Psychometrika (2025), **xx**:xx 1–35

CAMBRIDGE
UNIVERSITY PRESS

Theory and Methods

Robust Estimation of Polychoric Correlation

Max Welz[1,3], Patrick Mair[2] and Andreas Alfons[1]

[1]Department of Econometrics, Erasmus University Rotterdam, Rotterdam, 3062 PA, South Holland, The Netherlands.
E-mail: alfons@ese.eur.nl.
[2]Department of Psychology, Harvard University, Cambridge, MA 02138, Massachusetts, USA. E-mail: mair@fas.harvard.edu.
[3]Department of Psychology, University of Zurich, Zurich, CH-8050, Zurich, Switzerland. E-mail: max.welz@uzh.ch.

Keywords: Polychoric correlation, model misspecification, robust estimation, careless responding

Abstract
Polychoric correlation is often an important building block in the analysis of rating data, particularly for structural equation models. However, the commonly employed maximum likelihood (ML) estimator is highly susceptible to misspecification of the polychoric correlation model, for instance through violations of latent normality assumptions. We propose a novel estimator that is designed to be robust against partial misspecification of the polychoric model, that is, when the model is misspecified for an unknown fraction of observations, such as careless respondents. To this end, the estimator minimizes a robust loss function based on the divergence between observed frequencies and theoretical frequencies implied by the polychoric model. In contrast to existing literature, our estimator makes no assumption on the type or degree of model misspecification. It furthermore generalizes ML estimation, is consistent as well as asymptotically normally distributed, and comes at no additional computational cost. We demonstrate the robustness and practical usefulness of our estimator in simulation studies and an empirical application on a Big Five administration. In the latter, the polychoric correlation estimates of our estimator and ML differ substantially, which, after further inspection, is likely due to the presence of careless respondents that the estimator helps identify.

1. Introduction

Ordinal data are ubiquitous in psychology and related fields. With such data, e.g., arising from responses to rating scales, it is often recommended to estimate correlation matrices through polychoric correlation coefficients (e.g., Foldnes & Grønneberg, 2022; Garrido et al., 2013; Holgado–Tello et al., 2010). The resulting polychoric correlation matrix is an important building block in subsequent multivariate models like factor analysis models and structural equation models (SEMs), as well as in exploratory methods like principal component analysis, multidimensional scaling, and clustering techniques (see, e.g., Mair, 2018, for an overview). An individual polychoric correlation coefficient is the population correlation between two underlying latent variables that are postulated to have generated the observed categorical data through an unobserved discretization process. Traditionally, it is assumed that the two latent variables are standard bivariate normally distributed (Pearson, 1901) to estimate the polychoric correlation coefficient from observed ordinal data. Estimation of this latent normality model, called the *polychoric model*, is commonly carried out via maximum likelihood (Olsson, 1979). However, recent work has demonstrated that maximum likelihood (ML) estimation of polychoric correlation is highly sensitive to violations of the assumption of underlying normality. Violations of this assumption result in a misspecified polychoric model, which can lead to substantially biased estimates of its parameters and

https://doi.org/10.1017/psy.2025.10066 Published online by Cambridge University Press

those of subsequent multivariate models (Foldnes & Grønneberg, 2019, 2020; Grønneberg & Foldnes, 2022; Jin & Yang-Wallentin, 2017), particularly SEMs using *diagonally weighted least squares* (Foldnes & Grønneberg, 2022), where the latter is based on weights derived under latent normality.

Motivated by the recent interest in non-robustness of ML, we study estimation of the polychoric model under a misspecification framework stemming from the robust statistics literature (e.g., Huber & Ronchetti, 2009). In this setup, which we call *partial* misspecification here, the polychoric model is potentially misspecified for an unknown (and possibly zero-valued) fraction of observations. Heuristically, the model is misspecified such that the affected subset of observations contains little to no relevant information for the parameter of interest, the polychoric correlation coefficient. Examples of such uninformative observations include careless responses, misresponses, or responses due to item misunderstanding. Especially careless responding has been identified as a major threat to the validity of questionnaire-based research findings (e.g., Credé, 2010; Huang, Liu, & Bowling, 2015; Meade & Craig, 2012; Welz et al., 2024; Woods, 2006). We demonstrate that already a small fraction of uninformative observations (such as careless respondents) can result in considerably biased ML estimates.

As a remedy and our main contribution, we propose a novel way to estimate the polychoric model which is robust to partial model misspecification. Essentially, the estimator poses the question "What is the best fit that can be achieved with the polychoric model for (the majority of) the data at hand?" The estimator compares the observed frequency of each contingency table cell with its expected frequency under the polychoric model, and automatically downweights cells whose observed frequencies cannot be fitted sufficiently well. As such, our estimator generalizes the ML estimator, but, in contrast to ML, does not crucially rely on correct specification of the model. Specifically, our estimator allows the model to be misspecified for an unknown fraction of uninformative responses in a sample, but makes *no assumption* on the type, magnitude, or location of potential misspecification. The estimator is designed to identify such responses and to simultaneously reduce their influence so that the polychoric model can be accurately estimated from the remaining responses generated by latent normality. Conversely, if the polychoric model is correctly specified, that is, latent normality holds true for all observations, our estimator and ML estimation are asymptotically equivalent. As such, our proposed estimator can be thought of as a generalized ML estimator that is robust to potential partial model misspecification, due to, for instance (but not limited to) careless responding. We show that our robust estimator is consistent, asymptotically normal, and fully efficient under the polychoric model, while possessing similar asymptotic properties under misspecification, and it comes at no additional computational cost compared to ML.

The partial misspecification framework in this paper is fundamentally different to that considered in recent literature on misspecified polychoric models. In this literature (e.g., Foldnes & Grønneberg, 2019, 2020, 2022; Grønneberg & Foldnes, 2022; Jin & Yang-Wallentin, 2017; Lyhagen & Ornstein, 2023), the polychoric model is misspecified in the sense that *all* (unobserved) realizations of the latent continuous variables come from a distribution that is nonnormal. Under this framework, which is also known as *distributional misspecification*, the parameter of interest is the correlation coefficient of the latent nonnormal distribution, and all observations are informative for this parameter. While the distributional misspecification framework led to novel insights regarding (the lack of) robustness in ML estimation of polychoric correlation, the partial misspecification framework of this paper can provide complimentary insights regarding the effects of a fraction of uninformative observations in a sample (such as careless responses), which is our primary objective.

Nevertheless, while our estimator is designed to be robust to partial misspecification caused by some uninformative responses, it can in some situations also provide a robustness gain under distributional misspecification. It turns out that if a nonnormal latent distribution differs from a normal distribution

https://doi.org/10.1017/psy.2025.10066 Published online by Cambridge University Press

mostly in the tails, our estimator produces less biased estimates than ML because it can downweigh observations that are father from the center.

To enhance accessibility and adoption by empirical researchers, an implementation of our proposed methodology in R (R Core Team, 2024) is freely available in the package `robcat` (for "ROBust CATegorical data analysis"; Welz et al., 2025) on CRAN (the Comprehensive R Archive Network) at https://CRAN.R-project.org/package=robcat. Replication materials for all numerical results in this paper are provided on GitHub at https://github.com/mwelz/robust-polycor-replication. Online supplementary materials with proofs, derivations, and additional simulations can be found at https://github.com/mwelz/robust-polycor-replication/blob/main/WelzMairAlfons2025_SupplementaryMaterials.pdf.

This paper is structured as follows. We start with reviewing related literature (Section 2) followed by the polychoric correlation model and ML estimation thereof (Section 3). Afterwards, we elaborate on the partial misspecification framework (Section 4) and introduce our robust generalized ML estimator including its statistical properties (Section 5). These properties are then examined by a simulation study in which we vary the misspecification fraction systematically, and compare the result to the commonly employed standard ML estimator (Section 6). Subsequently, we demonstrate the practical usefulness in an empirical application on a Big Five administration (Goldberg, 1992) by Arias et al. (2020), where we find evidence of careless responding, manifesting in differences in polychoric correlation estimates of as much as 0.3 between our robust estimator and ML (Section 7). We then investigate the performance of the estimator under distributional misspecification (Section 8) and conclude with a discussion of the results and avenues for further research (Section 9).

2. Related literature

ML estimation of polychoric correlations was originally believed to be fairly robust to slight to moderate distributional misspecification (Coenders et al., 1997; Flora & Curran, 2004; Li, 2016; Maydeu-Olivares, 2006). This belief was based on simulations that generated data for nonnormal latent variables via the Vale-Maurelli (VM) method (Vale & Maurelli, 1983), which were then discretized to ordinal data. However, Grønneberg and Foldnes (2019) show that the distribution of ordinal data generated in this way is indistinguishable from that of ordinal data stemming from discretizing normally distributed latent variables.[1] In other words, simulation studies that ostensibly modeled nonnormality did in fact model normality. Simulating ordinal data in a way that ensures proper violations of latent normality (Grønneberg & Foldnes, 2017) reveals that polychoric correlation is in fact highly susceptible to distributional misspecification, resulting in possibly large biases (Foldnes & Grønneberg, 2020, 2022; Grønneberg & Foldnes, 2022; Jin & Yang-Wallentin, 2017). Consequently, it is recommended to test for the validity of the latent normality assumption, for instance by using the bootstrap test of Foldnes and Grønneberg (2020).

Another source of model misspecification occurs when the polychoric model is only misspecified for an uninformative subset of a sample (partial misspecification), where, in the context of this paper, the term "uninformative" refers to an absence of relevant information for polychoric correlation, for instance in careless responses. Careless responding *"occurs when participants are not basing their response on*

[1] A key reason for this finding is that the VM method produces a latent vector whose distribution corresponds either exactly or near-exactly to a Gaussian copula (Foldnes & Grønneberg, 2015) (except in regions where certain polynomials used in the VM transformation are non-monotonous). It follows that the VM transformation *"inherits a Gaussian-like property, indicating that simulation studies based on the VM approach might give overly optimistic impressions of finite-sample properties of estimators with non-Gaussian data"* (Foldnes & Grønneberg, 2015, p. 1078). To simulate non-normality, one should instead strive for a copula with distinctively non-Gaussian features, such as tail dependence and asymmetry. This is exactly what an alternative transformation method proposed by Grønneberg and Foldnes (2017) does. For instance, generating data with the tail-dependent and tail-asymmetric Clayton copula leads to markedly different behavior of normality-based estimators than with the VM method (e.g., Foldnes & Grønneberg, 2022).

https://doi.org/10.1017/psy.2025.10066 Published online by Cambridge University Press

the item content", for instance when a participant is *"unmotivated to think about what the item is asking"* (Ward & Meade, 2023). It has been shown to be a major threat to the validity of research results through a variety of psychometric issues, such as reduced scale reliability (Arias et al., 2020) and construct validity (Kam & Meyer, 2015), attenuated factor loadings, improper factor structure, and deteriorated model fit in factor analyses (Arias et al., 2020; Huang, Bowling, et al., 2015; Woods, 2006), as well as inflated type I or type II errors in hypothesis testing (Arias et al., 2020; Huang, Liu, & Bowling, 2015; Maniaci & Rogge, 2014; McGrath et al., 2010; Woods, 2006). Careless responding is widely prevalent (Bowling et al., 2016; Meade & Craig, 2012; Ward & Meade, 2023) with most estimates on its prevalence ranging from 10–15% of study participants (Curran, 2016; Huang, Liu, & Bowling, 2015; Huang et al., 2012; Meade & Craig, 2012), while already a prevalence 5–10% can jeopardize the validity of research findings (Arias et al., 2020; Credé, 2010; Welz et al., 2024; Woods, 2006). In fact, Ward and Meade (2023) conjecture that careless responding is likely present in all survey data. However, to the best of our knowledge, the effects of careless responding on estimates of the polychoric model have not yet been studied.

Existing model-based approaches to account for careless responding in various models typically explicitly model carelessness trough mixture models (e.g., Arias et al., 2020; Steinmann et al., 2022; Ulitzsch, Pohl, et al., 2022; Ulitzsch, Yildirim-Erbasli, et al., 2022; Van Laar & Braeken, 2022). In contrast, our method does not model carelessness since we refrain from making assumptions on how the polychoric model might be misspecified. Another way to address careless responding is to directly detect them through person-fit indices and subsequently remove them from the sample (e.g., Patton et al., 2019). As primary difference, our method simultaneously downweights aberrant observations during estimation rather than removing them. We refer to Alfons and Welz (2024) for a detailed overview of methods addressing careless responding in various settings.

Conceptually related to our approach, Itaya and Hayashi (2025) propose a way to robustly estimate parameters in item response theory (IRT) models. Their approach is conceptually similar to ours in the sense that it is based on minimizing a notion of divergence between an empirical density (from observed data) and a theoretical density of the IRT model. Like our approach, they achieve robustness by implicitly downweighting responses that the postulated model cannot fit well. Methodologically, our approach is different from Itaya and Hayashi (2025) because our method is based on C-estimation (Welz, 2024), which is designed specifically for categorical data, while they use *density power divergence estimation* (DPD) theory (Basu et al., 1998), which is not restricted to categorical data.[2] A relevant consequence is that our estimator is fully efficient, whereas DPD estimators lose efficiency as a price for gaining robustness. To the best of our knowledge, DPD estimators have not yet been studied for estimating polychoric correlation.

Another related branch of literature is that of outlier detection in contingency tables (see Sripriya et al., 2020, for a recent overview). In this literature, an outlier is a contingency table cell whose observed frequency is *"markedly deviant"* from those of the remaining cells (Sripriya et al., 2020). This literature is agnostic with respect to the observed contingency table and therefore does not impose a parameterization on each cell's probability. In contrast, the polychoric correlation model imposes such a parametrization through the assumption of latent bivariate normality. Another difference is that we are not primarily interested in outlier detection, but robust estimation of model parameters.

[2] Another slightly more subtle difference is that Itaya and Hayashi (2025) robustify *marginal* maximum ML, whereas we robustify *joint* ML estimation. Marginal ML is often used to estimate IRT models because joint ML is known to yield inconsistent parameter estimates in such models (e.g., Lindsay et al., 1991). The polychoric correlation model does not suffer from such problems and can therefore be estimated via joint ML. Furthermore, marginal ML in IRT is a *full information estimation* concept, whereas multivariate models that are being fitted to a given polychoric correlation matrix require *limited information estimation*. Hence, the work of Itaya and Hayashi (2025) demonstrates that model misspecification is also a relevant issue in full information estimation.

