Robust Estimation of the Polychoric Correlation Coefficient*

Max Welz
welz@ese.eur.nl
Erasmus University

Patrick Mair
mair@fas.harvard.edu
Harvard University

Andreas Alfons
alfons@ese.eur.nl
Erasmus University

March 5, 2024

Abstract

Structural equation models are typically fitted to a correlation matrix. When the variables are rating items, it is often recommended to use polychoric correlation coefficients estimated via maximum likelihood to calculate the correlation matrix. However, just like sample correlation, maximum likelihood (ML) is highly susceptible to model misspecification due to, for instance, inattentive/careless responding or other violations of the latent normality assumed by the polychoric model. We propose a novel estimator that is substantially more robust to model misspecification than ML. Crucially, and in contrast to previous literature, our estimator makes *no assumption whatsoever* on the type, magnitude, or location of potential misspecification, rendering it robust to an unlimited variety of possible variations of misspecification. It furthermore generalizes ML and is strongly consistent as well as asymptotically normally distributed, while being of the same time complexity as ML, meaning that it comes at no additional computational cost. In addition, we develop a novel diagnostic test that tests whether each individual cell in a contingency table can be fitted well by the polychoric model, thereby allowing to trace back potential sources of misspecification. We demonstrate the robustness and practical usefulness of our estimator in simulation studies and an empirical application on a Big-5 administration, where we find compelling evidence for the presence of inattentive respondents whose adverse influence our estimator withstands, unlike ML.

Keywords: Robust statistics, polychoric correlation, model misspecification, discrete data, asymptotic normality

*This work was supported by a grant from the Dutch Research Council (NWO), research program Vidi (Project No. VI.Vidi.195.141).

1 Introduction

Structural equation models, particularly factor models, are typically fitted to a correlation matrix that has been estimated a priori. When the modeled variables are discrete rating variables, it is often recommended to estimate the correlation matrix with polychoric correlation coefficients (e.g. Foldnes & Grønneberg, 2022; Garrido et al., 2013; Holgado-Tello et al., 2010). However, recent work has demonstrated that polychoric correlation is highly sensitive to violations of an assumption on underlying normality and therefore "cannot be considered a robust methodology" (Foldnes & Grønneberg, 2022, p. 566). Violations of a latent normal distribution can be caused by, for example but not limited to, careless or inattentive responding, misresponses, item misunderstanding, or heterogeneous subpopulations, resulting in possibly large biases in correlational estimates that in turn lead to biases in estimates of structural equation models.

We propose a novel way to estimate polychoric correlation coefficients that is robust to non-normality or, more generally, model misspecification. Crucially, we make *no assumption whatsoever* on the type, magnitude, or location of potential misspecification, a departure from existing literature where misspecification is often modeled explicitly. Hence, our estimator is robust to an unlimited and unspecified variety of possible variations of misspecification. Our estimator generalizes maximum likelihood estimation, is strongly consistent, asymptotically Gaussian, and comes at no additional computational cost compared to maximum likelihood. In addition, we develop a novel test that tests, for each individual cell in a contingency table, if the polychoric model provides an appropriate fit to that cell. The test rejects this null hypothesis when the cell in question cannot be fitted sufficiently well, that is, latent normality cannot be sustained for that cell, and thereby helps pinpoint potential sources of model misspecification.

We verify the attractive statistical and robustness properties of our estimator by means of extensive simulation studies and demonstrate its practical usefulness in an empirical application on a Big-5 administration (Goldberg, 1992), where we find strong evidence for the presence of inattentive responding. For instance, the polychoric correlation coefficient between two mutually contradictory unipolar markers in a *neuroticism* scale is estimated as -0.62 by maximum likelihood, whereas our estimator yields a substantially stronger correlation estimate of -0.93, which is, unlike the maximum likelihood result, in line with literature on this scale.

In the spirit of open and reproducible science and to enhance accessibility and adoption by empirical researchers, an R implementation of our proposed methodology is freely available in the package `robord` (Welz, 2024a, for "ROBust ORDinal data analysis"), and replication files are publicly available at [add URL when posting to arXiv]. To maximize speed and performance, the package is predominantly developed in C++ and integrated to R via `Rcpp` (Eddelbuettel, 2013).

This paper ties into a growing literature concerned with the validity of research findings when employed psychological models are misspecified due to responses that do not follow the assumed model and subsequently cannot be fitted well by that model. An ensuing

poor model fit may lead a researcher to doubt or even reject the theoretical model (Lai & Green, 2016). Poor model fit can occur for two reasons. First, the theory behind the model may simply be wrong, in case of which the model is also wrong and the theory is correctly rejected by the data. Second, which is the focus of this paper, theory and model may in fact be correct, yet a misfit occurs due to the presence of a limited number of observations from a different population, such as careless/inattentive responses or heterogeneous subgroups, which may lead one to incorrectly reject a theory (Arias et al., 2020). Indeed, already a small proportion of careless/inattentive responses can substantially deteriorate model fit (Arias et al., 2020; Huang et al., 2015a; Woods, 2006), and ultimately lead a researcher to reject a correct hypothesis or sustain an incorrect hypothesis (Arias et al., 2020; Huang et al., 2015b; Maniaci & Rogge, 2014; McGrath et al., 2010; Woods, 2006; Schmitt & Stults, 1985). Careless responding itself is widely prevalent (Ward & Meade, 2023; Bowling et al., 2016; Meade & Craig, 2012) with most estimates on its prevalence ranging from 10–15% of study participants (Curran, 2016; Huang et al., 2015b, 2012; Meade & Craig, 2012), while already a prevalence 5–10% can jeopardize the validity of research findings (Arias et al., 2020; Credé, 2010; Woods, 2006). In fact, Ward & Meade (2023) conjecture that careless responding is likely present in all survey data. In addition to poor model fit, we stress that careless responses also pose an issue for replication studies because different studies may yield different results solely due to different proportions of carelessness and not differences in actual effect sizes (Curran, 2016). Due to the damaging effects of careless responses, a large number of methods for their detection has emerged, ranging from consistency indicators such as psychometric antonyms/synonyms (Meade & Craig, 2012) over response times (e.g. Bowling et al., 2023) to model-based techniques, such as person-fit statistics (e.g. Drasgow et al., 1985), structural equation models (e.g. Kim et al., 2018), or mixture models (e.g. Arias et al., 2020). More recently, machine learning techniques have been proposed (Welz & Alfons, 2023; Schroeders et al., 2022). We refer to Alfons & Welz (2024) for a recent review of carelessness in survey data.

This paper is structured as follows. Section 2 reviews the polychoric model and maximum likelihood estimation thereof, Section 3 introduces the proposed methodology, Sections 4 and 5 respectively contain a simulation study and an empirical application on Big-5 data, and Section 6 concludes.

2 Polychoric correlation

The polychoric correlation model (Pearson & Pearson, 1922) models the association between two discrete ordinal variables by assuming that an observed pair of responses to two polytomous items is governed by an unobserved discretization process of latent variables that jointly follow a bivariate standard normal distribution.[1] It is an alternative to classic sample corre-

[1]The polychoric correlation model of Pearson & Pearson (1922) generalizes a previous model for dichotomous responses called *tetrachoric* correlation model (Pearson, 1901). Correspondingly, when both items are dichotomous, one may alternatively speak of tetrachoric correlation instead of polychoric correlation.

lation (also known as Pearson's correlation coefficient) when the data are discrete ratings to polytomous or dichotomous items since sample correlation is designed for continuous variables. When data are indeed discrete and ordinal, it is often recommended to use polychoric correlation over sample correlation, especially when the number of response options is relatively small, because sample correlation can be biased in such data (Foldnes & Grønneberg, 2022; Garrido et al., 2013; Holgado-Tello et al., 2010; Olsson, 1979b). However, their discrepancy may diminish with five or more response categories and when category thresholds are symmetric (Rhemtulla et al., 2012). As we will show later, achieving a robust estimation of the association between discrete ordinal variables necessitates the robust estimation of models specifically designed to capture the association between such variables, accounting for potential misspecifications of these models. In contrast, well-known robust alternatives to sample correlation such as minimum covariance determinant estimators (Rousseeuw, 1985) will not be effective when the data are discrete because they are designed for continuous variables. To robustly estimate the polychoric model, we leverage a recently developed general framework for robust estimation in discrete data (Welz, 2024b). Before we introduce our estimator in the next section, we briefly review the polychoric correlation model and maximum likelihood estimation thereof.

2.1 The polychoric model

Suppose we observe two ordinal discrete variables, X and Y, that take values in the sets $\mathcal{X} = \{1, 2, \ldots, K_x\}$ and $\mathcal{Y} = \{1, 2, \ldots, K_y\}$, respectively. The assumption that the sets contain adjacent integers is without loss of generality. One either observes realizations of X and Y directly or by means of a $K_x \times K_y$ contingency table that cross-tabulates observed frequencies. Further assume that there exist two latent random variables, ξ and η, that govern the observed discrete variables as follows:

$$X = \begin{cases} 1 & \text{if } \xi < a_1, \\ 2 & \text{if } a_1 \leq \xi < a_2, \\ 3 & \text{if } a_2 \leq \xi < a_3, \\ \vdots & \\ K_x & \text{if } a_{K_x - 1} \leq \xi, \end{cases} \quad \text{and} \quad Y = \begin{cases} 1 & \text{if } \eta < b_1, \\ 2 & \text{if } b_1 \leq \eta < b_2, \\ 3 & \text{if } b_2 \leq \eta < b_3, \\ \vdots & \\ K_y & \text{if } b_{K_y - 1} \leq \eta, \end{cases} \tag{1}$$

where the (unobserved) parameters $a_1 < a_2 < \cdots < a_{K_x - 1}$ and $b_1 < b_2 < \cdots < b_{K_y - 1}$ are called *thresholds*. Given a sample of (X, Y), the primary statistical problem is to estimate the correlation between the latent variables ξ and η, denoted $\rho = \mathbb{C}\text{or}\,[\xi,\ \eta]$. If one assumes that (ξ, η) are bivariate standard normally distributed with correlation $\rho \in (-1, 1)$, the correlation parameter ρ is called the *polychoric correlation coefficient* of X and Y, and process (1) gives rise the *polychoric model*. In this model, the process of mapping latent variables to discrete responses in (1) depends on $d = K_x + K_y - 1$ parameters, namely the polychoric correlation coefficient and two sets of thresholds, which are jointly collected in a

parameter vector

$$\boldsymbol{\theta} = \left(\rho, a_1, a_2, \ldots, a_{K_x-1}, b_1, b_2, \ldots, b_{K_y-1}\right)^\top.$$

Since the polychoric model assumes (ξ, η) to be standard normally distributed with correlation ρ, the process (1) implies that the probability of observing a contingency table cell $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is given by

$$p_{xy}(\boldsymbol{\theta}) = \mathbb{P}_{\boldsymbol{\theta}}\left[X = x, Y = y\right] = \int_{a_{x-1}}^{a_x} \int_{b_{y-1}}^{b_y} \phi_2\left(t, s; \rho\right) \mathrm{d}s \ \mathrm{d}t, \tag{2}$$

where we use the conventions $a_0 = b_0 = -\infty, a_{K_x} = b_{K_y} = +\infty$, and

$$\phi_2\left(u, v; \rho\right) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{u^2 - 2\rho uv + v^2}{2(1-\rho^2)}\right)$$

denotes the density of the bivariate standard normal distribution function with correlation $\rho \in (-1, 1)$ at some $u, v \in \mathbb{R}$, with corresponding distribution function

$$\Phi_2\left(u, v; \rho\right) = \int_{-\infty}^{u} \int_{-\infty}^{v} \phi_2\left(t, s; \rho\right) \mathrm{d}s \ \mathrm{d}t.$$

One collectively refers to the probabilities (2) as the *polychoric model*. This model is parametrized by the d-dimensional vector $\boldsymbol{\theta}$, so model fitting necessitates estimating the parameters in $\boldsymbol{\theta}$, including the object of primary interest, the polychoric correlation coefficient ρ. To distinguish between arbitrary parameter vectors $\boldsymbol{\theta}$ and the "true" parameter under which the polychoric model generates data, denote this true parameter by $\boldsymbol{\theta}_* = \left(\rho_*, a_{*,1}, \ldots, a_{*,K_x-1}, b_{*,1}, \ldots, b_{*,K_y-1}\right)^\top$. Correspondingly, the statistical problem is to estimate $\boldsymbol{\theta}_*$, which is typically done by maximum likelihood.

2.2 Maximum likelihood estimation

Suppose we observe a sample $\{(X_i, Y_i)\}_{i=1}^N$ of N independent copies of (X, Y) generated by process (1). Denote by

$$N_{xy} = \sum_{i=1}^{N} \mathbb{1}\left\{X_i = x, Y_i = y\right\}$$

the empirical frequency of a cell $(x, y) \in \mathcal{X} \times \mathcal{Y}$. We are interested in estimating the polychoric correlation coefficient $\rho = \mathbb{C}\mathrm{or}\left[\xi, \ \eta\right]$ as well as the thresholds $\left(a_1, \ldots, a_{K_x-1}\right)^\top$ and $\left(b_1, \ldots, b_{K_y-1}\right)^\top$ in the polychoric model (2). The maximum likelihood estimator (MLE) of the the true parameter vector $\boldsymbol{\theta}_*$ is then given by

$$\widehat{\boldsymbol{\theta}}_N^{\mathrm{MLE}} = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\{ \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} N_{xy} \log\left(p_{xy}(\boldsymbol{\theta})\right) \right\}, \tag{3}$$

where

$$\boldsymbol{\Theta} = \left(\left(\rho, \left(a_i\right)_{i=1}^{K_x-1}, \left(b_j\right)_{j=1}^{K_y-1} \right)^\top \;\middle|\; \rho \in (-1,1),\ a_1 < \cdots < a_{K_x-1},\ b_1 < \cdots < b_{K_y-1} \right)$$

is a set of parameters $\boldsymbol{\theta}$ that rules out degenerate cases that are not permitted by the polychoric model, like $\rho = \pm 1$ or thresholds that are not strictly monotonically increasing. This estimator, its computational details, as well as its statistical properties are derived in Olsson (1979a). In essence, if the polychoric model (2) is correctly specified—that is, (ξ, η) are indeed bivariate standard normal—and the sample $\{(X_i, Y_i)\}_{i=1}^N$ comprises independent and identical draws from this model, then estimator $\widehat{\boldsymbol{\theta}}_N^{\text{MLE}}$ is unbiased and consistent for the true $\boldsymbol{\theta}_*$, and is asymptotically normally distributed with the smallest possible variance.

As a computationally attractive alternative to estimating all parameters in $\boldsymbol{\theta}_*$ simultaneously, one may consider a "2-step-approach" where only the correlation coefficient ρ_* is estimated via maximum likelihood, but not the thresholds. In this approach, one estimates in a first step the thresholds as quantiles of the univariate standard normal distribution, evaluated at the observed cumulative marginal proportion of each cell. Formally, in the 2-step-approach, thresholds $a_{*,x}$ and $b_{*,y}$ are respectively estimated via

$$\widehat{a}_x = \Phi_1^{-1}\left(\frac{1}{N} \sum_{k=1}^{x} \sum_{y \in \mathcal{Y}} N_{ky} \right) \qquad \text{and} \qquad \widehat{b}_y = \Phi_1^{-1}\left(\frac{1}{N} \sum_{x \in \mathcal{X}} \sum_{l=1}^{y} N_{xl} \right),$$

for $x = 1, \ldots, K_x - 1$ and $y = 1, \ldots, K_y - 1$, where $\Phi_1^{-1}(\cdot)$ denotes the quantile function of the univariate standard normal distribution. Then, taking these threshold estimates as fixed in the polychoric model, one estimates in a second step the only remaining parameter, correlation coefficient ρ_*, via maximum likelihood. The main advantage of the 2-step approach is reduced computational time, while it comes at the cost of being theoretically non-optimal (Olsson, 1979a). As it turns out, another disadvantage of the 2-step-approach is that it cannot be made robust against nonnormality, so in our robust method, we need to estimate all parameters in $\boldsymbol{\theta}_*$ simultaneously.

