Quantifying effects of rating-scale response bias: Theory and implications for survey design

Max Welz
welz@ese.eur.nl

Aurore Archimbaud
archimbaud@ese.eur.nl

Andreas Alfons
alfons@ese.eur.nl

Econometric Institute
Erasmus School of Economics

June 6, 2023

Abstract

Responses to rating-scale items are often plagued by biases stemming from content-responsive faking (such as malingering or socially desirable responding) or content nonresponsitivity (particularly careless responding). While there is consensus that response biases can jeopardize the validity of survey measures through a variety of psychometric issues, their exact effects are yet to be statistically quantified. Leveraging robustness theory, we study the statistical properties of response biases in survey data. In particular, we derive bias curves and breakdown values of survey measures, with a focus on correlational measures due to their key role in factor analyses and structural equation models. Furthermore, we study how the adverse effects of response biases can be mitigated by survey design, for instance through the number of answer categories, number of items in a measure, and construct reliability. We find that already low prevalence of response biases can render survey measures fundamentally invalid. In addition, we show how comparatively short survey measures with a balanced number of negatively-worded items can enhance the robustness of survey measures against response biases. Furthermore, we provide freely available software in R for computation and visualization of bias curves in survey measures.

1 Introduction

Surveys are ubiquitous in scientific fields such as health, psychology, economics, and marketing. However, survey participants may not respond truthfully or accurately, for instance by inattentive responding (e.g. carelessness) or intentional misrepresentation of oneself (e.g. socially desirable responding). Such inaccurate responses are called *response biases*. While there is widespread empirical evidence that response biases can in some cases be a major threat to the validity of survey-based research (e.g. Huang et al., 2015; Meade & Craig, 2012; Paulhus et al., 1995; Topping & O'Gorman, 1997), their exact effects on survey measures are yet to be statistically quantified.

Leveraging statistical robustness theory, we study the theoretical properties of response biases in eating-scale survey data. Since many fundamental exploratory and confirmatory analyses of survey measures are either explicitly or implicitly correlational—such as validity, reliability, or factor structure—we focus on the statistical effects of response bias on Pearson's correlation measure, and additionally on mean and variance. In particular, for each of these three estimators, we derive bias curves that estimate how much a given level and type of response bias distorts the estimator of interest when evaluated at survey measures of given characteristics.

We find that questionnaire designs that are intended to yield high quality of measurement through high reliability, many answer categories, or many items are most prone to adverse effects of response bias. In particular, in lengthy questionnaires, already a low prevalence of careless respondents of about 5% can suffice to reverse research findings. We therefore argue that instead of maximizing measurement reliability in highly idealized scenarios, one should use measurements that are reasonably reliable across a large variety of potentially very noisy scenarios. Specifically, shorter scales are preferred because they have statistical advantages over lengthier scales when carelessness is present: When response bias is present, shorter scales yield reasonably accurate measurement, in contrast to lengthier scales. In addition, shorter scales offer practical advantages such as lower costs, higher convenience, and higher statistical power as a result of hiring more study subjects due to saved costs.

2 Literature

2.1 Response Bias as a Construct

McGrath et al. (2010) defines response bias as a *"consistent tendency to respond inaccurately to a substantive indicator, resulting in systematic error in prediction"*. Nichols et al. (1989) distinguish between two general types of response bias, *content nonresponsitivity* (CNR) and *content-responsive faking* (CRF).

CNR is defined as responding without regard to item content (Meade & Craig, 2012). Examples include careless responding (e.g. Meade & Craig, 2012), participant inattention (Maniaci & Rogge, 2014), and protocol invalidity (Johnson, 2005). CNR is typically charac-

terized by (near-)random responding (*response inconsistency*; Ward & Meade, 2022), such as choosing answer categories completely at random or a tendency to consistently choose extreme answer categories, and responding according to deterministic patterns, like straightlining or content-independent patterns like 1-2-3-1-2-3 (*response invariability*; Ward & Meade, 2022), and impossibly fast response times (e.g. Bowling et al., 2021b). There is a broad literature on causes of content-independent responding and the psychology behind it. One prominent cause is inattention, which may be due to fatigue or boredom in lengthy surveys (Bowling et al., 2021a; Gibson & Bowling, 2020; Galesic & Bosnjak, 2009). The phenomenon of inattention is studied extensively in behavioral economics (e.g., DellaVigna, 2009; Gabaix, 2019, and references therein). Another cause of CNR is confusion (e.g., Huang et al., 2012; Ward & Pond, 2015; Ward et al., 2017), for instance when a participant misunderstands an item due to ambiguous item wording or insufficient reading comprehension (Nichols et al., 1989).

Nichols et al.'s (1989) second category of response bias, CRF, refers to strategic response behavior with the goal of intentionally misrepresenting oneself. Examples include malingering ("faking bad"; e.g. Furnham & Henderson, 1982), denying problems to create an impression of being "normal" ("faking good"; e.g. Furnham & Henderson, 1982), socially desirable responding (Paulhus, 1984), impression management (Schlenker, 1980), and self-deception (Paulhus, 1986). Successful CRF requires careful, attentive and systematic content-dependent responding (cf. Holden & Book, 2011; Johnson & Hogan, 2006; Paulhus, 1993), whereas CNR is characterized by content-independent responding.

A third type of response bias that neither seems to match CRF nor CNR stems from item order, where responses depend on previously given responses (Dillman et al., 2014). For instance, respondents may use similar thought processes for answering two items that are seemingly related. We refer to Section 5.2 in Stantcheva (2022) for a detailed review on such item order effects.

2.2 Prevalence, Effects, and Identification of Response Biases

2.2.1 Prevalence

Content-independent responding due to carelessness is widely prevalent (Bowling et al., 2016; Curran, 2016; Meade & Craig, 2012; Ward & Pond, 2015; Ward et al., 2017) and suspected by Ward & Meade (2022) to be present in all survey data. Although estimates of the exact prevalence vary substantially (e.g., Arthur et al., 2021; Ward & Meade, 2022, and references therein), Curran (2016); Huang et al. (2012, 2015); Meade & Craig (2012) estimate prevalence to be generally between 10–15% of survey participants.

Studies on the prevalence of faking are generally focused on high-stake contexts, such as employment decisions. For instance, Griffith et al. (2007) estimate that between 30% and 50% of job applicants attempt to generate an inadequately positive impression of themselves. Griffith & Converse (2011) conclude from a literature review that a *"substantial portion of* [job] *applicants fakes personality measures"* and estimate the proportion of fakers

at roughly 30%.[1] In general, faking good can be expected in situations in where survey participants believe they can gain something of personal value from faking (Ellingson, 2011), particularly when stakes are high (Cao & Drasgow, 2019; Paulhus, 2002). Nevertheless, the prevalence, degree, reason, and type of faking may vary across populations (Zickar et al., 2004) and on the individual participant level (Ellingson, 2011; Griffith & McDaniel, 2006).

2.2.2 Effects

There is substantial empirical evidence that content-independent responding can jeopardize the validity of survey measures through, for instance, lower scale reliability, spurious variability, worse model fit, or type I or type II errors in hypothesis testing (Arias et al., 2020; Huang et al., 2015; Kam & Meyer, 2015; McGrath et al., 2010; Woods, 2006), and already a low presence of 5–10% of participants who respond carelessly or inattentively can be problematic for validity (Arias et al., 2020; Credé, 2010; Schmitt & Stults, 1985; Woods, 2006).

Studies on the effects of faking on the validity of survey measures have mixed conclusions, ranging from faking being a serious threat (Dunnette et al., 1962; Komar et al., 2008; Paulhus et al., 1995; Topping & O'Gorman, 1997) over being generally not a serious issue (Ellingson et al., 2001; Hough et al., 1990) to downright being a *"non-issue"* (Hogan & Hogan, 2007, in a reference to Hogan et al., 2007). Similarly, for correlational and factor structures, there is no consensus on whether faking is serious issue (e.g., Topping & O'Gorman, 1997; Schmit & Ryan, 1993) or not (e.g., Marshall et al., 2005; Smith & Ellingson, 2002). Holden & Book (2011) suspect that this inconclusiveness might be due to differences in assessment contexts, associated base rates for faking, criteria used, samples, operationalizations of faking, and analytical methods. In addition, effects may differ between high-stakes and low-stakes testing situations (White et al., 2008).

2.2.3 Detection

One can generally distinguish between a priori and post hoc methods for the detection of biased responses. Post hoc methods are based on statistical analyses on given responses and are popular in the detection of careless responding. Such methods include consistency indicators like psychometric antonyms and psychometric synonyms (Meade & Craig, 2012), longstring indices (Johnson, 2005), multivariate outlier analyses (e.g., Curran, 2016), threshold values for response times (Bowling et al., 2021b), and more recently machine learning based methods (Schroeders et al., 2022; Welz & Alfons, 2023).[2] A priori methods are based on specific items or even entire scales as part of the questionnaire. Popular items for the detection of careless responding are self-report items, instructed items, or attention checks

[1] We refer to Table 3.1 in Griffith & Converse (2011) for an overview of estimates on faking prevalence in employment decisions.

[2] Evaluating overviews of methods for the detection of carelessness are provided in Arthur et al. (2021), Curran (2016), DeSimone et al. (2015), and Ward & Pond (2015).

(Meade & Craig, 2012). Faking is typically detected via scales designed for measuring specific manifestations of faking.

3 Methodology

3.1 Robust Statistics

The field of robust statistics is primarily concerned with quantifying the effects of corrupted data on statistical estimators, and develop estimators that are less susceptible to the adverse effects of corrupted data.[3] An observation is said to be "corrupted" if it follows a different distribution than the distribution that the data are intended to be sampled from. For instance, if sampled data are intended to be normally distributed, either by assumption, modeling choices, or experimental design, an observation that is instead sampled from some heavy-tailed distribution is seen as corrupted. Hence, one may view corrupted data as a form of sampling error.

Corrupted data is typically assumed to follow a contaminated distribution F_ε, which is defined as

$$F_\varepsilon = (1-\varepsilon)F + \varepsilon H, \tag{1}$$

where F is the *model distribution* we intend to sample from, H is the *contaminating distribution* that causes sampling error, and constant $\varepsilon \in [0,1]$ is the contamination fraction. For distribution F_ε, we sample with probability $(1-\varepsilon)$ from the model distribution, and with probability ε from the contaminating distribution H. The contamination distribution model (1) is due to seminal work by Huber (1964), and the contaminating distribution H is traditionally seen as some unspecified outlier-generating distribution. One therefore often works with classes of contamination distributions rather than an individual contamination distribution. For model distribution F and contamination fraction ε, such a class is defined by

$$\mathcal{F}_\varepsilon = \left\{F_\varepsilon : F_\varepsilon = (1-\varepsilon)F + \varepsilon H \text{ for any distribution } H\right\}.$$

In robust statistics, one is often interested in how some statistical estimator T is affected by data sampled from a certain distribution, such as the contamination distribution F_ε. To study this, we view the estimator as a *statistical functional* of data-generating distribution F, resulting in a functional $T(F)$. For example, consider the example of the expectation of F for some random variable X distributed according to F, being

$$T(F) = \mathbb{E}_F\left[X\right].$$

Viewing an estimator as a functional enables us to simultaneously study the population version and empirical version of the estimator. Indeed, in the example of the expectation, let

[3] For detailed treatments of robust statistics, we refer the interested reader to classic textbooks by Hampel et al. (1986); Huber & Ronchetti (2009), and Maronna et al. (2018). As an example of a robust estimation method, Alfons et al. (2022) develop an estimator for mediation analyses that can resist the adverse effects of outliers.

$X_1, \ldots, X_N$ be an N-sized random sample from F, and define by $\widehat{F}_N(x) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{X_i \leq x\}$ the empirical distribution function obtained from the sample. then, evaluating expectation T at empirical distribution $\widehat{F}_N$ yields the sample mean

$$T(\widehat{F}_N) = \frac{1}{N}\sum_{i=1}^{N} X_i.$$

A derivation of this statement is given Appendix A.

One can use the functional representation of an estimator to study the estimator's theoretical and empirical behavior when evaluated at a contamination distribution F_ε (or an empirical version thereof), for a fixed contamination fraction $\varepsilon \in [0, 1]$. For example, a popular measure of robustness of an estimator is its *maximum bias curve*. The maximum bias curve of an estimator is defined by the maximum discrepancy between the estimator evaluated at an uncontaminated distribution F and any contaminated distribution $F_\varepsilon \in \mathcal{F}_\varepsilon$, for a fixed fraction of contamination ε. Hence, the maximum bias curve measures the maximum estimation bias caused by having a fraction ε of samples being corrupted. For example, the maximum bias curve of the expectation is unbounded for a standard normal model distribution, meaning that even tiny amounts of contamination can lead to infinite bias. A closely related measure of robustness is the *breakdown value*, which is the minimum contamination fraction required to completely destroy an estimator. What is meant by "completely destroying" depends on the nature of the estimator. In many classical estimators, it means sending the estimator's maximum bias to infinity, like in the example of the expectation at the standard normal distribution. We later return to the notations of maximum bias curve and breakdown value.

3.2 Setup

In this paper, we are interested in the the robustness properties of the Pearson correlation coefficient between two discrete ordinal random variables, X and Y. For instance, think about X and Y as responses to two Likert-type items in a Big 5 personality instrument. A functional representation of Pearson's correlation coefficient between X and Y, evaluated at a bivariate distribution F, is given by

$$T(F) := \operatorname{Cor}_F[X, Y] = \frac{\operatorname{Cov}_F[X, Y]}{\sqrt{\operatorname{Var}_F[X]\operatorname{Var}_F[Y]}} = \frac{\mathbb{E}_F[XY] - \mathbb{E}_F[X]\,\mathbb{E}_F[Y]}{\sqrt{\mathbb{E}_F[X^2] - \mathbb{E}_F[X]^2}\sqrt{\mathbb{E}_F[Y^2] - \mathbb{E}_F[Y]^2}}, \tag{2}$$

see Croux & Dehon (2010). To avoid cumbersome notation, $\mathbb{E}_F[X]$ represents the marginal expectation of X at a bivariate distribution F, and analogously for Y. For an N-sized bivariate sample $(X_1, Y_1), \ldots, (X_N, Y_N)$ from F, evaluating functional T at the sample's

empirical distribution $\widehat{F}_N$ recovers the well-known empirical version of Pearson correlation:

$$T\big(\widehat{F}_N\big) = \mathrm{Cor}_{\widehat{F}_N}[X,Y] = \frac{\frac{1}{N}\sum_{i=1}^N X_i Y_i - \overline{X}_N \overline{Y}_N}{\sqrt{\frac{1}{N}\sum_{i=1}^N X_i^2 - \overline{X}_N^2}\sqrt{\frac{1}{N}\sum_{i=1}^N Y_i^2 - \overline{Y}_N^2}},$$

where $\overline{X}_N = \frac{1}{N}\sum_{i=1}^N X_i$ and $\overline{Y}_N = \frac{1}{N}\sum_{i=1}^N Y_i$ denote the two marginal sample means. A derivation of this expression is provided in Appendix A.