An alternative way to gain robustness against violations of latent nonnormality is to assume a different latent distribution, for instance one with heavier tails. Examples from the SEM literature use elliptical distributions (Yuan et al., 2004) or skew-elliptical distributions (Asparouhov & Muthén, 2016), while Lyhagen and Ornstein (2023), Jin and Yang-Wallentin (2017), and Roscino and Pollice (2006) consider nonnormal distributions specifically in the context of polychoric correlation. Furthermore, it is worth pointing out the term "robustness" is used in different ways in the methodological literature. Here, it refers to robustness against model misspecification. A popular but different meaning is robustness against heteroskedastic standard errors and corrected goodness-of-fit test statistics (e.g., Li, 2016; Satorra & Bentler, 1994, 2001, and references therein), which is, for instance, how the popular software package `lavaan` (Rosseel, 2012) uses the term. We refer to Alfons and Schley (2025) for an overview of the different meanings of "robustness" and a more detailed discussion.

3. 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. If both items are dichotomous, the polychoric correlation model reduces to the tetrachoric correlation model of Pearson (1901). In the following, we first define the model and review maximum likelihood (ML) estimation thereof and then introduce a robust estimator in the next section.

3.1. *The polychoric model*

For ease of exposition, we restrict our presentation to the bivariate polychoric model. The model naturally generalizes to higher dimensions, see, e.g., Muthén (1984).

Let there be two ordinal random variables, X and Y, that take values in the sets $\mathcal{X} = \{1, 2, \dots, K_X\}$ and $\mathcal{Y} = \{1, 2, \dots, K_Y\}$, respectively. The assumption that the sets contain adjacent integers is without loss of generality. Suppose there exist two continuous latent random variables, ξ and η, that govern the ordinal variables through the discretization model

$$X = \begin{cases} 1 & \text{if } \xi < a_1, \\ 2 & \text{if } a_1 \leq \xi < a_2, \\ 3 & \text{if } a_2 \leq \xi < a_3, \\ \vdots & \\ K_X & \text{if } a_{K_X - 1} \leq \xi, \end{cases} \quad \text{and} \quad Y = \begin{cases} 1 & \text{if } \eta < b_1, \\ 2 & \text{if } b_1 \leq \eta < b_2, \\ 3 & \text{if } b_2 \leq \eta < b_3, \\ \vdots & \\ K_Y & \text{if } b_{K_Y - 1} \leq \eta, \end{cases} \tag{3.1}$$

where the fixed but unobserved parameters $a_1 < a_2 < \cdots < a_{K_X - 1}$ and $b_1 < b_2 < \cdots < b_{K_Y - 1}$ are called *thresholds*.

The primary object of interest is the population correlation between the two latent variables. To identify this quantity from the ordinal variables (X, Y), one assumes that the continuous latent variables follow a standard bivariate normal distribution with unobserved pairwise correlation coefficient $\rho \in (-1, 1)$, that is,

$$\begin{pmatrix} \xi \\ \eta \end{pmatrix} \sim \mathrm{N}_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right). \tag{3.2}$$

https://doi.org/10.1017/psy.2025.10066 Published online by Cambridge University Press

Combining the discretization model (3.1) with the latent normality model (3.2) yields the *polychoric model*. In this model, one refers to the correlation parameter $\rho = \mathbb{C}\text{or}\left[\xi,\ \eta\right]$ as the *polychoric correlation coefficient* of the ordinal X and Y. The polychoric model is subject to $d = K_X + K_Y - 1$ parameters, namely the polychoric correlation coefficient from the latent normality model (3.2) and the two sets of thresholds from the discretization model (3.1). These parameters are jointly collected in a d-dimensional vector

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

Under the polychoric model, the probability of observing an ordinal response $(x, y) \in \mathcal{X} \times \mathcal{Y}$ at a parameter vector $\boldsymbol{\theta}$ is given by

$$p_{xy}(\boldsymbol{\theta}) = \mathbb{P}_{\boldsymbol{\theta}}\left[X = x, Y = y\right] = \int_{a_{x-1}}^{a_x} \int_{b_{y-1}}^{b_y} \phi_2\left(t, s; \rho\right) \mathrm{d}s\, \mathrm{d}t, \tag{3.3}$$

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

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

Regarding identification, it is worth mentioning that in the case where both X and Y are dichotomous, the polychoric model is exactly identified by the standard bivariate normal distribution. If at least one of the ordinal variables has more than two response categories, the polychoric model is over-identified, so it could identify more parameters than those in $\boldsymbol{\theta}$.[3] We refer to Olsson (1979, Section 2) for a related discussion.

To distinguish arbitrary parameter values $\boldsymbol{\theta}$ from a specific value under which the polychoric model generates ordinal data, denote the latter by $\boldsymbol{\theta}_* = \left(\rho_*, a_{*,1}, \dots, a_{*,K_X-1}, b_{*,1}, \dots, b_{*,K_Y-1}\right)^\top$. Given a random sample of ordinal data generated by a polychoric model under parameter value $\boldsymbol{\theta}_*$, the statistical problem is to estimate the true $\boldsymbol{\theta}_*$, which is traditionally achieved by the maximum likelihood estimator of Olsson (1979).[4]

3.2. Maximum likelihood estimation

Suppose we observe a sample $\{(X_i, Y_i)\}_{i=1}^N$ of N independent copies of (X, Y) generated by the polychoric model under the true parameter $\boldsymbol{\theta}_*$. The sample may be observed directly or as a $K_X \times K_Y$ contingency

[3] An ordinal sample provides $K_X K_Y$ statistics, namely the observed frequencies for each response category $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Since the sum of the individual response frequencies must equal the sample size, one loses one degree of freedom so that only $K_X K_Y - 1$ of the frequency statistics are independent. Subsequently, the ordinal sample can identify at most $K_X K_Y - 1$ parameters. In particular, when X and Y are both dichotomous, only three parameters can be identified. In this case, the polychoric model (which reduces to the tetrachoric model here) depends on three parameters, namely a correlation parameter and two threshold parameters.

[4] Alternatives to the commonly used maximum likelihood estimator of Olsson (1979) have been proposed in the literature, see, for instance, Zhang et al. (2024), Jöreskog (1994), and Lau (1985).

https://doi.org/10.1017/psy.2025.10066 Published online by Cambridge University Press

table that cross-tabulates the observed frequencies. Denote by

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

the observed empirical frequency of a response $(x, y) \in \mathcal{X} \times \mathcal{Y}$, where the indicator function $\mathbb{1}\{E\}$ takes value 1 if an event E is true, and 0 otherwise. The maximum likelihood estimator (MLE) of $\boldsymbol{\theta}_*$ can be expressed as

$$\widehat{\boldsymbol{\theta}}_N^{\text{MLE}} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\{ \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} N_{xy} \log\left(p_{xy}(\boldsymbol{\theta})\right) \right\}, \tag{3.4}$$

where the $p_{xy}(\boldsymbol{\theta})$ are the response probabilities in (3.3), and

$$\boldsymbol{\Theta} = \left(\left(\rho, (a_i)_{i=1}^{K_X-1}, (b_j)_{j=1}^{K_Y-1}\right)^\top \,\middle|\, \rho \in (-1, 1),\ a_1 < \cdots < a_{K_X-1},\ b_1 < \cdots < b_{K_Y-1} \right) \tag{3.5}$$

is a set of parameters $\boldsymbol{\theta}$ that rules out degenerate cases such as $\rho = \pm 1$ or thresholds that are not strictly monotonically increasing. This estimator, its computational details, as well as its statistical properties are derived in Olsson (1979). In essence, if the polychoric model (3.1) is correctly specified—that is, the underlying latent variables (ξ, η) are indeed standard bivariate normal—then the estimator $\widehat{\boldsymbol{\theta}}_N^{\text{MLE}}$ is consistent for the true $\boldsymbol{\theta}_*$. In addition, $\widehat{\boldsymbol{\theta}}_N^{\text{MLE}}$ is asymptotically normally distributed with mean zero and covariance matrix being equal to the model's inverse Fisher information matrix, which makes it fully efficient.

As a computationally attractive alternative to estimating all parameters in $\boldsymbol{\theta}_*$ simultaneously in problem (3.4), one may consider a "2-step-approach" where only the correlation coefficient ρ_* is estimated via ML, 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 contingency table cell. Formally, in the 2-step-approach, thresholds $a_{*,x}$ and $b_{*,y}$ are respectively estimated via

$$\widehat{a}_x = \Phi_1^{-1}\left(\frac{1}{N}\sum_{k=1}^{x}\sum_{y\in\mathcal{Y}} N_{ky}\right) \qquad \text{and} \qquad \widehat{b}_y = \Phi_1^{-1}\left(\frac{1}{N}\sum_{x\in\mathcal{X}}\sum_{l=1}^{y} N_{xl}\right), \tag{3.6}$$

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 ML. The main advantage of the 2-step approach is reduced computational time, while it comes at the cost of being theoretically non-optimal because ML standard errors do not apply to the threshold estimators in (3.6) (Olsson, 1979). Using simulation studies, Olsson (1979) finds that the two approaches tend to yield similar results in practice—both in terms of correlation and variance estimation—for small to moderate true correlations, while there can be small differences for larger true correlations.

Software implementations of polychoric correlation vary with respect to their estimation strategy. For instance, the popular R packages `lavaan` (Rosseel, 2012) and `psych` (Revelle, 2024) only support the 2-step approach, while the package `polycor` (Fox, 2022) supports both the 2-step-approach and

https://doi.org/10.1017/psy.2025.10066 Published online by Cambridge University Press

simultaneous estimation of all model parameters, with the former being the default. Our implementation of ML estimation in package `robcat` also supports both strategies.

4. Conceptualizing model misspecification

To study the effects of partial model misspecification from a theoretical perspective, we first rigorously define this concept and explain how it differs from distributional misspecification.

4.1. Partial misspecification of the polychoric model

The polychoric model is partially misspecified when not all unobserved realizations of the latent variables (ξ, η) come from a standard bivariate normal distribution. Specifically, we consider a situation where only a fraction $(1-\varepsilon)$ of those realizations are generated by a standard bivariate normal distribution with true correlation parameter ρ_*, whereas a fixed but unknown fraction ε are generated by some different but unspecified distribution H. Note that H being unspecified allows its correlation coefficient to differ from ρ_* so that realizations generated by H may be uninformative for the true polychoric correlation coefficient ρ_*, such as, after discretization, careless responses, misresponses or responses stemming from item misunderstanding.

Formally, we say that the polychoric model is partially misspecified if the latent variables (ξ, η) are jointly distributed according to

$$(u, v) \mapsto G_\varepsilon(u, v) = (1 - \varepsilon)\Phi_2(u, v; \rho_*) + \varepsilon H(u, v), \tag{4.1}$$

for $u, v \in \mathbb{R}$. Conceptualizing model misspecification in such a manner is standard in the robust statistics literature, going back to the seminal work of Huber (1964).[5] We therefore adopt terminology from robust statistics and call ε the *contamination fraction*, the uninformative H the *contamination distribution* (or simply *contamination*), and G_ε the *contaminated distribution*. Observe that when the contamination fraction is zero, that is, $\varepsilon = 0$, there is no misspecification so that the polychoric model is correctly specified for all observations. However, neither the contamination fraction ε nor the contamination distribution H are assumed to be known. Thus, both quantities are left completely unspecified in practice and $\Phi_2(u, v; \rho_*)$ remains the distribution of interest. That is, we only aim to estimate the model parameters $\boldsymbol{\theta}$ of the polychoric model, while reducing the adverse effects of potential contamination in the observed ordinal data. The contaminated distribution G_ε, on the other hand, is never estimated. It serves as purely theoretical construct that we use to study the theoretical properties of estimators of the polychoric model when that model is partially misspecified due to contamination.

Leaving the contamination distribution H and contamination fraction ε unspecified in the partial misspecification model (4.1) means that we are not making any assumptions on the degree, magnitude, or type of contamination (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, or item misunderstanding.

Although we make no assumption on the specific value of the contamination fraction ε in the partial misspecification model (4.1), we require the identification restriction $\varepsilon \in [0, 0.5)$. That is, we require that the polychoric model is correctly specified for the majority of observations, which is standard in the robust statistics literature (e.g., Hampel et al., 1986, p. 67). While it is in principle possible to

[5] In the robust statistics literature, a distribution like (4.1) is known as *Huber contamination model*. For continuous random variables, this model is primarily used to model outliers and study the properties of outlier-robust estimators.

https://doi.org/10.1017/psy.2025.10066 Published online by Cambridge University Press

also consider a contamination fraction between 0.5 and 1, one would need to impose certain additional assumptions on the correct model to distinguish it from incorrect ones when the majority of observations are not generated by the correct model. Since we prefer refraining from imposing additional assumptions, we only consider $\varepsilon \in [0, 0.5)$. We discuss the link between identification and contamination fractions beyond 0.5 in more detail in Online Supplement E.1.

Furthermore, as another, more practical reason for considering $\varepsilon \in [0, 0.5)$, having more than half of all observations in a sample being not informative for the quantity of interest would be indicative of serious data quality issues. When data quality is unreasonably low, it is doubtful whether the data are suitable for modeling analyses in the first place.

4.2. Response probabilities under partial misspecification

Under contaminated distribution G_ε with contamination fraction $\varepsilon \in [0, 0.5)$, the probability of observing an ordinal response $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is given by

$$f_\varepsilon(x,y) = \mathbb{P}_{G_\varepsilon}[X = x, Y = y] = (1-\varepsilon) p_{xy}(\boldsymbol{\theta}_*) + \varepsilon \int_{a_{\varepsilon,x-1}}^{a_{\varepsilon,x}} \int_{b_{\varepsilon,y-1}}^{b_{\varepsilon,y}} \mathrm{d}H, \tag{4.2}$$

where the unobserved thresholds $a_{\varepsilon,x}, b_{\varepsilon,y}$ discretize the fraction ε of latent variables for which the polychoric model is misspecified. The thresholds $a_{\varepsilon,x}, b_{\varepsilon,y}$ may be different from the true $a_{*,x}, b_{*,y}$ and/or depend on contamination fraction ε. However, it turns out that from a theoretical perspective, studying the case where the $a_{\varepsilon,x}, b_{\varepsilon,y}$ are different from the $a_{*,x}, b_{*,y}$ is equivalent to a case where they are equal.[6]

The population response probabilities $f_\varepsilon(x, y)$ in (4.2) are unknown in practice because they depend on unspecified and unmodeled quantities, namely the contamination fraction ε, the contamination distribution H, and the discretization thresholds of the latter. Consequently, we do not attempt to estimate the population response probabilities $f_\varepsilon(x, y)$. We instead focus on estimating the true polychoric model probabilities $p_{xy}(\boldsymbol{\theta}_*)$ while reducing bias stemming from potential contamination in the observed data.