2.3 Non-robustness of maximum likelihood

Based on simulation studies, maximum likelihood estimation of the polychoric model was initially believed to be fairly robust to moderate violations of the latent normality assumption (e.g. Li, 2016; Flora & Curran, 2004; Coenders et al., 1997). However, the simulation design employed in these studies turned out to be equivalent to simulating exactly normally distributed data (Grønneberg & Foldnes, 2019). Using updated simulation designs that ensure proper violation of the normality assumption, Foldnes & Grønneberg (2022) conclude that maximum likelihood estimation of polychoric correlation is in fact highly sensitive to such violations, leading to potentially large biases. In the robust statistics literature, maximum likelihood estimation has been mathematically shown to be heavily biased even when the

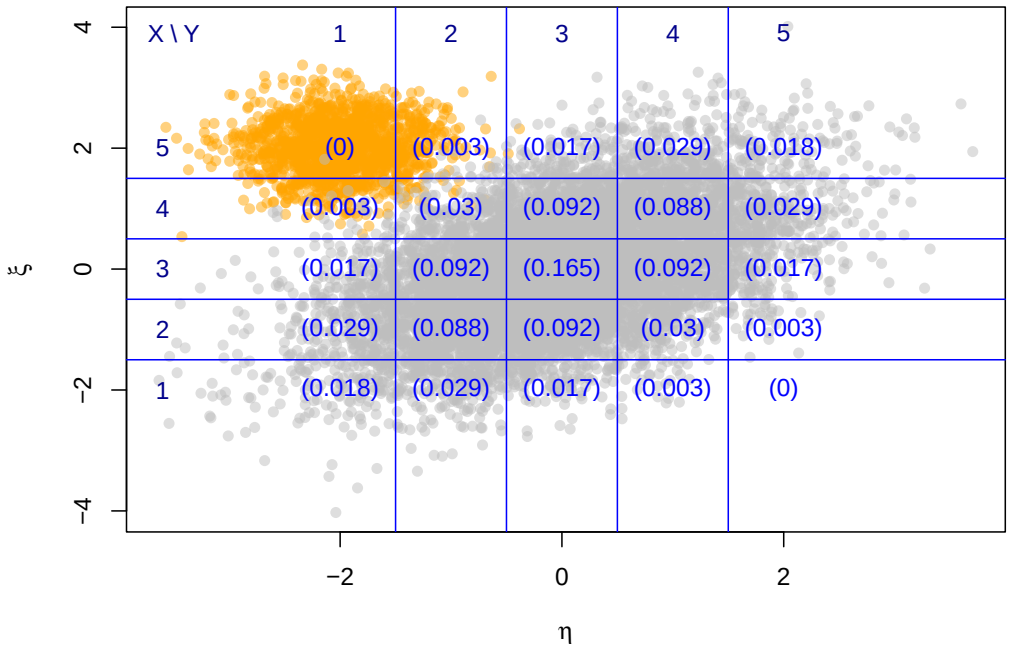


Figure 1: Simulated data with $K_x = K_y = 5$ response options where the polychoric model is misspecified due to a fraction $\varepsilon = 0.15$ of nonnormal data. The gray dots represent random draws of (ξ, η) from the polychoric model with $\rho_* = 0.5$, whereas the orange dots represent draws from a different distribution that inflates cell $(x, y) = (5, 1)$. The blue lines indicate the location of the thresholds. In each cell, the numbers in parentheses denote the population probability of that cell under the true polychoric model.

assumed model is only slightly misspecified (e.g. Maronna et al., 2018; Huber & Ronchetti, 2009; Hampel et al., 1986). Notwithstanding, we stress that also sample correlation is prone to violations of assumptions (Raymaekers & Rousseeuw, 2021), even when the data are discrete (Welz et al., 2023). Hence, robustly estimating association between discrete ordinal variables requires new estimation methods, which is the contribution of this paper.

3 Robust estimation of polychoric correlation

3.1 Conceptualizing model misspecification

We say that that the polychoric model is misspecified when at least some of the observed data have been generated by a different model. More specifically, we consider a situation where only a fraction $(1 - \varepsilon)$ of the latent realizations of (ξ, η) follow a bivariate standard

normal distribution with true correlation parameter ρ_*, whereas a fraction $\varepsilon \in [0, 1]$ come from some different unspecified distribution, G. It follows that in this situation, the latent variables (ξ, η) are jointly distributed according to the mixture distribution

$$(u, v) \mapsto F_\varepsilon(u, v) = (1 - \varepsilon)\Phi_2\left(u, v; \rho_*\right) + \varepsilon G(u, v), \tag{4}$$

for $u, v \in \mathbb{R}$. We call ε the *degree of misspecification*, G the *misspecifying distribution*, and F_ε the *misspecified distribution*. Neither ε nor G are assumed to be known, and subsequently both quantities are left completely unspecified in practice. Conceptualizing model misspecification through a misspecified distribution F_ε is standard in the robust statistics literature, and has been proposed in the seminal work of Huber (1964). Observe that when the degree of misspecification is zero, that is, $\varepsilon = 0$, then there is no misspecification so that the polychoric model is correctly specified.

Leaving the misspecifying distribution G and misspecification degree ε in mixture distribution (4) unspecified means that we are not making *any assumption whatsoever* on the degree, magnitude, or type of misspecification (which is possibly absent altogether). Hence, in our context of responses to rating items, the polychoric model can be misspecfied due to an unlimited variety of reasons, for instance but not limited to careless/inattentive responding (e.g., straightlining, pattern responding, random-like responding), misresponses, item misunderstanding, or accurate responses that are simply not generated by latent normality.

Figure 1 visualizes a simulated example of bivariate data drawn from misspecified distribution F_ε, where a fraction of $\varepsilon = 0.15$ of the data follow a (misspecifying) distribution with mean $(2, -2)^\top$ (orange dots), whereas the remaining data are generated by a bivariate standard normal distribution with correlation $\rho_* = 0.5$ (gray dots). In this example, the data from the misspecifying distribution will primarily inflate the cell $(x, y) = (5, 1)$, in the sense that this cell will have a larger empirical frequency than the polychoric model allows for, since the probability of this cell is nearly zero at the polychoric model yet many realized responses will populate it. Consequently, due to misspecification of the (polychoric) model, a maximum likelihood estimate of ρ_* on these data will not be consistent for ρ_* (Maronna et al., 2018; Huber & Ronchetti, 2009; Hampel et al., 1986). Indeed, calculating the MLE using the data plotted in Figure 1 yields an estimate of $\widehat{\rho}_N^{\text{MLE}} = -0.05$, which is far off from the true $\rho_* = 0.5$. In contrast, our proposed robust estimator, which is calculated from the exact same information as the MLE and is defined in the next section, yields a fairly accurate estimate of 0.45.

3.2 The estimator

Motivated by the MLE's inconsistency when the polychoric model is misspecified, we propose an estimator of $\boldsymbol{\theta}_*$ that is more robust to misspecification when present, but yields the asymptotically same estimate as the MLE when the polychoric model is correctly specified. For each cell $(x, y) \in \mathcal{X} \times \mathcal{Y}$, our estimator minimizes a certain disparity between empirical

relative cell frequencies,

$$\widehat{f}_N(x,y) = N_{xy}/N = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left\{X_i = x, Y_i = y\right\},$$

and the theoretical cell probabilities $p_{xy}(\boldsymbol{\theta})$ returned by the polychoric model (2) at a parameter $\boldsymbol{\theta}$. Specifically, the estimator minimizes with respect to $\boldsymbol{\theta}$ the loss function

$$L\left(\boldsymbol{\theta},\ \widehat{f}_N\right) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \varphi\left(\frac{\widehat{f}_N(x,y)}{p_{xy}(\boldsymbol{\theta})}\right) p_{xy}(\boldsymbol{\theta}), \tag{5}$$

where $\varphi : [0,\infty) \to \mathbb{R}$ is a prespecified function that will be defined momentarily. The proposed estimator $\widehat{\boldsymbol{\theta}}_N$ is given by the value minimizing the objective loss over $\boldsymbol{\Theta}$,

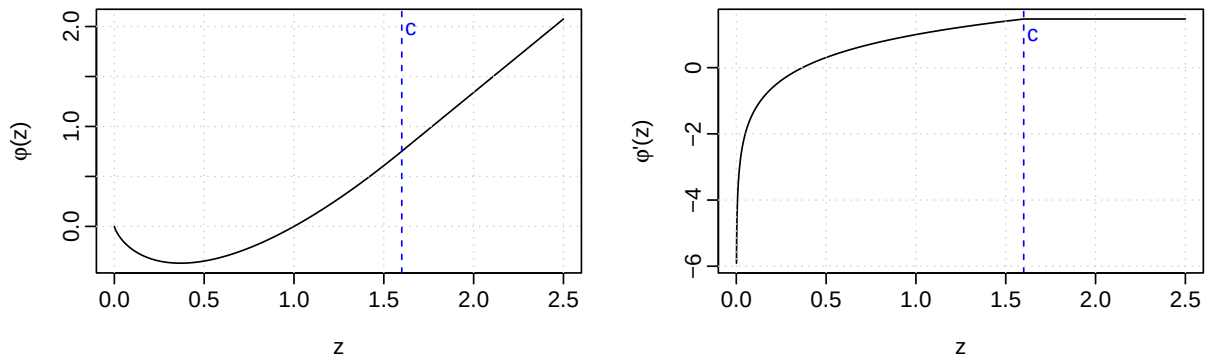
$$\widehat{\boldsymbol{\theta}}_N = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L\left(\boldsymbol{\theta}, \widehat{f}_N\right). \tag{6}$$

Estimators that minimize a loss function of the type in (5) are called *minimum disparity estimators* (Lindsay, 1994) because they minimize a certain disparity between empirical probabilities ($\widehat{f}_N(x,y)$ here) and theoretical probabilities ($p_{xy}(\boldsymbol{\theta})$ here). A small disparity indicates that the assumed model is able to fit observed data well. The disparity is governed by the choice of the function $\varphi(\cdot)$. Many well-known estimators can be written as minimum disparity estimators, including the MLE, through the choice of $\varphi(\cdot)$ (Victoria-Feser & Ronchetti, 1997; Lindsay, 1994). In the following, we motivate a specific choice of $\varphi(\cdot)$ that makes the estimator $\widehat{\boldsymbol{\theta}}_N$ less susceptible to misspecification of the polychoric model.

The fraction $\widehat{f}_N(x,y)/p_{xy}(\boldsymbol{\theta})$ in (5) is called a *Pearson residual*[2] (Lindsay, 1994) and can be interpreted as a goodness-of-fit measure for cell (x,y). Values close to 1 indicate a good fit between data and assumed model at $\boldsymbol{\theta}$, whereas values toward 0 or $+\infty$ indicate a poor fit. Indeed, cells whose observations are primarily generated by a nonnormal distribution will generally have a Pearson residual away from 1. Hence, to achieve robustness to such misspecifcation, cell frequencies that cannot be modeled well by the polychoric model, as indicated by their Pearson residual being away from 1, should be downweighted in the estimation procedure such that they do not over-proportionally affect the fit. Such robustness can be achieved by choosing an appropriate function for $\varphi(\cdot)$ in loss function (5). We propose to choose the following specification, suggested by Ruckstuhl & Welsh (2001) for robustly fitting the binomial model,

$$\varphi(z) = \begin{cases} z\log(z) & \text{if } z \in [0,c], \\ z(\log(c)+1) & \text{if } z > c, \end{cases} \tag{7}$$

[2]Technically, Lindsay (1994) defines Pearson residuals as $\widehat{f}_N(x,y)/p_{xy}(\boldsymbol{\theta}) - 1$. We renounce on subtracting the value 1 because it makes notation simpler in this paper.



(a) $\varphi(z)$ **(b)** $\varphi'(z)$

Figure 2: Visualization of the function $\varphi(z)$ in (7) (left panel) and its derivative (right panel), for $c = 1.6$ (vertical dashed blue lines).

where $c \in [1, \infty]$ is a prespecified tuning constant.[3] Figure 2 visualizes this function for the example choice $c = 1.6$. Note that function $\varphi(\cdot)$ is convex, rendering the estimator's optimization problem in (6) convex.

It is easy to see that for the choice $c = +\infty$ in function $\varphi(\cdot)$, minimizing the loss (5) is equivalent to maximizing the log-likelihood objective in (3), meaning that the estimator $\widehat{\boldsymbol{\theta}}_N$ is equal to $\widehat{\boldsymbol{\theta}}_N^{\text{MLE}}$ for this choice of c. More specifically, if a Pearson residual $z = \frac{\widehat{f}_N(x,y)}{p_{xy}(\boldsymbol{\theta})}$ of a cell $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is such that $z \in [0, c]$ for fixed $c \geq 1$, then the estimation procedure behaves at this cell like in maximum likelihood estimation. If this Pearson residual z equals 1, then its associated cell can be fitted perfectly with the polychoric model, so at this cell the estimation procedure will behave like maximum likelihood *regardless* of the choice of $c \geq 1$. In the absence of misspecification ($\varepsilon = 0$), $\widehat{f}_N(x, y) \xrightarrow{\text{a.s.}} p_{xy}(\boldsymbol{\theta}_*)$ as $N \to \infty$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$ (see e.g., Chapter 19.2 in Van der Vaart, 1998), meaning that all Pearson residuals are asymptotically equal to 1. In other words, if there is no misspecification, $\widehat{\boldsymbol{\theta}}_N$ is asymptotically equivalent to $\widehat{\boldsymbol{\theta}}_N^{\text{MLE}}$ no matter the choice of tuning constant c. On the other hand, if a cell's Pearson residual is far away from 1, it cannot be fitted well with the polychoric model, which is typically indicative of the polychoric model being misspecified. In this case, the cell should not be treated like in maximum likelihood estimation because maximum likelihood is not consistent under misspecification. Instead, the cell's influence on the final estimate should be downweighted to avoid that cells that cannot be fitted well dominate the fit, which happens in maximum likelihood. Such downweighting is employed by function $\varphi(z)$ in (7) whenever $z > c \geq 1$, that is, the Pearson residual is sufficiently far away from the ideal value 1, where the choice of c governs what is deemed "sufficiently far". Indeed, for values $z > c$, the function $\varphi(z)$ increases only linearly with z, as opposed to the non-linear exponential increase when $z \leq c$. The notion of requiring nonlinear effects for

[3] Equation (7) is actually just a special case of a more general formulation in Ruckstuhl & Welsh (2001).

the bulk of the data and linear effects in its tails is similar to classic robust estimation as in Huber (1964).

It is shown in Figure 2 how $\varphi(z)$ transitions from exponential growth to linear growth at $z = c$, as well as the boundedness of its first derivative when c is finite. Hence, if c is finite, any Pearson residual can only have a bounded effect on the final estimator, as opposed to unbounded effects in maximum likelihood estimation where $c = +\infty$. Thus, we achieve robustness against misspecification through the choice of c. The closer to 1 one chooses c, the more robust the estimator becomes. However, there is a well-known tradeoff between robustness and efficiency for robust estimators: the more robust an estimator, the more estimation variance is introduced (e.g. Huber & Ronchetti, 2009). Therefore, by choosing c, one is effectively choosing between robustness and efficiency. A characterization of the this tradeoff is work in progress.