Throughout this paper, we denote by F the true uncontaminated model distribution of (X,Y). We denote the true Pearson correlation between X and Y at their distribution F by $T(F) = \rho \in [-1,1]$. In addition, we assume that X and Y have support $\{-M_X, -(M_X - 1), \ldots, M_X - 1, M_X\}$ and $\{-M_Y, -(M_Y - 1), \ldots, M_Y - 1, M_Y\}$, respectively, for some finite $M_X, M_Y > 0$. This assumption is without loss of generality, as we demonstrate in the following. Suppose that a discrete ordinal variable X' has support $S' = \{1, 2, \ldots, K\}$. We can rescale the support be deducting the central answer category, C, from each element in S', resulting in a random variable X with support $S = \{1 - C, \ldots, K - C\}$, and now set $M_X = K - C$. For example, items with five Likert-type answer categories have original support region $S' = \{1, 2, 3, 4, 5\}$, which can be rescaled to $S = \{-2, -1, 0, 1, 2\}$ by deducting the central answer category, being 3 in this case.. In this case, the rescaled variable X has maximum value $M_X = 2$. Note that rescaling a variable's support region may affect its moments, which we discuss in detail in Appendix C.

3.3 Main Results

We are now ready to define the maximum bias curve of Pearson's correlation measure. Following Raymaekers & Rousseeuw (2021), we distinguish between a maximum upward and a maximum downward bias in Definition 1.

Definition 1 (Maximum bias curve)**.** At model distribution F and a contamination fraction $\varepsilon \in [0,1]$, the maximum upward and downward bias of Pearson correlation measure T are respectively defined as

$$B^+(\varepsilon, T, F) = \sup_{G \in \mathcal{F}_\varepsilon} \Big\{ T(G) - T(F) \Big\}, \quad \text{and}$$
$$B^-(\varepsilon, T, F) = \inf_{G \in \mathcal{F}_\varepsilon} \Big\{ T(G) - T(F) \Big\},$$

where $\mathcal{F}_\varepsilon = \big\{ G : G = (1-\varepsilon)F + \varepsilon H \text{ for any distribtion } H \text{ with same support as } F \big\}$.

The maximum upward bias in Definition 1 takes values in $[0,2]$, and is maximized by having contamination push the true correlation at F upwards towards value $+1$. Conversely, the downward bias takes values in $[-2,0]$ and is maximized by pushing correlation downwards towards value -1. Therefore, maximum upward bias and downward bias are intended for situations where the true correlation at F is negative and positive, respectively.

In Definition 1, it is not a necessity to restrict the contamination distributions H in $\mathcal{F}_\varepsilon$ to have the same support as the model distribution F. For instance, one could even consider continuous contamination distributions. However, because we are working with discrete data, we do not consider it meaningful to compare two distributions with different support regions.

It is useful to define the following functions. For fixed contamination fraction $\varepsilon \in [0, 0.5]$ and rating-scale variables X and Y that jointly follow model distribution F, let

$$\begin{aligned} m_\varepsilon(X,Y) &= (1-\varepsilon)\Big(\rho\sqrt{\mathrm{Var}_F\left[X\right]\mathrm{Var}_F\left[Y\right]} + \mathbb{E}_F\left[X\right]\mathbb{E}_F\left[Y\right]\Big) - (1-\varepsilon)^2\mathbb{E}_F\left[X\right]\mathbb{E}_F\left[Y\right], \\ m_\varepsilon(X) &= (1-\varepsilon)\big(\mathrm{Var}_F\left[X\right] + \mathbb{E}_F\left[X\right]^2\big) - (1-\varepsilon)^2\mathbb{E}_F\left[X\right]^2, \quad \text{and,} \\ m_\varepsilon(Y) &= (1-\varepsilon)\big(\mathrm{Var}_F\left[Y\right] + \mathbb{E}_F\left[Y\right]^2\big) - (1-\varepsilon)^2\mathbb{E}_F\left[Y\right]^2, \end{aligned} \tag{3}$$

where $\rho = T(F)$ is the Pearson correlation between X and Y at F.

The following proposition derives bias curves for a given type of contamination. The proof of this proposition and all other mathematical statements in this paper are given in the appendix.

Proposition 1. *For fixed contamination fraction $\varepsilon \in [0, 0.5]$ and contaminated distribution $G = (1-\varepsilon)F + \varepsilon H \in \mathcal{F}_\varepsilon$, the bias of Pearson correlation measure T at model distribution F is given by*

$$T(G) - T(F) = \frac{\mathrm{Cov}_G\left[X,Y\right]}{\sqrt{\mathrm{Var}_G\left[X\right]\mathrm{Var}_G\left[Y\right]}} - \rho,$$

where $\rho = T(F)$, and

$$\begin{aligned} &\mathrm{Cov}_G\left[X,Y\right] = m_\varepsilon(X,Y) + \\ &\quad \varepsilon\Big(-(1-\varepsilon)\mathbb{E}_F\left[X\right]\mathbb{E}_H\left[Y\right] - (1-\varepsilon)\mathbb{E}_F\left[Y\right]\mathbb{E}_H\left[X\right] + \mathbb{E}_H\left[XY\right] - \varepsilon\mathbb{E}_H\left[X\right]\mathbb{E}_H\left[Y\right]\Big), \end{aligned}$$

as well as

$$\begin{aligned} \mathrm{Var}_G\left[X\right] &= m_\varepsilon(X) + \varepsilon\Big(-2(1-\varepsilon)\mathbb{E}_F\left[X\right]\mathbb{E}_H\left[X\right] + \mathbb{E}_H\left[X^2\right] - \varepsilon\mathbb{E}_H\left[X\right]^2\Big), \quad \textit{and,} \\ \mathrm{Var}_G\left[Y\right] &= m_\varepsilon(Y) + \varepsilon\Big(-2(1-\varepsilon)\mathbb{E}_F\left[Y\right]\mathbb{E}_H\left[Y\right] + \mathbb{E}_H\left[Y^2\right] - \varepsilon\mathbb{E}_H\left[Y\right]^2\Big). \end{aligned}$$

Theorem 1. *Let X and Y be discrete ordinal random variables with support regions $\{-M_X, \ldots, M_X\}$ and $\{-M_Y, \ldots, M_Y\}$, respectively. For fixed contamination fraction $\varepsilon \in [0, 0.5]$, the maximum upward bias of Pearson correlation measure T at model distribution F is given by*

$$B^+(\varepsilon, T, F) = \frac{m_\varepsilon(X,Y) + \varepsilon M_X M_Y}{\sqrt{m_{F,\varepsilon}(X) + \varepsilon M_X^2}\sqrt{m_{F,\varepsilon}(Y) + \varepsilon M_Y^2}} - \rho,$$

and the maximum downward bias is

$$B^{-}(\varepsilon, T, F) = \frac{m_{\varepsilon}(X, Y) - \varepsilon M_X M_Y}{\sqrt{m_{F,\varepsilon}(X) + \varepsilon M_X^2}\sqrt{m_{F,\varepsilon}(Y) + \varepsilon M_Y^2}} - \rho,$$

where $\rho = T(F)$.

We stress that the denominator in either maximum bias curve in Theorem 1 is strictly positive; a proof of this claim is provided in a lemma in Appendix B.1.

This result of Theorem 1 is similar to Proposition 2 in Raymaekers & Rousseeuw (2021). However, our result does not require a symmetric and unimodal density of model distribution F, which is due to the boundedness of X and Y, nor do we require that X and Y are zero-mean and have the same variance at F.

We now turn to the breakdown value. Capéraà & Guillem (1997) define the breakdown value of a correlation estimator as the smallest contamination fraction required to render two perfectly correlated variables negatively correlated. Specifically,

Definition 2 (Breakdown value)**.** Let F be a bivariate distribution of (X, Y), where $X = Y$, and T be Pearson's correlation measure. The breakdown value of T is defined as

$$\varepsilon^*(T) = \inf\left\{\varepsilon > 0 : \inf_{G \in \mathcal{F}_\varepsilon} T(G) \leq 0\right\}.$$

The closer a breakdown value is to zero, the fewer contaminated observations are required to break Pearson's correlation. Consequently, a high breakdown value is desirable.

Proposition 2. *For a rating-scale variable* X *with support* $\{-M_X, \ldots, M_X\}$*, the breakdown value of Pearson correlation measure* T *at model distribution* F *is*

$$\varepsilon^*(T) = \begin{cases} \frac{\mathbb{E}_F[X]^2 - \mathrm{Var}_F[X] - M_X^2 + \sqrt{\left(\mathrm{Var}_F[X] - \mathbb{E}_F[X]^2 + M_X^2\right)^2 + 4\mathbb{E}_F[X]^2 \mathrm{Var}_F[X]}}{2\mathbb{E}_F[X]^2} & \text{if } \mathbb{E}_F[X] \neq 0, \\ \frac{\mathrm{Var}_F[X]}{\mathrm{Var}_F[X] + M_X^2} & \text{if } \mathbb{E}_F[X] = 0. \end{cases}$$

The breakdown value for the case $\mathbb{E}_F[X] = 0$ corresponds to Corollary 1 in Raymaekers & Rousseeuw (2021).

3.4 Results for Additive Variables

We now consider a case where we are interested in the correlation between additive responses. This is a common situation in psychology where one constructs trait scores by summing over an individual's responses and then computes the correlation between two trait scores. We formalize this setup in as follows.

Assumption 1 ((Mean) scores of traits)**.** *Let $X_1, \ldots, X_{J_X}$ and $Y_1, \ldots, Y_{J_Y}$ be two sets of responses to rating-scale items, where each X_j and Y_k have support regions $\{-M_X, \ldots, M_X\}$ and $\{-M_Y, \ldots, M_Y\}$, respectively, for all items $j = 1, \ldots, J_X$ and $k = 1, \ldots, J_Y$. Let $X_j \sim F_X$ and $Y_k \sim F_Y$ for all items $j = 1, \ldots, J_X$; $k = 1, \ldots, J_Y$, and put $\mu_X = \mathbb{E}_{F_X}[X_j], \sigma_X^2 = \mathrm{Var}_{F_X}[X_j], \mu_Y = \mathbb{E}_{F_Y}[Y_k]$, and $\sigma_Y^2 = \mathrm{Var}_{F_Y}[Y_k]$. In addition, for any two distinct items $i \neq j$, put $\rho_X = \mathrm{Cor}_{F_X}[X_i, X_j]$ and $\rho_Y = \mathrm{Cor}_{F_Y}[Y_i, Y_j]$. Denote the trait scores by*

$$\overline{X} = \frac{1}{J_X}\sum_{j=1}^{J_X} X_j \quad \text{and} \quad \overline{Y} = \frac{1}{J_Y}\sum_{j=1}^{J_Y} Y_j,$$

and denote by $F_{\overline{X},\overline{Y}}$ the joint distribution of $(\overline{X}, \overline{Y})$ (implied by F_X and F_Y), and by $\rho_{\overline{X},\overline{Y}} = T\left(F_{\overline{X},\overline{Y}}\right) = \mathrm{Cor}_{F_{\overline{X},\overline{Y}}}\left[\overline{X}, \overline{Y}\right]$ the Pearson correlation between trait scores $\overline{X}$ and $\overline{Y}$ at $F_{\overline{X},\overline{Y}}$.

The setup of trait scores allows us to calculate the population reliability for each construct. In particular, Cronbach's α reads under the setup of Assumption 1

$$\alpha = \frac{J_X}{J_X - 1}\left(1 - \frac{\sum_{j=1}^{J_X}\mathrm{Var}_{F_X}[X_j]}{\mathrm{Var}_{F_X}\left[\sum_{j=1}^{J_X} X_j\right]}\right) = \frac{J_X}{J_X - 1}\left(1 - \frac{1}{1 + (J_X - 1)\rho_X}\right), \tag{4}$$

which is a function of within-construct correlation, ρ_X, and construct size $J_X \geq 2$. A proof of statement (4) is provided in the appendix.

It is useful to define the following functions, which are equivalent to the functions in (3) evaluated at the mean scores under the setup of Assumption 1. For a fixed contamination fraction $\varepsilon \in [0, 0.5]$, let

$$\begin{aligned}
n_\varepsilon(\overline{X}, \overline{Y}) &= (1-\varepsilon)\left(\rho_{\overline{X},\overline{Y}}\sqrt{\frac{\sigma_X^2\sigma_Y^2}{J_X J_Y}\big(1 + (J_X - 1)\rho_X\big)\big(1 + (J_Y - 1)\rho_Y\big)} + \mu_X\mu_Y\right) - \\
&\qquad (1-\varepsilon)^2\mu_X\mu_Y \\
n_\varepsilon(\overline{X}) &= (1-\varepsilon)\left(\frac{\sigma_X^2}{J_X}\big(1 + (J_X - 1)\rho_X\big) + \mu_X^2\right) - (1-\varepsilon)^2\mu_X^2, \quad \text{and,} \\
n_\varepsilon(\overline{Y}) &= (1-\varepsilon)\left(\frac{\sigma_Y^2}{J_Y}\big(1 + (J_Y - 1)\rho_Y\big) + \mu_Y^2\right) - (1-\varepsilon)^2\mu_Y^2.
\end{aligned} \tag{5}$$

The following corollary refers to Proposition 1 and states the bias of Pearson correlation at a given contaminated distribution.