Figure 1 visualizes a simulated example of bivariate data drawn from contaminated distribution G_ε, where a fraction of $\varepsilon = 0.15$ of the data follow a bivariate normal contamination distribution H (orange dots) with mean $(2.5, -2.5)^\top$, variance $(0.25, 0.25)^\top$, and zero correlation, whereas the remaining data are generated by a standard bivariate normal distribution with correlation $\rho_* = 0.5$ (gray dots). In this example, the data from the contamination distribution H primarily inflate the cell $(x, y) = (5, 1)$ after discretization. That is, 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 (partial) misspecification of the polychoric model, a maximum likelihood estimate of ρ_* on these data might be substantially biased for ρ_*. Indeed, calculating the MLE using the data plotted in Figure 1 yields an estimate of $\widehat{\rho}_N^{\text{MLE}} = -0.10$, 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 Section 5, yields a fairly accurate estimate of 0.47.

It is worth addressing that there exist nonnormal distributions of the latent variables (ξ, η) that, after discretization with the same thresholds, result in the same response probabilities as under latent normality

[6] Let H' be an arbitrary contamination distribution of the latent variables. Let the thresholds that discretize these latent variables be given by arbitrary values $a'_{\varepsilon,x}, b'_{\varepsilon,y}$, for $x \in \mathcal{X}, y \in \mathcal{Y}$. Since one makes no assumption on the contamination distribution in (4.1), we can find another contamination distribution $H \neq H'$ that, when discretized with the true thresholds $a_{*,x}, b_{*,y}$, yields the same discretization as H' with thresholds $a'_{\varepsilon,x}, b'_{\varepsilon,y}$. Formally, $\forall a'_{\varepsilon,x}, b'_{\varepsilon,y}, H'$ $\exists H$ s.t. $\int_{a'_{\varepsilon,x-1}}^{a'_{\varepsilon,x}} \int_{b'_{\varepsilon,y-1}}^{b'_{\varepsilon,y}} \mathrm{d}H' = \int_{a_{*,x-1}}^{a_{*,x}} \int_{b_{*,y-1}}^{b_{*,y}} \mathrm{d}H$.

https://doi.org/10.1017/psy.2025.10066 Published online by Cambridge University Press

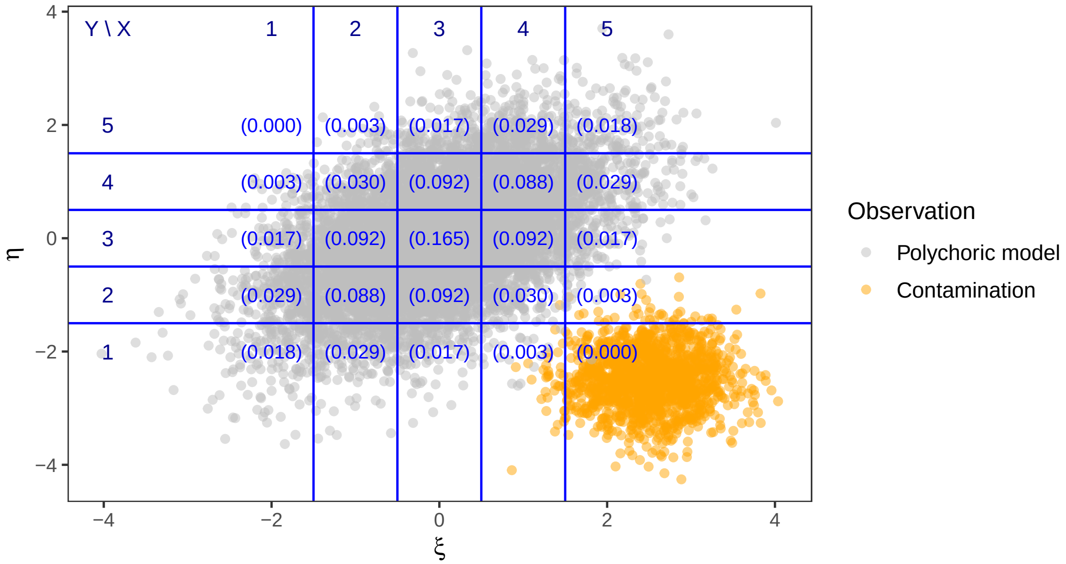

Figure 1: Simulated data with $K_X = K_Y = 5$ response options where the polychoric model is misspecified with contamination fraction $\varepsilon = 0.15$. The gray dots represent random draws of (ξ, η) from the polychoric model with $\rho_* = 0.5$, whereas the orange dots represent draws from a contamination distribution that primarily inflates the cell $(x, y) = (5, 1)$. The contamination distribution is bivariate normal with a mean $(2.5, -2.5)^\top$, variances $(0.25, 0.25)^\top$, and zero correlation. The blue lines indicate the locations of the thresholds. In each cell, the numbers in parentheses denote the population probability of that cell under the true polychoric model.

(Foldnes & Grønneberg, 2019). This implies that there may exist contamination distributions H and contamination fractions $\varepsilon > 0$ under which the population probabilities $f_\varepsilon(x, y)$ in (4.2) are equal to the true population probabilities of the polychoric model, $p_{xy}(\boldsymbol{\theta}_*)$, that is, $f_\varepsilon(x, y) = p_{xy}(\boldsymbol{\theta}_*)$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$. In this situation, the polychoric model is misspecified, but the misspecification does not have consequences because the response probabilities remain unaffected. To avoid cumbersome notation in the theoretical analysis of our robust estimator, we assume consequential misspecification throughout this paper, that is, $f_\varepsilon(x, y) \neq p_{xy}(\boldsymbol{\theta}_*)$ for some $(x, y) \in \mathcal{X} \times \mathcal{Y}$ whenever $\varepsilon > 0$. However, it is silently understood that misspecification need not be consequential, in which case there is no issue and both the MLE and our robust estimator are consistent for the true $\boldsymbol{\theta}_*$.

4.3. Distributional misspecification

A model is distributionally misspecified when all observations in a given sample are generated by a distribution that is different from the model distribution. In the context of the polychoric model, this means that all ordinal observations are generated by a latent distribution that is nonnormal. Let G denote the unknown nonnormal distribution that the latent variables (ξ, η) jointly follow under distributional misspecification. The object of interest is the population correlation between latent ξ and η under distribution G, for which the normality-based MLE of Olsson (1979) turns out to be substantially biased in many cases (e.g., Foldnes & Grønneberg, 2020, 2022; Jin & Yang-Wallentin, 2017; Lyhagen

& Ornstein, 2023). As such, distributional misspecification is fundamentally different from partial misspecification: In the former, one attempts to estimate the population correlation of the nonnormal and unknown distribution that generated a sample, instead of estimating the polychoric correlation coefficient (which is the correlation under standard bivariate normality). In the latter, one attempts to estimate the polychoric correlation coefficient with a contaminated sample that has only been partly generated by latent normality (that is, the polychoric model). The assumption that the polychoric model is only partially misspecified for some uninformative observations enables one to still estimate the polychoric correlation coefficient of that model, which would not be feasible under distributional misspecification (at least not without additional assumptions).

Despite not being designed for distributional misspecification, the robust estimator introduced in the next section can offer a robustness gain in some situations where the polychoric model is distributionally misspecified. We discuss this in more detail in Section 8.

5. Robust estimation of polychoric correlation

The behavior of ML estimates of any model crucially depends on correct specification of that model. Indeed, ML estimation can be severely biased even when the assumed model is only slightly misspecified (e.g., Hampel et al., 1986; Huber, 1964; Huber & Ronchetti, 2009). For instance, in many models of continuous variables like regression models, one single observation from a different distribution can be enough to make the ML estimator converge to an arbitrary value (Huber and Ronchetti, 2009; see also Alfons et al., 2022, for the special case of mediation analysis). The non-robustness of ML estimation of the polychoric model has been demonstrated empirically by Foldnes and Grønneberg (2020, 2022) and Grønneberg and Foldnes (2022) for the case of distributional misspecification. In this section, we introduce an estimator that is designed to be robust to partial misspecification when present, but remains (asymptotically) equivalent to the ML estimator of Olsson (1979) when misspecification is absent. We furthermore derive the statistical properties of the proposed estimator.

Throughout this section, let $\{(X_i, Y_i)\}_{i=1}^N$ be an observed ordinal sample of size N generated by discretizing latent variables (ξ, η) that follow the unknown and unspecified contaminated distribution G_ε in (4.1). Hence, the polychoric model is possibly misspecified for an unknown fraction ε of the sample.

5.1. *The estimator*

The proposed estimator is a special case of a class of robust estimators for general models of categorical data called C-estimators (Welz, 2024), and is based on the following idea. The empirical relative frequency of a response $(x, y) \in \mathcal{X} \times \mathcal{Y}$, denoted

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

is a consistent nonparametric estimator of the population response probability in (4.2),

$$f_\varepsilon\left(x, y\right) = \mathbb{P}_{G_\varepsilon}\left[X = x, Y = y\right] = (1-\varepsilon)p_{xy}(\boldsymbol{\theta}_*) + \varepsilon\int_{a_{\varepsilon,x-1}}^{a_{\varepsilon,x}}\int_{b_{\varepsilon,y-1}}^{b_{\varepsilon,y}} \mathrm{d}H,$$

https://doi.org/10.1017/psy.2025.10066 Published online by Cambridge University Press

as $N \longrightarrow \infty$ (see, e.g., Chapter 19.2 in Van der Vaart, 1998). If the polychoric model is correctly specified ($\varepsilon = 0$), then $\widehat{f}_N(x,y)$ will converge (in probability) to the true model probability $p_{xy}(\boldsymbol{\theta}_*)$ because

$$f_0(x,y) = p_{xy}(\boldsymbol{\theta}_*),$$

for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$. Conversely, if the polychoric model is misspecified ($\varepsilon > 0$), then $\widehat{f}_N(x,y)$ may *not* converge to the true $p_{xy}(\boldsymbol{\theta}_*)$ because

$$f_\varepsilon(x,y) \neq p_{xy}(\boldsymbol{\theta}_*)$$

for some $(x,y) \in \mathcal{X} \times \mathcal{Y}$, since we assume consequential misspecification.

It follows that if the polychoric model is misspecified, there exists no parameter value $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ for which the nonparametric estimates $\widehat{f}_N(x,y)$ converge pointwise to the associated model probabilities $p_{xy}(\boldsymbol{\theta})$ for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$. Hence, it is indicative of model misspecification if there exists at least one response $(x,y) \in \mathcal{X} \times \mathcal{Y}$ for which $\widehat{f}_N(x,y)$ does not converge to any polychoric model probability $p_{xy}(\boldsymbol{\theta})$, resulting in a discrepancy between $\widehat{f}_N(x,y)$ and $p_{xy}(\boldsymbol{\theta})$.[7] This observation can be exploited in model fitting by minimizing the discrepancy between the empirical relative frequencies, $\widehat{f}_N(x,y)$, and theoretical model probabilities, $p_{xy}(\boldsymbol{\theta})$, to find the most accurate fit that can be achieved with the polychoric model for the observed data. Specifically, our estimator minimizes with respect to $\boldsymbol{\theta}$ the loss function

$$L\left(\boldsymbol{\theta},\ \widehat{f}_N\right) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \varphi\left(\frac{\widehat{f}_N(x,y)}{p_{xy}(\boldsymbol{\theta})} - 1\right) p_{xy}(\boldsymbol{\theta}), \tag{5.1}$$

where $\varphi : [-1, \infty) \longrightarrow \mathbb{R}$ is a prespecified *discrepancy function* that will be defined momentarily. The proposed estimator $\widehat{\boldsymbol{\theta}}_N$ is given by the value minimizing the objective loss over parameter space $\boldsymbol{\Theta}$,

$$\widehat{\boldsymbol{\theta}}_N = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} L\left(\boldsymbol{\theta}, \widehat{f}_N\right). \tag{5.2}$$

For the choice of discrepancy function $\varphi(z) = \varphi^{\text{MLE}}(z) = (z+1)\log(z+1)$, it can be easily verified that $\widehat{\boldsymbol{\theta}}_N$ coincides with the MLE $\widehat{\boldsymbol{\theta}}_N^{\text{MLE}}$ in (3.4). In the following, we motivate a specific choice of discrepancy function $\varphi(\cdot)$ that makes the estimator $\widehat{\boldsymbol{\theta}}_N$ less susceptible to misspecification of the polychoric model while preserving equivalence with ML estimation in the absence of misspecification.

The fraction between empirical relative frequencies and model probabilities with value 1 deducted,

$$\frac{\widehat{f}_N(x,y)}{p_{xy}(\boldsymbol{\theta})} - 1,$$

is referred to as *Pearson residual* (PR) (Lindsay, 1994). It takes values in $[-1, +\infty)$ and can be interpreted as a goodness-of-fit measure. PR values close to 0 indicate a good fit between data and polychoric model at $\boldsymbol{\theta}$, whereas values toward -1 or $+\infty$ indicate a poor fit because empirical response probabilities disagree with their model counterparts. To achieve robustness to misspecification of the polychoric model, responses that cannot be modeled well by the polychoric model, as indicated by their PR being away from 0, should receive less weight in the estimation procedure such that they do not overproportionally affect the fit. Downweighting when necessary is achieved through a specific choice of

[7] Actually, if said convergence fails, it does so for at least two responses. Due to the relative frequencies summing up to 1, if the convergence fails for one response, it must also fail for at least one more.

https://doi.org/10.1017/psy.2025.10066 Published online by Cambridge University Press

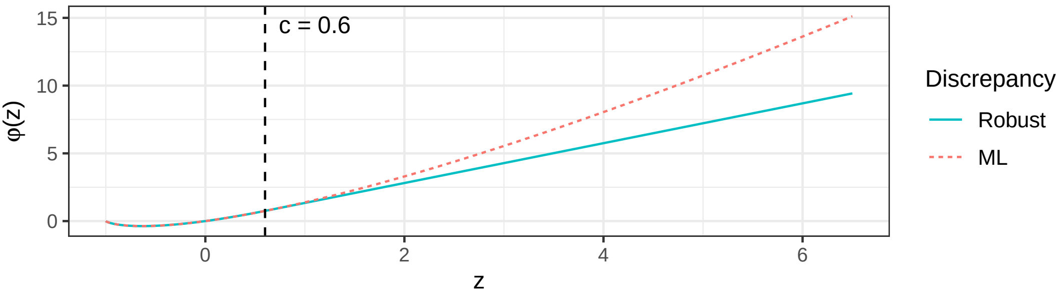

Figure 2: Visualization of the robust discrepancy function $\varphi(z)$ in (5.3) for $c = 0.6$ (solid line) and the ML discrepancy function $\varphi^{\text{MLE}}(z) = (z+1)\log(z+1)$ (dotted line).

discrepancy function $\varphi(\cdot)$ proposed by Welz (2024), which is a special case of a function suggested by Ruckstuhl and Welsh (2001). The discrepancy function reads

$$\varphi(z) = \begin{cases} (z+1)\log(z+1) & \text{if } z \in [-1, c], \\ (z+1)(\log(c+1)+1) - c - 1 & \text{if } z > c, \end{cases} \tag{5.3}$$

where $c \in [0, \infty]$ is a prespecified tuning constant that governs the estimator's behavior at the PR of each possible response. Figure 2 visualizes this function for the example choice $c = 0.6$ as well as the ML discrepancy function $\varphi^{\text{MLE}}(z) = (z+1)\log(z+1)$. Note that deducting 1 in (5.1) and adding it again in (5.3) is purely for keeping the interpretation that a PR close to 0 indicates a good fit. We further stress that although the discrepancy function (5.3) can be negative, the loss function (5.1) is always nonnegative (Welz, 2024).