With the proposed choice of $\varphi(\cdot)$, we stress that our estimator $\widehat{\boldsymbol{\theta}}_N$ in (6) has the same time complexity as maximum likelihood, that is, $O\big(K_x \times K_y\big)$, since one needs to calculate the Pearson residual of all $K_x \times K_y$ cells for every candidate parameter value. Consequently, our proposed estimator does not incur any additional computational cost compared to maximum likelihood, and therefore robustness can be achieved without having to pay a computational price.

3.3 Statistical properties

3.3.1 Estimand

Before we can turn to deriving the statistical properties of estimator $\widehat{\boldsymbol{\theta}}_N$ in (5) with (7) as the choice of function $\varphi(\cdot)$, we first require some additional notation. For unknown misspecification degree $\varepsilon \in [0, 1]$, denote by

$$f_\varepsilon\left(x, y\right) = (1 - \varepsilon) p_{xy}(\boldsymbol{\theta}_*) + \varepsilon \int_{a_{*,x-1}}^{a_{*,x}} \int_{b_{*,y-1}}^{b_{*,y}} \mathrm{d}G$$

the unobserved probability that cell $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is sampled under the possibly misspecified distribution F_ε in (4). Note that $f_0(x, y) = p_{xy}(\boldsymbol{\theta}_*)$ if there is no misspecification ($\varepsilon = 0$). When a sample is generated by distribution F_ε, the relative empirical cell frequency $\widehat{f}_N(x, y)$ is a (strongly) consistent estimator of $f_\varepsilon\left(x, y\right)$ when $N \to \infty$.

With this definition in mind, we now address what the estimator $\widehat{\boldsymbol{\theta}}_N$ in (6) actually estimates. Its estimand is given by

$$\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} L\Big(\boldsymbol{\theta}, f_\varepsilon\Big).$$

This minimization problem is simply the population analogue to the minimization problem in (6) that the sample-based $\widehat{\boldsymbol{\theta}}_N$ solves because the $f_\varepsilon\left(x, y\right)$ are the population analogues to the $\widehat{f}_N(x, y)$. In the absence of misspecification, $\boldsymbol{\theta}_0$ equals the true parameter $\boldsymbol{\theta}_*$. However, in

the presence of misspecification ($\varepsilon > 0$), it is generally different from $\boldsymbol{\theta}_*$. How much different it is depends on the degree of misspecification ε as well as the choice of tuning constant c in $\varphi(\cdot)$. In general, the larger ε (more severe misspecification) and c (less downweighting of hard-to-fit cells), the further $\boldsymbol{\theta}_0$ is away from $\boldsymbol{\theta}_*$. Hence, for fixed misspecification degree ε, the MLE ($c = +\infty$) will estimate a parameter that is farther or equally far away from the true $\boldsymbol{\theta}_*$ than for finite choices of c. Correspondingly, finite choices of c lead to an estimator that is at least as accurate as the MLE, and more accurate under misspecification of the polychoric model. A relevant question is whether the true parameter $\boldsymbol{\theta}_*$ can be identified when $\varepsilon > 0$ such that it can be estimated using $\widehat{\boldsymbol{\theta}}_N$ combined with a correction term. This question is closely related to point-identifying $\boldsymbol{\theta}_*$ under misspecification, which is work in progress.

3.3.2 Assumptions

Throughout this section, we assume that the number of response categories, K_x and K_y, are fixed and known, and that the sample $\{(X_i, Y_i)\}_{i=1}^N$ used to compute an estimate $\widehat{\boldsymbol{\theta}}_N$ has been generated by the process in (1) where the latent $\{(\xi_i, \eta_i)\}_{i=1}^N$ are draws from distribution F_ε in (4) with unobserved misspecification degree $\varepsilon \in [0, 1]$ and unknown misspecifying distribution G. In the following, we list a set of assumptions that will be entertained in the asymptotic analysis of $\widehat{\boldsymbol{\theta}}_N$. These assumptions are based on Welz (2024b).

Assumption Set A. *Suppose that the following assumptions hold true.*

A.1 $c \in [1, +\infty]$,

A.2 $\boldsymbol{\Theta} \subset \mathbb{R}^d$ *is compact, where* $d = K_x + K_y - 1$ *denotes the number of parameters in the polychoric model* (2),

A.3 $\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} L(\boldsymbol{\theta}, f_\varepsilon)$ *is a unique minimum, and* $\boldsymbol{\theta}_0$ *is an interior point of* $\boldsymbol{\Theta}$,

A.4 $p_{xy}(\boldsymbol{\theta}) > 0$ *for all* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ *and all* $(x, y) \in \mathcal{X} \times \mathcal{Y}$,

A.5 $\#\big\{(x, y) \in \mathcal{X} \times \mathcal{Y} : f_\varepsilon(x, y) > 0\big\} > d$,

A.6 $\frac{f_\varepsilon(x,y)}{p_{xy}(\boldsymbol{\theta}_0)} \neq c$ *for all* $(x, y) \in \mathcal{X} \times \mathcal{Y}$.

Assumption A.1 ensures that function $\varphi(\cdot)$ exhibits meaningful behavior when evaluated at Pearson residuals, such as the ideal residual value 1 being included in the interval $[0, c]$. Compactness of the parameter space (Assumption A.2) is primitive and required for a technicality when proving consistency of $\widehat{\boldsymbol{\theta}}_N$.[4] Uniqueness of a global minimum that is well separated from the boundary (Assumption A.3) is a common assumption in the literature on M-type estimators (e.g., Chapter 5.2 in Van der Vaart, 1998) like the one presented in

[4]This assumption can possibly be modified to $\boldsymbol{\Theta}$ being open by equipping it with a specific topological structure.

this paper. The assumption of strictly positive probabilities (A.4) is standard in the literature on minimum-disparity-type estimators (e.g., Cressie & Read, 1984; Victoria-Feser & Ronchetti, 1997) and rules out that one divides by zero when computing Pearson residuals. Assumption A.5 imposes that there are more populated (non-empty) cells than parameters in the polychoric model. In other words, to estimate all model parameters, there must be more sources of variation (populated cells) than parameters. One may view this assumption as a rank condition that ensures invertibility of the optimization problem's Hessian matrix. Finally, Assumption A.6 imposes that the population Pearson residual at the global minimum is not equal to tuning constant c. This is again a primitive assumption required for a technicality in the proof. This assumption does not have practical implications because such an event has probability zero.

We emphasize that no assumption restricts the type, source, or magnitude of potential misspecification of the polychoric model. In addition, most assumptions are not unique to our estimator. In fact, Assumptions A.2–A.5 are also required for consistency and asymptotic normality of the MLE. Only Assumptions A.1 and A.6 are specific to our proposed robust estimator because they concern tuning constant c.

3.3.3 Asymptotic analysis

The following theorem establishes almost sure (a.s.) convergence of $\widehat{\boldsymbol{\theta}}_N$ for $\boldsymbol{\theta}_0$.

Theorem 1 (Consistency)**.** *Under Assumptions A.1–A.4, it holds true that*

$$\widehat{\boldsymbol{\theta}}_N \xrightarrow{\text{a.s.}} \boldsymbol{\theta}_0,$$

as $N \to \infty$.

This theorem follows immediately from Theorem 1 in Welz (2024b).

We now study the limit distribution of the estimator. Doing so necessitates additional notation. For fixed tuning constant $c \geq 1$, let

$$w(z) = \mathbb{1}\left\{z \in [0, c]\right\} + c\mathbb{1}\left\{z > c\right\}/z \quad \text{for } z \geq 0,$$

with first derivative

$$w'(z) = 0\mathbb{1}\left\{z \in [0, c]\right\} - c\mathbb{1}\left\{z > c\right\}/z^2,$$

and further define the d-dimensional gradient of $\log\left(p_{xy}(\boldsymbol{\theta})\right)$ for cell $(x, y) \in \mathcal{X} \times \mathcal{Y}$ at parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ as

$$\boldsymbol{s}_{xy}(\boldsymbol{\theta}) = \frac{1}{p_{xy}(\boldsymbol{\theta})} \left(\frac{\partial}{\partial \boldsymbol{\theta}}\, p_{xy}(\boldsymbol{\theta})\right),$$

where one can write for the gradient of $p_{xy}(\boldsymbol{\theta})$

$$\frac{\partial}{\partial \boldsymbol{\theta}}\, p_{xy}(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \Phi_2\left(a_x, b_y; \rho\right) - \frac{\partial}{\partial \boldsymbol{\theta}} \Phi_2\left(a_{x-1}, b_y; \rho\right) - \frac{\partial}{\partial \boldsymbol{\theta}} \Phi_2\left(a_x, b_{y-1}; \rho\right) + \frac{\partial}{\partial \boldsymbol{\theta}} \Phi_2\left(a_{x-1}, b_{y-1}; \rho\right),$$

see e.g. Olsson (1979a, equation 4), as well as the $d \times d$ Hessian matrix of $\log\left(p_{xy}(\boldsymbol{\theta})\right)$ as

$$\boldsymbol{Q}_{xy}(\boldsymbol{\theta}) = \frac{1}{p_{xy}(\boldsymbol{\theta})}\left(\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}p_{xy}(\boldsymbol{\theta})\right) - \boldsymbol{s}_{xy}(\boldsymbol{\theta})\boldsymbol{s}_{xy}(\boldsymbol{\theta})^\top.$$

We derive closed-form expressions of all components in $\boldsymbol{Q}_{xy}(\boldsymbol{\theta})$ in Appendix A. In addition, for $K = K_x \times K_y$ the total number of cells and d-dimensional vectors

$$\boldsymbol{w}_{xy}\left(\boldsymbol{\theta}\right) = \boldsymbol{s}_{xy}(\boldsymbol{\theta})\mathbb{1}\left\{\frac{f_\varepsilon\left(x,y\right)}{p_{xy}(\boldsymbol{\theta})} \in [0,c]\right\},$$

define the $d \times K$ matrix

$$\boldsymbol{W}(\boldsymbol{\theta}) = \left(\boldsymbol{w}_{11}\left(\boldsymbol{\theta}\right),\cdots,\boldsymbol{w}_{1,K_y}\left(\boldsymbol{\theta}\right),\boldsymbol{w}_{21}\left(\boldsymbol{\theta}\right),\cdots,\boldsymbol{w}_{2,K_y}\left(\boldsymbol{\theta}\right),\cdots,\boldsymbol{w}_{K_x,1}\left(\boldsymbol{\theta}\right),\boldsymbol{w}_{K_x,2}\left(\boldsymbol{\theta}\right),\cdots,\boldsymbol{w}_{K_x,K_y}\left(\boldsymbol{\theta}\right)\right)$$

that row-binds all K vectors of $\boldsymbol{s}_{xy}(\boldsymbol{\theta})$ multiplied by an indicator that takes value 1 when associated population Pearson residual is in the MLE-part of the function $\varphi(\cdot)$ in (7) and 0 otherwise. In similar fashion, define the K-dimensional vector

$$\boldsymbol{f}_\varepsilon = \left(f_\varepsilon\left(1,1\right),\ldots,f_\varepsilon\left(1,K_y\right),f_\varepsilon\left(2,1\right),\ldots,f_\varepsilon\left(2,K_y\right),\ldots,f_\varepsilon\left(K_x,1\right),f_\varepsilon\left(K_x,2\right),\ldots,f_\varepsilon\left(K_x,K_y\right)\right)^\top$$

that holds all K evaluations of the function f_ε, and put

$$\boldsymbol{\Lambda} = \operatorname{diag}(\boldsymbol{f}_\varepsilon) - \boldsymbol{f}_\varepsilon\boldsymbol{f}_\varepsilon^\top.$$

The next theorem establishes root-N consistency and asymptotic normality of the estimator. This theorem follows immediately from Theorem 2 in Welz (2024b).

Theorem 2 (Asymptotic normality)**.** *Grant the assumptions of Assumption Set A. Then*

$$\sqrt{N}\left(\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_0\right) \xrightarrow{\text{d}} \text{N}_d\Big(\mathbf{0}, \boldsymbol{\Sigma}\left(\boldsymbol{\theta}_0\right)\Big),$$

as $N \to \infty$*, where*

$$\boldsymbol{\Sigma}\left(\boldsymbol{\theta}\right) = \boldsymbol{M}\left(\boldsymbol{\theta}\right)^{-1}\boldsymbol{U}\left(\boldsymbol{\theta}\right)\boldsymbol{M}\left(\boldsymbol{\theta}\right)^{-1},$$

with $d \times d$ *symmetric matrices*

$$\boldsymbol{U}\left(\boldsymbol{\theta}\right) = \boldsymbol{W}\left(\boldsymbol{\theta}\right)\boldsymbol{\Lambda}\boldsymbol{W}\left(\boldsymbol{\theta}\right)^\top \qquad \textit{and}$$

$$\boldsymbol{M}\left(\boldsymbol{\theta}\right) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} f_\varepsilon\left(x,y\right)\left(w'\left(\frac{f_\varepsilon\left(x,y\right)}{p_{xy}(\boldsymbol{\theta})}\right)\frac{f_\varepsilon\left(x,y\right)}{p_{xy}(\boldsymbol{\theta})}\boldsymbol{s}_{xy}(\boldsymbol{\theta})\boldsymbol{s}_{xy}(\boldsymbol{\theta})^\top - w\left(\frac{f_\varepsilon\left(x,y\right)}{p_{xy}(\boldsymbol{\theta})}\right)\boldsymbol{Q}_{xy}(\boldsymbol{\theta})\right).$$

A strongly consistent estimator of the unobserved asymptotic covariance matrix $\boldsymbol{\Sigma}\left(\boldsymbol{\theta}_0\right)$ can be constructed as follows. Replace all population class probabilities $f_\varepsilon\left(x,y\right)$ by their corresponding empirical counterparts $\widehat{f}_N(x,y)$ in matrices $\boldsymbol{W}\left(\boldsymbol{\theta}\right), \boldsymbol{M}\left(\boldsymbol{\theta}\right)$, and $\boldsymbol{\Lambda}$. Then exploit the plug-in principle and evaluate $\boldsymbol{U}\left(\boldsymbol{\theta}\right)$ and $\boldsymbol{M}\left(\boldsymbol{\theta}\right)$ at the point estimate $\widehat{\boldsymbol{\theta}}_N$. Denote the ensuing plug-in estimator by $\boldsymbol{\Sigma}\left(\widehat{\boldsymbol{\theta}}_N\right)$, which is strongly consistent for $\boldsymbol{\Sigma}\left(\boldsymbol{\theta}_0\right)$ by Theorem 1 and the continuous mapping theorem.