Corollary 1. *Assume the setup and notation of Assumption 1, where $\overline{X} = J_X^{-1}\sum_{j=1}^{J_X} X_j$ and $\overline{Y} = J_Y^{-1}\sum_{j=1}^{J_Y} Y_j$ have Pearson correlation $\rho_{\overline{X},\overline{Y}} = T(F_{\overline{X},\overline{Y}})$ under model distribution $F_{\overline{X},\overline{Y}}$. Let H_X and H_Y be distributions of the same support as F_X and F_Y, respectively,*

such that the individual variables X_j and Y_j are identically distributed under H_X and H_Y, respectively. Denote by ν_Z and τ_Z^2 the mean and variance, respectively, of a distribution H_Z, for $Z \in \{X, Y\}$. In addition, for any two distinct items $i \neq j$, put $\phi_X = \mathrm{Cor}_{H_X}[X_i, X_j]$ and $\phi_Y = \mathrm{Cor}_{H_Y}[Y_i, Y_j]$. Let $H_{\overline{X},\overline{Y}}$ be the contaminating joint distribution of scores $\overline{X}$ and $\overline{Y}$ that is implied by H_X and H_Y, and denote by $\phi_{\overline{X},\overline{Y}} = T(H_{\overline{X},\overline{Y}}) = \mathrm{Cor}_{H_{\overline{X},\overline{Y}}}[\overline{X}, \overline{Y}]$ the Pearson correlation measure T at $H_{\overline{X},\overline{Y}}$. For $\varepsilon \in [0, 0.5]$, let $G_{\overline{X},\overline{Y}} = (1-\varepsilon)F_{\overline{X},\overline{Y}} + \varepsilon H_{\overline{X},\overline{Y}}$ be the contamination distribution implied by $F_{\overline{X},\overline{Y}}$ and $H_{\overline{X},\overline{Y}}$, whose respective marginals are denoted by $G_{\overline{X}}$ and $G_{\overline{Y}}$. Then the bias at contamination distribution $G_{\overline{X},\overline{Y}}$ is given by

$$T(G_{\overline{X},\overline{Y}}) - T(F_{\overline{X},\overline{Y}}) = \frac{\mathrm{Cov}_{G_{\overline{X},\overline{Y}}}[\overline{X}, \overline{Y}]}{\sqrt{\mathrm{Var}_{G_{\overline{X}}}[\overline{X}]\,\mathrm{Var}_{G_{\overline{Y}}}[\overline{Y}]}} - \rho_{XY},$$

where

$$\mathrm{Cov}_{G_{\overline{X},\overline{Y}}}[\overline{X}, \overline{Y}] = n_\varepsilon(\overline{X}, \overline{Y}) + \varepsilon\Bigg((1-\varepsilon)\big(\nu_X\nu_Y - \mu_X\nu_Y - \mu_Y\nu_X\big) + \phi_{\overline{X},\overline{Y}}\sqrt{\frac{\tau_X^2\tau_Y^2}{J_X J_Y}\big(1 + (J_X-1)\phi_X\big)\big(1 + (J_Y-1)\phi_Y\big)}\Bigg),$$

as well as

$$\mathrm{Var}_{G_{\overline{X}}}[\overline{X}] = n_\varepsilon(\overline{X}) + \varepsilon\Big((1-\varepsilon)\nu_X(\nu_X - 2\mu_X) + \tau_X^2(1 + (J_X-1)\phi_X)/J_X\Big),$$

and

$$\mathrm{Var}_{G_{\overline{Y}}}[\overline{Y}] = n_\varepsilon(\overline{Y}) + \varepsilon\Big((1-\varepsilon)\nu_Y(\nu_Y - 2\mu_Y) + \tau_Y^2(1 + (J_Y-1)\phi_Y)/J_Y\Big).$$

Remark 1. It is possible to derive bias curves for contaminating distributions H under which the individual variables in a given construct are not identically distributed, but for the sake of simplicity and brevity, we do not consider such cases.

The next corollary refers to Theorem 1 and Proposition 2 and states the maximum bias curves as well as the breakdown value of Pearson's correlation measure when the variables of interest are additive.

Corollary 2. *Assume the setup and notation of Assumption 1, where $\overline{X} = J_X^{-1}\sum_{j=1}^{J_X} X_j$ and $\overline{Y} = J_Y^{-1}\sum_{j=1}^{J_Y} Y_j$. For fixed contamination fraction $\varepsilon \in [0, 0.5]$, the maximum upward bias of Pearson correlation measure T between $\overline{X}$ and $\overline{Y}$ at model distribution $F_{\overline{X},\overline{Y}}$ is given by*

$$B^+(\varepsilon, T, F_{\overline{X},\overline{Y}}) = \frac{n_\varepsilon(\overline{X}, \overline{Y}) + \varepsilon M_X M_Y}{\sqrt{n_\varepsilon(X) + \varepsilon M_X^2}\sqrt{n_\varepsilon(Y) + \varepsilon M_Y^2}} - \rho_{\overline{X},\overline{Y}},$$

and the maximum downward bias is given by

$$B^{-}(\varepsilon, T, F_{\overline{X},\overline{Y}}) = \frac{n_{\varepsilon}(\overline{X},\overline{Y}) - \varepsilon M_X M_Y}{\sqrt{n_{\varepsilon}(\overline{X}) + \varepsilon M_X^2}\sqrt{n_{\varepsilon}(\overline{Y}) + \varepsilon M_Y^2}} - \rho_{\overline{X},\overline{Y}}.$$

Corollary 3. *Assume the setup and notation of Assumption 1, where* $\overline{X} = J_X^{-1}\sum_{j=1}^{J_X} X_j$ *and* $\overline{Y} = J_Y^{-1}\sum_{j=1}^{J_Y} Y_j$*. The breakdown value of Pearson correlation measure* T *at model distribution* $F_{\overline{X},\overline{Y}}$ *is given by*

$$\varepsilon^*(T) = \begin{cases} \frac{\mu_X^2 + V_{\rho_X} - M_X^2 + \sqrt{\left(V_{\rho_X} - \mu_X^2 + M_X^2\right)^2 + 4\mu_X^2 V_{\rho_X}}}{2\mu_X^2} & \text{if } \mu_X \neq 0, \\ \frac{V_{\rho_X}}{V_{\rho_X} + M_X^2} & \text{if } \mu_X = 0, \end{cases}$$

where $V_{\rho_X} = \sigma_X^2\big(1 + (J_X - 1)\rho_X\big)/J_X$.

4 Evaluation at Various Distributions

In the previous section, we have derived maximum bias curves, breakdown values, and influence functions for Pearson's correlation measure between discrete ordinal variables at some model distribution. In this section, we evaluate these results at various model distributions that are commonly observed in questionnaire-based research.

An ordinal discrete variable X with K Likert-type answer category has support $\{1, \ldots, K\}$, where larger values correspond to stronger degrees of agreement. The probability mass function associated with X is given by $\mathbb{P}[X = k] = \pi_k$, such that $\sum_{k=1}^{K} \pi_k = 1$ for nonnegative response probabilities $\pi_k, k = 1, \ldots, K$. The response probabilities π_k are governed by the (marginal) response distribution of X, denoted F_X.[4] We focus on three types of response distributions F_X, namely *centered, agreeing*, and *polarizing* response distributions. First, *centered* response distributions emulate items to which respondents tend to have no strong sentiment and therefore prefer neutral response categories in the center of the set of Likert-type answer categories. Second, items with an *agreeing* response distribution are items to which respondents are likely to express agreement by choosing response categories toward the right end in the set of answer categories. Items to which disagreement is likely can be obtained by reversing an agreeing response distribution. Third, *polarizing* response distributions mimic items to which respondents are likely to have strong divergent sentiments by choosing response categories on either end of the response category set.

Table 1 lists each of the three types of response distribution F_X for three numbers of Likert-type answer categories, $K \in \{5, 7, 9\}$. For instance, the distributions for $K = 5$ answer categories have been used in Alfons & Welz (2022).

In the following, we evaluate our main findings on the Pearson correlation between two ordinal discrete random variables, X and Y. Each variable X and Y has one of the (marginal)

[4] That is, F_X is a categorical distribution, which is a multinomial distribution for one trial.

Type	π_1	π_2	π_3	π_4	π_5
Centered	0.15	0.20	0.30	0.20	0.15
Agreeing	0.10	0.15	0.20	0.25	0.30
Polarizing	0.30	0.175	0.05	0.175	0.30

(a) $K = 5$ Likert-type answer categories

Type	π_1	π_2	π_3	π_4	π_5	π_6	π_7
Centered	0.05	0.125	0.2	0.25	0.2	0.125	0.05
Agreeing	0.05	0.075	0.1	0.125	0.15	0.225	0.275
Polarized	0.25	0.15	0.075	0.05	0.075	0.15	0.25

(b) $K = 7$ Likert-type answer categories

Type	π_1	π_2	π_3	π_4	π_5	π_6	π_7	π_8	π_9
Centered	0.025	0.05	0.15	0.175	0.2	0.175	0.15	0.05	0.025
Agreeing	0.025	0.05	0.075	0.075	0.1	0.125	0.15	0.175	0.225
Polarized	0.24	0.15	0.05	0.05	0.02	0.05	0.05	0.15	0.24

(c) $K = 9$ Likert-type answer categories

Table 1: Response distributions F_X for given number of answer categories, K, where $\pi_k = \mathbb{P}[X = k]$ for the k-th response category of a discrete ordinal variable X. The set of answer categories is anchored at "1", expressing the strongest form of disapproval, and "K", expressing the strongest form of approval.

response distributions in Table 1. For simplicity, we restrict our analysis to variables with the same support regions, that is, both X and Y have Likert-type support $\{1, \ldots, K\}$. However, for fixed K, the corresponding individual marginal response distributions F_X and F_Y may be different. Furthermore, to be able to apply our theoretical results, we silently rescale the support to be of form $\{-M, \ldots, M\}$ for some $M > 0$ by following the steps outlined in Section 3.2.

4.1 Breakdown Value

Before we evaluate the breakdown value for different scenarios, it is useful to first evaluate construct reliability for various choices of construct size, J, and within-construct correlation ρ_X. Figure 1 plots Cronbach's α, as defined in (4), as a function of these two quantities. Construct reliability increases with construct size and within-construct correlation. For smaller constructs of size five or smaller, only high correlations of 0.6 or more achieve acceptable levels of α of at least 0.75 (Robinson, 2018). For larger constructs comprising more than five items, a weaker within-construct correlation of 0.4 can suffice for α-values of at least 0.75. Conversely, very weak within-correlations of 0.2 require excessively large construct sizes to achieve acceptable construct reliability.

Figure 2 evaluates the breakdown values for various choices of construct size J (Proposition 2 for $J = 1$ and Corollary 2 for $J > 1$), within-construct correlation ρ_X, number of answer categories K, and response distributions from Table 1. Breakdown values tend to

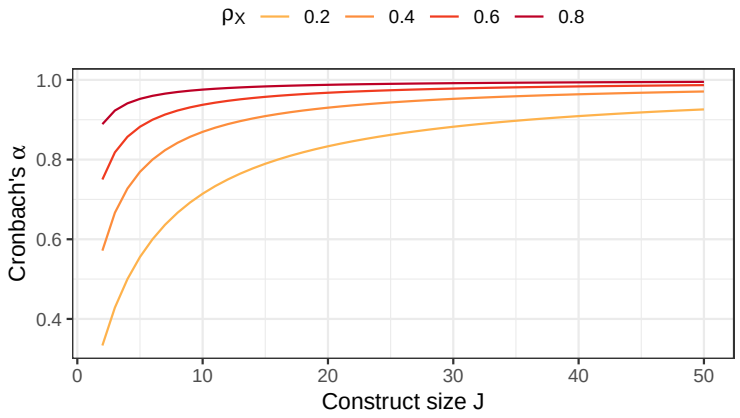


Figure 1: Cronbach's α (y-axis; ee Equation 4) for various choices of construct size $J \in \{2, \ldots, 50\}$ (x-axis) and within-construct correlation ρ_X (colors).

be lower for higher numbers of answer categories and weaker within-construct correlations. In addition, centered distributions consistently have lower breakdown values than "agreeing" distributions, where such distributions in turn consistently have lower breakdown values than polarized distributions. Furthermore, breakdown values diminish rapidly with growing construct sizes.

We conclude that constructs with fewer items are preferable from a robustness perspective for two reasons: First, having shorter constructs (and subsequently shorter questionnaires) drastically reduces the probability that respondents start responding carelessly Bowling et al. (2021a). Second, if there is careless responding, then a relatively large prevalence of careless respondents is required to break down Pearson's correlation measure (Figure 2). On the other hand, having smaller constructs may result in diminished construct reliability (Figure 1) and less accurate measurement.

Hence, we argue that in a discussion on construct size, one should take a more nuanced stance: While larger construct sizes are beneficial for reliability and accuracy in the absence of careless responding, such constructs are highly susceptible to the adverse effects of careless responding, and carelessness is more likely to occur in larger construct sizes (cf. Bowling et al., 2021a). It follows that larger constructs are only beneficial to measurement quality in an idealized, potentially overly optimistic setup. In fact, larger constructs may even be detrimental to measurement quality due to increased susceptibility to careless responding and overall higher likelihood of carelessness. In addition, questionnaires with lengthy constructs are expensive to administer due to, for instance, higher lab costs and more participant compensation.[5] Hence, we call for a more nuanced and realistic stance on measurement quality: Instead of striving for maximizing theoretical construct reliability by enlarging

[5]The budget saved on compensating participants for long questionnaire experiments could also be spent on recruiting a larger number of participants for shorter experiments, thereby enhancing statistical power of the experiment.

