

Discovering Careless Response Behavior in Psychometric Data

^{1,2}René Lehmann ²Paul Bengart ²Bodo Vogt
¹FOM University of Applied Science ²Otto von Guericke University ²Otto von Guericke University
rene.lehmann@fom.de paul.bengart@med.ovgu.de bodo.vogt@ovgu.de

Abstract

In psychological healthcare considering bipolar Likert scales data as compositional data can enhance statistical validity. Applying an isometric log-ratio transformation yields interval scaled real-valued data. It increases the normal approximation of item response means, reduces statistical biases and enhances the statistical power of the Pearson correlation test and two-sample t-tests (paired and unpaired) affecting linear regression, partial least squares path modeling and moderator analysis. Mental overload, missing attention, faking or social desirability can corrupt a test person's answers in a psychometric survey. As a result, the corresponding questionnaire data are useless affecting subsequent analyses and interpretations. Aiming to detect careless response behavior as statistical outliers we compare the well-known Mahalanobis-distance to a multivariate projection pursuit method. Performing outlier detections with traditional and with isometric log-ratio transformed data we point out the superiority of the compositional data interpretation of psychometric bipolar scales data.

Keywords: bipolar likert scale, isometric log-ratio transformation, careless response, mahalanobis-distance, projection pursuit

1. Introduction

A reduced data quality not only affects the statistical analysis. Biased parameter estimates can contribute to mischaracterizing patients and patient groups, potentially leading to increased treatment costs or economically unsustainable

decisions (Lehmann and Vogt, 2023b). From a microeconomic perspective, management disadvantages can also arise, for example when the psychometric profiles of patient groups are inaccurately estimated (Lehmann and Vogt, 2024b). In the context of psychometric scale development and validation, careless response behavior affects the assessment of convergent and discriminant validity (Huang et al., 2015; Lehmann and Vogt, 2023a; Maniaci and Rogge, 2014).

Even for surveys with $m = 20$ items and $N = 100$ participants, a dataset of $m \cdot N = 2000$ item responses is generated. Visually inspecting this dataset for patterns or careless response behavior is very time-consuming and potentially impossible. To detect careless response behavior, truth scales (Lanz et al., 2021) and scales measuring socially desirable behavior can be used (Satow, 2012). However, using additional scales to assess the quality of responses may impose additional cognitive burden on the test subjects. This can lead to mental fatigue effects, which could ultimately have a negative impact on the validity of the measurements. If one wishes to avoid additional scales, it is possible, for example, to consider the average response time per item. If this is too short, the items may have been clicked through too quickly and not read properly (Curran, 2016). Searching for patterns (Sjölander et al., 2014) or autocorrelated response behavior (Gottfried et al., 2022) could also be helpful.

We propose interpreting the m item responses of a test subject as an m -dimensional vector according to Maniaci and Rogge, 2014. Rapid clicking through or the use of a response pattern

could be indicated by an unusual combination of item responses compared to carefully chosen item responses. For example, rapid clicking could result in completely random item responses (Type A), a tendency towards the same value (Type B), or an arbitrary pattern in the responses (Type C). Using statistical outlier analysis, it is possible to detect such unusual combinations (Filzmoser and Hron, 2008; Filzmoser et al., 2005; Lehmann, 2012, 2014; Pena and Prieto, 2001b).

Lehmann and Vogt, 2023a, 2024a, 2024b, 2024c proposed to interpret bipolar psychometric data as compositional data. Using an isometric log-ratio (ilr) transformation, the data can be transferred to the real-valued interval scale Aitchison, 2003a. The question arises whether the detection of careless responses using outlier analytical statistical methods is more successful when the data analysis is conducted traditionally or with the ilr transformation. In addition to the Mahalanobis-distance (MD) proposed by Maniaci and Rogge, 2014 for detecting multivariate outliers in psychometric data, there are methods based on projection pursuit (PP) of multivariate data (Pan et al., 2000; Pena and Prieto, 2001b). We apply the MD (Filzmoser et al., 2005; Maniaci and Rogge, 2014) and the PP method of Pena and Prieto, 2001b to identify careless response behavior in the form of statistical outliers. Due to the traditionally discrete bipolar psychometric response scales, it sometimes happens that the estimated covariance matrix becomes singular and its inverse does not exist when the sample size N is not sufficiently large, making the calculation of the MD impossible. Unlike the MD, the covariance matrix is not required for the PP method (Lehmann, 2012, 2014).

Adding outliers of type A, B and C to a real Big-5 personality traits data set we compare the MD and the PP approach using both the traditional and the ilr transformed data.