For the choice $c = +\infty$, minimizing the loss (5.1) is equivalent to maximizing the log-likelihood objective in (3.4), 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 $z = \frac{\widehat{f}_N(x,y)}{p_{xy}(\boldsymbol{\theta})} - 1$ of a response $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is such that $z \in [-1, c]$ for fixed $c \geq 0$, then the estimation procedure behaves at this response like in classic ML estimation. As argued before, in the absence of misspecification, $\widehat{f}_N(x,y)$ converges to $p_{xy}(\boldsymbol{\theta}_*)$ for all responses $(x, y) \in \mathcal{X} \times \mathcal{Y}$, therefore all PRs are asymptotically equal to 0. Hence, if the polychoric model is correctly specified, then estimator $\widehat{\boldsymbol{\theta}}_N$ is asymptotically equivalent to the MLE $\widehat{\boldsymbol{\theta}}_N^{\text{MLE}}$ for any tuning constant value $c \geq 0$. On the other hand, if a response's PR z is larger than c, that is, $z > c \geq 0$, then the estimation procedure does not behave like in ML, but the response's contribution to loss (5.1) is linear rather than superlinear like in ML (Figure 2). It follows that the influence of responses that cannot be fitted well by the polychoric model is downweighted to prevent them from dominating the fit. The tuning constant $c \geq 0$ is the threshold beyond which a PR will be downweighted, so the choice thereof determines what is considered an insufficient fit. The closer to 0 the tuning constant c is chosen, the more robust the estimator is in theory. In Online Supplement C, we explore different values of c in simulations and motivate a specific choice that we use for all numerical results in this paper, namely $c = 0.6$.

Note that the discrepancy function $\varphi(\cdot)$ in (5.3) may only downweight *overcounts*, that is, the empirical probability $\widehat{f}_N(x,y)$ exceeding the theoretical probability $p_{xy}(\boldsymbol{\theta})$ for some cell $(x, y) \in \mathcal{X} \times \mathcal{Y}$. One might wonder why *undercounts*—$\widehat{f}_N(x,y)$ being smaller than $p_{xy}(\boldsymbol{\theta})$, resulting in negative Pearson residuals—are not downweighted as well. Indeed, the discrepancy function in (5.3) does not change its behavior compared to the MLE for Pearson residuals below 0. The empirical frequency $\widehat{f}_N(x,y)$

https://doi.org/10.1017/psy.2025.10066 Published online by Cambridge University Press

is a *relative* measure, so if a contingency table cell (x, y) has inflated counts, the other cells will have reduced values of $\widehat{f}_N$. If the discrepancy function would downweight undercounts, there is a risk of downweighting non-contaminated cells simply because these cells have reduced $\widehat{f}_N$ values if at least one cell is inflated due to contamination. Such behavior could result in bias since non-contaminated cells are not supposed to be downweighted. We refer to Ruckstuhl and Welsh (2001, p. 1128) for a related discussion.

With the proposed choice of $\varphi(\cdot)$, we stress that our estimator $\widehat{\boldsymbol{\theta}}_N$ in (5.2) has the same time complexity as ML, namely $O\big(K_X \cdot K_Y\big)$, since one needs to calculate the Pearson residual of all $K_X \cdot K_Y$ possible responses for every candidate parameter value. Consequently, our proposed estimator does not incur any additional computational cost compared to ML.

5.2. *Statistical properties*

We first address what quantity is estimated by the proposed estimator before discussing its asymptotic behavior.

5.2.1. Estimand

The estimand of the estimator $\widehat{\boldsymbol{\theta}}_N$ in (5.2) 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 (5.2) that the sample-based $\widehat{\boldsymbol{\theta}}_N$ solves because the probabilities $f_\varepsilon(x, y)$ are the population analogues to the empirical probabilities $\widehat{f}_N(x, y)$.

If the polychoric model is correctly specified, the estimand $\boldsymbol{\theta}_0$ equals the true parameter $\boldsymbol{\theta}_*$. Indeed, if $\varepsilon = 0$, then $f_0(x, y) = p_{xy}(\boldsymbol{\theta}_*)$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, so it follows that the loss

$$\begin{aligned}
L\Big(\boldsymbol{\theta}, f_0\Big) &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \varphi\left(\frac{p_{xy}(\boldsymbol{\theta}_*)}{p_{xy}(\boldsymbol{\theta})} - 1\right) p_{xy}(\boldsymbol{\theta}) \\
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{xy}(\boldsymbol{\theta}_*)\log\left(\frac{p_{xy}(\boldsymbol{\theta}_*)}{p_{xy}(\boldsymbol{\theta})}\right)\mathbb{1}\left\{\frac{p_{xy}(\boldsymbol{\theta}_*)}{p_{xy}(\boldsymbol{\theta})} - 1 \in [-1, c]\right\} \\
&\quad + \Big(p_{xy}(\boldsymbol{\theta}_*)\big(\log(c+1) + 1\big) - p_{xy}(\boldsymbol{\theta})(c+1)\Big)\mathbb{1}\left\{\frac{p_{xy}(\boldsymbol{\theta}_*)}{p_{xy}(\boldsymbol{\theta})} - 1 > c\right\}
\end{aligned}$$

attains its global minimum of zero if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}_*$, for any choice of $c \geq 0$. Thus, in the absence of contamination, our estimator estimates the same quantity as the MLE, namely the true $\boldsymbol{\theta}_*$. In other words, it obtains the true $\boldsymbol{\theta}_*$ in population when the model is correctly specified, a property known as *Fisher consistency*. We refer to Welz (2024) for details.

However, in the presence of misspecification ($\varepsilon > 0$), the sampling distribution differs from the model distribution such that the estimand $\boldsymbol{\theta}_0$—the parameter that minimizes the loss evaluated at the sampling distribution—is generally different from the true $\boldsymbol{\theta}_*$ (cf. White, 1982).[8] The population

[8] White (1982) is concerned with (quasi-)maximum likelihood estimation, but the same reasoning applies to our estimator. In the context of White (1982), the ML estimand of a misspecified model corresponds to the parameter value minimizing the Kullback-Leibler divergence between the sampling density and the model density. For $c = \infty$ (the ML case), the population analogue of our loss function (5.1) corresponds to the Kullback-Leibler divergence between the sampling density $f_\varepsilon(x, y)$ and the model density $p_{xy}(\cdot)$, that is, $\boldsymbol{\theta} \mapsto L(\boldsymbol{\theta}, f_\varepsilon)$. For finite choices

https://doi.org/10.1017/psy.2025.10066 Published online by Cambridge University Press

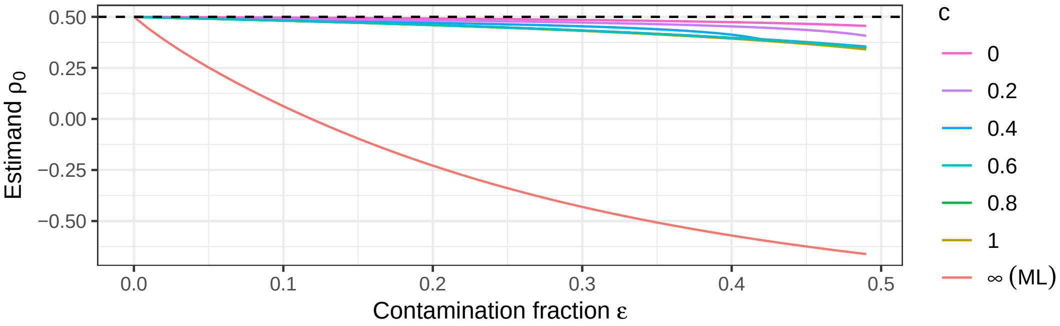

Figure 3: The population estimand ρ_0 of the polychoric correlation coefficient for various degrees of contamination fractions ε (x-axis) and tuning constants c (line colors), for the same contamination distribution as in Figure 1. The ML estimand corresponds to $c = +\infty$. There are $K_X = K_Y = 5$ response options and the true value corresponds to $\rho_* = 0.5$ (dashed line).

estimand being different from the true value translates to biased estimates of the latter as a consequence of the misspecification.

How much the estimand $\boldsymbol{\theta}_0$ differs from the true $\boldsymbol{\theta}_*$ depends on the unknown fraction of contamination ε, the unknown type of contamination H, as well as the choice of tuning constant c in $\varphi(\cdot)$. Mainly, the larger ε (more severe misspecification) and c (less downweighting of hard-to-fit responses), the further $\boldsymbol{\theta}_0$ is away from $\boldsymbol{\theta}_*$. Figure 3 illustrates this behavior for the polychoric correlation coefficient at an example misspecified distribution that is described in Section 4.2 and in which the true polychoric correlation under the correct model amounts to $\rho_* = 0.5$. For increasing contamination fractions, the MLE ($c = +\infty$) estimates a parameter value that is increasingly farther away from the true $\boldsymbol{\theta}_*$, where already a contamination fraction of less than $\varepsilon = 0.15$ suffices for a sign flip in the correlation coefficient. Conversely, choosing tuning constant c to be near 0 results in a much less severe bias. For instance, even at contamination fraction $\varepsilon = 0.4$, the difference between estimand and true value is approximately 0.1 or less.

Overall, 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, thereby gaining robustness to misspecification.

A relevant question is whether the true parameter $\boldsymbol{\theta}_*$ can be recovered when $\varepsilon > 0$ such that it can be estimated using $\widehat{\boldsymbol{\theta}}_N$ combined with a bias correction term. To derive such a bias correction term, one would need to impose assumptions on the contamination fraction and type of contamination. However, if one has strong prior beliefs about how the polychoric model is misspecified, modeling them explicitly rather than relying on the polychoric model seems more appropriate. Yet, one's beliefs about misspecification may not be accurate, so attempts to explicitly model the misspecification may themselves result in a misspecified model. Consequently, robust estimation traditionally refrains from making assumptions on how a model may potentially be misspecified by leaving ε and H unspecified in the contaminated distribution (4.1). Instead, one may use a robust estimator to identify data points that cannot be modeled with the model at hand, like the one presented in this paper.

of c, our loss function is designed to yield an estimand that is closer to the true $\boldsymbol{\theta}_*$ than that of the ML loss by downweighting responses that cannot be sufficiently well modeled.

https://doi.org/10.1017/psy.2025.10066 Published online by Cambridge University Press

5.2.2. Asymptotic analysis

It can be shown that under certain standard regularity conditions that do not restrict the degree or type of possible partial misspecification beyond $\varepsilon \in [0, 0.5)$, the robust estimator $\widehat{\boldsymbol{\theta}}_N$ is consistent for estimand $\boldsymbol{\theta}_0$ as well as asymptotically normally distributed. Specifically, under said regularity conditions and fixed tuning constant $c > 0$, Theorem A.1 in Online Supplement A establishes that

$$\widehat{\boldsymbol{\theta}}_N \xrightarrow{\mathbb{P}} \boldsymbol{\theta}_0,$$

as well as

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

as $N \longrightarrow \infty$, where "$\xrightarrow{\mathbb{P}}$" and "$\xrightarrow{\mathrm{d}}$" denote convergence in probability and distribution, respectively. The asymptotic covariance matrix has a sandwich-type construction

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

where the $d \times d$ matrices

$$\boldsymbol{M}\left(\boldsymbol{\theta}\right) = \frac{\partial^2 L\big(\boldsymbol{\theta}, f_\varepsilon\big)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \quad \text{and} \quad \boldsymbol{U}\left(\boldsymbol{\theta}\right) = \mathbb{V}\mathrm{ar}_{f_\varepsilon}\left[\frac{\partial \log\left(p_{XY}\left(\boldsymbol{\theta}\right)\right)}{\partial \boldsymbol{\theta}} \mathbb{1}\left\{\frac{f_\varepsilon\left(X, Y\right)}{p_{XY}\left(\boldsymbol{\theta}\right)} - 1 \in [-1, c]\right\}\right],$$

respectively, are the Hessian matrix of the population loss and the covariance matrix (evaluated at f_ε) of the likelihood score function—that is, the gradient of $\log(p_{XY}(\boldsymbol{\theta}))$—weighted by stochastic binary weights whether the Pearson residual is smaller than or equal to the tuning constant c.[9] We derive closed-form expressions of the matrices $\boldsymbol{M}\left(\boldsymbol{\theta}\right)$ and $\boldsymbol{U}\left(\boldsymbol{\theta}\right)$ as well as their properties in Online Supplement A..

The asymptotic covariance matrix $\boldsymbol{\Sigma}\left(\boldsymbol{\theta}_0\right)$ of our estimator is unobserved in practice because it depends on the unknown quantities $\boldsymbol{\theta}_0$ and f_ε. Yet, $\boldsymbol{\Sigma}\left(\boldsymbol{\theta}_0\right)$ can be consistently estimated by replacing $\boldsymbol{\theta}_0$ and f_ε by their corresponding consistent estimators $\widehat{\boldsymbol{\theta}}_N$ and $\widehat{f}_N$, respectively. Details are provided in Online Supplement A.

With this limit theory, one can construct standard errors and confidence intervals for every element in $\boldsymbol{\theta}_0$. Importantly, in the absence of contamination (such that $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_*$), the asymptotic covariance matrix $\boldsymbol{\Sigma}\left(\boldsymbol{\theta}_0\right)$ of the robust estimator reduces to that of the fully efficient MLE (i.e., inverse Fisher information matrix) as long as $c > 0$. It follows that there is no loss of statistical efficiency when there is no contamination.[10] Hence, if the polychoric model is correctly specified, the robust estimator and the MLE are asymptotically first and second order equivalent. We refer to Online Supplement A for a rigorous exposition and discussion of the robust estimator's asymptotic properties.

5.3. Implementation

We provide a free and open source implementation of our proposed methodology in a package for the statistical programming environment R (R Core Team, 2024). The package is called `robcat` (for "ROBust

[9] It is worth mentioning that for the MLE ($c = +\infty$), our sandwich-type covariance matrix $\boldsymbol{\Sigma}\left(\boldsymbol{\theta}_0\right)$ reduces to that of White (1982) and Huber (1967), who studied the limit distribution of ML in misspecified models.