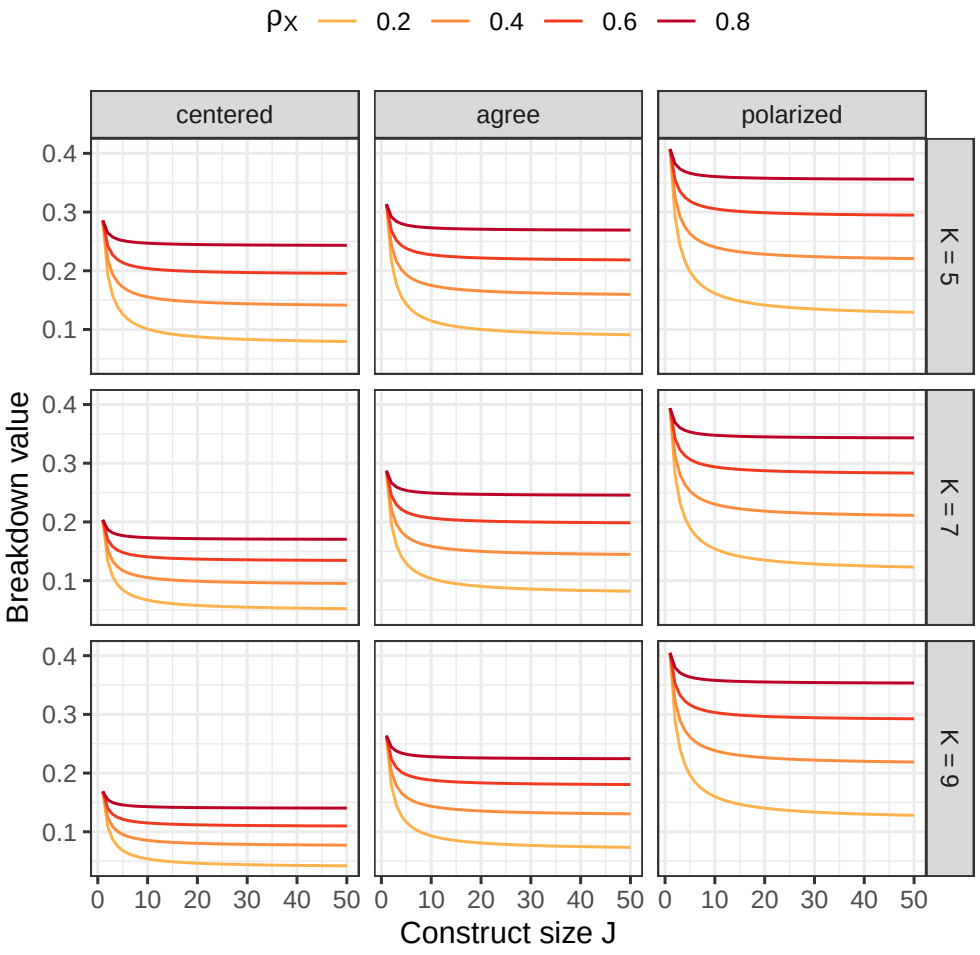


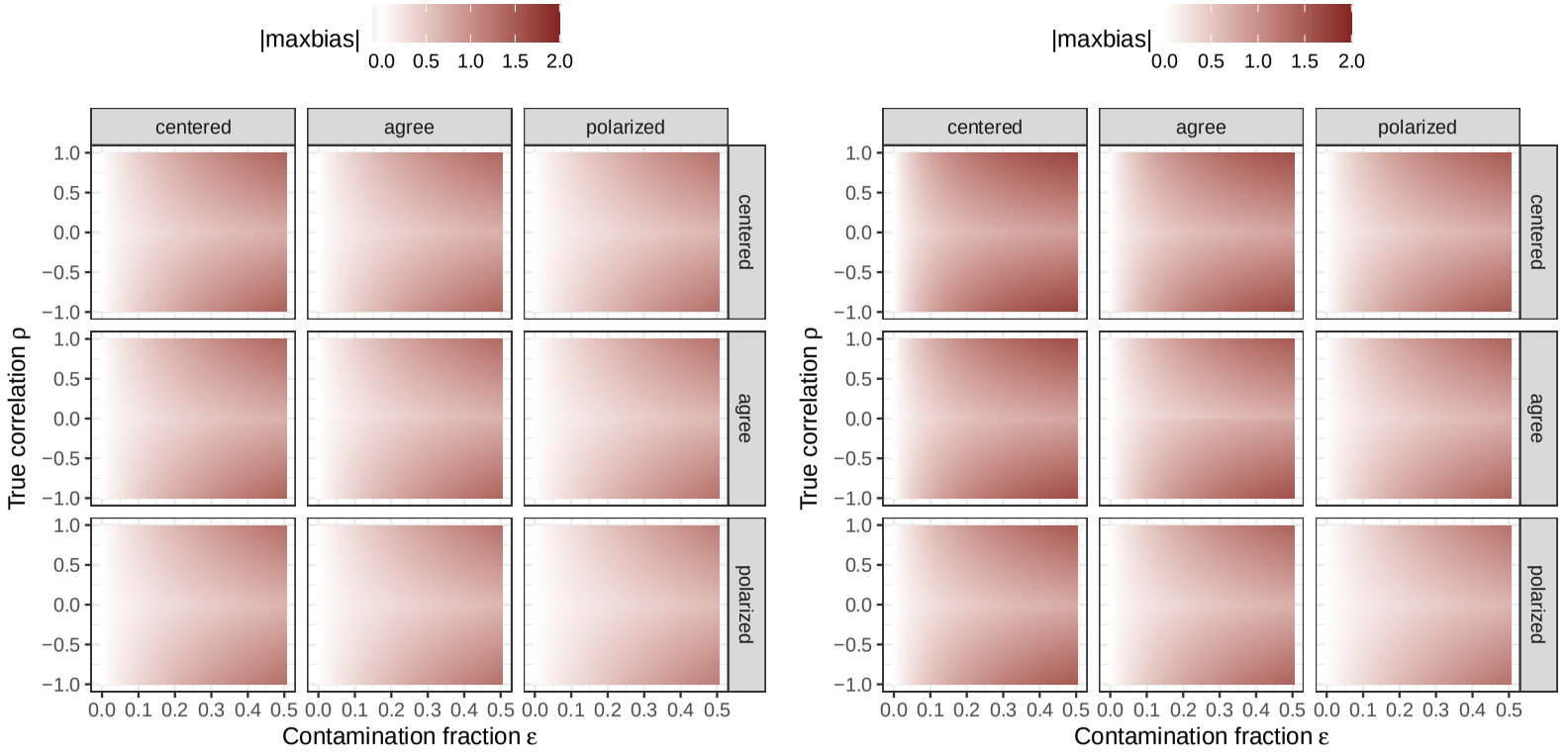
Figure 2: Breakdown values (y-axis; Proposition 2 and Corollary 2) for various choices of construct size $J \in \{1, \ldots, 50\}$ (x-axis), within-construct correlation ρ_X (colors), number of Likert-type answer categories K (rows), and response distributions (columns), where the latter ones are defined in Table 1.

construct size, one should strive for shorter constructs that still yield acceptable theoretical reliability. Doing so not only has statistical advantages due to enhanced robustness, but also practical ones due to lower costs and higher participant convenience. Overall, it might be preferable for the general scientific method to have instruments that work well in many, potentially very noisy scenarios, instead of instruments that work exceptionally well in a few, idealized scenarios, but can fail miserably in even slightly less favorable scenarios.

4.2 Maximum Bias Curves

The breakdown value represents a worst-case scenario in which perfect correlation gets dragged down to zero correlation due to careless responding. However, carelessness can already have unacceptably strong adverse effects when the correlation is not reduced to zero. To obtain a more nuanced picture of the effects of careless responding on correlation, we plot the maximum bias curves (Theorem 1 and Corollary 2) for various values of "true" correlation in the absence of carelessness, various model distributions (Table 1), and various contamination fractions.

Figure 3 contains heatmaps of the absolute value of maximum bias curves of Pearson's correlation measure T between ordinal discrete variables X and Y for various model distributions F. In particular, they visualize the absolute value of the maximum upward bias, $B^+(\varepsilon, T, F)$, if the true correlation $\rho = T(F)$ at model distribution F is negative, and the absolute value of maximum downward bias, $B^-(\varepsilon, T, F)$, if $\rho \geq 0$. The left panel (Figure 3a) shows maximum bias curves for single-item constructs ($J = 1$) for $K = 5$ Likert-type answer categories, while the right panel (Figure 3b) does so for $K = 9$ answer categories and a large construct size, $J = 10$, with strong within-construct correlations of $\rho_X = \rho_Y = 0.8$. Hence, the left panel reflects a situation with relatively low measurement quality under F (small construct and few answer categories), while the right panel reflects a situation with high measurement quality (α of > 0.95; Figure 1) under F.



(a) $J = 1, K = 5$.

(b) $J = 10, K = 9, \rho_X = \rho_Y = 0.8$.

Figure 3: Absolute value of maximum bias curves ("maxbias") of Pearson correlation measure T between X and Y for various model distributions F, where absolute maximum upward bias $|B^+(\varepsilon, T, F)|$ is plotted if $\rho = T(F) < 0$, and absolute maximum downward bias $|B^-(\varepsilon, T, F)|$ if $\rho \geq 0$. The x-axis denotes contamination fraction ε, and the y-axis denotes $\rho = T(F)$. The darker the red shade, the larger the absolute value of the maximum bias curves. Model distribution F is governed by the marginal distributions F_X and F_Y of the two constructs. The rows hold specifications for F_X, while the columns hold specifications for F_Y; see Table 1 for definitions. The left panel visualizes a scenario for single-item constructs ($J = 1$) and $K = 5$ Likert-type answer categories, and the right panel for $J = 10$ and $K = 9$, where within-construct correlation is 0.8 for both constructs, that is, $\rho_X = \rho_Y = 0.8$.

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. & Welz, M. (2022). Open science perspectives on machine learning for the identification of careless responding: A new hope or phantom menace? *psyArXiv preprint psyArXiv:10.31234*. https://doi.org/10.31234/osf.io/8t2cy

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

Arthur, W., Hagen, E., & George, F. (2021). The lazy or dishonest respondent: Detection and prevention. *Annual Review of Organizational Psychology and Organizational Behavior*, 8(1), 105–137. https://doi.org/10.1146/annurev-orgpsych-012420-055324

Bowling, N. A., Gibson, A. M., Houpt, J. W., & Brower, C. K. (2021a). Will the questions ever end? Person-level increases in careless responding during questionnaire completion. *Organizational Research Methods*, 24(4), 718–738. https://doi.org/10.1177/1094428120947794

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/10.1037/pspp0000085

Bowling, N. A., Huang, J. L., Brower, C. K., & Bragg, C. B. (2021b). The quick and the careless: The construct validity of page time as a measure of insufficient effort responding to surveys. *Organizational Research Methods*. https://doi.org/10.1177/10944281211056520. Forthcoming.

Cao, M. & Drasgow, F. (2019). Does forcing reduce faking? a meta-analytic review of forced-choice personality measures in high-stakes situations. *Journal of Applied Psychology*, 104(11), 1347–1368.

Capéraà, P. & Guillem, A. I. G. (1997). Taux de résistance des tests de rang d'indépendance. *Canadian Journal of Statistics*, 25(1), 113–124. https://doi.org/10.2307/3315361

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

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334. https://doi.org/doi.org/10.1007/BF02310555

Croux, C. & Dehon, C. (2010). Influence functions of the Spearman and Kendall correlation measures. *Statistical Methods & Applications*, 19, 497–515. https://doi.org/10.1007/s10260-010-0142-z

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

DellaVigna, S. (2009). Psychology and economics: Evidence from the field. *Journal of Economic Literature*, 47(2), 315–372. https://doi.org/10.1257/jel.47.2.315

DeSimone, J. A., Harms, P. D., & DeSimone, A. J. (2015). Best practice recommendations for data screening. *Journal of Organizational Behavior*, 36(2), 171–181. https://doi.org/10.1002/job.1962

Devlin, S. J., Gnanadesikan, R., & Kettenring, J. R. (1975). Robust estimation and outlier detection with correlation coefficients. *Biometrika*, 62(3), 531–545. https://doi.org/10.1093/biomet/62.3.531

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, phone, mail, and mixed-mode surveys: The tailored design method*. John Wiley & Sons.

Dunnette, M. D., McCartney, J., Carlson, H. C., & Kirchner, W. K. (1962). A study of faking behavior on a forced choice self-description checklist. *Personnel Psychology*, 15(1), 13–24. https://doi.org/10.1111/j.1744-6570.1962.tb01843.x

Ellingson, J. E. (2011). People fake only when they need to fake. *New Perspectives on Faking in Personality Assessment*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195387476.003.0014

Ellingson, J. E., Smith, D. B., & Sackett, P. R. (2001). Investigating the influence of social desirability on personality factor structure. *Journal of Applied Psychology*, 86(1), 122–133.

Furnham, A. & Henderson, M. (1982). The good, the bad and the mad: Response bias in self-report measures. *Personality and Individual Differences*, 3(3), 311–320. https://doi.org/10.1016/0191-8869(82)90051-4

Gabaix, X. (2019). Behavioral inattention. *Handbook of Behavioral Economics*, volume 2, 261–343. Elsevier. https://doi.org/10.1016/bs.hesbe.2018.11.001

Galesic, M. & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly*, 73(2), 349–360. https://doi.org/10.1093/poq/nfp031

Gibson, A. M. & Bowling, N. A. (2020). The effects of questionnaire length and behavioral consequences on careless responding. *European Journal of Psychological Assessment*, 36(2), 410–420. https://doi.org/10.1027/1015-5759/a000526

Griffith, R. L., Chmielowski, T., & Yoshita, Y. (2007). Do applicants fake? an examination of the frequency of applicant faking behavior. *Personnel Review*, 36(3), 341–355. https://doi.org/10.1108/00483480710731310

Griffith, R. L. & Converse, P. D. (2011). The rules of evidence and the prevalence of applicant faking. *New Perspectives on Faking in Personality Assessment*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195387476.003.0018

Griffith, R. L. & McDaniel, M. A. (2006). The nature of deception and applicant faking behavior. *A closer examination of applicant faking behavior*. Information Age Publishing.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346), 383–393. https://doi.org/10.1080/01621459.1974.10482962

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

Hogan, J., Barrett, P., & Hogan, R. (2007). Personality measurement, faking, and employment selection. *Journal of Applied Psychology*, 92(5), 1270–1285.

Hogan, R. & Hogan, J. (2007). *Hogan Personality Inventory Manual* (3rd ed.). Hogan Assessment Systems.

Holden, R. R. & Book, A. S. (2011). Faking does distort self-report personality assessment. *New Perspectives on Faking in Personality Assessment*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195387476.003.0026

Hough, L. M., Eaton, N. K., Dunnette, M. D., Kamp, J. D., & McCloy, R. A. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities. *Journal of Applied Psychology*, 75(5), 581–595. https://psycnet.apa.org/doi/10.1037/0021-9010.75.5.581

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. & Ronchetti, E. M. (2009). *Robust Statistics* (2nd ed.). John Wiley & Sons.

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality*, 39(1), 103–129. https://doi.org/10.1016/j.jrp.2004.09.009. Proceedings of the Association for Research in Personality

Johnson, J. A. & Hogan, R. (2006). A socioanalytic view of faking. *A closer examination of applicant faking behavior*. Information Age Publishing.

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

Komar, S., Brown, D. J., Komar, J. A., & Robie, C. (2008). Faking and the validity of conscientiousness: A Monte Carlo investigation. *Journal of Applied Psychology*, 93(1), 140–154. https://doi.org/10.1037/0021-9010.93.1.140

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

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

Marshall, M. B., De Fruyt, F., Rolland, J.-P., & Bagby, R. M. (2005). Socially desirable responding and the factorial stability of the neo pi-r. *Psychological Assessment*, 17(3), 379–384. https://doi.org/10.1037/1040-3590.17.3.379

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

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

Nichols, D. S., Greene, R. L., & Schmolck, P. (1989). Criteria for assessing inconsistent patterns of item endorsement on the MMPI: Rationale, development, and empirical trials. *Journal of Clinical Psychology*, 45(2), 239–250. https://doi.org/10.1002/1097-4679(198903)45:2%3C239::AID-JCLP2270450210%3E3.0.CO;2-1

Paulhus, D. (2002). Socially desirable responding: The evolution of a construct. *The role of constructs in psychological and educational measurement*, 49–69.