3.4 Goodness-of-fit test

Suppose we wish to test the null hypothesis that a cell in a $K_x \times K_y$ contingency table is outlying in the sense that it cannot be fitted well by the polychoric model, which is indicative of model misspecification. This notion can be conceptualized by means of Pearson residuals. Recall that a Pearson residual of value 1 indicates that the corresponding cell can be fitted well, whereas a Pearson residual significantly larger than 1 indicates poor fit. For a cell $(x, y) \in \mathcal{X} \times \mathcal{Y}$, this translates into the natural null hypothesis with one-sided alternative

$$\mathrm{H}_0 : \frac{\widehat{f}_N(x,y)}{p_{xy}(\boldsymbol{\theta}_0)} = 1 \qquad \text{vs.} \qquad \mathrm{H}_1 : \frac{\widehat{f}_N(x,y)}{p_{xy}(\boldsymbol{\theta}_0)} > 1,$$

which is equivalent to

$$\mathrm{H}_0 : p_{xy}(\boldsymbol{\theta}_0) = \widehat{f}_N(x,y) \qquad \text{vs.} \qquad \mathrm{H}_1 : p_{xy}(\boldsymbol{\theta}_0) < \widehat{f}_N(x,y). \tag{8}$$

Ideally, a test for such a hypothesis will reject H_0 if the polychoric model is misspecified for that cell, and sustain H_0 if it is correctly specified for that cell. It turns out that a test statistic that satisfies these two desirable properties is given by

$$Z_N = \frac{p_{xy}(\widehat{\boldsymbol{\theta}}_N) - \widehat{f}_N(x,y)}{\sqrt{\sigma^2_{xy}(\boldsymbol{\theta}_0)/N}}, \tag{9}$$

where

$$\sigma^2_{xy}(\boldsymbol{\theta}) = \boldsymbol{g}_{xy}(\boldsymbol{\theta})^\top \boldsymbol{\Sigma}(\boldsymbol{\theta})\, \boldsymbol{g}_{xy}(\boldsymbol{\theta})$$

for gradient

$$\boldsymbol{g}_{xy}(\boldsymbol{\theta}) = \frac{\partial p_{xy}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}.$$

By Theorem 1 and the continuous mapping theorem, the variance term $\sigma^2_{xy}(\boldsymbol{\theta}_0)$ can be consistently estimated by $\sigma^2_{xy}\left(\widehat{\boldsymbol{\theta}}_N\right)$. The following theorem, which follows immediately from Theorem 3 in Welz (2024b), establishes validity and exactness of test statistic Z_N for testing the null hypothesis $\mathrm{H}_0 : \widehat{f}_N(x,y) = p_{xy}(\boldsymbol{\theta}_0)$.

Theorem 3 (Limit distribution of test statistic)**.** *Grant the assumptions of Assumption Set A. Then, under the null hypothesis in* (8)*, the test statistic* Z_N *in* (9) *possesses the following limit distribution:*

$$Z_N \xrightarrow{\mathrm{d}} \mathrm{N}(0,1),$$

as $N \to \infty$*.*

It follows that the test is exact in the statistical sense. In practice, it is only approximately exact because the variance term $\sigma^2_{xy}(\boldsymbol{\theta}_0)$ is unobserved and must be estimated by $\sigma^2_{xy}\left(\widehat{\boldsymbol{\theta}}_N\right)$

in order to compute an approximate test statistic

$$\frac{p_{xy}(\widehat{\boldsymbol{\theta}}_N) - \widehat{f}_N(x,y)}{\sqrt{\sigma^2_{xy}(\widehat{\boldsymbol{\theta}}_N)/N}}.$$

In addition, it should be noted that using this test to test all $K_x \times K_y$ cells for being outlying creates a multiple testing problem. We therefore recommend to adjust for multiple comparisons, for instance through the procedure of Benjamini & Hochberg (1995), when testing multiple cells for outlyingness.

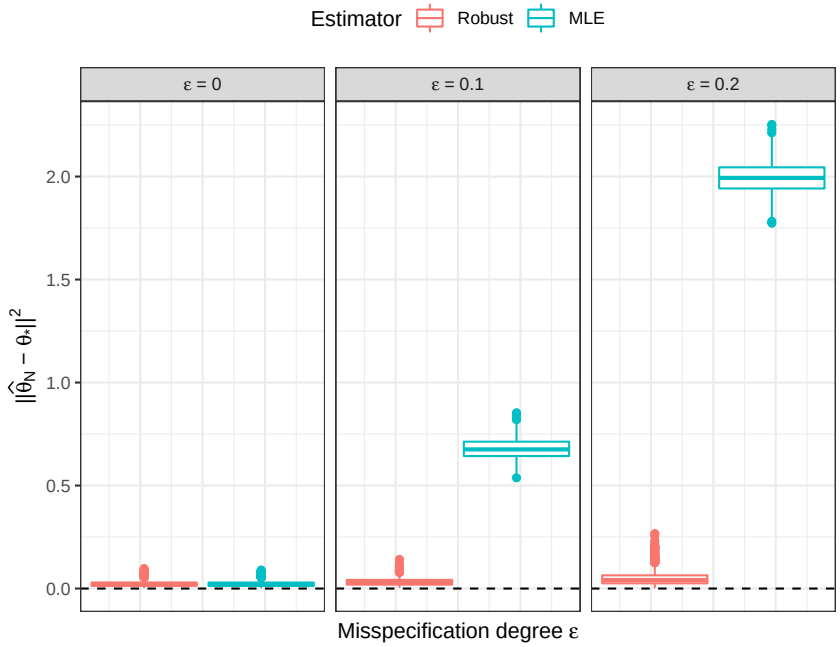
4 Simulation study

To verify the statistical properties of the proposed estimator and assess its performance in practice, we employ a simulation study. Let there be $K_x = K_y = 5$ response categories for each of the two rating variables and define the true thresholds in the polychoric model as

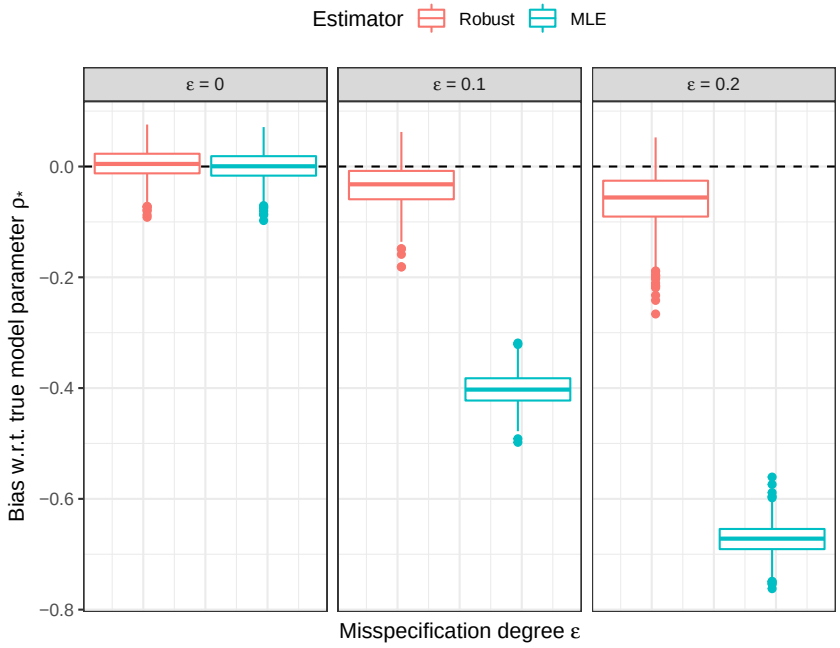
$$a_{*,1} = b_{*,1} = -1.5, \quad a_{*,2} = b_{*,2} = -0.5, \quad a_{*,3} = b_{*,3} = 0.5, \quad a_{*,4} = b_{*,4} = 1.5,$$

and let the true polychoric correlation coefficient be $\rho_* = 0.5$. To simulate misspecification of the polychoric model, we let a fraction ε of the data be generated by a misspecifying bivariate normal distribution with mean $(2, -2)^\top$, variances $(0.2, 0.2)^\top$, and zero covariance (and therefore zero correlation). This misspecifying distribution will inflate the empirical frequency of cells $(x, y) \in \{(5, 1), (4, 3), (5, 2)\}$, in the sense that they have a higher realization probability than under the true polychoric model. In fact, the data plotted in Figure 1 were generated by this process for misspecification degree $\varepsilon = 0.15$, and one can see in this figure that particularly cell $(x, y) = (5, 1)$ is sampled quite frequently although it only has a near-zero probability at the true polychoric model. The data points causing these three cells to be inflated are instances of *negative leverage points*. Here, such leverage points drag correlational estimates away from a positive value towards zero or, if there are sufficiently many of them, even a negative value.

For misspecification degrees $\varepsilon \in \{0, 0.1, 0.2\}$, we sample $N = 1,000$ responses from this data generating process and estimate the true parameter $\boldsymbol{\theta}_*$ with the MLE as well as our proposed estimator with tuning constant set to $c = 1.6$, since this choice yielded a good compromise between robustness and efficiency in further simulation studies. We repeat this procedure for 1,000 simulated datasets. As performance measures, we calculate the average bias, standard deviation across repetitions, coverage, and length of confidence intervals at significance level $\alpha = 0.05$. The coverage is defined as proportion of $(1 - \alpha)$-th confidence intervals $[\widehat{\rho}_N \mp q_{1-\alpha/2} \cdot \mathrm{SE}(\widehat{\rho}_N)]$ that contain the true ρ_*, where $q_{1-\alpha/2}$ is the $(1 - \alpha/2)$-th quantile of the standard normal distribution and $\mathrm{SE}(\widehat{\rho}_N)$ is the standard error of $\widehat{\rho}_N$, which is constructed using the limit theory developed in Theorem 2. The length of a confidence interval is given by $2 \cdot q_{1-\alpha/2} \cdot \mathrm{SE}(\widehat{\rho}_N)$.



(a) $\|\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_*\|^2$



(b) $\widehat{\rho}_N - \rho_*$

Figure 3: Boxplot visualization of the bias of two estimators, the MLE and the proposed robust estimator with $c = 1.6$, across 1,000 simulated datasets. The top panel shows the mean squared error for the whole vector $\boldsymbol{\theta}_*$ (that is, $\|\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_*\|^2$), whereas the bottom panel shows the bias of the polychoric correlation coefficient, $\widehat{\rho}_N - \rho_*$, which is the first coordinate in $\widehat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}_*$.

Misspecification	Estimator	$\widehat{\rho}_N$	Bias	StDev	Coverage	CI length
$\varepsilon = 0$	Robust	0.504	0.004	0.027	0.930	0.104
	MLE	0.500	0.000	0.026	0.943	0.102
$\varepsilon = 0.1$	Robust	0.466	−0.034	0.038	0.911	0.152
	MLE	0.097	−0.403	0.029	0.000	0.134
$\varepsilon = 0.2$	Robust	0.439	−0.061	0.051	0.951	0.220
	MLE	−0.172	−0.672	0.028	0.000	0.133

Table 1: Performance measures of the maximum likelihood estimator (MLE) and our robust estimator across 1,000 simulation repetitions with varying degrees of misspecification. The true polychoric correlation coefficient is $\rho_* = 0.5$. The performance measures are the average point estimate of the polychoric correlation coefficient, $\widehat{\rho}_N$, average bias ($\widehat{\rho}_N - \rho_*$), the standard deviation of the $\widehat{\rho}_N$ ("StDev"), the estimator's coverage with respect to the true ρ_* at significance level $\alpha = 0.05$, as well as the length of the estimator's confidence interval, again at level $\alpha = 0.05$.

Figure 3 visualizes the bias of each estimator with respect to the true $\boldsymbol{\theta}_*$ across the 1,000 simulated datasets. In the absence of misspecification, both MLE and the robust estimator yield accurate estimates of true $\boldsymbol{\theta}_*$ (in terms of mean squared error) and, in particular, the true polychoric correlation ρ_*. Both estimates are nearly equivalent to one another in the sense that their point estimates, standard deviation, and coverage at significance level $\alpha = 0.05$ are very similar (Table 1). However, when we introduce misspecification, the MLE and robust estimator yield noticeably different results. At misspecification degree $\varepsilon = 0.1$, the MLE is substantially biased with an average estimate of 0.097, corresponding to a bias of -0.403 and zero coverage, whereas the robust estimator maintains accuracy with an average estimate of 0.466, which corresponds to only a minor bias of -0.034 and a good coverage of 0.911 (Table 1). When the misspecification is increased to $\varepsilon = 0.2$, the contrast between the two estimators becomes even stronger. While the robust estimator is still remarkably close to the truth with a small bias of -0.061, the MLE produces estimates that are not only severely biased (bias of -0.672), but also sign-flipped: While the true correlation is strongly positive (0.5), the MLE's estimate is considerably negative (-0.172). It is worth noting that in the presence of misspecification, the confidence intervals of the robust estimator are wider than those of the MLE (see Table 1). This is expected because of the well-known trade-off between robustness and efficiency: An estimator that is designed to reduce bias, like a robust estimator, will inevitably have a larger estimation variance (e.g. Huber & Ronchetti, 2009). These wider confidence intervals furthermore explain why the robust estimator improves its coverage in Table 1 when the degree of misspecification is increased from 0.1 to 0.2.

This simulation study demonstrated that already a small degree of misspecificaion of the polychoric model can render the MLE unreliable, while our robust estimator retains good accuracy even in the presence of considerable misspecificaion. On the other hand, when the model is correctly specified, both estimators produce equivalent results.

5 Empirical application

5.1 Background and study design

We demonstrate our proposed method on empirical data by using a subset of the 100 unipolar markers of the Big-5 personality traits (Goldberg, 1992). Each marker is a an item comprising a single English adjective (such as "bold" or "timid") asking respondents to indicate how accurately the adjective describes their personality using a 5-point Likert-type rating scale (*very inaccurate, moderately inaccurate, neither accurate nor inaccurate, moderately accurate*, and *very accurate*). Here, each Big-5 personality trait is measured with six pairs of adjectives that are polar opposites to one another (such as "talkative" vs. "silent"), that is, twelve items in total for each trait. It seems implausible that an attentive respondent would choose to agree (or disagree) to *both* items in a pair of polar opposite adjectives. Consequently, one expects a strongly negative correlation between polar adjectives if all respondents respond attentively (Arias et al., 2020).

Arias et al. (2020) collect measurements of three Big-5 traits in this way, namely *extroversion, neuroticism*, and *conscientiousness*.[5] The sample that we shall use, Sample 1 in Arias et al. (2020), consists of $N = 725$ online respondents who are all U.S. citizens, native English speakers, and tend to have relatively high levels of reported education (about 90% report to hold an undergraduate or higher degree). Concerned about respondent inattention in their data, Arias et al. (2020) construct a factor mixture model for detecting inattentive participants. Their model crucially relies on response inconsistencies to polar opposite adjectives and is designed to primarily detect inattentive straightlining responding. They find that inattentive responding is a sizable problem in their data. Their model estimates that the proportion of inattentive participants amounts to 4.7% in the *conscientiousness*, 6% in the *neuroticism*, and 7.3% in the *extroversion* scale. After some further analyses, the authors conclude that if unaccounted for, inattentive responses can substantially deteriorate the fit of theoretical models, produce spurious variance, and overall jeopardize the validity of research results.

Due to the suspected presence of inattentive respodents, we apply our proposed method to estimate the polychoric correlation coefficients between all $\binom{12}{2} = 66$ unique item pairs in the *neuroticism* scale. The results of the remaining two scales are qualitatively similar and are reported in Appendix B. For each item pair, we estimate the polychoric correlation coefficient twice: via classic maximum likelihood and via our proposed robust alternative, with tuning parameter $c = 1.6$. The results remain qualitatively similar for different finite choices of c.

[5] Arias et al. (2020) synonymously refer to *neuroticism* as *emotional stability*. Furthermore, in addition to the three listed traits, Arias et al. (2020) collect measurements of the trait *dispositional optimism* by using a different instrument, and another scale that is designed to not measure any construct. We do not consider these scales in this empirical demonstration.