2. Literature Review

Uncovering careless response behavior is a methodologically broad research field in psychometric statistics. For example, Gottfried et al., 2022 examine item responses for autocorrelation, while Sjölander et al., 2014 choose a partial least squares (PLS) approach to find patterns in the m -dimensional vectors of individual item responses. Buchanan and Scofield, 2018 detect low-quality item responses using

a complex algorithm based on the number of utilized choices, page response times, click counts, and the distribution of survey responses. The long string analysis interprets each test subject's responses as an m -dimensional vector and looks for the longest sequence of identical responses within a vector to identify careless response behavior (Huang et al., 2011; Meade and Craig, 2012). Unusual combinations of responses can be identified as statistical outliers based on the MD (Maniaci and Rogge, 2014). Curran, 2016 recommend using multiple analytical tools to uncover careless response behavior.

Recently, Lehmann and Vogt, 2023a proposed to consider psychological bipolar scales data as compositional data assuming that any order of magnitude of agreement (OMA) towards an item assertion implies a complementary order of magnitude of disagreement (OMD). Traditionally, the statistical analysis of bivariate psychometric Likert scale data focuses on the OMA data (Likert, 1932). Lehmann and Vogt, 2023a proposed to consider both the OMA and the OMD in order to reduce statistical biases (Filzmoser and Hron, 2009; Filzmoser et al., 2009) and increase the normal approximation of item response means (Lehmann and Vogt, 2024a, 2024c).

According to Aitchison, 1986; Aitchison, Mateu-Figueras, and Ng, 2003, the bivariate compositional OMA-OMD-pairs can be transferred towards the univariate real-valued interval scale using an ilr transformation. The central limit theorem (CLT) of statistics in its various versions postulates the approximate normal distribution of OMA means (Fischer, 2011). However, Lehmann and Vogt, 2024a, 2024c provided evidence that means of ilr transformed OMA-OMD-pairs are more likely normally distributed than means of untransformed item responses. As a result of the increased convergence towards normality and the ilr-induced reduction of bias the statistical power of the well-known correlation test and the two-samples t-test (paired and unpaired) increases (Lehmann and Vogt, 2024b; Lehmann and Vogt, 2023b). Lehmann and Vogt, 2024d show that the ilr approach increases the statistical power of the correlation test if the distribution of item response means is heavy-tailed pointing out the robustness of the ilr approach. Depending on the normal distribution the MD should yield more reliable results when applied to the ilr transformed data.

Many statistical procedures suffer from outliers (e.g. the computation of (item response) means

and correlation tests) but the term "outlier" is not clearly defined in statistics Lehmann, 2012. There are various qualitatively different approaches. For example, outliers are interpreted as values that are far from the center of the data (Filzmoser et al., 2005). On the other hand, the concept of outlyingness is used to describe how far a data point is from the rest of the data (Gather et al., 2004). The MD is based on the Euclidean metric. However, compositional data are subject to the Aitchison metric (Aitchison and Egozcue, 2005). Filzmoser and Hron, 2008 show that the MD is only suitable for outlier analysis of compositional data when applied to ilr-transformed data. Considering that Lehmann and Vogt, 2023a; Lehmann and Vogt, 2023b have demonstrated the compositional data structure of bipolar psychometric data, this contradicts the suggestion by Maniaci and Rogge, 2014 to apply the MD to untransformed item responses. The PP method also relies on the Euclidean metric and should therefore be applied to ilr-transformed item responses (Lehmann, 2012, 2014). The use of statistical methods based on the Euclidean metric always introduces bias when applied to compositional data Filzmoser et al., 2009. Therefore, it is expected that outlier analysis using the MD and the PP method will provide higher quality results if bipolar questionnaire data have been subjected to an ilr transformation beforehand.

3. The new compositional data approach in brief

For proper understanding of the ilr approach consider the different types of psychological scales. It is necessary to distinguish between statements (i.e., items of a questionnaire) and their corresponding RS as well as a Likert scale (LS, i.e., a set of items represented by the sum or mean value of their corresponding responses) and the scale of a personality trait or state (TS, i.e., the continuum of all possible manifestations of a trait or state). Associating verbal responses (e.g., ranging from "not at all" to "very much") with numerical values (e.g., 1, ..., 5) is common practice (Pennycook et al., 2021). The RS directly quantifies the OMA and indirectly quantifies the OMD of a test person's agreement and disagreement towards a statement. The LS represents a model of the TS for estimating the order of magnitude of a personality trait or state (OMT) (Likert, 1932). In

the following, if not otherwise stated, the term scale refers to a bipolar scale. Fig. 1 provides an illustration.

The items of a questionnaire (e.g., the BFI-10 inventory of Rammstedt et al., 2014; Rammstedt and John, 2007) cover specific aspects of a psychological construct. Considering an overall value of the item responses (e.g., the arithmetic mean) provides an individual estimate of the order of magnitude of the psychological construct. Due to imperfect knowledge, uncertainty about situations and a complex environment the psychometric scale cannot cover all individual manifestations of the psychological construct (see James and Wood, 1988; Loke, 1989; Romano et al., 2016) implying the existence of a limit of quantification (LOQ) (Lehmann and Vogt, 2023a).