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*, 46(3), 598—-609.

Paulhus, D. L. (1986). *Self-Deception and Impression Management in Test Responses*, 143–165. Springer. https://doi.org/10.1007/978-3-642-70751-3_8

Paulhus, D. L. (1993). Bypassing the will: The automatization of affirmations. *Handbook of Mental Control*. Prentice-Hall.

Paulhus, D. L., Bruce, M. N., & Trapnell, P. D. (1995). Effects of self-presentation strategies on personality profiles and their structure. *Personality and Social Psychology Bulletin*, 21(2), 100–108. https://doi.org/10.1177/0146167295212001

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

Robinson, M. A. (2018). Using multi-item psychometric scales for research and practice in human resource management. *Human Resource Management*, 57(3), 739–750.

Schlenker, B. R. (1980). *mpression management: The self-concept, social identity, and interpersonal relations*. Brooks/Cole.

Schmit, M. J. & Ryan, A. M. (1993). The big five in personnel selection: Factor structure in applicant and nonapplicant populations. *Journal of Applied Psychology*, 78(6), 966–974. https://doi.org/10.1037/0021-9010.78.6.966

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

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

Smith, D. B. & Ellingson, J. E. (2002). Substance versus style: A new look at social desirability in motivating contexts. *Journal of Applied Psychology*, 87(2), 211–219. https://doi.org/10.1037/0021-9010.87.2.211

Stantcheva, S. (2022). How to run surveys: A guide to creating your own identifying variation and revealing the invisible. Working Paper 30527, National Bureau of Economic Research. https://doi.org/10.3386/w30527

Topping, G. D. & O'Gorman, J. (1997). Effects of faking set on validity of the NEO-FFI. *Personality and Individual Differences*, 23(1), 117–124. https://doi.org/10.1016/S0191-8869(97)00006-8

Ward, M. & Meade, A. W. (2022). Dealing with careless responding in survey data: Prevention, identification, and recommended best practices. *Annual Review of Psychology*. https://doi.org/10.1146/annurev-psych-040422-045007. Forthcoming.

Ward, M., Meade, A. W., Allred, C. M., Pappalardo, G., & Stoughton, J. W. (2017). Careless response and attrition as sources of bias in online survey assessments of personality traits and performance. *Computers in Human Behavior*, 76, 417–430. https://doi.org/10.1016/j.chb.2017.06.032

Ward, M. & Pond, S. B. (2015). Using virtual presence and survey instructions to minimize careless responding on internet-based surveys. *Computers in Human Behavior*, 48, 554–568. https://doi.org/10.1016/j.chb.2015.01.070

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

White, L. A., Young, M. C., Hunter, A. E., & Rumsey, M. G. (2008). Lessons learned in transitioning personality measures from research to operational settings. *Industrial and Organizational Psychology*, 1(3), 291–295. https://doi.org/10.1111/j.1754-9434.2008.00049.x

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

Zickar, M. J., Gibby, R. E., & Robie, C. (2004). Uncovering faking samples in applicant, incumbent, and experimental data sets: An application of mixed-model item response theory. *Organizational Research Methods*, 7(2), 168–190. https://doi.org/10.1177/1094428104263674

A Statistical Functionals

In this section, we briefly introduce the concept of (statistical) functionals, and derive the sample Pearson correlation coefficient expressed as a functional.

Suppose X is a real-valued random variable with distribution F. The population mean of X is defined by the integral

$$T(F) := \mathbb{E}_F[X] = \int x \, \mathrm{d}F(x), \tag{A.1}$$

which we express as a functional T of distribution F. Recall that for a given N-sized random sample $X_1, \ldots, X_N$ from F, the empirical distribution function of F at some $x \in \mathbb{R}$ is given by

$$\widehat{F}_N(x) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{X_i \leq x\} = \frac{1}{N} \sum_{i=1}^{N} \Delta_{X_i}(x),$$

where $\Delta_y(x) = \mathbb{1}\{y \leq x\}$ is a point mass (Dirac) measure at some point y.

To obtain the empirical counterpart of the population mean $T(F)$ in (A.1), we simply evaluate T at the empirical distribution function $\widehat{F}_N$:

$$\begin{aligned} T(\widehat{F}_N) = \mathbb{E}_{\widehat{F}_N}[X] &= \int x \, \mathrm{d}\widehat{F}_N(x) \\ &= \int x \, \mathrm{d}\left(\frac{1}{N} \sum_{i=1}^{n} \Delta_{X_i}(x)\right) \\ &= \frac{1}{N} \sum_{i=1}^{N} \int x \, \mathrm{d}\Delta_{X_i}(x) \\ &= \frac{1}{N} \sum_{i=1}^{N} X_i \\ &= \overline{X}_N, \end{aligned} \tag{A.2}$$

where we have used the property $\int f \mathrm{d}\Delta_x = f(x)$ for any function f.

In similar fashion, we can show that $\mathbb{E}_{\widehat{F}_N}[X^2] = \frac{1}{N}\sum_{i=1}^{N} X_i^2$. It follows that the empirical counterpart of population variance $\mathrm{Var}_F[X] = \mathbb{E}_F[X^2] - \mathbb{E}_F[X]^2$ can be obtained by

$$\mathrm{Var}_{\widehat{F}_N}[X] = \mathbb{E}_{\widehat{F}_N}\left[X^2\right] - \mathbb{E}_{\widehat{F}_N}[X]^2 = \sum_{i=1}^{N} X_i^2 - \overline{X}_N^2, \tag{A.3}$$

and similarly for $\mathrm{Var}_{\widehat{F}_N}[Y]$.

Now, suppose that F is a joint bivariate distribution of random variables (X, Y). The empirical distribution function of an N-sized random sample $\{(X_i, Y_i)\}_{i=1}^{N}$ from F is given

by $\widehat{F}_N(x,y) = \frac{1}{N}\sum_{i=1}^N \Delta_{(X_i,Y_i)}(x,y)$, for $x,y \in \mathbb{R}$. Repeating the steps in (A.2), we obtain the empirical counterpart of population mean $\mathbb{E}_F[XY]$ by

$$\mathbb{E}_{\widehat{F}_N}[XY] = \frac{1}{N}\sum_{i=1}^N X_i Y_i. \tag{A.4}$$

Finally, the population covariance of X and Y is defined as

$$\mathrm{Cov}_F[X,Y] = \mathbb{E}_F[XY] - \mathbb{E}_F[X]\,\mathbb{E}_F[Y],$$

and its empirical counterpart is subsequently given by

$$\begin{aligned}\mathrm{Cov}_{\widehat{F}_N}[X,Y] &= \mathbb{E}_{\widehat{F}_N}[XY] - \mathbb{E}_{\widehat{F}_N}[X]\,\mathbb{E}_{\widehat{F}_N}[Y] \\ &= \frac{1}{N}\sum_{i=1}^N X_i Y_i - \overline{X}_N \overline{Y}_N,\end{aligned} \tag{A.5}$$

where we have used (A.2) and (A.4).

The population correlation between X and Y is defined as

$$\mathrm{Cor}_F[X,Y] = \frac{\mathrm{Cov}_F[X,Y]}{\sqrt{\mathrm{Var}_F[X]}\sqrt{\mathrm{Var}_F[Y]}}.$$

Plugging in $\widehat{F}_N$ yields its empirical counterpart,

$$\begin{aligned}\mathrm{Cor}_{\widehat{F}_N}[X,Y] &= \frac{\mathrm{Cov}_{\widehat{F}_N}[X,Y]}{\sqrt{\mathrm{Var}_{\widehat{F}_N}[X]}\sqrt{\mathrm{Var}_{\widehat{F}_N}[Y]}} \\ &= \frac{\frac{1}{N}\sum_{i=1}^N X_i Y_i - \overline{X}_N \overline{Y}_N}{\sqrt{\frac{1}{N}\sum_{i=1}^N X_i^2 - \overline{X}_N^2}\sqrt{\frac{1}{N}\sum_{i=1}^N Y_i^2 - \overline{Y}_N^2}},\end{aligned}$$

where we have used (A.3) and (A.5). This expression is exactly the definition of Pearson's correlation measure. Hence, writing Pearson's correlation measure as statistical functional allows us to express both its population and empirical version in one term.

B Proofs

B.1 Useful Lemmata

Lemma B.1. *Let X be a discrete ordinal random variable with an arbitrary distribution F of support $\{-M_X, \ldots, M_X\}$, where $M_X > 0$. Then the term*

$$m_{F,\varepsilon}(X) + \varepsilon M_X^2$$

is strictly positive for all $\varepsilon \in [0,1)$.

Proof. Fix $\varepsilon \in [0, 1)$. Then

$$\begin{aligned}
0 &\leq \mathbb{E}_F\left[X^2\right] - \mathbb{E}_F\left[X\right]^2 \\
&< \mathbb{E}_F\left[X^2\right] - (1-\varepsilon)^2\mathbb{E}_F\left[X\right]^2 \\
&= (1-\varepsilon)\mathbb{E}_F\left[X^2\right] - (1-\varepsilon)^2\mathbb{E}_F\left[X\right]^2 + \varepsilon\mathbb{E}_F\left[X^2\right] \\
&\leq (1-\varepsilon)\mathbb{E}_F\left[X^2\right] - (1-\varepsilon)^2\mathbb{E}_F\left[X\right]^2 + \varepsilon M_X^2 \\
&= (1-\varepsilon)\left(\mathrm{Var}_F\left[X\right] + \mathbb{E}_F\left[X\right]^2\right) - (1-\varepsilon)^2\mathbb{E}_F\left[X\right]^2 + \varepsilon M_X^2 \\
&= m_{F,\varepsilon}(X) + \varepsilon M_X^2,
\end{aligned}$$

where the first line follows from Jensen's inequality and the fourth line follows from the assumption that $|X|$ is bounded from above by M_X. □

This lemma demonstrates that the denominators in either maximum bias curve in Theorem 1 are strictly positive and real-valued.

The next lemma provides quantities that will be useful in proofs of the corollaries of Proposition 1, 2, and Theorem 1 in the setup of Assumption 1.

Lemma B.2. *Assume the setup and notation of Assumption 1, where* $\overline{X} = J_X^{-1}\sum_{j=1}^{J_X} X_j$. *Denote by* $F_{\overline{X}}$ *the (marginal) distribution of mean score* $\overline{X}$. *It holds true that*

$$\mathbb{E}_{F_{\overline{X}}}\left[\overline{X}\right] = \mu_X, \tag{B.6a}$$

$$\mathrm{Var}_{F_{\overline{X}}}\left[\overline{X}\right] = \frac{\sigma_X^2}{J_X}\big(1 + (J_X - 1)\rho_X\big), \tag{B.6b}$$

and, for distinct items $i, j \in \{1, \ldots, J_X\}, i \neq j$,

$$\mathbb{E}_{F_X}\left[X_i X_j\right] = \rho_X\sigma_X^2 + \mu_X^2, \tag{B.6c}$$

$$\mathrm{Cov}_{F_X}\left[X_i, X_j\right] = \rho_X\sigma_X^2. \tag{B.6d}$$

Proof. For (B.6a), we have by the assumption that the X_j are identically distributed with distribution F_X that

$$\mathbb{E}_{F_{\overline{X}}}\left[\overline{X}\right] = \frac{1}{J_X}\sum_{j=1}^{J_X}\mathbb{E}_{F_X}\left[X_j\right] = \mu_X.$$

For (B.6c) and distinct items $i \neq j$, we have by the assumption of identical distribution that

$$\begin{aligned}
\mathbb{E}_{F_X}\left[X_i X_j\right] &= \mathrm{Cov}_{F_X}\left[X_i, X_j\right] + \mathbb{E}_{F_X}\left[X_i\right]\mathbb{E}_{F_X}\left[X_j\right] \\
&= \rho_X\sqrt{\mathrm{Var}_{F_X}\left[X_i\right]\mathrm{Var}_{F_X}\left[X_j\right]} + \mu_X^2 \\
&= \rho_X\sigma_X^2 + \mu_X^2.
\end{aligned}$$

Statement (B.6d) trivially follows from the definition of correlation and identical distribution of the individual item responses. For (B.6b), we again have by identical distribution that

$$\begin{aligned}\mathrm{Var}_{F_{\overline{X}}}\left[\overline{X}\right] &= \frac{1}{J_X^2}\left(\sum_{j=1}^{J_X}\mathrm{Var}_{F_X}\left[X_j\right] + 2\overbrace{\sum\sum_{i<j}}^{J_X(J_X-1)/2 \text{ terms}}\mathrm{Cov}_{F_X}\left[X_i, X_j\right]\right)\\ &= \frac{1}{J_X}\Big(\sigma_X^2 + 2(J_X-1)/2\rho_X\sigma_X^2\Big)\\ &= \frac{\sigma_X^2}{J_X}\big(1+(J_X-1)\rho_X\big),\end{aligned}$$

where we have used (B.6d) in the second line. □

The next lemma establishes Equation (4).

Lemma B.3. *Assume the setup and notation of Assumption 1, where* $\overline{X} = J_X^{-1}\sum_{j=1}^{J_X} X_j$. *Denote by* $F_{\overline{X}}$ *the (marginal) distribution of mean score* $\overline{X}$. *Cronbach's* α *(Cronbach, 1951), defined by,*

$$\alpha = \frac{J_X}{J_X-1}\left(1 - \frac{\sum_{j=1}^{J_X}\mathrm{Var}_{F_X}\left[X_j\right]}{\mathrm{Var}_{F_X}\left[\sum_{j=1}^{J_X}X_j\right]}\right), \tag{B.7}$$

is then equal to

$$\alpha = \frac{J_X}{J_X-1}\left(1 - \frac{1}{1+(J_X-1)\rho_X}\right).$$

Proof. This statement follows immediately from Equation (B.6b) in Lemma B.2 and the assumption that the individual item responses X_j are identically distributed with distribution F_X. □

B.2 Proof of Proposition 1

Let $G = (1-\varepsilon)F + \varepsilon H \in \mathcal{F}_\varepsilon$ for a fixed $\varepsilon \in [0, 0.5]$.

