of person misfit from patterns of summed polytomous item scores. *Applied Psychological Measurement*, *33*(8), 599-619.

Ferrando, P. J. (2004). Person reliability in personality measurement: An item response theory analysis. *Applied Psychological Measurement*, *28*(2), 126-140.

Ferrando, P. J. (2010). Some statistics for assessing person-fit based on continuous-response models. *Applied Psychological Measurement*, *34*(4), 219-237.

Ferrando, P. J. (2012). Assessing inconsistent responding in E and N measures: An application of person-fit analysis in personality. *Personality and Individual Differences*, *52*(6), 718-722.

Ferrando, P. J. (2015). Assessing person fit in typical-response measures. *Handbook of item response theory modeling: Applications to typical performance assessment*, 128-155.

Ferrando, P. J., & Anguiano-Carrasco, C. (2013). A structural model-based optimal person-fit procedure for identifying faking. *Educational and Psychological Measurement*, *73*(2), 173-190.

Falk, C.F., & Cai, L. (2016). A flexible full-information approach to the modeling of response styles. *Psychological Methods, 21*, 328-347. http://dx.doi.org/10.1037/met0000059

First, M. B. (2014). Structured clinical interview for the DSM (SCID). *The Encyclopedia of Clinical Psychology*, 1-6.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*(4), 466.

Fox, J. (1991). *Regression diagnostics: An introduction* (Vol. 79). Sage.

Green, P. J. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society. Series B (Methodological)*, 149-192.

Hoijtink, H. (1987). Rasch schaal constructie met behulp van een passingsindex voor personen [Rasch scale construction using a person-fit index]. *Kwantitatieve Methoden, 25,* 101–110.

Holland, P. W., & Welsch, R. E. (1977). Robust regression using iteratively reweighted least squares. *Communications in Statistics-Theory and Methods,6*(9), 813-827.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1-55.

Jin, S., & Yang-Wallentin, F. (2016). Asymptotic robustness study of the polychoric correlation estimation. *Psychometrika*, *82*(1), 1-19.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, *16*(4), 277-298.

Lawley, D. N., & Maxwell, A. E. (1971). *Factor analysis as a statistical method* (Vol. 18). London: Butterworths.

Lee, S. Y., & Lam, M. L. (1988). Estimation of polychoric correlation with elliptical latent variables. *Journal of Statistical Computation and Simulation*, *30*(3), 173-188.

Lee, S. Y., Poon, W. Y., & Bentler, P. M. (1990). A three-stage estimation procedure for structural equation models with polytomous variables. *Psychometrika*, *55*(1), 45-51.

Levine, M. V., & Drasgow, F. (1983). Appropriateness measurement: Validating studies and variable ability models. *New horizons in testing: Latent trait test theory and computerized adaptive testing*, 109-131.

Mansolf, M., & Reise, S. P. (2018). Case diagnostics for factor analysis of ordered categorical data with applications to person-fit measurement. *Structural Equation Modeling: A Multidisciplinary Journal, 25*(1), 86-100.

Mardia, K. V. (1962). Multivariate pareto distributions. *The Annals of Mathematical Statistics, 33*, 1008–1015.

Meijer, R. R. (1996). The influence of the presence of deviant item score patterns on the power of a person-fit statistic. *Applied Psychological Measurement*, *20*(2), 141-154.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25*, 107–135.

Mislevy, R. J. (1986). Recent developments in the factor analysis of categorical variables. *Journal of Educational and Behavioral Statistics*, *11*(1), 3-31.

Molenaar, I. W., & Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika*, *55*(1), 75-106.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*(1), 115-132.

Muthén , B.O, du Toit, S., & Spisic,, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. Unpublished manuscript.

Muthén, L.K. and Muthén, B.O. (1998-2017). *Mplus User's Guide. Eighth Edition*. Los Angeles, CA: Muthén & Muthén.

Nering, M. L., & Meijer, R. R. (1998). A comparison of the person response function and the l_z person fit statistic. *Applied Psychological Measurement, 22*, 53–69.

O'Leary, D. P. (1990). Robust regression computation using iteratively reweighted least squares. *SIAM Journal on Matrix Analysis and Applications, 11*(3), 466–480.

Oshima, T. C. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement, 31*, 200–219.

Pastore, M. & Altoe, G. (2018). influence.SEM: Case Influence in Structural Equation Models. R package version 2.2. https://CRAN.R-project.org/package=influence.SEM

Patton, J. M., Cheng, Y., Hong, M., & Diao, Q. (2019). Detection and treatment of careless responses to improve item parameter estimation. *Journal of Educational and Behavioral Statistics*.

Pek, J., MacCallum, R.C. (2011). Sensitivity analysis in structural equation models: Cases and their influence. *Multivariate Behavioral Research, 46*, 202-228.

Phillips, S. E. (1986). The effects of deletion of misfitting persons on vertical equating via the Rasch model. *Journal of Educational Measurement, 23,* 107–118.