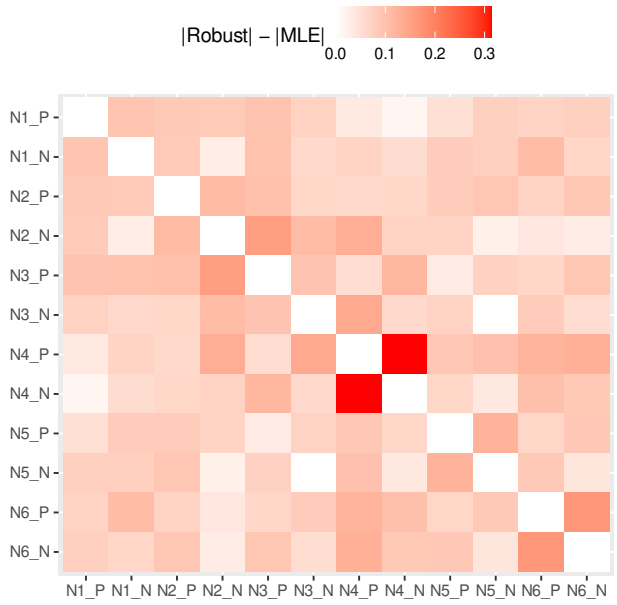


Figure 4: Difference between absolute estimates for the polychoric correlation coefficient of our robust estimator and the MLE for each item pair in the *neuroticism* scale, using the data of Arias et al. (2020). The items are "calm" (N1_P), "angry" (N1_N), "relaxed" (N2_P), "tense" (N2_N), "at ease" (N3_P), "nervous" (N3_N), "not envious" (N4_P), "envious" (N4_N), "stable" (N5_P), "unstable" (N5_N), "contented" (N6_P), and "discontented" (N6_N). For the item naming given in parentheses, items with identical identifier (the integer after the first "N") are polar opposites, where a last character "P" refers to the positive opposite and "N" to the negative opposite. The individual estimates of each method are provided in Table B.2 in Appendix B.

5.2 Results

Figure 4 visualizes the difference in absolute estimates for the polychoric correlation coefficient between all 66 unique item pairs in the *neuroticism* scale. For all unique pairs, our method estimates a stronger correlation coefficient than maximum likelihood. The differences in absolute estimates on average amount to 0.083, ranging from only marginally larger than zero to a substantive 0.314. For correlations between polar opposite adjectives, the average absolute difference between our robust method and MLE is 0.151. The fact that a robust method consistently yields stronger correlation estimates than the MLE, particularly between polar opposite adjectives, is indicative of the presence of negative leverage points, which drag negative correlational estimates towards zero, that is, they attenuate the estimated strength of correlation. Here, such negative leverage points could be the responses of inattentive participants who report agreement or disagreement to both items in item pairs that are designed to be negatively correlated. For instance, recall that it is implausible that an attentive respondent would choose to agree (or disagree) to *both* adjectives in the pair "envious" and "not envious" (cf. Arias et al., 2020). If sufficiently many such respondents ex-

	MLE		Robust	
Parameter	Estimate	SE	Estimate	SE
ρ	-0.618	0.025	-0.925	0.062
a_1	-1.370	0.061	-1.570	0.276
a_2	-0.476	0.043	-0.560	0.203
a_3	0.121	0.042	0.109	0.187
a_4	1.060	0.054	1.080	0.105
b_1	-0.857	0.049	-0.905	0.073
b_2	-0.004	0.041	-0.040	0.091
b_3	0.608	0.045	0.640	0.364
b_4	1.580	0.071	1.171	0.811

Table 2: Parameter estimates with standard errors (SEs) of the polychoric model for the *neuroticism* adjective pair "envious" and "not envious" in the data of Arias et al. (2020), using maximum likelihood (MLE) and our robust estimator with tuning constant $c = 1.6$. Each adjective item has five ordinal answer categories.

ist, then the presumably strongly negative correlation between these two opposite adjectives will be estimated to be weaker than it actually is.

To further investigate the presence of inattentive respondents who attenuate correlational estimates, we study in detail the adjective pair "not envious" and "envious", which featured the largest discrepancy between the maximum likelihood estimate and robust estimate in Figure 4, with an absolute difference of 0.314. The results of the two estimators are summarized in Table 2. The maximum likelihood estimate of -0.618 for the polychoric correlation coefficient seems remarkably weak considering that the two adjectives in question are polar opposites.[6] In contrast, its robust estimate estimate is given by -0.925, which seems much more in line with what one would expect if all participants responded accurately and attentively.

To study the potential presence of inattentive responses in each cell (x, y) for item pair "envious" and "not envious", Figure 5 visualizes the empirical relative frequencies, $\widehat{f}_N(x, y)$, through dot size, as well as the associated Pearson residual at the robust estimate, $\widehat{f}_N(x, y)/p_{xy}(\widehat{\boldsymbol{\theta}}_N)$, through dot color (the darker the blue shade, the larger). Importantly, the color of cells whose Pearson residual exceed 5.48 has been fixed to red.[7] This truncation value is equal to the value of the smallest Pearson residual that is substantially larger than the ideal value 1. We consider cells whose Pearson residual exceeds this truncation value to have a poor fit at the polychoric model. This applies to a total of 12 cells, some of which have enormous Pearson residuals. The Pearson residuals of the remaining 13

[6] For reference, the Pearson correlation coefficient between these two items is given by -0.562.

[7] The truncation of the color gradient in Figure 5 prevents that the color gradient is dominated by single cells with extreme Pearson residuals, which would blur the distinction between well fitted and poorly fitted cells.

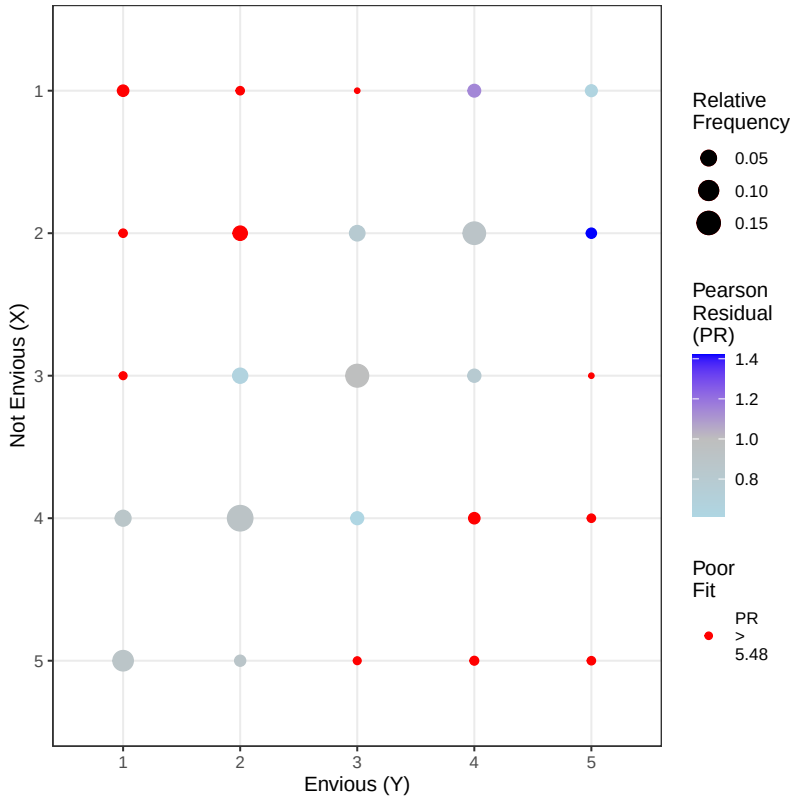


Figure 5: Dot plot of cells for the *neuroticism* item adjective pair "envious" and "not envious" in the data of Arias et al. (2020), where each item has five Likert-type response options, anchored by "very inaccurate" (= 1) and "very accurate" (= 5). Each dot's size is proportional to the relative empirical frequency of its associated cell, $\widehat{f}_N(x, y)$, whereas its color varies by the value of the cell's Pearson residual, $\widehat{f}_N(x, y)/p_{xy}(\widehat{\boldsymbol{\theta}}_N)$, at robust parameter estimate with tuning constant $c = 1.6$: The darker in blue a dot, the larger the value of the Pearson residual of its associated cell. The color of cells that could not be fitted well is fixed to red, where we deem a fit poor if the Pearson residual exceeds value 5.48 (which is the value of the smallest Pearson residual that is substantially larger than ideal value 1).

		"Envious"				
	$X \setminus Y$	1	2	3	4	5
	1	$\mathbf{< 0.0001}$	**0.0002**	0.8429	0.8274	0.8429
	2	$\mathbf{< 0.0001}$	0.0033	0.8429	0.9146	0.8274
"Not Envious"	3	0.8274	0.9896	0.8429	0.8429	0.8429
	4	0.8429	0.9079	0.8863	0.5251	**0.0003**
	5	0.8429	0.8429	0.5251	$\mathbf{< 0.0001}$	$\mathbf{< 0.0001}$

Table 3: p-values, adjusted for multiple comparisons by the procedure of Benjamini & Hochberg (1995), of the cellwise goodness-of-fit test in Theorem 3 for the *neuroticism* adjective pair "envious" and "not envious" in the data of Arias et al. (2020), where each item has five Likert-type response options, anchored by "1 = very inaccurate" and "5 = very accurate". The test statistics were computed with robust estimates using tuning constant $c = 1.6$. Cells in boldface are those for which the null hypothesis of unit Pearson residual is rejected at significance level $\alpha = 0.001$ in favor of the alternative of it being larger than one.

cells are reasonably close to ideal value 1, ranging from 0.65 to 1.42 with average 0.89. The Pearson residuals as well as relative empirical frequencies of all cells can be found in Table B.5 in the Appendix. It stands out that all poorly fitted cells are those whose responses might be viewed as inconsistent. Indeed, response cells $(x, y) = (1, 1), (1, 2), (2, 1), (1, 2)$ indicate that a participant reports that *neither* "envious" nor "not envious" characterizes them accurately, which are mutually contradicting responses, while for response cells $(x, y) = (4, 4), (4, 5), (5, 4), (5, 5)$ *both* adjectives characterize them accurately, which is again contradicting. As discussed previously, such responses are likely due to inattentiveness. The robust estimator suggests that such responses cannot be fitted well by the polychoric model and subsequently downweighs their influence in the estimation procedure by mapping their Pearson residual with the linear part of the $\varphi(\cdot)$ function in (7). Notably, also cells $(x, y) = (1, 3), (3, 1), (3, 5), (5, 3)$ are classified as poorly fitted. These responses report (dis)agreement to one opposite adjective, while being neutral about the other opposite. It is beyond the scope of this paper to assess whether such response patterns are indicative of inattentive responding, but the robust estimator suggests that such responses at least cannot be fitted well by the polychoric model with the data of Arias et al. (2020).

Next, we perform the goodness-of-fit test derived in Theorem 3 for each response cell to assess for which cells the polychoric model achieves a statistically significantly poor fit. Table 3 presents the p-values for the hypothesis test in (8), adjusted for multiple comparisons via the procedure of Benjamini & Hochberg (1995). Values for which the null hypothesis is rejected at significance level $\alpha = 0.001$ are in boldface. This choice of significance level is deliberately extremely conservative because the literature on inattentive responding recommends overwhelming evidence in favor of inattention before one should label responses as such (cf. Huang et al., 2012). At this significance level, we reject the null hypothesis of a good fit for six cells, namely $(x, y) = (1, 1), (2, 1), (1, 2), (5, 4), (4, 5), (5, 5)$. These six cells comprise 5.52% of the entire sample. As discussed in the previous section, it seems likely that

these responses are due to inattention because of inconsistent and contradictory responding. Either way, our test offers strong empirical evidence that these cells are outlying in the sense that they cannot be fitted well by the polychoric model and therefore lead to deteriorated model fit. This is consistent with Arias et al. (2020), who find that even a relatively small proportion of inconsistent responses can drastically reduce a model's fit. In their analyses, they estimate that 6% of all respondents in the *neuroticism* scale have been inattentive. Yet, albeit similar, we emphasize that our estimate of 5.52% can be, if at all, understood as a lower bound for the proportion of inattentive responding because of the extremely conservative significance level we chose for our analyses. For instance, the null hypothesis of good model fit was *not* rejected for the seemingly inconsistent response cell $(x, y) = (2, 2)$ (relative empirical frequency of about 4%) with a p-value of approximately 0.003, but would have been rejected at a slightly more liberal significance level. In addition, it is worth noting that the null hypothesis was also not rejected for one more seemingly inconsistent response cell, namely $(x, y) = (4, 4)$, despite a relatively large Pearson residual of 12.66. This non-rejection is likely due to low statistical power stemming from a small empirical frequency of this cell, since it only counted 14 responses (out of 725). Similar reasoning applies to the remaining four cells that were highlighted in red in Figure 4 but for which the null hypothesis of good fit was not rejected, namely those who indicate (dis)agreement to one adjective, while being neutral about its opposite. These four cells, $(x, y) = (1, 3), (3, 1), (3, 5), (5, 3)$, only count empirical frequencies of 2, 4, 2, and 4, respectively.

Overall, leveraging our robust estimator, we find strong evidence for the presence of inattentive respondents in the data of Arias et al. (2020). While they substantially affect the correlational estimate of the MLE, amounting to about -0.62, which is much weaker than one would expect for polar opposite items, our robust estimator can withstand their influence with an estimate of about -0.93 and also identify them by means of our proposed test.

6 Conclusion

Motivated by the susceptibility of maximum likelihood estimation to model misspecification, we propose a robust estimator of the polychoric model. Our estimator generalizes maximum likelihood estimation, does not make any assumption on the magnitude or type of potential misspecification, comes at no additional computational cost, and is consistent as well as asymptotically normally distributed. In addition, we propose a novel exact test that tests whether each individual cell in a contingency table can be fitted well by the polychoric model, allowing one to trace back potential sources of model misspecification. The methodology proposed in this paper is implemented in the free open source package `robord` (Welz, 2024a) in the statistical programming environment R, although it is primarily developed in C++ to maximize speed and computational performance.

We verify the enhanced robustness and theoretical properties of our estimator in simulation studies and demonstrate its practical usefulness in an empirical application on a Big-5

administration. We find compelling evidence for the presence of inattentive respondents as source of model misspecification. For instance, in a rating item pair with polar opposite content where a strong negative correlation is expected in the literature, our estimator yields a correlational estimate of -0.93, whereas maximum likelihood yields only -0.62; It follows that the robust estimate is more in line with the literature on this scale. Utilizing our test, we find that the maximum likelihood's lower-than-expected estimate is likely due to a few possibly inattentive participants who gave mutually contradictory responses, while our robust estimator can resist their influence.

Being substantially more robust to model misspecification than maximum likelihood, in particular to inattentive or careless responses, our estimator allows for robust estimation of correlation matrices of rating variables, which in turn allows for robust fitting of structural equation models. Since our estimator does not require assumptions on the type of misspecification, it could be especially useful in rating questionnaires without negatively worded items or attention checks. We leave a detailed investigation to future research.

References

Alfons, A. & Welz, M. (2024). Open science perspectives on machine learning for the identification of careless responding: A new hope or phantom menace? *Social and Personality Psychology Compass*. https://doi.org/10.1111/spc3.12941. In press.