We first derive the variance expressions under G. We have for the first two moments of X under G that

$$\begin{aligned}\mathbb{E}_G\left[X\right] &= (1-\varepsilon)\mathbb{E}_F\left[X\right] + \varepsilon\mathbb{E}_H\left[X\right]\\ \mathbb{E}_G\left[X^2\right] &= (1-\varepsilon)\mathbb{E}_F\left[X^2\right] + \varepsilon\mathbb{E}_H\left[X^2\right],\end{aligned}$$

where we have used the definition of $m_\varepsilon(X)$ in (3). It follows from squaring the first moment that

$$\mathbb{E}_G\left[X\right]^2 = (1-\varepsilon)^2\mathbb{E}_F\left[X\right]^2 + 2(1-\varepsilon)\varepsilon\mathbb{E}_F\left[X\right]\mathbb{E}_H\left[X\right] + \varepsilon^2\mathbb{E}_H\left[X\right]^2,$$

and therefore

$$\begin{aligned}\mathrm{Var}_G[X] &= \mathbb{E}_G\left[X^2\right] - \mathbb{E}_G[X]^2 \\ &= m_\varepsilon(X) + \varepsilon\Big(- 2(1-\varepsilon)\mathbb{E}_F[X]\,\mathbb{E}_H[X] + \mathbb{E}_H\left[X^2\right] - \varepsilon\mathbb{E}_H[X]^2\Big).\end{aligned}$$

Repeating these steps for Y proves the variance expressions under G.

We now turn to deriving the expression for $\mathrm{Cov}_G[X,Y]$. We have that

$$\mathbb{E}_G[XY] = (1-\varepsilon)\mathbb{E}_F[XY] + \varepsilon\mathbb{E}_H[XY]$$

and that

$$\begin{aligned}&\mathbb{E}_G[X]\,\mathbb{E}_G[Y] = \\ &\quad(1-\varepsilon)^2\mathbb{E}_F[X]\,\mathbb{E}_F[Y] + (1-\varepsilon)\varepsilon\mathbb{E}_F[X]\,\mathbb{E}_H[Y] + (1-\varepsilon)\varepsilon\mathbb{E}_F[Y]\,\mathbb{E}_H[X] + \varepsilon^2\mathbb{E}_H[X]\,\mathbb{E}_H[Y].\end{aligned}$$

The desired expression follows from substituting into the identity

$$\mathrm{Cov}_G[X,Y] = \mathbb{E}_G[XY] - \mathbb{E}_G[X]\,\mathbb{E}_G[Y],$$

and the definition of $m_\varepsilon(X,Y)$ in (3). Substituting these (co)variance expressions into the definition of the bias curve at G for Pearson correlation (2) completes the proof. □

B.3 Proof of Theorem 1

We prove the expression for the maximum upward bias curve $B^+(\varepsilon,T,F)$. The proof for the maximum downward bias curve $B^+(\varepsilon,T,F)$ follows by replacing Y by $-Y$ in the steps below.

The Pearson correlation measure T evaluated at any contaminating distribution $G = (1-\varepsilon)F + \varepsilon H \in \mathcal{F}_\varepsilon$ is by definition given by

$$T(G) = \frac{\mathrm{Cov}_G[X,Y]}{\sqrt{\mathrm{Var}_G[X]}\sqrt{\mathrm{Var}_G[Y]}}. \tag{B.8}$$

We aim at deriving an upper bound for this expression that holds uniformly over all contaminated distributions $G \in \mathcal{F}_\varepsilon$ for fixed $\varepsilon \in [0, 0.5]$.

We derive an upper bound for $T(G)$ in (B.8) by deriving lower bounds for its denominator and upper bounds for its numerator. By Proposition 1, the denominator is identified by the square root of variance expressions

$$\mathrm{Var}_G[X] = m_\varepsilon(X) + \varepsilon\Big(- 2(1-\varepsilon)\mathbb{E}_F[X]\,\mathbb{E}_H[X] + V_{H,\varepsilon}(X)\Big), \tag{B.9}$$

where

$$V_{H,\varepsilon}(X) = \mathbb{E}_H\left[X^2\right] - \varepsilon\mathbb{E}_H[X]^2.$$

By Proposition 1, the numerator in (B.8) equals

$$\begin{aligned}\mathrm{Cov}_G\left[X,Y\right] &= m_\varepsilon(X,Y) + \\ &\varepsilon\Big(-(1-\varepsilon)\mathbb{E}_F\left[X\right]\mathbb{E}_H\left[Y\right] - (1-\varepsilon)\mathbb{E}_F\left[Y\right]\mathbb{E}_H\left[X\right] + \mathbb{E}_H\left[XY\right] - \varepsilon\mathbb{E}_H\left[X\right]\mathbb{E}_H\left[Y\right]\Big).\end{aligned} \tag{B.10}$$

We distinguish between two cases for the cross term $\mathbb{E}_H\left[X\right]\mathbb{E}_H\left[Y\right]$.

Case 1: $\mathbb{E}_{\boldsymbol{H}}\left[\boldsymbol{X}\right]\mathbb{E}_{\boldsymbol{H}}\left[\boldsymbol{Y}\right] \leq \mathbf{0}$. Because $\varepsilon\mathbb{E}_H\left[X\right]^2 \leq \mathbb{E}_H\left[X\right]^2$, it follows that (B.9) can be upper bounded as

$$\begin{aligned}\sqrt{\mathrm{Var}_G\left[X\right]} &\geq \sqrt{m_\varepsilon(X) + \varepsilon\Big(-2(1-\varepsilon)\mathbb{E}_F\left[X\right]\mathbb{E}_H\left[X\right] + \mathbb{E}_H\left[X^2\right] - \mathbb{E}_H\left[X\right]^2\Big)} \\ &= \sqrt{m_\varepsilon(X) + \varepsilon\Big(-2(1-\varepsilon)\mathbb{E}_F\left[X\right]\mathbb{E}_H\left[X\right] + \mathrm{Var}_H\left[X\right]\Big)}.\end{aligned}$$

An analogous bound can be obtained for $\sqrt{\mathrm{Var}_G\left[Y\right]}$.

Following the same steps as in the proof of Proposition 2 in Raymaekers & Rousseeuw (2021), the numerator (B.10) is bounded from above by

$$m_\varepsilon(X,Y) + \varepsilon\Big(-(1-\varepsilon)\mathbb{E}_F\left[X\right]\mathbb{E}_H\left[Y\right] - (1-\varepsilon)\mathbb{E}_F\left[Y\right]\mathbb{E}_H\left[X\right] + \sqrt{\mathrm{Var}_H\left[X\right]}\sqrt{\mathrm{Var}_H\left[Y\right]}\Big).$$

Combining the previous two displays, Pearson correlation $T(G)$ is bounded from above by

$$\frac{m_\varepsilon(X,Y) + \varepsilon\Big(-(1-\varepsilon)\mathbb{E}_F\left[X\right]\mathbb{E}_H\left[Y\right] - (1-\varepsilon)\mathbb{E}_F\left[Y\right]\mathbb{E}_H\left[X\right] + \sqrt{\mathrm{Var}_H\left[X\right]}\sqrt{\mathrm{Var}_H\left[Y\right]}\Big)}{\sqrt{m_\varepsilon(X) + \varepsilon\Big(-2(1-\varepsilon)\mathbb{E}_F\left[X\right]\mathbb{E}_H\left[X\right] + \mathrm{Var}_H\left[X\right]\Big)}\sqrt{m_\varepsilon(Y) + \varepsilon\Big(-2(1-\varepsilon)\mathbb{E}_F\left[Y\right]\mathbb{E}_H\left[Y\right] + \mathrm{Var}_H\left[Y\right]\Big)}}. \tag{B.11}$$

This bound increases with $\mathrm{Var}_H\left[X\right] = \mathbb{E}_H\left[X^2\right] - \mathbb{E}_H\left[X\right]^2$ and $\mathrm{Var}_H\left[Y\right] = \mathbb{E}_H\left[Y^2\right] - \mathbb{E}_H\left[Y\right]^2$, so we wish to make these two terms as large as possible. These terms are maximized when $\mathbb{E}_H\left[X^2\right] = \sup_{H'}\mathbb{E}_{H'}\left[X^2\right] = M_X^2$ and $\mathbb{E}_H\left[X\right] = 0$, as well as $\mathbb{E}_H\left[Y^2\right] = M_Y^2$ and $\mathbb{E}_H\left[Y\right] = 0$. Combining this with the intermediate bound (B.11), it follows that Pearson correlation $T(G)$ is bounded from above by

$$T(G) \leq \frac{m_\varepsilon(X,Y) + \varepsilon M_X M_Y}{\sqrt{m_\varepsilon(X) + \varepsilon M_X^2}\sqrt{m_\varepsilon(Y) + \varepsilon M_Y^2}}.$$

We now turn to the complementary case.

Case 2: $\mathbb{E}_H[X]\,\mathbb{E}_H[Y] > 0$. In the proof of Proposition 2 of Raymaekers & Rousseeuw (2021) it is shown that for any distribution H,

$$\mathbb{E}_H[XY] - \varepsilon\mathbb{E}_H[X]\,\mathbb{E}_H[Y] \leq \sqrt{V_{H,\varepsilon}(X)}\sqrt{V_{H,\varepsilon}(Y)}.$$

Applying this to the numerator in (B.8), it follows that $T(G)$ is bounded from above by

$$\frac{m_\varepsilon(X,Y) + \varepsilon\Big(-(1-\varepsilon)\mathbb{E}_F[X]\,\mathbb{E}_H[Y] - (1-\varepsilon)\mathbb{E}_F[Y]\,\mathbb{E}_H[X] + \sqrt{V_{H,\varepsilon}(X)}\sqrt{V_{H,\varepsilon}(Y)}\Big)}{\sqrt{m_\varepsilon(X) + \varepsilon\Big(-2(1-\varepsilon)\mathbb{E}_F[X]\,\mathbb{E}_H[X] + V_{H,\varepsilon}(X)\Big)}\sqrt{m_\varepsilon(Y) + \varepsilon\Big(-2(1-\varepsilon)\mathbb{E}_F[Y]\,\mathbb{E}_H[Y] + V_{H,\varepsilon}(Y)\Big)}}.$$

Juts like in the previous case, this bound increases with $V_{H,\varepsilon}(X) = \mathbb{E}_H[X^2] - \varepsilon\mathbb{E}_H[X]^2$ and $V_{H,\varepsilon}(Y) = \mathbb{E}_H[Y^2] - \varepsilon\mathbb{E}_H[Y]^2$. Likewise, these terms are maximized when $\mathbb{E}_H[X^2] = M_X^2$ and $\mathbb{E}_H[X] = 0$, as well as $\mathbb{E}_H[Y^2] = M_Y^2$ and $\mathbb{E}_H[Y] = 0$. Combining this with the previous display, it follows that Pearson correlation $T(G)$ can be bounded from above by

$$\begin{aligned} T(G) &\leq \frac{m_\varepsilon(X,Y) + \varepsilon M_X M_Y}{\sqrt{m_\varepsilon(X) + \varepsilon M_X^2}\sqrt{m_\varepsilon(Y) + \varepsilon M_Y^2}} \\ &=: C(\varepsilon). \end{aligned}$$

We conclude that in either case, $T(G)$ is upper bounded by $C(\varepsilon)$.

It remains to be shown that the bound $C(\varepsilon)$ is sharp. For the maximum upward bias, the two worst-placed contamination points in the (X,Y) space are $(-M_X, -M_Y)$ and (M_X, M_Y) because the upward bias is maximal when Pearson correlation $\rho = T(F)$ under model distribution F is negative. The distribution corresponding to the worst placed contamination points is therefore

$$H' := \frac{1}{2}\Delta_{(-M_X,-M_Y)} + \frac{1}{2}\Delta_{(M_X,M_Y)}, \tag{B.12}$$

where $(s,t) \mapsto \Delta_{(x,y)}(s,t) = \mathbb{1}\{x \leq s, y \leq t\}$ is a bivariate distribution function that puts all its mass at a point (x,y). For the relevant moments of X and Y under contaminating distribution H' it holds that

$$\mathbb{E}_{H'}[XY] = M_X M_Y, \quad \mathbb{E}_{H'}[X] = \mathbb{E}_{H'}[Y] = 0, \quad \mathbb{E}_{H'}\left[X^2\right] = M_X^2, \quad \text{and} \quad \mathbb{E}_{H'}\left[Y^2\right] = M_Y^2.$$

Therefore, for $G' := (1-\varepsilon)F + \varepsilon H'$, it follows from (B.8) that Pearson correlation T evaluated at contaminated distribution G' is given by

$$T(G') = \frac{m_\varepsilon(X,Y) + \varepsilon M_X M_Y}{\sqrt{m_\varepsilon(X) + \varepsilon M_X^2}\sqrt{m_\varepsilon(Y) + \varepsilon M_Y^2}},$$

which is equal to the previously derived upper bound $C(\varepsilon)$, thereby proving that this bound is indeed sharp. This concludes the proof for the maximum upward bias. The proof for the maximum downward bias follows by repeating the above steps with Y replaced by $-Y$, and with contaminating distribution H' in (B.12) replaced by $H' := \frac{1}{2}\Delta_{(-M_X,M_Y)} + \frac{1}{2}\Delta_{(M_X,-M_Y)}$. □

B.4 Proof of Proposition 2

The case $\mathbb{E}_F[X] = 0$ corresponds to Corollary 1 in Raymaekers & Rousseeuw (2021), so it suffices to prove the statement for the complementary case $\mathbb{E}_F[X] \neq 0$.

Let $X = Y$ so that $\rho = T(F) = 1$. From the maximum downward bias (Theorem 1) we know that $T(G)$ is bounded from below by

$$\frac{m_\varepsilon(X, X) - \varepsilon M_X^2}{m_\varepsilon(X) + \varepsilon M_X^2},$$

for any $G \in \mathcal{F}_\varepsilon$ and fixed $\varepsilon \in [0, 0.5]$. Lemma B.1 implies that this bound is nonpositive if

$$(1-\varepsilon)\left(\mathrm{Var}_F[X] + \mathbb{E}_F[X]^2\right) - (1-\varepsilon)^2\mathbb{E}_F[X]^2 - \varepsilon M_X^2 \leq 0,$$

which is equivalent to

$$\varepsilon^2\mathbb{E}_F[X]^2 + \varepsilon\Big(\mathrm{Var}_F[X] - \mathbb{E}_F[X]^2 + M_X^2\Big) - \mathrm{Var}_F[X] \geq 0.$$

This quadratic expression is satisfied with equality by

$$\varepsilon = \frac{\mathbb{E}_F[X]^2 - \mathrm{Var}_F[X] - M_X^2 + \sqrt{\left(\mathrm{Var}_F[X] - \mathbb{E}_F[X]^2 + M_X^2\right)^2 + 4\mathbb{E}_F[X]^2\,\mathrm{Var}_F[X]}}{2\mathbb{E}_F[X]^2},$$

which completes the proof. □

B.5 Proof of Corollary 1

To simplify notation, put $F = F_{\overline{X},\overline{Y}}$, $H = H_{\overline{X},\overline{Y}}$, $\rho = \rho_{\overline{X},\overline{Y}}$, and $G = (1-\varepsilon)F + \varepsilon H$. In addition, suppose that contaminating distribution H or model distribution F used as a subscript of an operator such as $\mathbb{E}_H$ is to be understood as either joint or marginal distribution; the type of which will be clear from the context.