[10] This property does not contradict standard theory on the Cramér-Rao lower bound (CRLB) for the variance of unbiased estimators (e.g., Theorem 7.3.9 in Casella & Berger, 2002). In brief, if a statistical model is correctly specified, the variance of the MLE asymptotically attains this lower bound such that no unbiased estimator can be asymptotically more efficient than the MLE. Nevertheless, this does not imply that no other estimators can (asymptotically) attain the CRLB. When the polychoric model is correctly specified (i.e., no contamination), our estimator has the same asymptotic variance as the MLE, which satisfies the Cramér-Rao lower bound with equality.

https://doi.org/10.1017/psy.2025.10066 Published online by Cambridge University Press

CATegorical data analysis"; Welz et al., 2025), and it is available from CRAN (the Comprehensive R Archive Network) at https://CRAN.R-project.org/package=robcat. To maximize speed and performance, the package is predominantly developed in C++ and integrated to R via Rcpp (Eddelbuettel, 2013). All numerical results in this paper were obtained with this package.

The estimator's minimization problem in (5.2) can be solved with standard algorithms for numerical optimization. In our experience, using an unconstrained version of the quasi-Newton algorithm L-BFGS-B of Byrd et al. (1995) works fine. However, additional stability might be gained from imposing the boundary constraint on the correlation coefficient and the monotonicity constraints on the thresholds, see (3.5), for which the the simplex algorithm of Nelder and Mead (1965) for constrained optimization can be used. In our implementation in package robcat, the default behavior is to first try unconstrained optimization via L-BFGS-B. If numerical instability is encountered or a monotonicity constraint is violated, the constrained optimization algorithm of Nelder and Mead (1965) is used instead. While this is the default behavior, the package allows users to freely specify any supported optimization routine.

An important user choice is that of the tuning constant c in discrepancy function (5.3). The closer c is to 0, the more robust the estimator will be to possible misspecification of the polychoric model (see, e.g., Figure 3). On the other hand, in the presence of model misspecification, the more robust the estimator is made, the larger its estimation variance becomes. Moreover, if the model is correctly specified, then Welz (2024) shows that the most robust choice, $c = 0$, is associated with two drawbacks, namely asymptotic nonnormality as well as certain finite sample issues. We therefore suggest choosing a value slightly larger than 0. In simulation experiments (see Online Supplement C), we find that the estimator is relatively insensitive to the specific choice of $c > 0$, as long as it is reasonably small (for robustness) yet sufficiently far away from 0 (to avoid the aforementioned issues). The choice $c = 0.6$ thereby yields a good compromise so that we use this value for all applications in this paper. However, we acknowledge that a detailed study, preferably founded in statistical theory, is necessary to provide guidelines on the choice of c. We will explore this in future work.

Furthermore, a two-step estimation procedure like in (3.6) is not recommended for robust estimation. The possible presence of responses that have not been generated by the polychoric model can inflate the empirical cumulative marginal proportion of some responses, which may result in a sizable bias of threshold estimates (3.6), possibly translating into biased estimates of polychoric correlation coefficients in the second stage. Our robust estimator therefore estimates all model parameters (thresholds and polychoric correlation) simultaneously.

6. Simulation studies on partial misspecification

In this section, we employ two simulation studies to demonstrate the robustness gain of our proposed estimator under partial misspecification of the polychoric model. The first simulation design (Section 6.1) is a simplified setting with respect to the partial misspecification, chosen specifically to illustrate the effects of a particular type of contamination with high leverage affecting only a small number of contingency table cells. The second design (Section 6.2) is more involved and considers estimation of a polychoric correlation matrix with a contamination type that scatters in many directions so that nonnormal data points can occur in every contingency table cell. Section 6.3 summarizes findings from additional simulations in the appendix. For all simulation designs, we perform $5,000$ repetitions.

6.1. Individual polychoric correlation coefficient

Let there be $K_X = K_Y = 5$ response categories for each of the two rating variables and define the true thresholds in the discretization process (3.1) 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 in latent normality model (3.2) be $\rho_* = 0.5$. To simulate partial misspecification of the polychoric model according to (4.1), we let a fraction ε of the data be generated by a particular contamination distribution H—which is left unspecified and therefore not explicitly modeled by our estimator—namely a bivariate normal distribution with mean $(2.5, -2.5)^\top$, variances $(0.25, 0.25)^\top$, and zero covariance (and therefore zero correlation). We discretize the realizations of the contamination distribution according to the same thresholds $a_{*,1}, \ldots, a_{*,4}, b_{*,1}, \ldots, b_{*,4}$ as the uncontaminated realizations. This contamination distribution will inflate the empirical frequency of contingency table cells $(x, y) \in \{(5,1), (4,1), (5,2)\}$, in the sense that they have a higher realization probability than under the true polychoric model.[11] In fact, the data plotted in Figure 1 were generated by this process for contamination fraction $\varepsilon = 0.15$, and one can see in this figure that particularly cell $(x, y) = (5, 1)$ is sampled 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 intuition, one may think of such points as the responses of careless or inattentive participants whose responses are not based on item content. Although careless responding is only one special case of the unlimited and unrestricted variety of uninformative responses generated by H, we use careless responding as an illustrative running example throughout our simulations.

For contamination fraction $\varepsilon \in \{0, 0.01, 0.05, 0.1, 0.15, 0.2, 0.3, 0.4, 0.49\}$, we sample $N = 1{,}000$ ordinal responses from this data generating process. We estimate the true parameter $\boldsymbol{\theta}_*$ with our proposed estimator with tuning constant set to $c = 0.6$, the MLE (Olsson, 1979), and, for comparison, the Pearson sample correlation calculated on the integer-valued responses.

Let $\widehat{\rho}_N$ be the estimate on a simulated dataset and $\widehat{\text{SE}}\left(\widehat{\rho}_N\right)$ the estimated standard error of $\widehat{\rho}_N$, which is constructed using the limit theory developed in Theorem A.1 in Online Supplement A. As performance measures, we calculate the average bias of the correlation estimates, the average bias of the standard error estimates (using the standard deviation of the correlation estimates across repetitions as an approximation of the true standard error), as well as coverage and average length of confidence intervals at significance level $\alpha = 0.05$. The coverage is defined as the proportion (across repetitions) of confidence intervals $\left[\widehat{\rho}_N \mp q_{1-\alpha/2} \cdot \widehat{\text{SE}}\left(\widehat{\rho}_N\right)\right]$ that contain the true ρ_*, where $q_{1-\alpha/2}$ is the $(1 - \alpha/2)$ quantile of the standard normal distribution. The length of a confidence interval is given by $2 \cdot q_{1-\alpha/2} \cdot \widehat{\text{SE}}\left(\widehat{\rho}_N\right)$.

Figure 4 visualizes the bias of each estimator with respect to the true polychoric correlation ρ_* across the 5,000 simulated datasets. An analogous plot for the whole parameter $\boldsymbol{\theta}_*$ can be found in Online Supplement D.1; the results are similar to those of ρ_*. Additional performance measures are shown in Table 1. For all considered contamination fractions, the estimates of the MLE and sample correlation are somewhat similar, which is expected because these two estimators are known to yield similar results when there are five or more rating options and the discretization thresholds are symmetric (cf., Rhemtulla et al., 2012). In the absence of contamination, the MLE and the robust estimator yield accurate estimates.

[11] Conceptually, contamination in the cell $(5, 1)$ would correspond to a straightlining careless respondent in $(5, 5)$ or $(1, 1)$ if one of the items is negatively keyed. Thus, after recoding, cell $(5, 1)$ will be inflated.

https://doi.org/10.1017/psy.2025.10066 Published online by Cambridge University Press

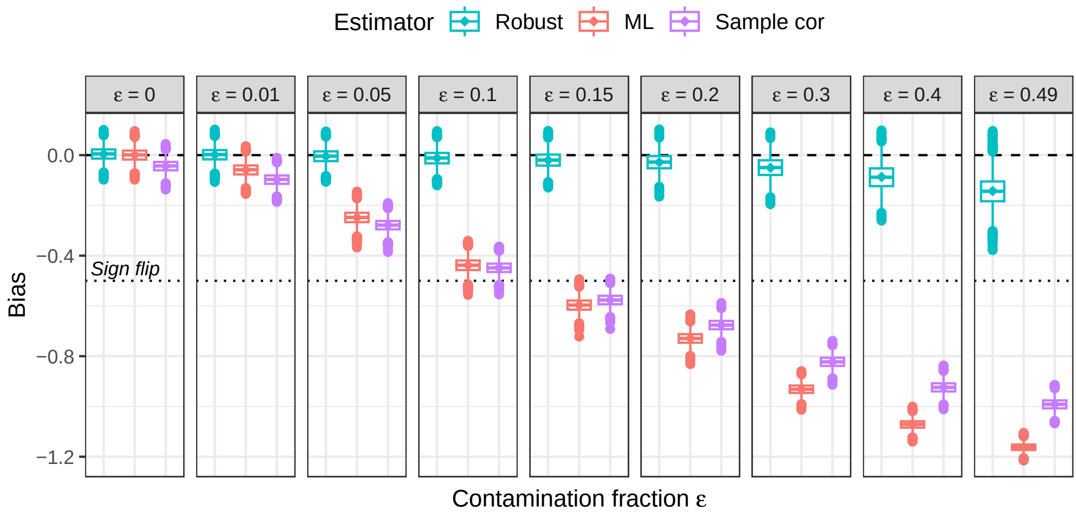

Figure 4: Boxplot visualization of the bias of three estimators of the polychoric correlation coefficient, $\widehat{\rho}_N - \rho_*$, for various contamination fractions in the misspecified polychoric model across 5,000 repetitions. The estimators are the robust estimator with $c = 0.6$ (left), the MLE (center), and the Pearson sample correlation (right). Diamonds represent the respective average bias. The dashed line denotes value 0 and the dotted line $-\rho_* = -0.5$, the latter of which indicating a sign flip in the correlation estimate.

Both estimators 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. However, when contamination is introduced, MLE, sample correlation, and the robust estimator yield noticeably different results. Already at the small contamination fraction $\varepsilon = 0.01$ (corresponding to only 10 observations), MLE and sample correlation are noticeably biased, resulting in poor coverage of only about 0.45 and 0.04, respectively. Increasing the contamination fraction to the still relatively small value of $\varepsilon = 0.05$, MLE and sample correlation start to be substantially biased with average biases of −0.25 and −0.28, respectively, leading to zero coverage. The biases of these two methods further deteriorate as the contamination fraction is gradually increased. At $\varepsilon \geq 0.15$, MLE and sample correlation produce estimates that are not only severely biased but also sign-flipped: while the true correlation is positive (0.5), both estimates are negative. In stark contrast, the proposed robust estimator remains accurate throughout nearly all considered contamination fractions. At the small $\varepsilon = 0.01$, the robust estimator is nearly unaffected, while at $\varepsilon = 0.15$, it only exhibits a minor bias of about −0.02. Even at extreme contamination $\varepsilon = 0.4$, its bias amounts to less than 0.1. In addition, coverage of the robust method remains above or close to 0.9 for contamination fractions $\varepsilon \leq 0.2$.

It is worth noting that the confidence intervals of the robust estimator grow wider with increasing contamination fraction ε. We also observe that the standard deviation of the robust estimates over the repetitions grow similarly. This indicates that the derived asymptotic distribution used to estimate standard errors matches well with the simulated distribution of the estimator across the repetitions. Indeed, the bias of standard error estimation remains at near-zero for the robust estimator, except

https://doi.org/10.1017/psy.2025.10066 Published online by Cambridge University Press

		Point estimate			Standard error		Confidence interval	
Contamination	Estimator	$\widehat{\rho}_N$	Bias	SE	$\widehat{\text{SE}}$	Bias	Coverage	Length
	Robust	0.504	0.004	0.027	0.027	−0.001	0.931	0.104
$\varepsilon = 0$	ML	0.500	0.000	0.027	0.026	−0.001	0.939	0.102
	Sample cor	0.456	−0.044	0.025	0.025	0.003	0.690	0.110
	Robust	0.501	0.001	0.028	0.027	−0.001	0.938	0.107
$\varepsilon = 0.01$	ML	0.441	−0.059	0.027	0.028	0.001	0.446	0.110
	Sample cor	0.402	−0.098	0.025	0.029	0.004	0.043	0.114
	Robust	0.495	−0.005	0.029	0.029	−0.001	0.939	0.112
$\varepsilon = 0.05$	ML	0.252	−0.248	0.028	0.033	0.005	0.000	0.128
	Sample cor	0.221	−0.279	0.025	0.031	0.006	0.000	0.121
	Robust	0.488	−0.012	0.031	0.031	0.000	0.933	0.120
$\varepsilon = 0.1$	ML	0.062	−0.438	0.028	0.035	0.006	0.000	0.135
	Sample cor	0.052	−0.448	0.025	0.032	0.007	0.000	0.124
	Robust	0.480	−0.020	0.033	0.033	0.000	0.907	0.130
$\varepsilon = 0.15$	ML	−0.097	−0.597	0.027	0.035	0.007	0.000	0.135
	Sample cor	−0.076	−0.576	0.025	0.032	0.007	0.000	0.124
	Robust	0.472	−0.028	0.036	0.036	0.001	0.882	0.142
$\varepsilon = 0.2$	ML	−0.229	−0.729	0.026	0.033	0.008	0.000	0.131
	Sample cor	−0.176	−0.676	0.025	0.031	0.007	0.000	0.122
	Robust	0.450	−0.050	0.043	0.045	0.002	0.789	0.176
$\varepsilon = 0.3$	ML	−0.431	−0.931	0.022	0.030	0.008	0.000	0.118
	Sample cor	−0.322	−0.822	0.024	0.030	0.007	0.000	0.117
	Robust	0.413	−0.087	0.053	0.067	0.014	0.595	0.261
$\varepsilon = 0.4$	ML	−0.571	−1.071	0.019	0.026	0.008	0.000	0.103
	Sample cor	−0.424	−0.924	0.024	0.029	0.005	0.000	0.112
	Robust	0.357	−0.143	0.062	0.071	0.009	0.247	0.278
$\varepsilon = 0.49$	ML	−0.662	−1.162	0.016	0.024	0.008	0.000	0.095
	Sample cor	−0.492	−0.992	0.023	0.028	0.004	0.000	0.108

Table 1: Results for the robust estimator with $c = 0.6$, the MLE, and the Pearson sample correlation, for various contamination fractions across 5,000 simulated datasets. The true polychoric correlation coefficient is $\rho_* = 0.5$. We compute the average of point estimates $\widehat{\rho}_N$ of the polychoric correlation coefficient, the average bias ($\widehat{\rho}_N - \rho_*$), the standard deviation of the $\widehat{\rho}_N$ (SE; an approximation of the true standard error), the average standard error estimate $\widehat{\text{SE}}$, the corresponding average bias ($\widehat{\text{SE}} - \text{SE}$), confidence interval coverage with respect to the true ρ_* at nominal level 95%, and the average length of the respective confidence intervals.

for extremely large contamination fractions ($\varepsilon \geq 0.4$). We investigate this in more detail in Online Supplement D.1.