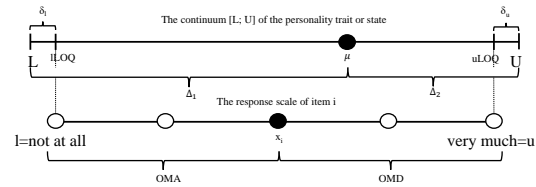


Figure 1: Illustration of the different types of scales used in psychometrics. The upper continuum [L; U] represents the TS. The lower scale represents the RS. Figure according to Lehmann and Vogt, 2024d

The continuum [L; U] contains all possible individual manifestations of a construct ranging from a minimum value L (e.g., non-openness to anything) to a maximum value U (e.g., openness to everything). A person's order of magnitude of the construct (say, μ) is located within these bounds. Moreover, the complements Δ_1 and Δ_2 , both represent the order of magnitude of the construct. We have $\Delta_1 + \Delta_2 = U - L$. For example, set $L=0, U=100, \mu = 70, \Delta_1 = 70$ and $\Delta_2 = 30$.

The psychometric scale consists of different items $i = 1, \dots, I$ associated with a RS, e.g., ranging from l="not at all" to u="very much". As the items cannot cover all aspects of the construct the lower (l) and upper (u) limit of the RS are different from L and U reflecting the lower (ILOQ) and upper (uLOQ) limit of quantification. The edge area of the construct scale which is not covered by the items and their respective RS are named δ_l and δ_u .

Any response x_i towards an item assertion

reflects an OMA and an OMD towards the item assertion. For example, let $l=ILOQ=2.5$, $u=uLOQ=97.5$, $x_i = 73.75 = OMA$, $OMD = 26.25$, $\delta_l = [0; 2.5)$ and $\delta_u = (97.5; 100]$). That is, x_i estimates the unknown value of μ and the pair $(73.75, 26.25)^T$ denotes a so-called (bivariate) compositional data point.

According to Lehmann and Vogt, 2024c and Lehmann and Vogt, 2024a we set $lLOQ = 100 \cdot (p/2)$ and $uLOQ = 100 \cdot (1 - p/2)$ with $p = LOQ = 0.1$. Lehmann and Vogt, 2023a; Lehmann and Vogt, 2023b provide a detailed introduction to the compositional data approach in psychometrics.

The present compositional data space is defined as $S := \{x = (OMA, OMD)^T \in \mathbb{R}^2 \mid OMA + OMD = 100, OMA, OMD > 0\}$ (Aitchison and Egozcue, 2005; Aitchison, Mateu-Figueras, and Ng, 2003; Aitchison, 2003b; Aitchison, Mateu-Figueras, and Ng, 2003; Lehmann and Vogt, 2023a). Fig. 2 presents an illustration of the Simplex of bipolar scales data.

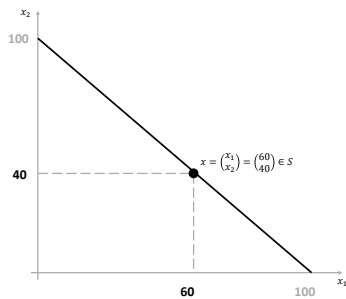


Figure 2: The black line illustrates the Simplex of bipolar scales data. x_1 (x_2) represents the OMA (OMD) towards the item assertion, respectively. The exemplary point $x = (60, 40)^T$ illustrates an OMA of 60 and an OMD of 40. Figure according to Lehmann and Vogt, 2023b

The ilr transformation is defined as $ilr((OMA, OMD)^T) = \sqrt{0.5} \ln \frac{OMA}{OMD}$. Please note that the ilr transformation is nearly identical with the additive log-ratio, the centered log-ratio and the logit transformation. Lehmann and Vogt, 2024d provide a deeper introduction to the different types of data transformation and discuss advantages and disadvantages.

4. Outlier detection

This subsection provides a brief introduction to statistical outlier detection via the MD (Filzmoser and Hron, 2008; Filzmoser et al., 2005) and the PP method of Pena and Prieto, 2001b. Details

are provided by Filzmoser et al., 2005; Lehmann, 2012; Pena and Prieto, 2001a, 2001b.

To answer the central question whether or not the ilr approach improves outlier detection in bipolar psychometric questionnaire data the initial data sets must be free of outliers. Then, imputing artificial outliers we can compute the correct discovery rate (CDR) of the MD and the PP with untransformed and ilr transformed data allowing us to assess and compare both approaches (traditional vs. ilr) and methods (MD vs. PP).