Applying Lemma B.2 to distribution H yields

$$\mathbb{E}_H\left[\overline{X}\right] = \nu_X, \qquad \mathrm{Var}_H\left[\overline{X}\right] = \tau_X^2\big(1 + (J_X - 1)\phi_X\big)/J_X,$$

and, for distinct items $i \neq j$,

$$\mathbb{E}_H[X_i X_j] = \phi_X\tau_X^2 + \nu_X^2, \qquad \mathrm{Cov}_H[X_i, X_j] = \phi_X\tau_X^2.$$

Therefore,

$$\mathbb{E}_H\left[\overline{X}^2\right] = \mathrm{Var}_H\left[\overline{X}\right] + \mathbb{E}_H\left[\overline{X}\right]^2 = \tau_X^2\big(1 + (J_X - 1)\phi_X\big)/J_X + \tau_X^2,$$

and analogously for $\overline{Y}$ under H. Thus,

$$\begin{aligned}\mathbb{E}_H\left[\overline{XY}\right] &= \mathrm{Cov}_H\left[\overline{X},\overline{Y}\right] + \mathbb{E}_H\left[\overline{X}\right]\mathbb{E}_H\left[\overline{Y}\right] \\ &= \phi_{\overline{X},\overline{Y}}\sqrt{\mathrm{Var}_H\left[\overline{X}\right]\mathrm{Var}_H\left[\overline{Y}\right]} + \nu_X\nu_Y \\ &= \phi_{\overline{X},\overline{Y}}\sqrt{\frac{\tau_X^2\tau_Y^2}{J_XJ_Y}\big(1+(J_X-1)\phi_X\big)\big(1+(J_Y-1)\phi_Y\big)} + \nu_X\nu_Y.\end{aligned}$$

Applying Proposition 1 now yields, using the quantities just derived,

$$\begin{aligned}\mathrm{Cov}_G\left[\overline{X},\overline{Y}\right] &= m_\varepsilon(\overline{X},\overline{Y}) + \varepsilon\Bigg(-(1-\varepsilon)\mathbb{E}_F\left[\overline{X}\right]\mathbb{E}_H\left[\overline{Y}\right] - (1-\varepsilon)\mathbb{E}_F\left[\overline{Y}\right]\mathbb{E}_H\left[\overline{X}\right] + \\ &\qquad \mathbb{E}_H\left[\overline{XY}\right] - \varepsilon\mathbb{E}_H\left[\overline{X}\right]\mathbb{E}_H\left[\overline{Y}\right]\Bigg) \\ &= n_\varepsilon(\overline{X},\overline{Y}) + \varepsilon\Bigg((1-\varepsilon)\big(\nu_X\nu_Y - \mu_X\nu_Y - \mu_Y\nu_X\big) + \\ &\qquad \phi_{\overline{X},\overline{Y}}\sqrt{\frac{\tau_X^2\tau_Y^2}{J_XJ_Y}\big(1+(J_X-1)\phi_X\big)\big(1+(J_Y-1)\phi_Y\big)}\Bigg),\end{aligned}$$

and

$$\begin{aligned}\mathrm{Var}_G\left[\overline{X}\right] &= m_\varepsilon(\overline{X}) + \varepsilon\Big(-2(1-\varepsilon)\mathbb{E}_F\left[\overline{X}\right]\overline{HX} + \mathbb{E}_H\left[\overline{X}^2\right] - \varepsilon\mathbb{E}_H\left[\overline{X}\right]^2\Big) \\ &= n_\varepsilon(\overline{X}) + \varepsilon\Big((1-\varepsilon)\nu_X(\nu_X-2\mu_X) + \tau_X^2(1+(J_X-1)\phi_X)/J_X\Big),\end{aligned}$$

and analogously for $\mathrm{Var}_G\left[\overline{Y}\right]$. The result now follows from Proposition 1. □

B.6 Proof of Corollary 2

Evaluating the function $m_\varepsilon(X,Y)$ in (3) at scores $(\overline{X},\overline{Y})$ yields

$$\begin{aligned}m_\varepsilon(\overline{X},\overline{Y}) &= (1-\varepsilon)\Big(\rho_{\overline{X},\overline{Y}}\sqrt{\mathrm{Var}_{F_{\overline{X}}}\left[\overline{X}\right]\mathrm{Var}_{F_{\overline{Y}}}\left[\overline{Y}\right]} + \mathbb{E}_{F_{\overline{X}}}\left[\overline{X}\right]\mathbb{E}_{F_{\overline{Y}}}\left[\overline{X}\right]\Big) - \mathbb{E}_{F_{\overline{X}}}\left[\overline{X}\right]\mathbb{E}_{F_{\overline{Y}}}\left[\overline{X}\right], \\ &= (1-\varepsilon)\Bigg(\rho_{\overline{X},\overline{Y}}\sqrt{\frac{\sigma_X^2\sigma_Y^2}{J_XJ_Y}\big(1+(J_X-1)\rho_X\big)\big(1+(J_Y-1)\rho_Y\big)} + \mu_X\mu_Y\Bigg) - \\ &\qquad (1-\varepsilon)^2\mu_X\mu_Y,\end{aligned}$$

and, evaluating $m_\varepsilon(X)$ at $\overline{X}$,

$$\begin{aligned}m_\varepsilon(\overline{X}) &= (1-\varepsilon)\big(\mathrm{Var}_{F_{\overline{X}}}\left[\overline{X}\right] + \mathbb{E}_{F_{\overline{X}}}\left[\overline{X}\right]^2\big) - (1-\varepsilon)^2\mathbb{E}_{F_{\overline{X}}}\left[\overline{X}\right]^2 \\ &= (1-\varepsilon)\left(\frac{\sigma_X^2}{J_X}\big(1+(J_X-1)\rho_X\big) + \mu_X^2\right) - (1-\varepsilon)^2\mu_X^2,\end{aligned}$$

where we have used Equations (B.6a) and (B.6b) in Lemma B.2. To simplify notation, we denote the previous displays by $n_\varepsilon(\overline{X}, \overline{Y})$ and $n_\varepsilon(\overline{X})$, respectively, which motivates the definitions in Equation (5). Expression $n_\varepsilon(\overline{Y})$ is derived in analogue to $n_\varepsilon(\overline{X})$.

The result for the maxbias curves of Pearson correlation T evaluated at $F_{\overline{X},\overline{Y}}$ now follows from a direct application of Theorem 1, upon realizing that $\overline{X}$ and $\overline{Y}$ have support regions $\{-M_X, -(M_X - 1/J_X), \ldots, M_X - 1/J_X, M_X\}$ and $\{-M_Y, -(M_Y - 1/J_Y), \ldots, M_Y - 1/J_Y, M_Y\}$, respectively. □

B.7 Proof of Corollary 3

The result follows directly from Proposition 2 from the fact that the support region of $\overline{X}$ is given by $\{-M_X, \ldots, M_X\}$, that $\mu_X = 0$ implies that $\mathbb{E}_{F_{\overline{X}}}\left[\overline{X}\right] = 0$ and that $\mu_X \neq 0$ implies that $\mathbb{E}_{F_{\overline{X}}}\left[\overline{X}\right] \neq 0$ in the assumed setup. □

C Rescaling Support Regions

TODO

D Results for Mean and Variance

In this section, we derive (maximum) bias curves for the population mean and population variance of a rating scale variable X with finite support region of the form $\{-M_X, \ldots, M_X\}$. Like before, the possible realizations in the support region may be non-integer valued. Also like before, we denote by F the uncontaminated model distribution of X and by $\mathcal{F}_\varepsilon = \{G : G = (1-\varepsilon)F + \varepsilon H$ for any distribtion H with same support as $F\}$ the class of contaminated distributions associated with model distribution F.

Notably, we do not derive breakdown values for mean or variance in our setup with rating-scale variables. The reason is that breakdown values are traditionally defined as the smallest contamination fraction required for an arbitrarily large bias, that is, infinite maximum bias (e.g., Huber & Ronchetti, 2009; Maronna et al., 2018). While this definition is natural in continuous random variables (which are typically unbounded), it does not apply to rating-scale variables. Due to rating-scale variables being bounded by construction, the maximum bias of any statistic will be bounded as well. In the special case of Pearson correlation where the statistic is bounded by construction (namely by ± 1), the maximum bias will be bounded as well, which necessitates a refined definition of the breakdown value. Such a refined definition—originally proposed by Capéraà & Guillem (1997)—is given in Definition 1. Proposing similar refined definitions of breakdown values for population mean and population variance when they are known to be bounded by construction (like when variables are bounded) is beyond the scope of this paper and we therefore cannot derive breakdown values for means and variances in our setting with rating-scale variables.

For (maximum) bias curves of mean and variance in our setting, we first establish general propositions, and thereupon derive useful corollaries for additive rating-scale variables, like a (mean) score of a personality trait.

We start with bias curves at a specific contaminated distribution.

Proposition D.1 (Bias curves at specific contaminated distribution)**.** *Let X be a discrete ordinal variable with support region $\{-M_X, \ldots, M_X\}$. For fixed contamination fraction $\varepsilon \in [0, 0.5]$ and contaminated distribution $G = (1-\varepsilon)F + \varepsilon H \in \mathcal{F}_\varepsilon$, the bias of the population mean "$\mathbb{E}$" of X at model distribution F is given by*

$$\mathbb{E}_G[X] - \mathbb{E}_F[X] = \varepsilon\Big(\mathbb{E}_H[X] - \mathbb{E}_F[X]\Big),$$

and the bias of population variance "Var" *at F is given by*

$$\begin{aligned}\mathrm{Var}_G[X] - \mathrm{Var}_F[X] = m_\varepsilon(X) + \\ \varepsilon\bigg((1-\varepsilon)\mathbb{E}_H[X]\Big(\mathbb{E}_H[X] - 2\mathbb{E}_F[X]\Big) + \mathrm{Var}_H[X]\bigg) - \mathrm{Var}_F[X],\end{aligned}$$

where the function $m_\varepsilon(X)$ is defined in (3).

Proof. For the mean, the result follows immediately from $\mathbb{E}_F[G] = (1-\varepsilon)\mathbb{E}_F[X] + \varepsilon\mathbb{E}_H[X]$. For the variance, simple algebra shows that

$$\begin{aligned}\mathbb{E}_G\left[X^2\right] &= (1-\varepsilon)\mathbb{E}_F\left[X^2\right] + \varepsilon\mathbb{E}_H\left[X^2\right], \\ \mathbb{E}_G[X]^2 &= (1-\varepsilon)^2\mathbb{E}_F[X]^2 + \varepsilon^2\mathbb{E}_H[X]^2 + 2(1-\varepsilon)\varepsilon\mathbb{E}_F[X]\,\mathbb{E}_H[X].\end{aligned} \tag{D.13}$$

Then

$$\begin{aligned}\mathrm{Var}_G[X] &= \mathbb{E}_G\left[X^2\right] - \mathbb{E}_G[X]^2 \\ &= m_\varepsilon(X) + \varepsilon\Big(-2(1-\varepsilon)\varepsilon\mathbb{E}_F[X]\,\mathbb{E}_H[X] + \mathbb{E}_H\left[X^2\right] - \varepsilon\mathbb{E}_H[X]^2\Big) \\ &= m_\varepsilon(X) + \varepsilon\bigg((1-\varepsilon)\mathbb{E}_H[X]\Big(\mathbb{E}_H[X] - 2\mathbb{E}_F[X]\Big) + \mathrm{Var}_H[X]\bigg),\end{aligned}$$

thereby concluding the proof. □

Next, we derive maximum bias curves in the following proposition.

Proposition D.2 (Maximum bias curves)**.** *Let X be a discrete ordinal variable with support region $\{-M_X, \ldots, M_X\}$. For fixed contamination fraction $\varepsilon \in [0, 0.5]$, the maximum bias upward of the mean at model distribution F is given by*

$$B^+(\varepsilon, \mathbb{E}, F) = \varepsilon\Big(M_X - \mathbb{E}_F[X]\Big),$$

and the maximum downward bias is

$$B^{-}(\varepsilon, \mathbb{E}, F) = -\varepsilon\Big(M_X + \mathbb{E}_F\left[X\right]\Big).$$

The maximum bias of the variance at model distribution F is given by

$$B(\varepsilon, \mathrm{Var}, F) = m_\varepsilon(X) + \varepsilon M_X^2 - \mathrm{Var}_F\left[X\right].$$

Proof. Let $G = (1-\varepsilon)F + \varepsilon H \in \mathcal{F}_\varepsilon$ be arbitrary.

Maximum bias for expectation. Denote by $T(F) = \mathbb{E}_F\left[X\right]$ the functional form of the mean at distribution F. We prove the result for maximum upward bias in the following. The result for maximum downward bias follows by replacing X in by $-X$ in the steps below. We have that

$$\begin{aligned}\mathbb{E}_G\left[X\right] &= (1-\varepsilon)\mathbb{E}_F\left[X\right] + \varepsilon\mathbb{E}_H\left[X\right] \\ &\leq (1-\varepsilon)\mathbb{E}_F\left[X\right] + \varepsilon\mathbb{E}_H\left[|X|\right] \\ &\leq (1-\varepsilon)\mathbb{E}_F\left[X\right] + \varepsilon M_X.\end{aligned}$$

We now verify that this bound is sharp. The contamination distribution corresponding to the worst-placed contamination point is given by $H' := \Delta_{x}$, at which the expectation takes value $\mathbb{E}_{H'}\left[X\right] = M_X$ and. For $G' := (1-\varepsilon)F + \varepsilon H'$, we have that $\mathbb{E}_{G'}\left[X\right] = (1-\varepsilon)\mathbb{E}_F\left[X\right] + \varepsilon M_X$, which verifies sharpness of the bound derived in the preceding display and thereby the proof for the maximum upward bias. The proof for the maximum downward bias follows by repeating these steps for X replaced by $-X$ and H' replaced by $H' = \Delta_{(-M_X)}$.

We proceed by proving the claimed maximum bias curve for variances.