Variable	1	2	3	4	5
1	1.00				
2	0.56	1.00			
3	0.48	0.42	1.00		
4	0.40	0.35	0.30	1.00	
5	0.32	0.28	0.24	0.20	1.00

Table 2: Correlation matrix of $r = 5$ latent variables as in Foldnes and Grønneberg (2020). In line with the multivariate polychoric correlation model (e.g., Muthén, 1984), the latent variables are jointly normally distributed with mean zero and this correlation matrix as covariance matrix.

6.2. *Polychoric correlation matrix*

The goal of this simulation study is to robustly estimate a polychoric correlation matrix, that is, a matrix comprising of pairwise polychoric correlation coefficients. The simulation design is based on Foldnes and Grønneberg (2020).

Let there be r observed ordinal random variables and assume that a latent variable underlies each ordinal variable. In accordance with the multivariate polychoric model (e.g., Muthén, 1984), the latent variables are assumed to jointly follow an r-dimensional normal distribution with mean zero and a covariance matrix with unit diagonal elements so that the covariance matrix is a correlation matrix. Each individual latent variable is discretized to its corresponding observed ordinal variable akin to discretization process (3.1).

Following the five-dimensional simulation design in Foldnes and Grønneberg (2020), there are $r = 5$ ordinal variables with polychoric correlation matrix as in Table 2 such that the pairwise correlations vary from a low 0.2 to a moderate 0.56.[12] For all latent variables, the discretization thresholds are set to, in ascending order, $\Phi_1^{-1}(0.1) = -1.28$, $\Phi_1^{-1}(0.3) = -0.52$, $\Phi_1^{-1}(0.7) = 0.56$, and $\Phi_1^{-1}(0.9) = 1.28$, such that each ordinal variable can take five possible values. A visualization of the implied distribution of each ordinal variable can be found in Figure 5 in Foldnes and Grønneberg (2020).

As contamination distribution, we choose an r-dimensional Gumbel distribution comprising of mutually independent Gumbel marginal distributions, each with location and scale parameters equal to 0 and 3, respectively. To obtain ordinal observations, the unobserved realizations from this distribution are discretized via the same threshold values as realizations from the model (normal) distribution. As such, the uninformative ordinal observations generated by this contaminated distribution emulate the erratic behavior of a careless respondent. Unlike in the previous simulation design (Section 6.1), the uninformative responses are not concentrated around a few response options, but may occur in every response option.

For contamination fraction $\varepsilon \in \{0, 0.01, 0.05, 0.1, 0.15, 0.2, 0.3, 0.4, 0.49\}$, we sample $N = 1,000$ ordinal five-dimensional responses from this data generating process and use them to estimate the polychoric correlation matrix in Table 2 via our robust estimator (again with tuning constant $c = 0.6$) as well as the MLE.

Figure 5 visualizes the absolute average bias as well as confidence interval coverage at 95% nominal level (calculated over the 5,000 repetitions) of the robust estimator and the MLE for each pairwise polychoric correlation coefficient. As expected, when the model is correctly specified ($\varepsilon = 0$), both estimators coincide with accurate estimates. However, in the presence of contamination ($\varepsilon > 0$), the two estimators deviate. The MLE exhibits a notable bias for all correlation coefficients, which increases

[12] The correlation matrix in Table 2 has the additional interpretation of being the covariance matrix of a factor model for a single factor with loadings vector $(0.8, 0.7, 0.6, 0.5, 0.4)^\top$.

https://doi.org/10.1017/psy.2025.10066 Published online by Cambridge University Press

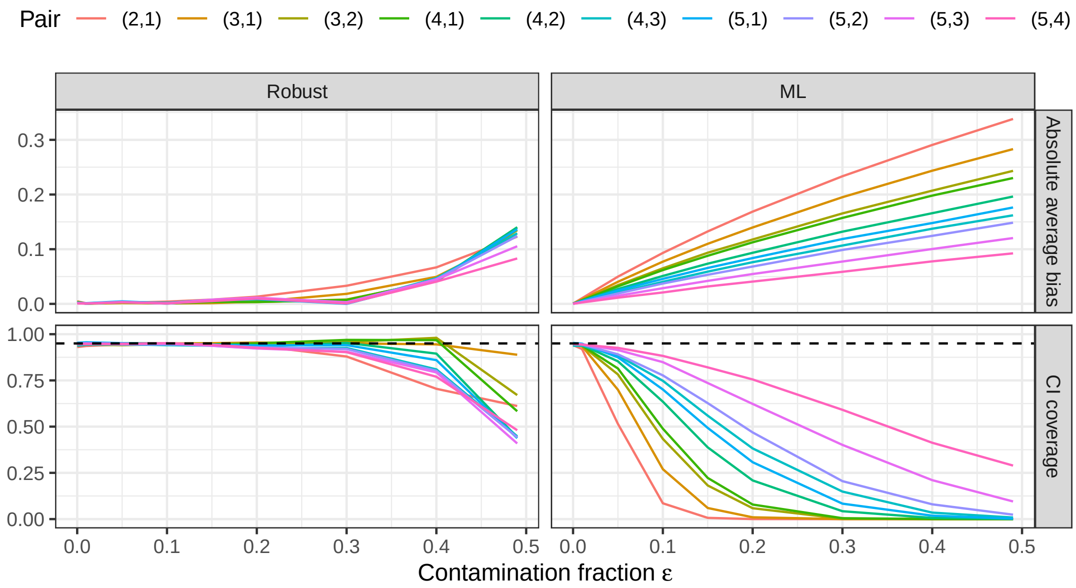

Figure 5: Absolute average bias (top) and confidence interval coverage (bottom) at nominal level 95% (dashed horizontal lines) of the robust estimator with $c = 0.6$ (left) and the MLE (right) for each unique pairwise polychoric correlation coefficient in the true correlation matrix (Table 2), expressed as a function of the contamination fraction ε (x-axis). Results are aggregated over 5,000 repetitions.

gradually with increasing contamination fraction. The magnitude of the bias tends to be larger for pairs with larger true correlation, such as 0.56 for pair $(2, 1)$, than for pairs with weaker true correlation, such as 0.20 for pair $(5, 4)$. In addition, coverage of the MLE drops quickly for many pairs and gradually for the remaining ones. Conversely, the robust estimator remains nearly unaffected for a broad range of contamination fractions for each correlation coefficient ($\varepsilon \leq 0.2$ or $\varepsilon \leq 0.3$ depending on the variable pair), with bias only somewhat increasing afterwards. Furthermore, its coverage remains close to 0.9 or higher even at the high contamination fraction of $\varepsilon = 0.3$. This reflects excellent performance of the proposed estimator with respect to robustness to uninformative responses. Online Supplement D.2 contains additional evaluations. In essence, the robust estimator yields accurate standard errors, but its confidence intervals tend to be wider than those of the MLE in the presence of contamination.

We stress that polychoric correlation matrices need not be positive definite (e.g., Mair, 2018, p. 22), although all estimated polychoric correlation matrices in this simulation turned out positive definite. If an estimated polychoric correlation matrix is not positive definite, one may opt to apply a smoothing procedure like in Yuan et al. (2011) or Bock et al. (1988).

6.3. *Discussion and additional simulation experiments*

The two simulation studies above demonstrate that already a small degree of partial misspecification due to uninformative responses, such as careless responses, can render the commonly employed MLE unreliable, while the proposed robust estimator retains good accuracy and coverage even in the presence

https://doi.org/10.1017/psy.2025.10066 Published online by Cambridge University Press

of a considerable number of uninformative responses. On the other hand, when the polychoric model is correctly specified, the MLE and the robust estimator produce equivalent estimates.

To further evaluate our robust estimator and investigate its limitations, we conduct additional simulation experiments in Online Supplement E.

The first experiment, described in Online Supplement E.2, is a generalization of the design in Section 6.1 with different *mean shifts* in the contamination distribution H. For small mean shifts, the proposed estimator does not improve upon the MLE, but the bias of both estimators remains reasonable. The larger the mean shift, the larger the detrimental effect on the MLE and the higher the robustness gain of our proposed estimator.

The second experiment, described in Online Supplement E.3, focuses on *correlation shifts* in the contamination distribution H. Specifically, the contamination distribution is the same as the true model distribution except for a sign-flipped correlation coefficient $-\rho_*$. For moderate correlation ρ_*, the proposed estimator does not improve upon the MLE due to substantial overlap between the true model distribution and the contamination. However, the gain in robustness increases substantially for higher correlation coefficients ρ_*. We expect the gain in robustness to increase for a higher number of response options and decrease for fewer response options. In the most extreme case of two dichotomous rating variables, no improvement can be expected.

7. Empirical application

We now demonstrate our proposed method on empirical data by using a subset of the 100 unipolar markers of the Big Five personality traits (Goldberg, 1992).

7.1. *Background and study design*

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 Five 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 would expect 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 Five traits in this way, namely *extroversion, neuroticism*, and *conscientiousness*.[13] 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/careless participants. Their model crucially relies on response inconsistencies to polar opposite adjectives and is designed to primarily detect careless straightlining responding. They find that careless responding is a sizable problem in their data. Their model finds evidence of straightliners, and the authors conclude that if unaccounted for, they can substantially

[13] 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. Furthermore, Arias et al. (2020) have made their data publicly available at https://osf.io/n6krb.

https://doi.org/10.1017/psy.2025.10066 Published online by Cambridge University Press

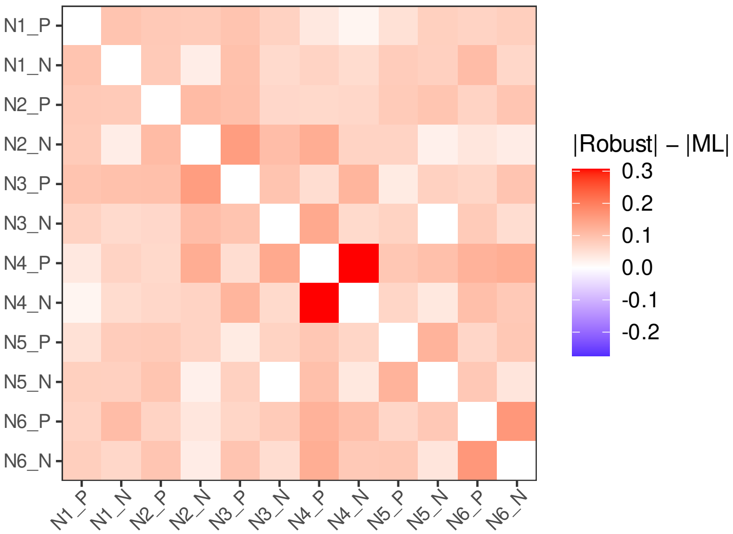

Figure 6: Difference between absolute estimates for the polychoric correlation coefficient of the robust estimator with $c = 0.6$ and the MLE for each item pair in the *neuroticism* scale, using the data of Arias et al. (2020). The items are "calm" (N1_P), "angry" (N1_N), "relaxed" (N2_P), "tense" (N2_N), "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 F.2 in Online Supplement F.

deteriorate the fit of theoretical models, produce spurious variance, and overall jeopardize the validity of research results.

Due to the suspected presence of careless respondents, we apply our proposed method to estimate the polychoric correlation coefficients between all $\binom{12}{2} = 66$ unique item pairs in the *neuroticism* scale to obtain an estimate of the scale's (polychoric) correlation matrix. The results of the remaining two scales are qualitatively similar and are reported in Online Supplement F. We estimate the polychoric correlation matrix via the MLE and via our proposed robust alternative with tuning parameter $c = 0.6$. As a robustness check, we further investigate the effect of the choice of c.

7.2. Results

Figure 6 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 the MLE. 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 the 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 leverage points, which drag negative correlation estimates towards zero, that is, they attenuate the estimated strength of correlation. Here, such

https://doi.org/10.1017/psy.2025.10066 Published online by Cambridge University Press

	Robust		ML		Sample cor	
Parameter	Estimate	$\widehat{\text{SE}}$	Estimate	$\widehat{\text{SE}}$	Estimate	$\widehat{\text{SE}}$
ρ	−0.925	0.062	−0.618	0.025	−0.562	0.031
a_1	−1.567	0.276	−1.373	0.061		
a_2	−0.560	0.203	−0.476	0.043		
a_3	0.110	0.187	0.121	0.042		
a_4	1.076	0.105	1.059	0.054		
b_1	−0.905	0.073	−0.857	0.049		
b_2	−0.040	0.091	−0.004	0.041		
b_3	0.640	0.364	0.608	0.044		
b_4	1.171	0.811	1.583	0.071		

Table 3: Parameter estimates and standard error estimates ($\widehat{\text{SE}}$) for the correlation between the *neuroticism* adjective pair "not envious" and "envious" in the data of Arias et al. (2020), using the robust estimator with $c = 0.6$, the MLE, and the Pearson sample correlation. Each adjective item has five answer categories. Note that the Pearson sample correlation does not model thresholds.

leverage points could be the responses of careless respondents 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 are present, 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 careless respondents who attenuate correlational estimates, we study in detail the adjective pair "not envious" and "envious", which featured the largest discrepancy between the ML estimate and the robust estimate in Figure 6, with an absolute difference of 0.314. The results of the two estimators and, for completeness, the sample correlation, are summarized in Table 3. The ML estimate of −0.618 and sample correlation estimate of −0.562 for the (polychoric) correlation coefficient seem remarkably weak considering that the two adjectives in question are polar opposites. In contrast, the robust correlation estimate is given by −0.925, which seems much more in line with what one would expect if all participants responded accurately and attentively (cf., Arias et al., 2020).