4.1. The Mahalanobis-distance

Mahalanobis, 1936 developed a measure to determine the distance between two multidimensional points. Let $x, y \in \mathbb{R}^m$ be realizations of stochastically independent and identically multivariately distributed random variables X and Y . The covariance matrix of the distribution of X is denoted as S . The distance between x and y is then given by Mahalanobis

as $d(x, y) = \sqrt{(x - y)' S^{-1} (x - y)}$ where x' denotes the transpose. If data $x_1, \dots, x_n \in \mathbb{R}^m$ are given data points, they are considered as realizations of stochastically independent and identically multivariately distributed random variables X_1, \dots, X_n . The MD of a data point $x_i \in x_1, \dots, x_n$ to the center of the data is defined by

$$MD(x_i) = \sqrt{(x_i - \hat{\mu})' \hat{S}^{-1} (x_i - \hat{\mu})} \quad (1)$$

where $\hat{\mu}$ and \hat{S} are estimates of location and scale based on the data x_1, \dots, x_n . The MD is obviously only robust against outliers when robust estimates of location and scale are used. Filzmoser and Hron, 2008 propose the MCD estimator. In the case of multivariately normally distributed data, the squared MD is approximately χ_m^2 distributed (Filzmoser et al., 2005; Hardin and Rocke, 2005). If the squared MD of a data point x_i is greater than $\chi_{1-\alpha, m}^2$ it is considered an outlier (Filzmoser and Hron, 2008). Here, $\chi_{1-\alpha, m}^2$ represents the $(1 - \alpha)$ quantile of the χ_m^2 distribution. To calculate the robust MD using the MCD estimator, we use the R package chemometrics (Version 1.4.4) (Varmuza and Filzmoser, 2009). In the case of not multivariately normally distributed data Filzmoser et al., 2005 and Lehmann, 2012 provide critical values.

4.2. The PP method

Pena and Prieto, 2001b suggest identifying multivariate outliers using projection methods. The projection directions are to be chosen based on the kurtosis coefficient of the multivariate sample and the resulting test statistic is asymptotically χ^2 -distributed. Pena and Prieto, 2001b argue that a low contamination of data by outliers would increase the kurtosis coefficient. This would be the case regardless of whether the contamination is symmetric or asymmetric, meaning whether there is contamination that manifests itself more in the "upper or lower range" of the data, or whether there is contamination by both high and low values equally. Therefore, it makes sense to choose the projection direction of multivariate data in such a way that the resulting univariate data exhibit a large kurtosis coefficient when detecting statistical outliers. On the other hand, it is possible that in univariate data, a large asymmetric contamination by outliers (e.g., a cluster of statistical outliers) can lead to a very small kurtosis coefficient. This suggests choosing the projection direction of multivariate data such that the kurtosis coefficient of the projected data is as small as possible. Furthermore, it is possible to detect outliers by choosing a projection direction that is perpendicular to the direction maximizing or minimizing the kurtosis coefficient (Pena and Prieto, 2001b). The projected univariate data are further examined using a conventional robust univariate method (e.g., the Hampel identifier, see Hampel et al., 2005) to identify and iteratively clean them from statistical outliers. Lehmann, 2012 provides a detailed algorithm for determining the projection directions.

5. Data collection and evaluation

We conducted the data collection using the BFI-10 scale (Rammstedt et al., 2014; Rammstedt and John, 2007). A total of 413 individuals started the survey, out of which 293 completed it (169 females; $M_{age} = 37.00$, $SD_{age} = 10.26$). The BFI-10 measures the Big five personality traits using two items per trait. Traditionally, it is given by the response scale $\{1, \dots, 5\}$ with 1="not at all" to 5="fully agree". It is well-known that the number of responses k of a response scale $\{1, \dots, k\}$ ($k \in \mathbb{N}$) does not affect the validity of a psychometric scale (Weijters and Baumgartner, 2012) but increasing k can enhance the reliability

of the measurements (Preston and Colman, 2000). According to Lehmann and Vogt, 2024b, 2024c the number of responses k of the response scale $\{1, \dots, k\}$ can affect the results of the statistical analyses. Controlling for possible effects, we chose the common values $k \in \{5, 6, 7\}$.

Please note that the BFI-10 contains negatively formulated items. We do not invert the corresponding item responses because recoding could mask a tendency towards the same response (type B outlier) or create an alternating pattern (type C outlier).

According to section 4 we apply both the MD and PP methods (alternating) to the untransformed data repeatedly excluding statistical outliers, until no more outliers are detected. The resulting datasets are called "initial datasets". Applying the MD and the PP to the ilr transformed initial datasets we find that there are no outliers. That is, both the untransformed and the ilr transformed initial datasets are free of statistical outliers. The corresponding sample sizes are $n_5^{untransformed} = n_5^{ilr} = 265$, $n_6^{untransformed} = n_6^{ilr} = 213$ and $n_7^{untransformed} = n_7^{ilr} = 223$.