Arias, V. B., Garrido, L., Jenaro, C., Martínez-Molina, A., & Arias, B. (2020). A little garbage in, lots of garbage out: Assessing the impact of careless responding in personality survey data. *Behavior Research Methods*, 52(6), 2489–2505. https://doi.org/https://doi.org/10.3758/s13428-020-01401-8

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289–300. https://doi.org/https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bowling, N. A., Huang, J. L., Bragg, C. B., Khazon, S., Liu, M., & Blackmore, C. E. (2016). Who cares and who is careless? Insufficient effort responding as a reflection of respondent personality. *Journal of Personality and Social Psychology*, 111(2), 218–229. https://psycnet.apa.org/doi/10.1037/pspp0000085

Bowling, N. A., Huang, J. L., Brower, C. K., & Bragg, C. B. (2023). The quick and the careless: The construct validity of page time as a measure of insufficient effort responding to surveys. *Organizational Research Methods*, 26(2), 323–352. https://doi.org/10.1177/10944281211056520

Coenders, G., Satorra, A., & Saris, W. E. (1997). Alternative approaches to structural modeling of ordinal data: A Monte Carlo study. *Structural Equation Modeling: A Multidisciplinary Journal*, 4(4), 261–282. https://doi.org/10.1080/10705519709540077

Credé, M. (2010). Random responding as a threat to the validity of effect size estimates in correlational research. *Educational and Psychological Measurement*, 70(4), 596–612. https://doi.org/10.1177/0013164410366686

Cressie, N. & Read, T. R. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society: Series B (Methodological)*, 46(3), 440–464.

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, 66, 4–19. https://doi.org/10.1016/j.jesp.2015.07.006

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. https://doi.org/10.1111/j.2044-8317.1985.tb00817.x

Drezner, Z. & Wesolowsky, G. O. (1990). On the computation of the bivariate normal integral. *Journal of Statistical Computation and Simulation*, 35(1-2), 101–107. https://doi.org/10.1080/00949659008811236

Eddelbuettel, D. (2013). *Seamless R and C++ Integration with Rcpp*. Springer. https://doi.org/10.1007/978-1-4614-6868-4. ISBN 978-1-4614-6867-7

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), 466–491. https://doi.org/10.1037/1082-989X.9.4.466

Foldnes, N. & Grønneberg, S. (2022). The sensitivity of structural equation modeling with ordinal data to underlying non-normality and observed distributional forms. *Psychological Methods*, 27(4), 541–567. https://doi.org/10.1037/met0000385

Garrido, L. E., Abad, F. J., & Ponsoda, V. (2013). A new look at Horn's parallel analysis with ordinal variables. *Psychological Methods*, 18(4), 454–474. https://doi.org/10.1037/a0030005

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment*, 4(1), 26–42. https://doi.org/https://doi.org/10.1037/1040-3590.4.1.26

Grønneberg, S. & Foldnes, N. (2019). A problem with discretizing Vale–Maurelli in simulation studies. *Psychometrika*, 84(2), 554–561. https://doi.org/10.1007/s11336-019-09663-8

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P., & Stahel, W. A. (1986). *Robust statistics: The approach based on influence functions*. Wiley series in probability and mathematical statistics. Wiley.

Holgado-Tello, F. P., Chacón-Moscoso, S., Barbero-García, I., & Vila-Abad, E. (2010). Polychoric versus pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. *Quality & Quantity*, 44, 153–166. https://doi.org/10.1007/s11135-008-9190-y

Huang, J. L., Bowling, N. A., Liu, M., & Li, Y. (2015a). Detecting insufficient effort responding with an infrequency scale: Evaluating validity and participant reactions. *Journal of Business and Psychology*, 30(2), 299–311. https://doi.org/10.1007/s10869-014-9357-6

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, 27(1), 99–114. https://doi.org/10.1007/s10869-011-9231-8

Huang, J. L., Liu, M., & Bowling, N. A. (2015b). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology*, 100(3), 828–845. https://doi.org/10.1037/a0038510

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35(1), 73–101. https://doi.org/10.1214/aoms/1177703732

Huber, P. J. & Ronchetti, E. M. (2009). *Robust Statistics*. Wiley. https://doi.org/10.1002/9780470434697

Kim, D. S., Reise, S. P., & Bentler, P. M. (2018). Identifying aberrant data in structural equation models with IRLS-ADF. *Structural Equation Modeling: A Multidisciplinary Journal*, 25(3), 343–358. https://doi.org/10.1080/10705511.2017.1379881

Lai, K. & Green, S. B. (2016). The problem with having two watches: Assessment of fit when rmsea and cfi disagree. *Multivariate Behavioral Research*, 51(2-3), 220–239. https://doi.org/10.1080/00273171.2015.1134306

Li, C.-H. (2016). The performance of ML, DWLS, and ULS estimation with robust corrections in structural equation models with ordinal variables. *Psychological Methods*, 21(3), 369–387. https://doi.org/10.1037/met0000093

Lindsay, B. G. (1994). Efficiency versus robustness: The case for minimum Hellinger distance and related methods. *Annals of Statistics*, 22(2), 1081–1114. https://doi.org/10.1214/aos/1176325512

Maniaci, M. R. & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, 48, 61–83. https://doi.org/10.1016/j.jrp.2013.09.008

Maronna, R. A., Martin, R. D., Yohai, V. J., & Salibián-Barrera, M. (2018). *Robust Statistics: Theory and Methods* (2nd ed.). John Wiley & Sons.

McGrath, R. E., Mitchell, M., Kim, B. H., & Hough, L. (2010). Evidence for response bias as a source of error variance in applied assessment. *Psychological Bulletin*, 136(3), 450–470. https://doi.org/10.1037/a0019216

Meade, A. W. & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, 17(3), 437–455. https://psycnet.apa.org/doi/10.1037/a0028085

Olsson, U. (1979a). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, 44(4), 443–460. https://doi.org/10.1007/BF02296207

Olsson, U. (1979b). On the robustness of factor analysis against crude classification of the observations. *Multivariate Behavioral Research*, 14(4), 485–500. https://doi.org/10.1207/s15327906mbr1404_7

Pearson, K. (1901). I. mathematical contributions to the theory of evolution, vii: On the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London, Series A*, 195(262-273), 1–47. https://doi.org/10.1098/rsta.1900.0022

Pearson, K. & Pearson, E. S. (1922). On polychoric coefficients of correlation. *Biometrika*, 14(1–2), 127–156. https://doi.org/10.1093/biomet/14.1-2.127

Raymaekers, J. & Rousseeuw, P. J. (2021). Fast robust correlation for high-dimensional data. *Technometrics*, 63(2), 184–198. https://doi.org/10.1080/00401706.2019.1677270

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), 354–373. https://doi.org/10.1037/a0029315

Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. *Mathematical Statistics and Applications*, 283–297. Reidel.

Ruckstuhl, A. F. & Welsh, A. H. (2001). Robust fitting of the binomial model. *Annals of Statistics*, 29(4), 1117–1136. https://doi.org/10.1214/aos/1013699996

Schmitt, N. & Stults, D. M. (1985). Factors defined by negatively keyed items: The result of careless respondents? *Applied Psychological Measurement*, 9, 367–373. https://doi.org/10.1177/014662168500900405

Schroeders, U., Schmidt, C., & Gnambs, T. (2022). Detecting careless responding in survey data using stochastic gradient boosting. *Educational and Psychological Measurement*, 82(1), 29–56. https://doi.org/10.1177/00131644211004708

Tallis, G. M. (1962). The maximum likelihood estimation of correlation from contingency tables. *Biometrics*, 18(3), 342–353. https://doi.org/10.2307/2527476

Van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Victoria-Feser, M.-P. & Ronchetti, E. (1997). Robust estimation for grouped data. *Journal of the American Statistical Association*, 92(437), 333–340. https://doi.org/10.1080/01621459.1997.10473631

Ward, M. & Meade, A. W. (2023). Dealing with careless responding in survey data: Prevention, identification, and recommended best practices. *Annual Review of Psychology*, 74(1), 577–596. https://doi.org/10.1146/annurev-psych-040422-045007

Welz, M. (2024a). *robcat: Robust Categorical Data Analysis*. https://github.com/mwelz/robcat. R package version 0.0.1

Welz, M. (2024b). Robust estimation and inference in categorical data. *Working paper*. https://mwelz.github.io/assets/pdf/mwelz_jmp.pdf.

Welz, M. & Alfons, A. (2023). *I don't care anymore: Identifying the onset of careless responding*. https://doi.org/10.48550/arXiv.2303.07167. arXiv:2303.07167

Welz, M., Archimbaud, A., & Alfons, A. (2023). Quantifying effects of rating-scale response bias: Theory and implications for survey design. Draft available upon request.

Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment*, 28(3), 186–191. https://psycnet.apa.org/doi/10.1007/s10862-005-9004-7

A Expressions of first and second order derivatives

A.1 First order terms

For cell $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, the gradient of $p_{xy}(\boldsymbol{\theta})$ can be expressed as

$$\frac{\partial p_{xy}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{\partial}{\partial \boldsymbol{\theta}} \Phi_2 \left(a_x, b_y; \rho\right) - \frac{\partial}{\partial \boldsymbol{\theta}} \Phi_2 \left(a_{x-1}, b_y; \rho\right) - \frac{\partial}{\partial \boldsymbol{\theta}} \Phi_2 \left(a_x, b_{y-1}; \rho\right) + \frac{\partial}{\partial \boldsymbol{\theta}} \Phi_2 \left(a_{x-1}, b_{y-1}; \rho\right), \tag{A.1}$$

see e.g. Olsson (1979a, equation 4). To characterize this gradient, we provide expressions for individual partial derivatives of $p_{xy}(\boldsymbol{\theta})$, that is,

$$\frac{\partial p_{xy}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \left(\frac{\partial p_{xy}(\boldsymbol{\theta})}{\partial \rho}, \frac{\partial p_{xy}(\boldsymbol{\theta})}{\partial a_1}, \ldots, \frac{\partial p_{xy}(\boldsymbol{\theta})}{\partial a_{K_x - 1}}, \frac{\partial p_{xy}(\boldsymbol{\theta})}{\partial b_1}, \ldots, \frac{\partial p_{xy}(\boldsymbol{\theta})}{\partial b_{K_y - 1}} \right)^\top.$$

First, for any $u, v \in \mathbb{R}$, it can be shown (e.g. Drezner & Wesolowsky, 1990) that

$$\frac{\partial}{\partial \rho} \Phi_2 \left(u, v; \rho\right) = \phi_2 \left(u, v; \rho\right),$$

as well as (e.g. Tallis, 1962)

$$\frac{\partial}{\partial u} \Phi_2 \left(u, v; \rho\right) = \phi_1 \left(u\right) \Phi_1 \left(\frac{v - \rho u}{\sqrt{1 - \rho^2}} \right),$$

where $\phi_1 \left(\cdot\right)$ and $\Phi_1 \left(\cdot\right)$ denote the density and distribution function, respectively, of the *univariate* standard normal distribution. The complementary partial derivative with respect to v follows analogously by symmetry.

It now follows immediately from (A.1) that the partial derivative of $p_{xy}(\boldsymbol{\theta})$ with respect to ρ is given by

$$\frac{\partial p_{xy}(\boldsymbol{\theta})}{\partial \rho} = \phi_2 \left(a_x, b_y; \rho\right) - \phi_2 \left(a_{x-1}, b_y; \rho\right) - \phi_2 \left(a_x, b_{y-1}; \rho\right) + \phi_2 \left(a_{x-1}, b_{y-1}; \rho\right),$$

whereas the partial derivatives with respect to the individual thresholds are characterized by

$$\frac{\partial p_{xy}(\boldsymbol{\theta})}{\partial a_k} = \begin{cases} \dfrac{\partial}{\partial a_x} \Phi_2 \left(a_x, b_y; \rho\right) - \dfrac{\partial}{\partial a_x} \Phi_2 \left(a_x, b_{y-1}; \rho\right) & \text{if } k = x, \\ -\dfrac{\partial}{\partial a_{x-1}} \Phi_2 \left(a_{x-1}, b_y; \rho\right) + \dfrac{\partial}{\partial a_{x-1}} \Phi_2 \left(a_{x-1}, b_{y-1}; \rho\right) & \text{if } k = x - 1, \\ 0 & \text{otherwise,} \end{cases}$$

for $k = 1, \ldots, K_x - 1$. An expression for $\frac{\partial p_{xy}(\boldsymbol{\theta})}{\partial b_k}$ can be derived analogously.

A.2 Second order terms

In this section, we provide expressions for the individual coordinates of the $d \times d$ symmetric Hessian matrix of $p_{xy}(\boldsymbol{\theta})$, that is,

$$\frac{\partial^2 p_{xy}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} = \begin{pmatrix} \frac{\partial^2 p_{xy}(\boldsymbol{\theta})}{\partial \rho^2} & \frac{\partial^2 p_{xy}(\boldsymbol{\theta})}{\partial \rho \partial a_1} & \cdots & \frac{\partial^2 p_{xy}(\boldsymbol{\theta})}{\partial \rho \partial a_{K_x-1}} & \frac{\partial^2 p_{xy}(\boldsymbol{\theta})}{\partial \rho \partial b_1} & \cdots & \frac{\partial^2 p_{xy}(\boldsymbol{\theta})}{\partial \rho \partial b_{K_y-1}} \\ \frac{\partial^2 p_{xy}(\boldsymbol{\theta})}{\partial a_1 \partial \rho} & \frac{\partial^2 p_{xy}(\boldsymbol{\theta})}{\partial {a_1}^2} & \cdots & \frac{\partial^2 p_{xy}(\boldsymbol{\theta})}{\partial a_1 \partial a_{K_x-1}} & \frac{\partial^2 p_{xy}(\boldsymbol{\theta})}{\partial a_1 \partial b_1} & \cdots & \frac{\partial^2 p_{xy}(\boldsymbol{\theta})}{\partial a_1 \partial b_{K_y-1}} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 p_{xy}(\boldsymbol{\theta})}{\partial a_{K_x-1} \partial \rho} & \frac{\partial^2 p_{xy}(\boldsymbol{\theta})}{\partial a_{K_x-1} \partial a_1} & \cdots & \frac{\partial^2 p_{xy}(\boldsymbol{\theta})}{\partial {a_{K_x-1}}^2} & \frac{\partial^2 p_{xy}(\boldsymbol{\theta})}{\partial a_{K_x-1} \partial b_1} & \cdots & \frac{\partial^2 p_{xy}(\boldsymbol{\theta})}{\partial a_{K_x-1} \partial b_{K_y-1}} \\ \frac{\partial^2 p_{xy}(\boldsymbol{\theta})}{\partial b_1 \partial \rho} & \frac{\partial^2 p_{xy}(\boldsymbol{\theta})}{\partial b_1 \partial a_1} & \cdots & \frac{\partial^2 p_{xy}(\boldsymbol{\theta})}{\partial b_1 \partial a_{K_x-1}} & \frac{\partial^2 p_{xy}(\boldsymbol{\theta})}{\partial {b_1}^2} & \cdots & \frac{\partial^2 p_{xy}(\boldsymbol{\theta})}{\partial b_1 \partial b_{K_y-1}} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 p_{xy}(\boldsymbol{\theta})}{\partial b_{K_y-1} \partial \rho} & \frac{\partial^2 p_{xy}(\boldsymbol{\theta})}{\partial b_{K_y-1} \partial a_1} & \cdots & \frac{\partial^2 p_{xy}(\boldsymbol{\theta})}{\partial b_{K_y-1} \partial a_{K_x-1}} & \frac{\partial^2 p_{xy}(\boldsymbol{\theta})}{\partial b_{K_y-1} \partial b_1} & \cdots & \frac{\partial^2 p_{xy}(\boldsymbol{\theta})}{\partial {b_{K_y-1}}^2} \end{pmatrix}.$$

This Hessian matrix can alternatively be expressed as follows, which follows by (A.1):

$$\frac{\partial^2 p_{xy}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} = \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \Phi_2\left(a_x, b_y; \rho\right) - \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \Phi_2\left(a_{x-1}, b_y; \rho\right) - \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \Phi_2\left(a_x, b_{y-1}; \rho\right) + \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \Phi_2\left(a_{x-1}, b_{y-1}; \rho\right). \tag{A.2}$$

First, by means of repeated applications of the product rule and chain rule it can be shown that for any $u, v \in \mathbb{R}$,

$$\frac{\partial^2}{\partial \rho^2} \Phi_2\left(u, v; \rho\right) = \frac{\partial}{\partial \rho} \phi_2\left(u, v; \rho\right) = \frac{\phi_2\left(u, v; \rho\right)}{(1-\rho^2)^2} \Big((1-\rho^2)(\rho + uv) - \rho\big(u^2 - 2\rho uv + v^2\big) \Big),$$

as well as

$$\frac{\partial^2}{\partial u^2} \Phi_2\left(u, v; \rho\right) = \phi_1'(u) \Phi_1\left(\frac{v - \rho u}{\sqrt{1-\rho^2}}\right) - \frac{\rho}{\sqrt{1-\rho^2}} \phi_1\left(u\right) \phi_1\left(\frac{v - \rho u}{\sqrt{1-\rho^2}}\right),$$

where

$$\phi_1'(u) = -\frac{u}{\sqrt{2\pi}} \exp\left(-u^2/2\right),$$

which follows immediately by the chain rule.