To study the potential presence of careless responses in each contingency table cell (x, y) for item pair "not envious" and "envious", Table 4 lists the PRs at the robust estimate, alongside the associated model probabilities and empirical relative frequencies.[14] A total of six cells have extremely large PR values of higher than 1,000, and, in addition, five cells have a PR of higher than 9, and one cell has a PR of higher than 4. Such PR values are far away from ideal value 0 at which the model would fit perfectly, thereby indicating a poor fit of the polychoric model for such responses. It stands out that all such poorly fitted cells are those whose responses might be viewed as inconsistent. Indeed, response cells $(x, y) \in \{(1, 1), (1, 2), (2, 1), (2, 2)\}$ indicate that a participant reports that *neither* "not envious" nor "envious" characterizes them accurately, which are mutually contradicting responses, while for response cells $(x, y) \in \{(4, 4), (4, 5), (5, 4), (5, 5)\}$ *both* adjectives characterize them accurately, which is again contradicting. As discussed previously, such responses might be due to careless responding. The robust estimator suggests that such responses cannot be fitted well by the polychoric model and subsequently

[14] A visualization of Table 4 is provided in Figure F.4 in the Online Supplement.

https://doi.org/10.1017/psy.2025.10066 Published online by Cambridge University Press

$X\backslash Y$	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

(a) 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

(b) Estimated response probabilities $p_{xy}(\widehat{\boldsymbol{\theta}}_N)$

$X\backslash Y$	1	2	3	4	5
1	$> 1,000$	$> 1,000$	10.81	0.14	−0.35
2	$> 1,000$	9.06	−0.20	−0.10	0.42
3	4.48	−0.35	−0.01	−0.20	76.11
4	−0.12	−0.08	−0.39	11.66	$> 1,000$
5	−0.11	−0.12	34.98	$> 1,000$	$> 1,000$

(c) Pearson residuals $\widehat{f}_N(x, y)/p_{xy}(\widehat{\boldsymbol{\theta}}_N) - 1$

Table 4: Empirical relative frequency (top), estimated response probability (center), and Pearson residual (PR) (bottom) of each response (x, y) for the item pair "not envious" (X) and "envious" (Y) in the measurements of Arias et al. (2020) of the *neuroticism* scale. Estimate $\widehat{\boldsymbol{\theta}}_N$ was computed via the robust estimator with tuning constant $c = 0.6$. The complete PR values are provided in Table F.5 in the Online Supplement.

downweighs their influence in the estimation procedure by mapping their Pearson residual with the linear part of the discrepancy function $\varphi(\cdot)$ in (5.3). Notably, also cells $(x, y) \in \{(1, 3), (3, 1), (3, 5), (5, 3)\}$ are 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 also indicative of careless 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).

As a robustness check on the role of tuning constant c, Figure 7 visualizes the point estimate $\widehat{\rho}_N$ of the polychoric correlation between the item pair "not envious" and "envious" for various values of c. The point estimate stays relatively constant for c between 0 and 0.75, with $\widehat{\rho}_N$ ranging between −0.95 and −0.92. Just after $c = 0.75$, $\widehat{\rho}_N$ abruptly jumps to about −0.85, before it stabilizes again and slowly transitions to the ML estimate of −0.62 (see Table 3) for $c \longrightarrow \infty$. Since $\widehat{\rho}_N$ very slowly approaches the value of the ML estimate, we only visualize choices of c up to 2 in Figure 7. The instability of the estimate around $c = 0.75$ suggests that c should be chosen below this value to avoid disproportionate influence of poorly fitting cells. For the broad range of $c \leq 0.75$, we obtain a robust finding of "not envious" and "envious" having a very strong negative correlation after accounting for likely careless responding.

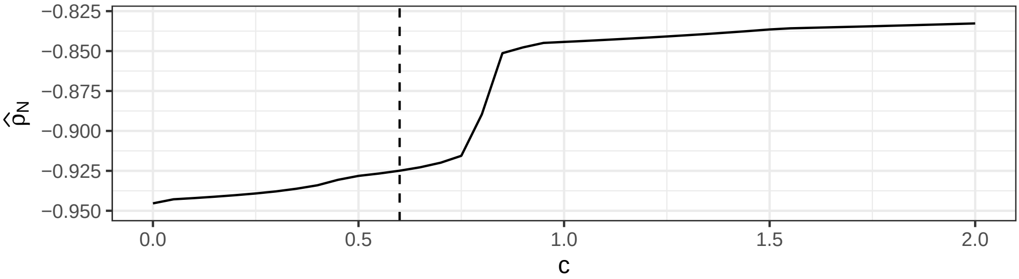

Figure 7: Estimates of the polychoric correlation between the items "not envious" and "envious" in the data of Arias et al. (2020) for various choices of the tuning constant c (x-axis). The dashed vertical line marks the default value of $c = 0.6$.

Overall, leveraging our robust estimator, we find evidence for the presence of careless respondents in the data of Arias et al. (2020). While they substantially affect the correlation 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 help identify them through unreasonably large PR values. On a final note, our findings from the empirical application align remarkably well with those of the simulation from Online Supplement E.3, therefore strengthening their validity. We provide a detailed discussion of this similarity in Online Supplement E.3.3.

8. Estimation under distributional misspecification

The simulation studies in Section 6 have been concerned with a situation in which the polychoric model is misspecified for a subset of a sample (partial misspecification). However, as outlined in Section 4.3, another misspecification framework of interest is that of *distributional misspecification* where the model is misspecified for the entire sample. Suppose that instead of a bivariate standard normal distribution, the latent variables (ξ, η) jointly follow an unknown and unspecified distribution G. In this framework, the object of interest is the population correlation between ξ and η under distribution G, that is, $\rho_G = \mathbb{C}\text{or}_G [\xi,\ \eta]$, rather than a polychoric correlation coefficient. Estimators for such situations where the population distribution G is nonnormal have been proposed by Lyhagen and Ornstein (2023), Jin and Yang-Wallentin (2017), and Roscino and Pollice (2006).

8.1. *Distributional vs. partial misspecification*

Huber and Ronchetti (2009, p. 4) note that, although conceptually distinct, robustness to distributional misspecification and partial misspecification are *"practically synonymous notions"*. Hence, despite distributional misspecification not being covered by the partial misspecification framework under which we study the theoretical properties of our proposed estimator, our estimator could still offer a gain in robustness compared to the MLE of the polychoric model under distributional misspecification. Specifically, the robust estimator may still be useful if the central part of the nonnormal distribution G is not too different from a standard bivariate normal distribution. Intuitively, if the difference between G and a standard normal distribution is mainly in the tails, G can be approximated by a contaminated distribution as in (4.1), with the standard normal distribution covering the central part and some contamination distribution H covering the tails. The polychoric MLE tries to treat influential observations

https://doi.org/10.1017/psy.2025.10066 Published online by Cambridge University Press

from the tails—which cannot be fitted well by the polychoric model—as if they were normally distributed, resulting in a possibly large estimation bias. In contrast, the robust estimator uses the normal distribution only for observations from the central part—which may fit the polychoric model well enough—and downweights observations from the tails. Thus, as long as such a contaminated normal distribution is a decent approximation of the nonnormal distribution G, the robust estimator should perform reasonably well. However, if G cannot be approximated by such a contaminated normal distribution, neither the polychoric MLE nor our estimator can be expected to perform well. Overall, though, our estimator could offer an improvement in terms of robustness to distributional misspecification.

In the following, we perform a simulation study to investigate the performance of our estimator when the polychoric model is distributionally misspecified.

8.2. Simulation study

To simulate ordinal variables that were generated by a nonnormal latent distribution G, we employ the VITA simulation method of Grønneberg and Foldnes (2017). For a pre-specified value of the population correlation ρ_G, the VITA method models the latent random vector (ξ, η) such that the individual variables ξ and η both possess standard normal marginal distributions with population correlation set equal to $\rho_G = \mathbb{C}\text{or}_G\,[\xi,\ \eta]$, but are *not* jointly normally distributed. Instead, their joint distribution G is equal to a pre-specified nonnormal copula distribution, such as the Clayton or Gumbel copula. Grønneberg and Foldnes (2017) show that discretizing such VITA latent variables yields ordinal observations that could not have been generated by a standard bivariate normal distribution, thereby ensuring proper violation of the latent normality assumption.

To investigate the robustness of our estimator to distributional misspecification, we use the VITA implementation in package `covsim` (Grønneberg et al., 2022) to generate draws of the latent variables (ξ, η) such that the latent variables are jointly distributed according to a Clayton copula G with population correlation $\rho_G \in \{0.9, 0.3\}$ (see Figure 8 for visualizations). Following the discretization process (3.1), we discretize both latent variables via discretization thresholds

$$\begin{aligned} a_1 &= -1.5, & a_2 &= 0, & a_3 &= 0.5, & a_4 &= 1, & &\text{and}\\ b_1 &= -1, & b_2 &= 1, & b_3 &= 1.5, & b_4 &= 2, \end{aligned}$$

such that both resulting ordinal variables have five response options each. We generate $N = 1,000$ ordinal responses according to this data generation process and compute across 5,000 repetitions the same estimators and performance measures as in the simulations in Section 6.

Figure 9 visualizes the bias of the robust estimator and the polychoric MLE under both Clayton copulas across the repetitions.[15] For correlation 0.9, the polychoric MLE exhibits a noteworthy bias, whereas the robust estimator remains accurate, albeit with a larger estimation variance as compared to most simulation configurations of partial misspecification (cf. Section 6). Conversely, for the weaker correlation 0.3, both estimators are fairly accurate with average biases of about −0.015. Table 5 contains additional performance measures regarding inference. For $\rho_G = 0.3$, the average standard error estimate of the robust method is accurate, but for $\rho_G = 0.9$, it notably overestimates. We therefore also computed the median of the standard error estimates in the latter case: at 0.017, it is fairly close to the true standard error of 0.014. It turns out that there is a small number of simulated datasets with a large majority of empty cells in the contingency table, resulting in numerical instability of the standard error estimates

[15] In 32 of the 5,000 repetitions for the Clayton copula, the robust estimator experienced numerical instability and subsequently did not converge to a solution. The polychoric MLE failed to converge 56 times. Consequently, these unconverged estimates are omitted from Figure 9. We explain this numerical issue in detail in Online Supplement B. Note that due to the shape of the Clayton copula with high correlation, stability issues with the robust estimator are in our experience amplified if the true thresholds are not well distributed over the domain of the distribution.

https://doi.org/10.1017/psy.2025.10066 Published online by Cambridge University Press

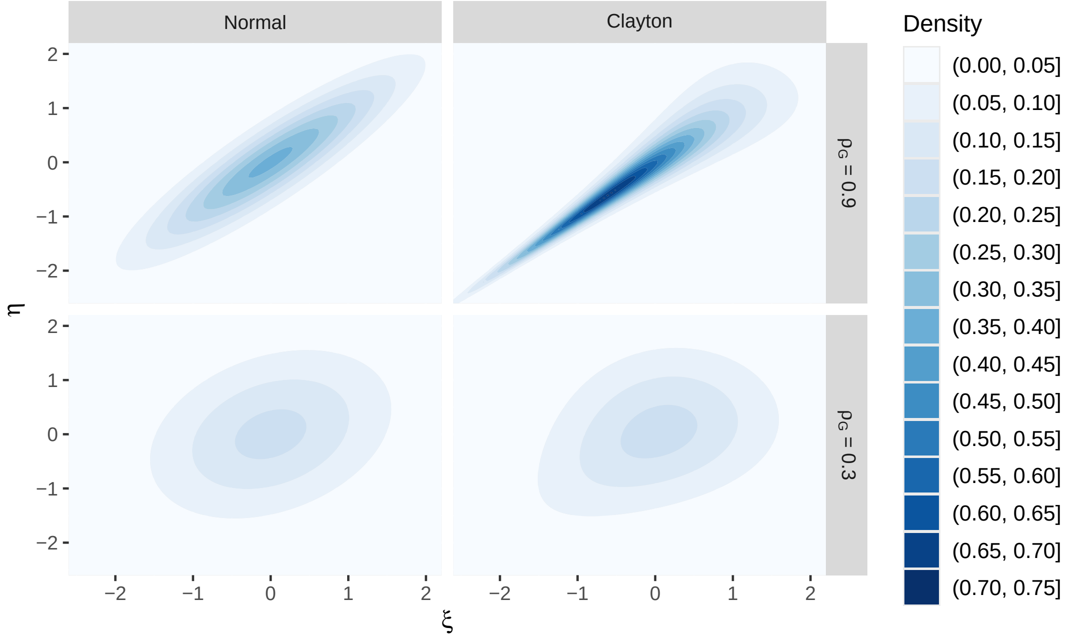

Figure 8: Bivariate density plots of the standard normal distribution (left) and Clayton copula with standard normal marginals (right), for population correlations 0.9 (top) and 0.3 (bottom).

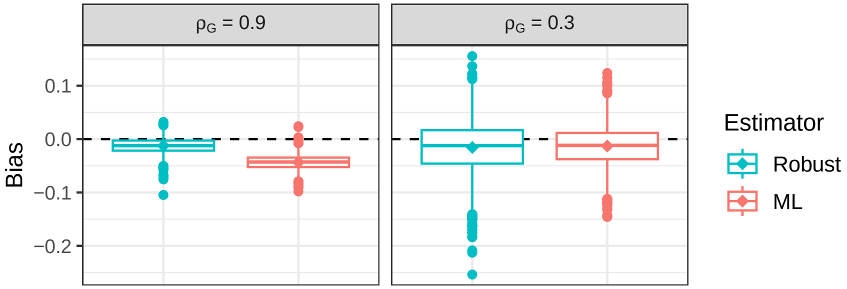

Figure 9: Boxplot visualization of the bias of the robust estimator and the polychoric MLE, $\widehat{\rho}_N - \rho_G$, under distributional misspecification via a Clayton copula with correlation $\rho_G = 0.9$ (left) and $\rho_G = 0.3$ (right), across 5,000 repetitions. Diamonds represent the respective average bias. The tuning constant of the robust estimator is set to $c = 0.6$.

and inflating their average. As discussed in Section 4.3, distributional misspecification is not covered by our partial misspecification framework, so it is not surprising that standard errors derived under partial misspecification are not always valid. Bootstrap inference may therefore be an attractive alternative. Nevertheless, unlike the polychoric MLE with coverage of only about 13% for $\rho_G = 0.9$, the robust estimator maintains high coverage of over 90%.

https://doi.org/10.1017/psy.2025.10066 Published online by Cambridge University Press

		Point estimate			Standard error		Confidence interval	
Correlation	Estimator	$\widehat{\rho}_N$	Bias	SE	$\widehat{\text{SE}}$	Bias	Coverage	Length
$\rho_G = 0.9$	Robust	0.888	−0.012	0.014	0.037	0.023	0.915	0.146
	ML	0.857	−0.043	0.013	0.016	0.002	0.132	0.061
$\rho_G = 0.3$	Robust	0.284	−0.016	0.047	0.052	0.006	0.938	0.205
	ML	0.287	−0.013	0.036	0.036	−0.001	0.932	0.141

Table 5: Results for the robust estimator with $c = 0.6$ and the polychoric MLE across 5,000 simulated datasets under distributional misspecification via a Clayton copula with true population correlation $\rho_G \in \{0.9, 0.3\}$. See Table 1 for explanatory notes on the performance measures.