Table 1: Type A outliers generated using the `sample()` R-function. k =# responses of the RS $\{1, \dots, k\}$. i1...i10 refer to item response 1...10.

k	i1	i2	i3	i4	i5	i6	i7	i8	i9	i10
5	3	4	2	2	1	1	1	3	2	2
	2	5	4	1	2	2	1	4	3	4
	3	5	3	4	1	1	2	5	2	5
6	2	6	5	4	4	2	3	1	1	1
	6	2	3	1	5	4	2	3	1	5
	3	5	6	3	5	2	3	5	6	1
7	3	7	7	2	6	2	5	2	4	6
	2	6	2	6	6	7	6	1	1	1
	4	1	4	3	1	3	6	4	7	4

The exclusion of statistical outliers prior to the imputation of artificial outliers yields identical starting conditions for both the traditional and the ilr approach. Moreover, it avoids masking and swamping effects of other outliers preventing the detection of the artificially imputed outliers (Rambold, 1999).

Next, we add artificially generated outliers. In detail we have three type A outliers generated completely at random using the `sample()` R-function. We have k type B outliers showing

a tendency towards the same value, that is, each item response equals j with $j = 1, \dots, k$ ($k \in \{5, 6, 7\}$). Finally, there are three type C outliers representing arbitrary patterns, say C_1, C_2, C_3 . The arbitrary patterns are " C_1 : alternating between 1 and k ", " C_2 : starting from 1 and increasing in steps of 1 until k is reached and then decreasing in steps of 1" and " C_3 : like C_2 but in steps of 2". Table 1 and Table 2 provide an overview of the outliers. In total, we add 3 type A + k type B + 3 type C = $6 + k$ outliers to each dataset ($k \in 5, 6, 7$). The final samples sizes including outliers are $n_5^{final} = 276$, $n_6^{final} = 225$ and $n_7^{final} = 236$.

Table 2: Type B and C outliers (manually defined). k =# responses of the RS $\{1, \dots, k\}$. i1...i10 refer to item response 1...10. $j = 1, \dots, k$.

k	i1	i2	i3	i4	i5	i6	i7	i8	i9	i10
5,6,7	j	j	j	j	j	j	j	j	j	j
	1	k	1	k	1	k	1	k	1	k
5	1	3	5	3	1	3	5	3	1	3
	1	2	3	4	5	5	4	3	2	1
6	1	3	5	6	4	2	1	3	5	6
	1	2	3	4	5	6	5	4	3	2
7	1	3	5	7	5	3	1	3	5	7
	1	2	3	4	5	6	7	6	5	4

Using the MD and PP methods, we check for outliers in both the raw data and the ilr transformed datasets. Since the ilr transformed values depend on the LOQ, we choose $p \in 0.05, 0.1, 0.2$ to assess the influence of the LOQ on the quality of outlier analysis, setting $lLOQ = p/2$ and $uLOQ = 100 - p/2$.

6. Results

The outliers identified using MD and PP partially correspond to the artificially generated outliers from Tab. 1 and Tab. 2. Tab. 3 presents the overall numbers of outliers detected and the number correctly detected artificial outliers.

As Tab. 3 indicates the MD and the PP also identified other data points as outliers (i.e., false positives). Therefore, it is necessary to consider the correct discovery rate (CDR). The CDR indicates the proportion of correctly identified outliers among all outliers identified, see Eq. 2. Tab. 4 presents the CDR of the MD and the PP using untransformed and ilr transformed data.

Table 3: Numbers of outliers detected. O=# outliers detected (including artificial outliers). AO=# artificial outliers detected. k =# responses of the RS $\{1, \dots, k\}$. DT=data type (raw or ilr transformed). $p \in \{0.05, 0.1, 0.2\}$ represents the LOQ.

k	DT	O_{MD}	AO_{MD}	O_{PP}	AO_{PP}
5	raw	0	0	0	0
	$ilr_{p=0.05}$	0	0	17	2
	$ilr_{p=0.1}$	0	0	10	3
	$ilr_{p=0.2}$	0	0	0	0
6	raw	13	9	9	8
	$ilr_{p=0.05}$	91	9	45	6
	$ilr_{p=0.1}$	73	8	25	6
	$ilr_{p=0.2}$	49	9	12	5
7	raw	6	6	3	3
	$ilr_{p=0.05}$	73	9	72	7
	$ilr_{p=0.1}$	68	9	45	5
	$ilr_{p=0.2}$	53	9	8	4

$$CDR = \frac{\text{number of artificial outliers detected}}{\text{number of outliers detected}} \quad (2)$$

Table 4: CDR of the MD and PP with raw data and ilr transformed data. k =# responses of the RS $\{1, \dots, k\}$. DT=data type (raw or ilr transformed). $p \in \{0.05, 0.1, 0.2\}$ represents the LOQ.

k	DT	CDR_{MD}	CDR_{PP}
5	raw	-	-
	$ilr_{p=0.05}$	-	11.76%
	$ilr_{p=0.1}$	-	30%
	$ilr_{p=0.2}$	-	-
6	raw	69.23%	88.89%
	$ilr_{p=0.05}$	9.89%	13.33%
	$ilr_{p=0.1}$	10.96%	24%
	$ilr_{p=0.2}$	18.37%	41.67%
7	raw	100%	100%
	$ilr_{p=0.05}$	12.33%	9.72%
	$ilr_{p=0.1}$	13.24%	11.11%
	$ilr_{p=0.2}$	16.98%	50%