Next, for the second order cross-derivatives, it can be shown that

$$\frac{\partial^2}{\partial u \partial \rho} \Phi_2\left(u, v; \rho\right) = \phi_1\left(u\right) \phi_1\left(\frac{v - \rho u}{\sqrt{1-\rho^2}}\right) \frac{\rho v - u}{(1-\rho^2)^{3/2}}$$

and

$$\frac{\partial^2}{\partial u \partial v}\Phi_2\left(u, v; \rho\right) = \phi_1\left(u\right)\phi_1\left(\frac{v - \rho u}{\sqrt{1-\rho^2}}\right)\frac{1}{\sqrt{1-\rho^2}},$$

both by applications of the chain rule and product rule.

It now follows by (A.2) combined with these second order cross-derivatives that

$$\frac{\partial^2 p_{xy}(\boldsymbol{\theta})}{\partial a_k \partial \rho} = \begin{cases} \dfrac{\partial^2}{\partial a_x \partial \rho}\Phi_2\left(a_x, b_y; \rho\right) - \dfrac{\partial^2}{\partial a_x \partial \rho}\Phi_2\left(a_x, b_{y-1}; \rho\right) & \text{if } k = x, \\ -\dfrac{\partial^2}{\partial a_{x-1} \partial \rho}\Phi_2\left(a_{x-1}, b_y; \rho\right) + \dfrac{\partial^2}{\partial a_{x-1} \partial \rho}\Phi_2\left(a_{x-1}, b_{y-1}; \rho\right) & \text{if } k = x-1, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\frac{\partial^2 p_{xy}(\boldsymbol{\theta})}{\partial a_k \partial b_l} = \begin{cases} \dfrac{\partial^2}{\partial a_k \partial b_l}\Phi_2\left(a_x, b_y; \rho\right) & \text{if } (k,l) \in \big\{(x,y),(x-1,y-1)\big\}, \\ -\dfrac{\partial^2}{\partial a_k \partial b_l}\Phi_2\left(a_x, b_y; \rho\right) & \text{if } (k,l) \in \big\{(x-1,y),(x,y-1)\big\}, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\frac{\partial^2 p_{xy}(\boldsymbol{\theta})}{\partial a_k \partial a_l} = \begin{cases} \dfrac{\partial^2 p_{xy}(\boldsymbol{\theta})}{\partial {a_k}^2} & \text{if } k = l, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\frac{\partial^2 p_{xy}(\boldsymbol{\theta})}{\partial {a_k}^2} = \begin{cases} \dfrac{\partial^2}{\partial {a_x}^2}\Phi_2\left(a_x, b_y; \rho\right) - \dfrac{\partial^2}{\partial {a_x}^2}\Phi_2\left(a_x, b_{y-1}; \rho\right) & \text{if } k = x, \\ -\dfrac{\partial^2}{\partial {a_{x-1}}^2}\Phi_2\left(a_{x-1}, b_y; \rho\right) + \dfrac{\partial^2}{\partial {a_{x-1}}^2}\Phi_2\left(a_{x-1}, b_{y-1}; \rho\right) & \text{if } k = x-1, \\ 0 & \text{otherwise,} \end{cases}$$

and, finally,

$$\frac{\partial^2 p_{xy}(\boldsymbol{\theta})}{\partial \rho^2} = \\ \frac{\partial^2}{\partial \rho^2}\Phi_2\left(a_x, b_y; \rho\right) - \frac{\partial^2}{\partial \rho^2}\Phi_2\left(a_{x-1}, b_y; \rho\right) - \frac{\partial^2}{\partial \rho^2}\Phi_2\left(a_x, b_{y-1}; \rho\right) + \frac{\partial^2}{\partial \rho^2}\Phi_2\left(a_{x-1}, b_{y-1}; \rho\right).$$

		Adjective marker pairs	
Construct	Number	Positive (P)	Negative (N)
Extroversion (E)	1	extraverted	introverted
	2	energetic	unenergetic
	3	talkative	silent
	4	bold	timid
	5	assertive	unassertive
	6	adventurous	unadventurous
Conscientiousness (C)	1	organized	disorganized
	2	responsible	irresponsible
	3	conscientious	negligent
	4	practical	impractical
	5	thorough	careless
	6	hardworking	lazy
Neuroticism (N)	1	calm	angry
	2	relaxed	tense
	3	at ease	nervous
	4	not envious	envious
	5	stable	unstable
	6	contented	discontented

Table B.1: Unipolar markers of three Big-5 personality traits (Goldberg, 1992). Each trait is measured by six pairs of items, where each item is a single English adjective. Each item pair consists of a positive and negative item. We explain item identifiers by means of the following example. Item "C3_P" refers to the positive (P) item in the 3rd pair of the conscientiousness (C) scale, that is, adjective "conscientious", whereas "N1_N" would refer to "angry".

B Empirical application: Additional results

This appendix contains additional results of the empirical application from Section 5. Table B.1 lists the unipolar markers of the three Big-5 scales used by Arias et al. (2020), namely *extroversion, conscientiousness,* and *neuroticism*. Tables B.2–B.4 contain the (polychoric) correlation matrices of the items in each scale, estimated by maximum likelihood and our robust estimator, while Figures B.1–B.3 visualize the absolute difference between the two estimators for each pairwise correlation. Table B.5 contains the cellwise Pearson residuals, empirical frequencies, as well as estimated model probabilities for the "envious"–"not envious" item pair in the *neuroticism* scale.

	N1_P	N1_N	N2_P	N2_N	N3_P	N3_N	N4_P	N4_N	N5_P	N5_N	N6_P	N6_N
N1_P	1.00	-0.37	0.71	-0.50	0.69	-0.49	0.27	-0.24	0.58	-0.47	0.42	-0.32
N1_N	-0.37	1.00	-0.40	0.55	-0.39	0.47	-0.19	0.40	-0.39	0.60	-0.32	0.56
N2_P	0.71	-0.40	1.00	-0.55	0.75	-0.54	0.26	-0.26	0.55	-0.41	0.53	-0.47
N2_N	-0.50	0.55	-0.55	1.00	-0.54	0.65	-0.24	0.42	-0.41	0.57	-0.31	0.52
N3_P	0.69	-0.39	0.75	-0.54	1.00	-0.53	0.29	-0.28	0.63	-0.44	0.52	-0.48
N3_N	-0.49	0.47	-0.54	0.65	-0.53	1.00	-0.28	0.43	-0.44	0.58	-0.29	0.47
N4_P	0.27	-0.19	0.26	-0.24	0.29	-0.28	1.00	-0.61	0.26	-0.20	0.18	-0.20
N4_N	-0.24	0.40	-0.26	0.42	-0.28	0.43	-0.61	1.00	-0.33	0.46	-0.22	0.44
N5_P	0.58	-0.39	0.55	-0.41	0.63	-0.44	0.26	-0.33	1.00	-0.69	0.53	-0.46
N5_N	-0.47	0.60	-0.41	0.57	-0.44	0.58	-0.20	0.46	-0.69	1.00	-0.35	0.57
N6_P	0.42	-0.32	0.53	-0.31	0.52	-0.29	0.18	-0.22	0.53	-0.35	1.00	-0.58
N6_N	-0.32	0.56	-0.47	0.52	-0.48	0.47	-0.20	0.44	-0.46	0.57	-0.58	1.00

(a) Maximum likelihood estimates

	N1_P	N1_N	N2_P	N2_N	N3_P	N3_N	N4_P	N4_N	N5_P	N5_N	N6_P	N6_N
N1_P	1.00	-0.47	0.80	-0.58	0.79	-0.56	0.30	-0.26	0.63	-0.54	0.49	-0.39
N1_N	-0.47	1.00	-0.48	0.58	-0.49	0.54	-0.26	0.45	-0.47	0.68	-0.43	0.63
N2_P	0.80	-0.48	1.00	-0.66	0.85	-0.60	0.32	-0.32	0.64	-0.50	0.60	-0.56
N2_N	-0.58	0.58	-0.66	1.00	-0.70	0.76	-0.37	0.49	-0.48	0.60	-0.35	0.55
N3_P	0.79	-0.49	0.85	-0.70	1.00	-0.62	0.35	-0.39	0.66	-0.52	0.59	-0.57
N3_N	-0.56	0.54	-0.60	0.76	-0.62	1.00	-0.42	0.49	-0.52	0.58	-0.37	0.53
N4_P	0.30	-0.26	0.32	-0.37	0.35	-0.42	1.00	-0.92	0.35	-0.30	0.30	-0.33
N4_N	-0.26	0.45	-0.32	0.49	-0.39	0.49	-0.92	1.00	-0.39	0.50	-0.33	0.53
N5_P	0.63	-0.47	0.64	-0.48	0.66	-0.52	0.35	-0.39	1.00	-0.82	0.59	-0.55
N5_N	-0.54	0.68	-0.50	0.60	-0.52	0.58	-0.30	0.50	-0.82	1.00	-0.44	0.61
N6_P	0.49	-0.43	0.60	-0.35	0.59	-0.37	0.30	-0.33	0.59	-0.44	1.00	-0.75
N6_N	-0.39	0.63	-0.56	0.55	-0.57	0.53	-0.33	0.53	-0.55	0.61	-0.75	1.00

(b) Robust estimates

Table B.2: Estimated correlation matrices of the items in the *neuroticism* scale from the data in Arias et al. (2020, Sample 1; $N = 725$) using MLE (top panel) and our robust estimator with tuning constant $c = 1.6$ (bottom panel). The items are "calm" (N1_P), "angry" (N1_N), "relaxed" (N2_P), "tense" (N2_N), "at ease" (N3_P), "nervous" (N3_N), "not envious" (N4_P), "envious" (N4_N), "stable" (N5_P), "unstable" (N5_N), "contented" (N6_P), and "discontented" (N6_N). For the item naming given in parentheses, items with identical identifier (the integer after the first "N") are polar opposites, where a last character "P" refers to the positive opposite and "N" to the negative opposite.

	E1_P	E1_N	E2_P	E2_N	E3_P	E3_N	E4_P	E4_N	E5_P	E5_N	E6_P	E6_N
E1_P	1.00	-0.77	0.50	-0.26	0.70	-0.50	0.56	-0.42	0.51	-0.40	0.51	-0.32
E1_N	-0.77	1.00	-0.38	0.34	-0.59	0.61	-0.45	0.54	-0.47	0.50	-0.35	0.37
E2_P	0.50	-0.38	1.00	-0.65	0.43	-0.27	0.49	-0.28	0.47	-0.38	0.54	-0.39
E2_N	-0.26	0.34	-0.65	1.00	-0.24	0.34	-0.30	0.40	-0.32	0.48	-0.38	0.49
E3_P	0.70	-0.59	0.43	-0.24	1.00	-0.59	0.44	-0.36	0.46	-0.40	0.41	-0.25
E3_N	-0.50	0.61	-0.27	0.34	-0.59	1.00	-0.27	0.56	-0.35	0.45	-0.24	0.37
E4_P	0.56	-0.45	0.49	-0.30	0.44	-0.27	1.00	-0.41	0.64	-0.49	0.54	-0.34
E4_N	-0.42	0.54	-0.28	0.40	-0.36	0.56	-0.41	1.00	-0.49	0.60	-0.27	0.40
E5_P	0.51	-0.47	0.47	-0.32	0.46	-0.35	0.64	-0.49	1.00	-0.71	0.39	-0.23
E5_N	-0.40	0.50	-0.38	0.48	-0.40	0.45	-0.49	0.60	-0.71	1.00	-0.34	0.45
E6_P	0.51	-0.35	0.54	-0.38	0.41	-0.24	0.54	-0.27	0.39	-0.34	1.00	-0.68
E6_N	-0.32	0.37	-0.39	0.49	-0.25	0.37	-0.34	0.40	-0.23	0.45	-0.68	1.00

(a) Maximum likelihood estimates

	E1_P	E1_N	E2_P	E2_N	E3_P	E3_N	E4_P	E4_N	E5_P	E5_N	E6_P	E6_N
E1_P	1.00	-0.87	0.55	-0.34	0.75	-0.62	0.58	-0.58	0.54	-0.45	0.55	-0.39
E1_N	-0.87	1.00	-0.40	0.36	-0.67	0.63	-0.52	0.62	-0.52	0.51	-0.36	0.37
E2_P	0.55	-0.40	1.00	-0.84	0.50	-0.32	0.56	-0.38	0.55	-0.43	0.57	-0.44
E2_N	-0.34	0.36	-0.84	1.00	-0.30	0.35	-0.40	0.43	-0.41	0.54	-0.45	0.53
E3_P	0.75	-0.67	0.50	-0.30	1.00	-0.71	0.50	-0.51	0.52	-0.50	0.42	-0.28
E3_N	-0.62	0.63	-0.32	0.35	-0.71	1.00	-0.38	0.62	-0.47	0.47	-0.30	0.37
E4_P	0.58	-0.52	0.56	-0.40	0.50	-0.38	1.00	-0.55	0.73	-0.64	0.61	-0.48
E4_N	-0.58	0.62	-0.38	0.43	-0.51	0.62	-0.55	1.00	-0.61	0.66	-0.33	0.44
E5_P	0.54	-0.52	0.55	-0.41	0.52	-0.47	0.73	-0.61	1.00	-0.85	0.44	-0.29
E5_N	-0.45	0.51	-0.43	0.54	-0.50	0.47	-0.64	0.66	-0.85	1.00	-0.41	0.47
E6_P	0.55	-0.36	0.57	-0.45	0.42	-0.30	0.61	-0.33	0.44	-0.41	1.00	-0.83
E6_N	-0.39	0.37	-0.44	0.53	-0.28	0.37	-0.48	0.44	-0.29	0.47	-0.83	1.00

(b) Robust estimates

Table B.3: Estimated correlation matrices of the items in the *extroversion* scale from the data in Arias et al. (2020, Sample 1; $N = 725$) using MLE (top panel) and our robust estimator with tuning constant $c = 1.6$ (bottom panel). The items are "extraverted" (E1_P), "introverted" (E1_N), "energetic" (E2_P), "unenergetic" (E2_N), "talkative" (E3_P), "silent" (E3_N), "bold" (E4_P), "timid" (E4_N), "assertive" (E5_P), "unassertive" (E5_N), "adventurous" (E6_P), and "unadventurous" (E6_N). For the item naming given in parentheses, items with identical identifier (the integer after the first "N") are polar opposites, where a last character "P" refers to the positive opposite and "N" to the negative opposite.