Quiroga, A. M. (1992). *Studies of the polychoric correlation and other correlation measures for ordinal variables.* Doctoral dissertation, Uppsala University.

Reise, S.P., & Due, A.M. (1991). The influence of test characteristics on the detection of aberrant response patterns. *Applied Psychological Measurement, 22*, 53-69.

Reise, S. P., & Flannery, P. (1996). Assessing person-fit on measures of typical performance. *Applied Measurement in Education*, *9*(1), 9-26.

Reise, S. P., & Waller, N. G. (1993). Traitedness and the assessment of response pattern scalability. *Journal of Personality and Social Psychology*, *65*(1), 143.

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*(1), 3.

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17*(3), 354.

Roscino, A., & Pollice, A. (2006). A generalization of the polychoric correlation coefficient. In *Data Analysis, Classification and the Forward Search* (pp. 135-142). Springer, Berlin, Heidelberg.

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection*. New York, NY: Wiley.

Rousseeuw, P. J., & van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points (with discussion). *Journal of the American Statistical Association, 85*, 633– 651.

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

Satorra, A., & Bentler, P. M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika*, *75*(2), 243-248.

Schmitt, N., Chan, D., Sacco, J. M., McFarland, L. A., & Jennings, D. (1999). Correlates of person fit and effect of person fit on test validity. *Applied Psychological Measurement, 23,* 41–53.

Sinharay, S. (2016). Asymptotically correct standardization of person-fit statistics beyond dichotomous items. *Psychometrika*, *81*(4), 992-1013.

Snijders, T. A. B. (2001). Asymptotic null distribution of person-fit statistics with estimated person parameter. *Psychometrika, 66*, 331–334.

Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*(3), 393-408.

Tellegen, A. (1988). The analysis of consistency in personality assessment. *Journal of Personality, 56*, 622-663.

van Barneveld, C. (2007). The effect of respondent motivation on test construction within an IRT framework. *Applied Psychological Measurement, 31*, 31–46.

van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.

van Krimpen-Stoop, E. M., & Meijer, R. R. (1999). The null distribution of person-fit statistics for conventional and adaptive tests. *Applied Psychological Measurement*, *23*(4), 327-345.

Waller, N. G., & Reise, S. P. (1992). Genetic and environmental influences on item response pattern scalability. *Behavior Genetics*, *22*(2), 135-152.

Wilcox, R. R. (2001). *Fundamentals of modern statistical methods: Substantially improving power and accuracy*. New York, NY: Springer.

Wise, S. L., Kingsbury, G. G., Thomason, J. T.,& Kong, X. (2004).An investigation of motivation filtering in a statewide achievement testing program. Paper Presented at the April, 2004 annual meeting of the National Council of Measurement in Education, San Diego, CA.

Yang-Wallentin, F., Jöreskog, K. G., & Luo, H. (2010). Confirmatory factor analysis of ordinal variables with misspecified models. *Structural Equation Modeling, 17*(3), 392–423.

Yuan, K.-H., & Bentler, P. M. (1998). Structural equation modeling with robust covariances. *Sociological Methodology, 28*, 363–396. doi: 10.1111/0081-1750.00052

Yuan, K. H., & Bentler, P. M. (2000). Robust mean and covariance structure analysis through iteratively reweighted least squares. *Psychometrika, 65*(1), 43-58.

Yuan, K. H., Fung, W. K., & Reise, S. P. (2004). Three Mahalanobis distances and their role in assessing unidimensionality. *British Journal of Mathematical and Statistical Psychology*, *57*(1), 151-165.

Yuan, K.-H., & Hayashi, K. (2010). Fitting data to model: Structural equation modeling diagnosis using two scatter plots. *Psychological Methods, 15*(4), 335–351.

Yuan, K.-H., Marshall, L. L., & Weston, R. (2002). Cross-validation by downweighting influential cases in structural equation modelling. *British Journal of Mathematical and Statistical Psychology, 55*, 125–143.

Yuan, K.-H., & Zhang, Z. (2012). Structural equation modeling diagnostics using R package semdiag and EQS. *Structural Equation Modeling: A Multidisciplinary Journal, 19*(4), 683-702.

Yuan, K.-H., & Zhong, X. (2008). Outliers, leverage observations, and influential cases in factor analysis: Using robust procedures to minimize their effect. *Sociological Methodology, 38*, 329–368. doi:10.1111/j .1467-9531.2008.00198.x

Zhong, X., & Yuan, K.-H. (2011). Bias and efficiency in structural equation modeling: Maximum likelihood versus robust methods. *Multivariate Behavioral Research, 46*, 229–265. doi:10.1080/00273171.2011 .558736

Zickar, M. J., & Drasgow, F. (1996). Detecting faking on a personality instrument using appropriateness measurement. *Applied Psychological Measurement*, *20*(1), 71-87.