Additionally, we consider the discovery rate (DR). This rate indicates the proportion of identified outliers among the set of artificial outliers $k + 6$ ($k \in \{5, 6, 7\}$), see Eq. 3.

$$DR = \frac{\text{number of artificial outliers detected}}{\text{number of artificial outliers}} \quad (3)$$

The results of Tables 3-5 are contradictory. The interpretation heavily depends on the aims advantages or disadvantages of the different approaches (traditional vs. ilr) and methods (MD vs. PP). On the one hand, if $k = 5$ Tab. 3 indicates the PP method applied to ilr transformed data superior to the MD and raw data analysis. However, $k \in \{6,7\}$ Tab. 3 suggest that the traditional approach is superior to the ilr approach.

For example, on the one hand, the ilr approach seems to produce a large number of false positives (see Tab. 2 and 3). However, as p increases the number of false positives decreases. That is, choosing large p seems to improve the ilr approach in terms of the CDR. On the other hand, Tables 4 and 5 suggest that outlier analysis tends to be more reliable when using the raw data instead of the ilr transformed data. The proportion of artificial outliers detected within the set of all outliers detected is largest if the raw data are analyzed (Tab. 4). Again, a large value of p seems preferable when using the ilr approach. The PP seems superior to the MD because the number of false positives is smaller (see Tab. 3) and the CDR_{PP} tends to be equal or larger than the CDR_{MD} (see Tab. 4).

Table 5: DR of the MD and PP with raw and ilr transformed data. k =number of responses of the RS $\{1, \dots, k\}$. DT=data type (raw or ilr transformed). $p \in \{0.05, 0.1, 0.2\}$ represents the LOQ.

k	DT	DR_{MD}	DR_{PP}
5	raw	0%	0%
	$ilr_{p=0.05}$	0%	18.18%
	$ilr_{p=0.1}$	0%	27.27%
	$ilr_{p=0.2}$	0%	0%
6	raw	75%	66.67%
	$ilr_{p=0.05}$	75%	50%
	$ilr_{p=0.1}$	66.67%	50%
	$ilr_{p=0.2}$	75%	41.67%
7	raw	46.15%	23.08%
	$ilr_{p=0.05}$	69.23%	53.85%
	$ilr_{p=0.1}$	69.23%	38.46%
	$ilr_{p=0.2}$	69.23%	30.77%

According to Tab. 3 the MD tends to discover more (less) of the artificial outliers than the PP if $k \in \{6,7\}$ ($k = 5$). The superiority of either method always results from the larger overall number of outliers detected. As the number of outliers detected increases both the DR and the number of false positives increases.

Overall, without predefined aims a superior approach cannot be identified. A way out is to identify and answer the most relevant question which will be discussed in the next section. What is the greater evil: incorrectly classifying a good data point as an outlier, or overlooking an actual outlier? The answer to this question determines the objective and thus the evaluation of the quality of the methodological approaches (ilr vs. traditional; MD vs. PP).

7. Discussion and Limitations

Before the imputation of artificial outliers we applied the MD and PP to detect outliers. During this initial process masking and swamping effects may have already occurred. Rambold, 1999 describes the masking effect as a group of outliers potentially concealing the presence of other outliers. The term swamping describes the phenomenon where non-outliers surrounded by true outliers are mistakenly identified as outliers as well. Both effects may have occurred during the initial outlier cleaning of the data. On the other hand, the MD is the multivariate analogue to the Hampel identifier (Hampel et al., 2005), which is comparatively robust against masking and swamping effects. Additionally, no outliers were detected using the PP method. Finally, we cannot ultimately exclude the presence of masking and swamping effects. Together, the robustness of the MD and the results of the PP suggest that the initial datasets were either free of outliers or had a low proportion of outliers. The outliers detected by MD and PP, which were not artificially generated beforehand, must therefore be attributed to the swamping effect. The fact that some artificially generated outliers were not found could indicate masking effects. However, it is also possible that both swamping and masking effects occurred. That is, labelling and treating good data as outliers (swamping) could cover up real outliers (masking). This cannot be conclusively determined.

Both the MD and the PP are based on the Euclidean metric of \mathbb{R}^n . The ilr transformed data

are subject to the Euclidean metric, while the raw data are subject to the Aitchison metric (Aitchison and Egozcue, 2005; Filzmoser and Hron, 2009). In this light, the better results using the raw data are surprising. As seen in Lehmann and Vogt, 2024a, 2024b, 2024c, a value of $p = 0.1$ has been found to be practical and effective. Tab. 3 and 4 $p \geq 0.2$. A value of $p \geq 0.2$, however, indicates a psychometric scale of low quality not covering a large edge area of the construct considered.