These results suggest that at least for point estimation, the Clayton copula with correlation 0.9 might be reasonably approximable by a contaminated normal distribution where the normal distribution covers the center of the probability mass and some contamination distribution covers the tails (cf. Section 8.1). Indeed, Figure 8 indicates that the normal distribution and Clayton copula at correlation 0.9 seem to behave similarly in the center, but deviate from one another towards the tails. On the other hand, at correlation 0.3, the two densities do not appear to be drastically different from one another, which may explain why *both* the polychoric MLE and the robust estimator work reasonably well for such distributional misspecification. If the latent distribution is not too different from a normal distribution, then the polychoric model may offer a satisfactory fit despite being technically misspecified.

Overall, this simulation study demonstrates that in some cases of distributional misspecification, robustness can be gained with our estimator, compared to the polychoric MLE. However, it also demonstrates that there are cases of distributional misspecification for which the polychoric MLE still works quite well such that the robust estimator offers little gain. Nevertheless, the fact that in some situations robustness can be gained under distributional misspecification represents an overall gain in robustness.

9. Discussion and conclusion

We consider a situation where the polychoric correlation model is potentially misspecified for a subset of responses in a sample, that is, a set of uninformative observations not generated by a latent standard normal distribution. This model misspecification framework, called *partial misspecification* here, stems from the robust statistics literature, and a relevant special case is that of careless respondents in questionnaire studies. We demonstrate that maximum likelihood estimation is highly susceptible to the presence of such uninformative responses, resulting in possibly large estimation biases and low coverage of confidence intervals.

As a remedy, we propose an estimator based on the C-estimation framework of Welz (2024) that is designed to be robust to partial model misspecification. 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 possesses attractive statistical guarantees such as asymptotic normality. It furthermore allows to pinpoint the sources of potential model misspecification through the notion of Pearson residuals (PRs). Each possible response option is assigned a PR, where values substantially larger than the ideal value 0 imply that the response in question cannot be fitted well by the polychoric correlation model. In addition, the methodology proposed in this paper is implemented in

https://doi.org/10.1017/psy.2025.10066 Published online by Cambridge University Press

the free open source package robcat (Welz et al., 2025) for the statistical programming environment R and is publicly available at https://CRAN.R-project.org/package=robcat.

Although not covered by our partial misspecification framework, we also discuss how and when our estimator can offer a robustness gain (compared to the polychoric maximum likelihood estimator) when the polychoric model is misspecified for *all* observations, which has been a subject of interest in recent literature. In essence, there can be a robustness gain if the latent nonnormal distribution that generated the data can be reasonably well approximated by a contaminated normal distribution where the normal distribution reflects the central part and some unspecified contamination distribution reflects the tails.

We verify the enhanced robustness and theoretical properties of our robust estimator in simulation studies. Furthermore, we demonstrate the estimator's practical usefulness in an empirical application on a Big Five administration, where we find compelling evidence for the presence of careless respondents as a source of partial model misspecification.

However, our estimator depends on a user-specified choice of a tuning constant c, which governs a tradeoff between robustness and efficiency in case of misspecification. While simulation experiments suggest that the choice $c = 0.6$ provides a good tradeoff and estimates do not change considerably for a broad range of finite choices of c, a detailed investigation on this tuning constant needs to be carried out in future work. As an alternative, one could consider other discrepancy functions that do not depend on tuning constants, like the ones discussed in Welz (2024). As practical guidelines, we recommend always comparing robust estimates to that of ML and running the robust estimator for various choices of c, like in Figure 7. Doing so not only helps assess the estimates' stability, but also the severity of (partial) model misspecification. If the ML and robust estimates strongly differ, one may want to opt for choices of $c > 0$ not too far from 0 to achieve larger robustness gains.

A practical consideration of polychoric correlation is computing time. Our estimator simultaneously estimates all model parameters (i.e., correlation coefficient and thresholds) for robustness reasons, hence it is computationally more intensive than the non-robust two-stage approach of Olsson (1979). To alleviate the computational burden, our implementation in package robcat is written in fast and efficient C++ code. Furthermore, its default behavior first tries a fast unconstrained numerical optimization routine, which in our experience almost always suffices and executes in about half a second for five-point rating variables on a regular laptop. For estimating polychoric correlation matrices, package robcat supports parallel computing to keep computation time low. Since our method generalizes maximum likelihood and has the same time complexity, these functionalities also provide a fast implementation of the maximum likelihood estimator.

The methodology proposed in this paper suggests a number of extensions. For instance, one could use a robustly estimated polychoric correlation matrix in structural equation models to robustify such models and their fit indices against misspecification. Similar robustification could by achieved in, for instance but not limited to, principal component analyses, multidimensional scaling, or clustering. In addition, the theory of Welz (2024) may allow to pinpoint possible sources of model misspecification on the individual response level. That is, it may enable the derivation of statistically sound cutoff values for Pearson residuals in order to detect whether a given response can be fitted well by the polychoric correlation model. We leave these avenues to further research.

Overall, our novel robust estimator could open the door for a new line of research that is concerned with making the correlation-based analysis of rating data more reliable by reducing dependence on modeling assumptions.

Acknowledgments. We thank Patrick Groenen, Steffen Grønneberg, and Jonas Moss for valuable feedback. We further thank the associate editor and three anonymous reviewers, whose comments led to a substantially improved paper. This paper is based on Chapter 3 of M.W.'s PhD thesis.

https://doi.org/10.1017/psy.2025.10066 Published online by Cambridge University Press

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

Competing Interests. None.

Data Availability Statement. Replication files are publicly available on GitHub at https://github.com/mwelz/robust-polycor-replication. Online supplementary materials can be found at https://github.com/mwelz/robust-polycor-replication/blob/main/WelzMairAlfons2025_SupplementaryMaterials.pdf.

Ethical Standards. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

Author Contributions. M.W.: Conceptualization, Formal Analysis, Methodology, Software, Visualization, Writing - Original Draft, Writing - Review & Editing. P.M.: Conceptualization, Writing - Review & Editing. A.A.: Conceptualization, Funding Acquisition, Visualization, Writing - Review & Editing. All authors approved the final submitted draft.

References

Alfons, A., Ateş, N., & Groenen, P. (2022). A robust bootstrap test for mediation analysis. *Organizational Research Methods*, *25*(3), 591–617. https://doi.org/10.1177/1094428121999096

Alfons, A., & Schley, D. R. (2025). Robust mediation analysis: What we talk about when we talk about robustness [accepted for publication]. *WIREs Computational Statistics*. https://doi.org/10.1002/wics.70051

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*, *18*(2), e12941. https://doi.org/10.1111/spc3.12941

Arias, V. B., Garrido, L., Jenaro, C., Martinez-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

Asparouhov, T., & Muthén, B. (2016). Structural Equation Models and mixture models with continuous nonnormal skewed distributions. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(1), 1–19. https://doi.org/10.1080/10705511.2014.947375

Basu, A., Harris, I. R., Hjort, N. L., & Jones, M. C. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika*, *85*(3), 549–559. https://doi.org/10.1093/biomet/85.3.549

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied psychological measurement*, *12*(3), 261–280. https://doi.org/10.1177/014662168801200305

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://doi.org/doi/10.1037/pspp0000085

Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, *16*(5), 1190–1208. https://doi.org/10.1137/0916069

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd). Cengage Learning.

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

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

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

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. (2015). How general is the Vale-Maurelli simulation approach? *Psychometrika*, *80*(4), 1066–1083. https://doi.org/10.1007/s11336-014-9414-0

Foldnes, N., & Grønneberg, S. (2019). On identification and non-normal simulation in ordinal covariance and item response models. *Psychometrika*, *84*(4), 1000–1017. https://doi.org/10.1007/s11336-019-09688-z

Foldnes, N., & Grønneberg, S. (2020). Pernicious polychorics: The impact and detection of underlying non-normality. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(4), 525–543. https://doi.org/10.1080/10705511.2019.1673168

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

Fox, J. (2022). `polycor`*: Polychoric and polyserial correlations* [R package version 0.8-1]. https://CRAN.R-project.org/package=polycor

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. (2017). Covariance model simulation using regular vines. *Psychometrika*, *82*, 1035–1051. https://doi.org/10.1007/s11336-017-9569-6

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

Grønneberg, S., & Foldnes, N. (2022). Factor analyzing ordinal items requires substantive knowledge of response marginals [In press.]. *Psychological Methods*. https://doi.org/10.1037/met0000495

Grønneberg, S., Foldnes, N., & Marcoulides, K. M. (2022). `covsim`: An R package for simulating non-normal data for structural equation models using copulas. *Journal of Statistical Software*, *102*(3), 1–45. https://doi.org/10.18637/jss.v102.i03

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust statistics: The approach based on influence functions*. 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. (2015). 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. (2015). 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. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In L. M. Le Cam & J. Neyman (Eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (pp. 221–234, Vol. 5.1). University of California Press.

Huber, P. J., & Ronchetti, E. M. (2009). *Robust statistics* (2nd). Wiley. https://doi.org/10.1002/9780470434697

Itaya, Y., & Hayashi, K. (2025). Robust estimation of item parameters via divergence measures in Item Response Theory [arXiv:2502.10741]. https://doi.org/10.48550/arXiv.2502.10741

Jin, S., & Yang-Wallentin, F. (2017). Asymptotic robustness study of the polychoric correlation estimation. *Psychometrika*, *82*(1), 67–85. https://doi.org/10.1007/s11336-016-9512-2

Jöreskog, K. G. (1994). On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika*, *59*(3), 381–389. https://doi.org/10.1007/BF02296131

Kam, C. C. S., & Meyer, J. P. (2015). How careless responding and acquiescence response bias can influence construct dimensionality: The case of job satisfaction. *Organizational Research Methods*, *18*(3), 512–541. https://doi.org/10.1177/1094428115571894

Lau, S.-K. (1985). *The generalized least square estimation of polychoric correlation* [Doctoral dissertation, The Chinese University of Hong Kong].

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

Lindsay, B. G., Clogg, C. C., & Grego, J. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association*, *86*(413), 96–107. https://doi.org/10.1080/01621459.1991.10475008

https://doi.org/10.1017/psy.2025.10066 Published online by Cambridge University Press

Lyhagen, J., & Ornstein, P. (2023). Robust polychoric correlation. *Communications in Statistics - Theory and Methods*, *52*(10), 3241–3261. https://doi.org/10.1080/03610926.2021.1970770

Mair, P. (2018). *Modern psychometrics with R*. Springer. https://doi.org/10.1007/978-3-319-93177-7

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

Maydeu-Olivares, A. (2006). Limited information estimation and testing of discretized multivariate normal structural models. *Psychometrika*, *71*, 57–77. https://doi.org/10.1007/s11336-005-0773-4

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://doi.org/10.1037/a0028085

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*(1), 115–132. https://doi.org/10.1007/BF02294210

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, *7*(4), 308–313. https://doi.org/10.1093/comjnl/7.4.308

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

Patton, J. M., Cheng, Y., Hong, M., & Diao, Q. (2019). Detection and treatment of careless responses to improve item parameter estimation. *Journal of Educational and Behavioral Statistics*, *44*(3), 309–341. https://doi.org/10.3102/1076998618825116

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

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Revelle, W. (2024). `psych`*: Procedures for psychological, psychometric, and personality research* [R package version 2.4.6]. Northwestern University. Evanston, Illinois. https://CRAN.R-project.org/package=psych

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

Roscino, A., & Pollice, A. (2006). A generalization of the polychoric correlation coefficient. In S. Zani, A. Cerioli, M. Riani, & M. Vichi (Eds.), *Data analysis, classification and the forward search* (pp. 135–142). https://doi.org/10.1007/3-540-35978-8_16

Rosseel, Y. (2012). `lavaan`: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

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

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research*. Sage.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*(4), 507–514. https://doi.org/10.1007/BF02296192

Sripriya, T. P., Gallo, M., & Srinivasan, M. R. (2020). Detection of outlying cells in contingency tables using model based diagnostics. *Austrian Journal of Statistics*, *49*(5), 59–67. https://doi.org/10.17713/ajs.v49i5.938

Steinmann, I., Strietholt, R., & Braeken, J. (2022). A constrained factor mixture analysis model for consistent and inconsistent respondents to mixed-worded scales. *Psychological Methods*, *27*(4), 667–702. https://doi.org/10.1037/met0000392

Ulitzsch, E., Pohl, S., Khorramdel, L., Kroehne, U., & von Davier, M. (2022). A response-time-based latent response mixture model for identifying and modeling careless and insufficient effort responding in survey data. *Psychometrika*, *87*(2), 593–619. https://doi.org/10.1007/s11336-021-09817-7

Ulitzsch, E., Yildirim-Erbasli, S. N., Gorgun, G., & Bulut, O. (2022). An explanatory mixture IRT model for careless and insufficient effort responding in self-report measures. *British Journal of Mathematical and Statistical Psychology*, *75*(3), 668–698. https://doi.org/10.1111/bmsp.12272

Vale, C. D., & Maurelli, V. A. (1983). Simulating multivariate nonnormal distributions. *Psychometrika*, *48*, 465–471. https://doi.org/10.1007/BF02293687

https://doi.org/10.1017/psy.2025.10066 Published online by Cambridge University Press

Van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge University Press.

Van Laar, S., & Braeken, J. (2022). Random responders in the TIMSS 2015 student questionnaire: A threat to validity? *Journal of Educational Measurement*, *59*(4), 470–501. https://doi.org/10.1111/jedm.12317

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. (2024). Robust estimation and inference for categorical data [arXiv:2403.11954]. https://doi.org/10.48550/arXiv.2403.11954

Welz, M., Alfons, A., & Mair, P. (2025). `robcat`*: Robust categorical data analysis* [R package version 0.1.0]. https://CRAN.R-project.org/package=robcat

Welz, M., Archimbaud, A., & Alfons, A. (2024). How much carelessness is too much? quantifying the impact of careless responding [PsyArXiv:8fj6p]. https://doi.org/10.31234/osf.io/8fj6p

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*(1), 1–26. https://doi.org/10.2307/1912526

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://doi.org/10.1007/s10862-005-9004-7

Yuan, K.-H., Bentler, P. M., & Chan, W. (2004). Structural Equation Modeling with heavy tailed distributions. *Psychometrika*, *69*(3), 421–436. https://doi.org/10.1007/BF02295644

Yuan, K.-H., Wu, R., & Bentler, P. M. (2011). Ridge structural equation modelling with correlation matrices for ordinal and continuous data. *British Journal of Mathematical and Statistical Psychology*, *64*(1), 107–133. https://doi.org/10.1348/000711010X497442

Zhang, P., Liu, B., & Pan, J. (2024). Iteratively reweighted least squares method for estimating polyserial and polychoric correlation coefficients. *Journal of Computational and Graphical Statistics*, *33*(1), 316–328. https://doi.org/10.1080/10618600.2023.2257251

https://doi.org/10.1017/psy.2025.10066 Published online by Cambridge University Press