Maximum bias for variance. We have that

$$\begin{aligned}\mathrm{Var}_G\left[X\right] &= \mathbb{E}_G\left[X^2\right] - \mathbb{E}_G\left[X\right]^2 \\ &= m_\varepsilon(X) + \varepsilon\Big(-2(1-\varepsilon)\varepsilon\mathbb{E}_F\left[X\right]\mathbb{E}_H\left[X\right] + \mathbb{E}_H\left[X^2\right] - \varepsilon\mathbb{E}_H\left[X\right]^2\Big),\end{aligned}$$

where we have used (D.13). This term increases with $\mathbb{E}_H\left[X^2\right] - \varepsilon\mathbb{E}_H\left[X\right]^2$, which is maximized when $\mathbb{E}_H\left[X^2\right] = M_X^2$ and $\mathbb{E}_H\left[X\right] = 0$. It follows that $\mathrm{Var}_G\left[X\right]$ is bounded from above by

$$m_\varepsilon(X) + \varepsilon M_X^2.$$

To verify sharpness of this bound, consider the contaminating distribution at the worst-placed point, being

$$H' := \frac{1}{2}\Delta_{M_X} + \frac{1}{2}\Delta_{(-M_X)}.$$

Note that $\mathbb{E}_{H'}\left[X\right] = 0$ and $\mathbb{E}_{H'}\left[X^2\right] = M_X^2$. Then, evaluating the variance at contaminated distribution $G' := (1-\varepsilon)F + \varepsilon H'$ yields $\mathrm{Var}_{G'}\left[X\right] = m_\varepsilon(X) + \varepsilon M_X^2$, thereby verifying sharpness of the derived bound. This completes the proof. □

We now derive corollaries for the additive scores. Specifically, Corollary D.1 provides bias curves at a specific contaminated distribution and Corollary D.2 the maximum bias curves.

Corollary D.1. *Assume the setup and notation of Assumption 1, where* $\overline{X} = J_X^{-1}\sum_{j=1}^{J_X} X_j$ *obeys model distribution* $F_{\overline{X}}$ *and the individual variables* X_j *are identically distributed according to model distribution* F_X *. Let* H_X *be a distribution of the same support as* F_X *such that the individual variables* X_j *are identically distributed under* H_X*. Denote by* ν_X *and* τ_X^2 *the mean and variance, respectively, of distribution* H_X*. In addition, for any two distinct items* $i \neq j$*, put* $\phi_X = \mathrm{Cor}_{H_X}[X_i, X_j]$*. Let* $H_{\overline{X}}$ *be the contaminating joint distribution of scores* $\overline{X}$ *that is implied by* H_X*. For contamination fraction* $\varepsilon \in [0, 0.5]$*, let* $G_{\overline{X}} = (1-\varepsilon)F_{\overline{X}} + \varepsilon H_{\overline{X}}$ *be the contamination distribution implied by* $F_{\overline{X}}$ *and* $H_{\overline{X}}$*. Then at contamination distribution* $G_{\overline{X}}$*, the bias of the population mean is given by*

$$\mathbb{E}_{G_{\overline{X}}}\left[\overline{X}\right] - \mathbb{E}_{F_{\overline{X}}}\left[\overline{X}\right] = \varepsilon(\nu_X - \mu_X),$$

and the bias of the population variance is given by

$$\begin{aligned}\mathrm{Var}_{G_{\overline{X}}}\left[\overline{X}\right] - \mathrm{Var}_{F_{\overline{X}}}\left[\overline{X}\right] &= n_\varepsilon(\overline{X}) + \\ &\varepsilon\left((1-\varepsilon)\nu_X(nu_X - 2\mu_X) + \frac{\tau_X^2}{J_X}\big(1 + (J_X-1)\phi_X\big)\right) - \frac{\sigma_X^2}{J_X}\big(1 + (J_X-1)\rho_X\big).\end{aligned}$$

Proof. To simplify notation, put $F = F_{\overline{X}}, H = H_{\overline{X}}$, and $G = G_{\overline{X}}$. Applying Lemma B.2 to distributions F and H yields

$$\begin{aligned}\mathbb{E}_F\left[\overline{X}\right] &= \mu_X, \\ \mathbb{E}_H\left[\overline{X}\right] &= \nu_X, \\ \mathrm{Var}_F\left[\overline{X}\right] &= \sigma_X^2\big(1 + (J_X-1)\rho_X\big)/J_X, \qquad \text{and} \\ \mathrm{Var}_H\left[\overline{X}\right] &= \tau_X^2\big(1 + (J_X-1)\phi_X\big)/J_X.\end{aligned} \tag{D.14}$$

Then, applying Proposition D.1,

$$\mathbb{E}_G\left[\overline{X}\right] - \mathbb{E}_F\left[\overline{X}\right] = \varepsilon\big(\mathbb{E}_H\left[\overline{X}\right] - \mathbb{E}_F\left[\overline{X}\right]\big) = \varepsilon\big(\nu_X - \mu_X\big),$$

and

$$\begin{aligned}&\mathrm{Var}_G\left[\overline{X}\right] - \mathrm{Var}_F\left[\overline{X}\right] \\ &\quad= m_\varepsilon(\overline{X}) + \varepsilon\Big((1-\varepsilon)\mathbb{E}_H\left[\overline{X}\right]\big(\mathbb{E}_H\left[\overline{X}\right] - 2\mathbb{E}_F\left[\overline{X}\right]\big) + \mathrm{Var}_H\left[\overline{X}\right]\Big) - \mathrm{Var}_F\left[\overline{X}\right] \\ &\quad= n_\varepsilon(\overline{X}) + \varepsilon\left((1-\varepsilon)\nu_X(\nu_X - 2\mu_X) + \frac{\tau_X^2}{J_X}\big(1 + (J_X-1)\phi_X\big)\right) - \frac{\sigma_X^2}{J_X}\big(1 + (J_X-1)\rho_X\big).\end{aligned}$$

This concludes the proof. □

Corollary D.2. *Assume the setup and notation of Assumption 1, where* $\overline{X} = J_X^{-1} \sum_{j=1}^{J_X} X_j$ *obeys model distribution* $F_{\overline{X}}$ *and the individual variables* X_j *are identically distributed according to model distribution* F_X*. For fixed contamination fraction* $\varepsilon \in [0, 0.5]$*, the maximum upward bias and maximum downward bias of the population mean of* $\overline{X}$ *at model distribution* $F_{\overline{X}}$ *are respectively given by*

$$B^+(\varepsilon, \mathbb{E}, F_{\overline{X}}) = \varepsilon\big(M_X - \mu_X\big) \qquad \textit{and}$$
$$B^-(\varepsilon, \mathbb{E}, F_{\overline{X}}) = -\varepsilon\big(M_X + \mu_X\big),$$

and the maximum bias of the population variance is given by

$$B(\varepsilon, \mathrm{Var}, F_{\overline{X}}) = n_\varepsilon(\overline{X}) + \varepsilon M_X^2 - \sigma^2\big(1 + (J_X - 1)\rho_X\big)/J_X.$$

Proof. The result follows from a direct application of Proposition D.2 on $\overline{X}$ (which has support region $\{-M_X, \ldots, M_X\}$) and (D.14). □

E Influence Functions

Definition 3 (Influence Function (Hampel, 1974))**.** The *influence function* of Pearson correlation measure T at model distribution F is defined as

$$\mathrm{IF}\big((x,y), T, F\big) = \lim_{\varepsilon \downarrow 0} \frac{T\big((1-\varepsilon)F + \varepsilon\Delta_x\big) - T(F)}{\varepsilon} = \frac{\partial}{\partial\varepsilon} T\big((1-\varepsilon)F + \varepsilon\Delta_{(x,y)}\big)\Big|_{\varepsilon=0},$$

where (x, y) are points in the support of F, and $(s,t) \mapsto \Delta_{(x,y)}(s,t) = \mathbb{1}\{x \le s, y \le t\}$ is a bivariate distribution function that puts all its mass at point (x, y).

The influence function is a Gâteaux derivative (which in turn is a generalization of directional derivatives) and measures how $T(F)$ reacts when an infinitesimally small amount of contamination is added in point (x, y). A textbook treatment on influence functions is given by Hampel et al. (1986).

A concept closely related to the influence function is the gross-error sensitivity, which is defined as the maximum value of the influence function for a given estimator and distribution.

Definition 4 (Gross-error sensitivity)**.** The gross-error sensitivity of Pearson correlation measure T at model distribution F is defined as

$$\gamma^*(T, F) = \sup_{(x,y)} \Big|\mathrm{IF}\big((x,y), T, F\big)\Big|,$$

where the supremum is taken over the support of F.

We now state the influence function and gross-error sensitivity of Pearson correlation for discrete ordinal variables.

Theorem 2. *Let X and Y be discrete ordinal random variables with support regions $S_X = \{-M_X, \ldots, M_X\}$ and $S_Y = \{-M_Y, \ldots, M_Y\}$, respectively. The influence function of Pearson correlation measure T at model distribution F is given by*

$$\mathrm{IF}\big((x,y),T,F\big) = \frac{\big(x - \mathbb{E}_F[X]\big)\big(y - \mathbb{E}_F[Y]\big)}{\sqrt{\mathrm{Var}_F[X]}\sqrt{\mathrm{Var}_F[Y]}} - \rho\Bigg(1 + \frac{1}{2}\big(x - \mathbb{E}_F[X]\big)^2\mathrm{Var}_F[Y] + \frac{1}{2}\big(y - \mathbb{E}_F[Y]\big)^2\mathrm{Var}_F[X] - \mathrm{Var}_F[X]\,\mathrm{Var}_F[Y]\Bigg),$$

where $\rho = T(F)$ and $x \in S_X$ as well as $y \in S_Y$.

Remark 2. If $\mathbb{E}_F[X] = \mathbb{E}_F[Y] = 0$ and $\mathrm{Var}_F[X] = \mathrm{Var}_F[Y] = 1$, then $\mathrm{IF}\big((x,y),T,F\big) = xy - (x^2+y^2)\rho/2$, which is a classic result derived in Devlin et al. (1975) for the influence function of Pearson's correlation measure for standardized random variables.

Proof. For arbitrary $\varepsilon \in [0, 0.5]$, let $F_\varepsilon = (1-\varepsilon)F + \varepsilon\Delta_{(x,y)} \in \mathcal{F}_\varepsilon$ be the contaminated distribution of F associated with point mass contamination distribution $\Delta_{(x,y)}$ at $x \in S_X, y \in S_Y$. □

Proposition E.1 (Influence functions of covariance)**.** *Let X and Y be arbitrary random variables with support regions S_X and S_Y, respectively, and let $C(F) = \mathrm{Cov}_F[X,Y]$ be a population covariance between X and Y at joint model distribution F, expressed as a statistical functional. The influence function of covariance C at F is given by*

$$\mathrm{IF}\big((x,y),C,F\big) = -C(F) + (x - \mathbb{E}_F[X])(y - \mathbb{E}_F[Y]).$$

Proof. For arbitrary $\varepsilon \in [0, 0.5]$, let $F_\varepsilon = (1-\varepsilon)F + \varepsilon\Delta_{(x,y)} \in \mathcal{F}_\varepsilon$ be the contaminated distribution of F associated with point mass contamination distribution $\Delta_{(x,y)}$ at $x \in S_X, y \in S_Y$. Let

$$C(F) := \mathrm{Cov}_F[X,Y] = \mathbb{E}_F[XY] - \mathbb{E}_F[X]\,\mathbb{E}_F[Y]$$

denote a functional version of the population covariance at F. Then, evaluating covariance at contaminated distribution F_ε yields

$$\begin{aligned} C(F_\varepsilon) &= \int st\ \mathrm{d}F_\varepsilon(s,t) - \left(\int s\ \mathrm{d}F_\varepsilon(s,t)\right)\left(\int t\ \mathrm{d}F_\varepsilon(s,t)\right) \\ &= (1-\varepsilon)\mathbb{E}_F[XY] + \varepsilon xy - \big((1-\varepsilon)\mathbb{E}_F[X] + \varepsilon x\big)\big((1-\varepsilon)\mathbb{E}_F[Y] + \varepsilon y\big) \\ &= (1-\varepsilon)\mathbb{E}_F[XY] + \varepsilon xy - \Big((1-\varepsilon)^2\mathbb{E}_F[X]\,\mathbb{E}_F[Y] + (1-\varepsilon)\varepsilon y\mathbb{E}_F[X] + (1-\varepsilon)\varepsilon x\mathbb{E}_F[Y] + \varepsilon^2 xy\Big), \end{aligned}$$

where we have used that $\int \mathrm{d}\Delta_{(x,y)}(s,t) = \int \mathrm{d}\Delta_{(x)}(s) = x$ when t is fixed. Differentiating with with respect to ε and evaluating at $\varepsilon = 0$,

$$\begin{aligned}
&\frac{\partial}{\partial\varepsilon} C(F_\varepsilon)\Big|_{\varepsilon=0} \\
&\quad = -\mathbb{E}_F\left[XY\right] + xy - \Big(-2(1-\varepsilon)\mathbb{E}_F\left[X\right]\mathbb{E}_F\left[Y\right] + (1-2\varepsilon)y\mathbb{E}_F\left[X\right] + (1-2\varepsilon)x\mathbb{E}_F\left[Y\right] + 2\varepsilon xy\Big)\Big|_{\varepsilon=0} \\
&\quad = -\mathbb{E}_F\left[XY\right] + xy + \mathbb{E}_F\left[X\right]\mathbb{E}_F\left[Y\right] + \mathbb{E}_F\left[X\right]\mathbb{E}_F\left[Y\right] - y\mathbb{E}_F\left[X\right] - x\mathbb{E}_F\left[Y\right] \\
&\quad = -C(F) + (x - \mathbb{E}_F\left[X\right])(y - \mathbb{E}_F\left[Y\right]),
\end{aligned}$$

which concludes the proof because the left-hand side corresponds to the definition of the influence function of the population covariance. □

Corollary E.1. *Let X be an arbitrary random variable with support region S_X. At some model distribution F, the influence function of variance $V(F) = \mathrm{Var}_F\left[X\right]$ is given by*

$$\mathrm{IF}(x, V, F) = -V(F) + (x - \mathbb{E}_F\left[X\right])^2,$$

and the influence function of the corresponding standard deviation $S(F) = \sqrt{V(F)}$ is given by

$$\mathrm{IF}(x, S, F) = \frac{1}{\sqrt{V(F)}}\Big(\big(x - \mathbb{E}_F\left[X\right]\big)^2 - V(F)\Big).$$

Proof. The result for the influence function of the variance follows immediately from evaluating Proposition E.1 at $Y = X$ because the covariance of two identical variables X and X equals the variance of X. □