Outliers distort statistical analysis results. The reduction in the quality of standard is just as relevant as the incorrect assessment of correlations within the context of examining convergent validity or discriminant validity. A reduction in sample size, on the other hand, does not cause systematic errors. It simply reduces the convergence of estimates to population parameters. For example, the unbiasedness of the arithmetic mean is independent of the sample size. With sufficiently large samples, a reduction in sample size is therefore more acceptable than the presence of a systematic estimation error. From this perspective, the results regarding DR outweigh those of CDR. It can thus be stated that the ilr approach is preferable to the analysis of untransformed data and we suggest using the MD instead of PP.

While some authors believe that the proportion of outliers in real data samples is empirically small (less than 1%) others suggest larger proportions (e.g. 10% and more) (Hampel et al., 2005; Huber and Ronchetti, 2009; Lehmann, 2012). That is, a clear reference proportion of outliers cannot be obtained. Adding $6+k \in 11, 12, 13$ outliers we chose a small proportion to reduce additional masking and swamping effects and obtain a first impression of the reliability of the MD and PP results.

Due to the limited data and restricted number and types of artificially generated outliers, the generalizability of the results is not readily given. More real data and additional approaches to generate outliers are needed to ultimately assess the impact of the ilr approach on the quality of the identified outliers. The very way in which outliers are generated presents a non-trivial problem. How many outliers are necessary to adequately represent and detect masking and swamping effects? We have attempted, in an initial approach, to identify possible outlier structures and add them to the data. However, it remains unclear what the typical proportions of each type of outlier are and how many can be expected in

total. This fundamental information is necessary to design future studies more precisely and better estimate the effects of the ilr approach in this context. It is also unclear what types of outliers, aside from those considered here, still exist and how they could be artificially generated.

References

- Aitchison, J. (1986). *The statistical analysis of compositional data*. Chapman; Hall.
- Aitchison, J. (2003a). *The statistical analysis of compositional data* (Reprint of 1986 containing additional material). Blackburn Press.
- Aitchison, J., & Egozcue, J. J. (2005). Compositional data analysis: Where are we and where should we be heading? *Mathematical Geology*, 37, 829–850.
- Aitchison, J., Mateu-Figueras, G., & Ng, K. W. (2003). Characterization of distributional forms for compositional data and associated distributional tests. *Mathematical Geology*, 35, 667–680.
- Aitchison, J. (2003b). *A concise guide to compositional data analysis*. Department of Statistics University of Glasgow.
- Aitchison, J., Mateu-Figueras, G., & Ng, K. (2003). Characterization of distributional forms for compositional data and associated distributional tests. *Mathematical Geology*, 35, 667–680.
- Buchanan, E. M., & Scofield, J. E. (2018). Methods to detect low quality data and its implication for psychological research. *Behavior Research Methods*, 50(6), 2586–2596. <https://doi.org/10.3758/s13428-018-1035-6>
- 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>
- Filzmoser, P., Garrett, R. G., & Reimann, C. (2005). Multivariate outlier detection in exploration geochemistry. *Computational Geosciences*, 31, 579–587.
- Filzmoser, P., & Hron, K. (2008). Outlier detection for compositional data using robust methods. *Mathematical Geosciences*, 40, 233–248.