	C1_P	C1_N	C2_P	C2_N	C3_P	C3_N	C4_P	C4_N	C5_P	C5_N	C6_P	C6_N
C1_P	1.00	-0.76	0.56	-0.42	0.34	-0.35	0.37	-0.26	0.51	-0.40	0.43	-0.43
C1_N	-0.76	1.00	-0.51	0.58	-0.24	0.55	-0.32	0.44	-0.43	0.61	-0.44	0.55
C2_P	0.56	-0.51	1.00	-0.69	0.42	-0.55	0.57	-0.42	0.51	-0.54	0.65	-0.55
C2_N	-0.42	0.58	-0.69	1.00	-0.39	0.75	-0.48	0.67	-0.42	0.70	-0.53	0.63
C3_P	0.34	-0.24	0.42	-0.39	1.00	-0.32	0.39	-0.34	0.44	-0.33	0.38	-0.25
C3_N	-0.35	0.55	-0.55	0.75	-0.32	1.00	-0.37	0.59	-0.38	0.71	-0.44	0.53
C4_P	0.37	-0.32	0.57	-0.48	0.39	-0.37	1.00	-0.51	0.36	-0.38	0.39	-0.31
C4_N	-0.26	0.44	-0.42	0.67	-0.34	0.59	-0.51	1.00	-0.38	0.59	-0.31	0.43
C5_P	0.51	-0.43	0.51	-0.42	0.44	-0.38	0.36	-0.38	1.00	-0.43	0.54	-0.39
C5_N	-0.40	0.61	-0.54	0.70	-0.33	0.71	-0.38	0.59	-0.43	1.00	-0.43	0.53
C6_P	0.43	-0.44	0.65	-0.53	0.38	-0.44	0.39	-0.31	0.54	-0.43	1.00	-0.61
C6_N	-0.43	0.55	-0.55	0.63	-0.25	0.53	-0.31	0.43	-0.39	0.53	-0.61	1.00

(a) Maximum likelihood estimates

	C1_P	C1_N	C2_P	C2_N	C3_P	C3_N	C4_P	C4_N	C5_P	C5_N	C6_P	C6_N
C1_P	1.00	-0.89	0.57	-0.56	0.36	-0.46	0.43	-0.35	0.54	-0.56	0.49	-0.52
C1_N	-0.89	1.00	-0.58	0.64	-0.31	0.60	-0.38	0.47	-0.48	0.69	-0.52	0.61
C2_P	0.57	-0.58	1.00	-0.86	0.45	-0.68	0.62	-0.54	0.55	-0.65	0.69	-0.64
C2_N	-0.56	0.64	-0.86	1.00	-0.44	0.80	-0.57	0.74	-0.50	0.76	-0.61	0.66
C3_P	0.36	-0.31	0.45	-0.44	1.00	-0.43	0.42	-0.46	0.17	-0.41	0.40	-0.26
C3_N	-0.46	0.60	-0.68	0.80	-0.43	1.00	-0.48	0.70	-0.52	0.78	-0.55	0.59
C4_P	0.43	-0.38	0.62	-0.57	0.42	-0.48	1.00	-0.68	0.39	-0.47	0.44	-0.33
C4_N	-0.35	0.47	-0.54	0.74	-0.46	0.70	-0.68	1.00	-0.47	0.66	-0.42	0.45
C5_P	0.54	-0.48	0.55	-0.50	0.17	-0.52	0.39	-0.47	1.00	-0.54	0.60	-0.45
C5_N	-0.56	0.69	-0.65	0.76	-0.41	0.78	-0.47	0.66	-0.54	1.00	-0.59	0.61
C6_P	0.49	-0.52	0.69	-0.61	0.40	-0.55	0.44	-0.42	0.60	-0.59	1.00	-0.69
C6_N	-0.52	0.61	-0.64	0.66	-0.26	0.59	-0.33	0.45	-0.45	0.61	-0.69	1.00

(b) Robust estimates

Table B.4: Estimated correlation matrices of the items in the *conscientiousness* scale from the data in Arias et al. (2020, Sample 1; $N = 725$) using MLE (top panel) and our robust estimator with tuning constant $c = 1.6$ (bottom panel). The items are "calm" (C1_P), "angry" (C1_N), "relaxed" (C2_P), "tense" (C2_N), "at ease" (C3_P), "nervous" (C3_N), "not envious" (C4_P), "envious" (C4_N), "stable" (C5_P), "unstable" (C5_N), "contented" (C6_P), and "discontented" (C6_N). For the item naming given in parentheses, items with identical identifier (the integer after the first "N") are polar opposites, where a last character "P" refers to the positive opposite and "N" to the negative opposite.

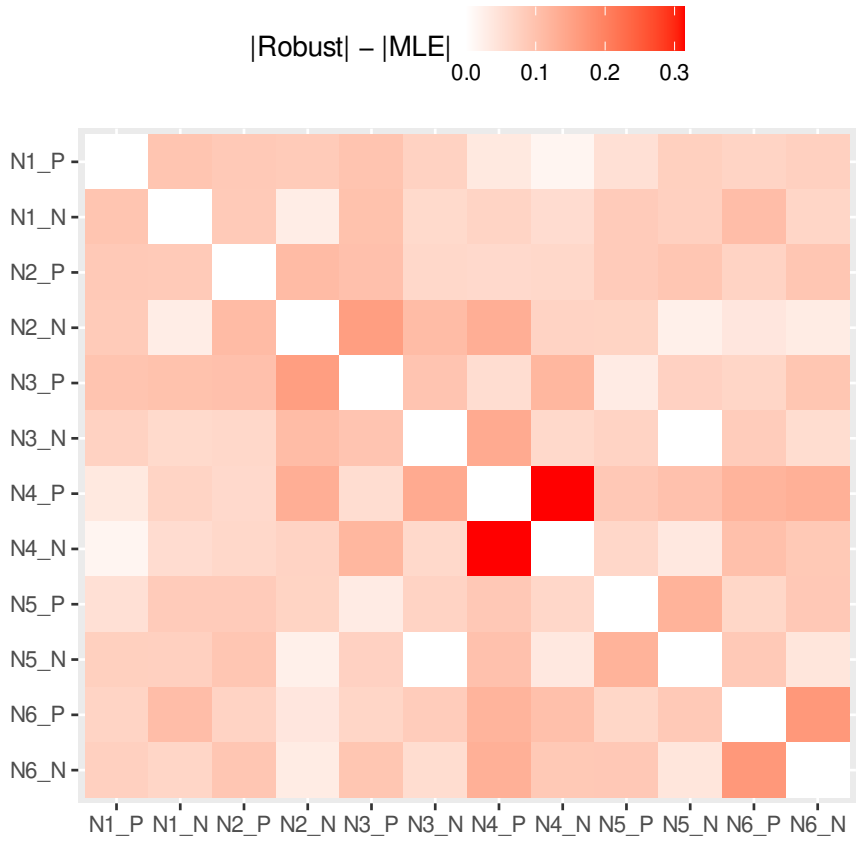


Figure B.1: Difference between absolute estimates for the polychoric correlation coefficient of our robust estimator and the MLE for each item pair in the *neuroticism* scale, using the data of Arias et al. (2020). The items are "calm" (N1_P), "angry" (N1_N), "relaxed" (N2_P), "tense" (N2_N), "at ease" (N3_P), "nervous" (N3_N), "not envious" (N4_P), "envious" (N4_N), "stable" (N5_P), "unstable" (N5_N), "contented" (N6_P), and "discontented" (N6_N). For the item naming given in parentheses, items with identical identifier (the integer after the first "N") are polar opposites, where a last character "P" refers to the positive opposite and "N" to the negative opposite.

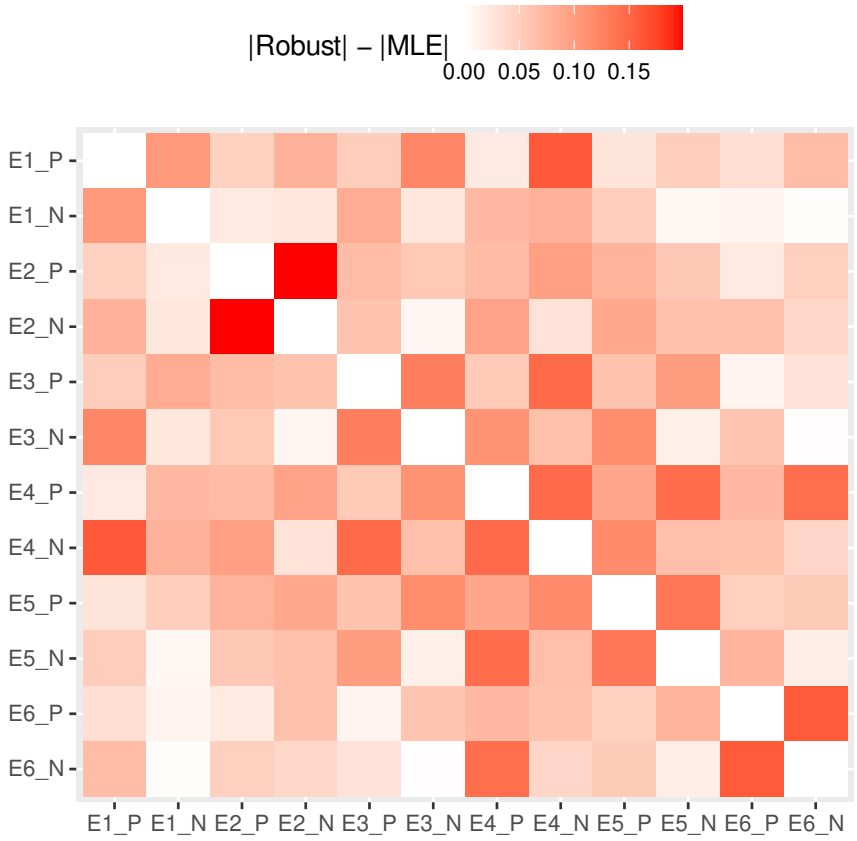


Figure B.2: Difference between absolute estimates for the polychoric correlation coefficient of our robust estimator and the MLE for each item pair in the *extroversion* scale, using the data of Arias et al. (2020). The items are "extraverted" (E1_P), "introverted" (E1_N), "energetic" (E2_P), "unenergetic" (E2_N), "talkative" (E3_P), "silent" (E3_N), "bold" (E4_P), "timid" (E4_N), "assertive" (E5_P), "unassertive" (E5_N), "adventurous" (E6_P), and "unadventurous" (E6_N). For the item naming given in parentheses, items with identical identifier (the integer after the first "N") are polar opposites, where a last character "P" refers to the positive opposite and "N" to the negative opposite.

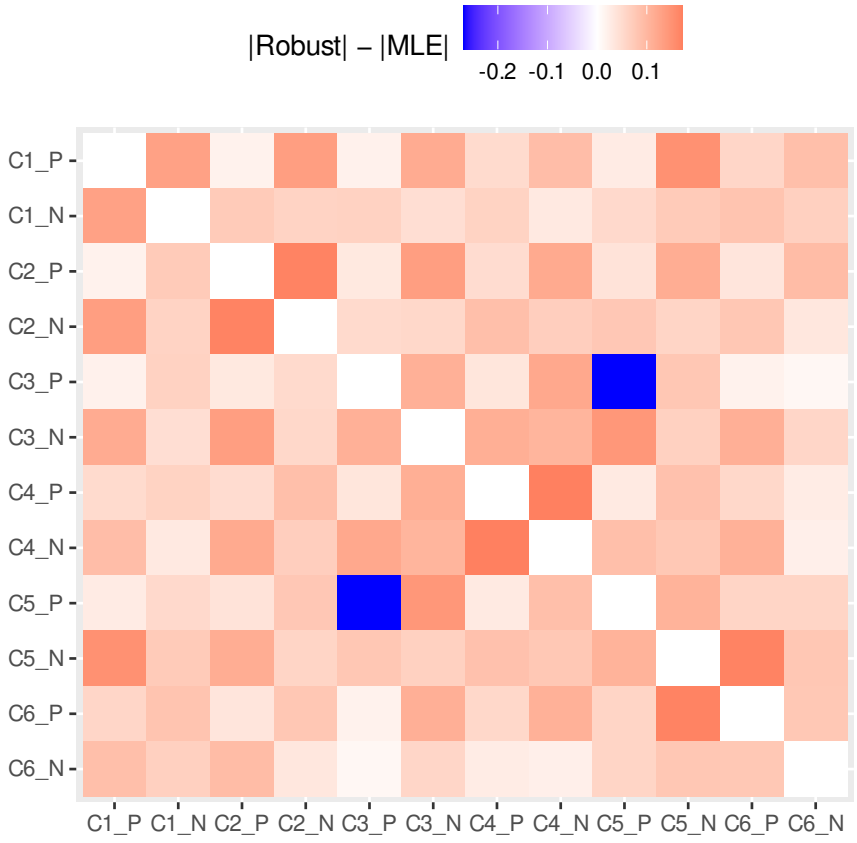


Figure B.3: Difference between absolute estimates for the polychoric correlation coefficient of our robust estimator and the MLE for each item pair in the *conscientiousness* scale, using the data of Arias et al. (2020). The items are "calm" (C1_P), "angry" (C1_N), "relaxed" (C2_P), "tense" (C2_N), "at ease" (C3_P), "nervous" (C3_N), "not envious" (C4_P), "envious" (C4_N), "stable" (C5_P), "unstable" (C5_N), "contented" (C6_P), and "discontented" (C6_N). For the item naming given in parentheses, items with identical identifier (the integer after the first "N") are polar opposites, where a last character "P" refers to the positive opposite and "N" to the negative opposite.

$X \backslash Y$	1	2	3	4	5
1	9,814,457,557.73	16,011.33	11.82	1.14	0.65
2	2,424.07	10.07	0.80	0.90	1.42
3	15.48	0.65	0.99	0.80	77.14
4	0.88	0.92	0.61	12.66	222,528.08
5	0.89	0.88	36.01	55,420.33	995,017,243,197.60

(a) Pearson residuals $\widehat{f}_N(x,y)/p_{xy}(\widehat{\boldsymbol{\theta}}_N)$

$X \backslash Y$	1	2	3	4	5
1	0.019	0.007	0.003	0.028	0.022
2	0.007	0.040	0.050	0.138	0.014
3	0.006	0.047	0.143	0.030	0.003
4	0.054	0.189	0.029	0.019	0.007
5	0.108	0.018	0.006	0.008	0.007

(b) Empirical relative frequencies $\widehat{f}_N(x,y)$

$X \backslash Y$	1	2	3	4	5
1	< 0.001	< 0.001	< 0.001	0.024	0.034
2	< 0.001	0.004	0.062	0.153	0.010
3	0.001	0.072	0.145	0.038	< 0.001
4	0.061	0.205	0.047	0.002	< 0.001
5	0.120	0.020	< 0.001	< 0.001	< 0.001

(c) Estimated cell probabilities $p_{xy}(\widehat{\boldsymbol{\theta}}_N)$

Table B.5: Pearson residual (top), empirical relative frequency (center), and estimated cell probability (bottom) of each cell for the "not envious" (X)–"envious" (Y) item pair in the measurements of Arias et al. (2020) of the *neuroticism* scale. Estimate $\widehat{\boldsymbol{\theta}}_N$ was computed with tuning constant $c = 1.6$.