- Filzmoser, P., & Hron, K. (2009). Correlation analysis for compositional data. *Mathematical Geosciences*, 41, 905–919.
- Filzmoser, P., Hron, K., & Reimann, C. (2009). Univariate statistical analysis of environmental (compositional) data: Problems and possibilities. *Science of the Total Environment*, 407, 6100–6108.
- Fischer, H. (2011). *A history of the central limit theorem*. Springer. <https://doi.org/10.1007/978-0-387-87857-7>
- Gather, U., Kuhnt, S., & Pawlitschko, J. (2004). *Concepts of outlyingness for various data structures*. University of Dortmund.
- Gottfried, J., Jeek, S., Králová, M., & ?ihá?ek, T. (2022). Autocorrelation screening: A potentially efficient method for detecting repetitive response patterns in questionnaire data. *Practical Assessment, Research, and Evaluation*, 27. <https://doi.org/10.7275/VYXB-GT24>
- Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (2005). *Robust statistics: The approach based on influence functions*. John Wiley & Sons.
- Hardin, J., & Rocke, D. M. (2005). The distribution of robust distances. *Journal of Computational Graphics and Statistics*, 14, 928–946.
- Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2011). 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., & Ronchetti, E. M. (2009). *Robust statistics*. John Wiley & Sons.
- James, J., & Wood, G. (1988). The effects of incomplete information on the formation of attitudes toward behavioral alternatives. *Journal of Personality and Social Psychology*, 54(4), 580–591. <https://doi.org/10.1037/0022-3514.54.4.580>
- Lanz, L., Thielmann, I., & Gerpott, F. H. (2021). Are social desirability scales desirable? a meta-analytic test of the validity of social desirability scales in the context of prosocial behavior. *Journal of Personality*, 90(2), 203–221. <https://doi.org/10.1111/jopy.12662>
- Lehmann, R., & Vogt, B. (2023a). Reconsidering bipolar scales data as compositional data improves psychometric healthcare data analytics. *Proceedings of the 56th Hawaii International Conference on System Sciences*, 2380–2389.
- Lehmann, R., & Vogt, B. (2024a). Increasing normal approximation in psychometric health care data analyses using a compositional data approach. *Proceedings of the 57th Hawaii International Conference on System Sciences*.
- Lehmann, R., & Vogt, B. (2024b). Compositional data statistics improves smart tourism data analytics: Profound managerial decisions through reduced statistical bias and increased power. *Proceedings of the 57th Hawaii International Conference on System Sciences*.
- Lehmann, R., & Vogt, B. (2024c). Shifting psychometric bipolar scales data towards the normal distribution. *Proceedings of the 57th Hawaii International Conference on System Sciences*.
- Lehmann, R. (2012). *Der einfluss statistischer ausreißer auf die schätzung der natürlichen variabilität in daten zu biota* [Doctoral dissertation, RWTH Aachen].
- Lehmann, R. (2014). A new approach for assessing the state of environment using isometric log-ratio transformation and outlier detection for computation of mean pcdd/f patterns in biota. *Environmental Monitoring and Assessment*, 187(1), 4149. <https://doi.org/10.1007/s10661-014-4149-z>
- Lehmann, R., & Vogt, B. (2023b). Increasing the power of two-sample t-tests in health psychology using a compositional data approach. In F. Liu, Y. Zhang, H. Kuai, E. P. Stephen, & H. Wang (Eds.), *Brain informatics* (pp. 333–347). Springer Nature Switzerland.
- Lehmann, R., & Vogt, B. (2024d). Improving likert scale big data analysis in psychometric health economics: Reliability of the new compositional data approach. *Brain Informatics*, 11(1). <https://doi.org/10.1186/s40708-024-00232-z>

- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 22(140), 5–55.
- Loke, W. H. (1989). The effects of framing and incomplete information on judgments. *Journal of Economic Psychology*, 10(3), 329–341. [https://doi.org/10.1016/0167-4870\(89\)90028-7](https://doi.org/10.1016/0167-4870(89)90028-7)
- Mahalanobis, P. C. (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2, 49–55.
- 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>
- Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, 17(3), 437–455. <https://doi.org/10.1037/a0028085>
- Pan, J.-X., Fung, W.-K., & Fang, K.-T. (2000). Multiple outlier detection in multivariate data using projection pursuit techniques. *Journal of Statistical Planning and Inference*, 1, 153–167.
- Pena, D., & Prieto, F. J. (2001a). Cluster identification using projections. *Journal of the American Statistical Association*, 96, 456.
- Pena, D., & Prieto, F. J. (2001b). Multivariate outlier detection and robust covariance matrix estimation. *Technometrics*, 43(3), 286–310.
- Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, 592(7855), 590–595. <https://doi.org/10.1038/s41586-021-03344-2>
- Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, 104(1), 1–15. [https://doi.org/10.1016/s0001-6918\(99\)00050-5](https://doi.org/10.1016/s0001-6918(99)00050-5)
- Rambold, A. (1999). *Ausgewählte verfahren zur identifikation von ausreißern und einflußreichen beobachtungen in multivariaten daten und verfahren*. Herbert Utz Verlag.
- Rammstedt, B., Kemper, C. J., Klein, M. C., Beierlein, C., & Kovaleva, A. (2014). Big five inventory (bfi-10). *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS)*. <https://doi.org/10.6102/ZIS76>
- Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. *Journal of Research in Personality*, 41, 203–212. <https://doi.org/10.1016/j.jrp.2006.02.001>
- Romano, A., Mosso, C., & Merlone, U. (2016). The role of incomplete information and others' choice in reducing traffic: A pilot study. *Frontiers in Psychology*, 7, 135. <https://doi.org/10.3389/fpsyg.2016.00135>
- Satow, L. (2012). Sea - skala zur erfassung von testverfälschung durch positive selbstdarstellung und sozial erwünschte antwortendenzen. <https://doi.org/10.23668/PSYCHARCHIVES.417>
- Sjölander, P., Lindström, N., Ericsson, A., & Kjellström, S. (2014). A pattern recognition method for disclosing different levels of value system from questionnaire data. *Behavioral Development Bulletin*, 19(3), 114–127. <https://doi.org/10.1037/h0100596>
- Varmuza, K., & Filzmoser, P. (2009). *Introduction to multivariate statistical analysis in chemometrics*. CRC Press, Boca Raton.
- Weijters, B., & Baumgartner, H. (2012). Misresponse to reversed and negated items in surveys: A review. *Journal of Marketing Research*, 49(5), 737–747. <https://doi.org/10.1509/jmr.11.0368>