

Reconsidering Bipolar Scales Data As Compositional Data Improves Psychometric Healthcare Data Analytics

^{1,2}René Lehmann
¹FOM University of Applied Science,
 Essen, Germany.
rene.lehmann@fom.de

²Bodo Vogt
²Otto-von-Guericke-University,
 Magdeburg, Germany.
bodo.vogt@ovgu.de

Abstract

*Correct psychometric profiling and the choice of adequate therapeutic measures are the basis of any psychotherapeutic treatment. The preparation of a correct psychological profile benefits the patient and saves time and costs. Regarding psychometric questionnaires it is common practice to consider data of bipolar scales as interval scaled. This paper reveals the true compositional data structure (namely the Simplex) with respect to the psychometric limit of quantification of bipolar traits and constructs. The Simplex heavily affects the set of statistical procedures applicable. Disregarding the Simplex causes serious bias and results in erroneous standards and standard deviations, biased correlations, reduced convergent validity and a loss of statistical power. In this paper, the isometric log-ratio (ilr) transformation is suggested. It transforms Simplex data towards the interval scale and provides unbiased results, e.g., standards. By means of a simulation study, this paper shows that up to an 18% increase in the statistical power of the well-known correlation test based on Student's *t*-distribution can be achieved. As the statistical power increases the sample size of psychometric studies can be reduced resulting in lower data collection costs. Besides economic and psychotherapeutic aspects, the results of the simulation study generalize from correlation analysis towards a larger set of standard statistical procedures. For example, testing the hypothesis of equality of the two means of independent samples using a *t*-test based on Student's *t* distribution is equivalent to testing the hypothesis of a null-correlation between the binary grouping variable and the dependent variable. Furthermore, the coefficient of correlation contributes to the slope of a regression line. Thus, the ilr approach also affects linear regression techniques.*

1. Introduction

In health care reducing unnecessary pain of the patients, reducing the cost of a treatment and saving time are important goals for ethical and economic reasons. A proper data analysis should help to achieve such goals. We provide a method that allows to draw conclusions from smaller samples sizes and with higher accuracy to go into this direction.

Discussion about the true nature of Likert type data is ubiquitous (e.g., [1, 2, 3]). [4] notes that "the response categories have a rank order but the intervals between values cannot be presumed equal". However, [3] argues that statistical procedures based on normally distributed or interval scaled data can be applied nonetheless because of their robustness towards violations of the underlying assumptions (e.g., normal distribution). Furthermore, [5] and [6] argue rather intuitively that Likert scales can be considered approximately interval scaled.

For proper understanding of the different types of scales, it is necessary to distinguish between statements (i.e., items of a questionnaire) and their corresponding response scale as well as Likert scales (i.e., a set of items represented by the sum or mean value of their corresponding responses) and the scale of a personality trait (e.g., openness). The trait scale can be considered a continuum ranging from a minimum value (e.g., 0) to a maximum value (e.g., 100) of all possible manifestations of a trait. The response scale measures the order of magnitude of a person's agreement or disagreement towards a statement. Associating verbal responses (e.g., ranging from "strongly disagree" to "strongly agree") with numerical values (e.g., 1, ..., 5) is common practice, see [7, 8]). An aggregate value of the item response values (e.g., the mean or sum) is used to estimate a test person's order of magnitude of a trait. Thus, the Likert scale represents a model of the trait scale for estimating the order of magnitude of a personality trait [9]. In the following, if not otherwise stated, the term scale refers to a bipolar scale.

Following [10] and [11], this article treats variables as continuous. However, the assumption of an interval scale cannot be justified for any of the scales. Instead, Likert scale data should be considered compositional data underlying the Aitchison metric. The compositional data space is also called the Simplex (see, e.g., [12]). In section 3.4 the compositional structure of data obtained using bipolar scales is revealed.

2. Literature review

As noted by [13] compositional data structures in psychometric measure scales can be overseen, e.g., regarding Thurstonian scales. To date, the Simplex of bipolar scales data has remained unconsidered.

Simplex data must not be evaluated using methods designed for interval data [12]. For example, Pearson correlations r are biased estimates of the true correlation ρ if the compositional structure is ignored by considering the data as interval scaled [14]. Criterion-related validity cannot be measured properly affecting the quality of psychometric evaluations. Mean values and standard deviations cannot be computed either (see [15]) causing biased psychometric standards. Overall, ignoring the Simplex can cause biased psychometric profiling possibly leading to suboptimal psychotherapeutic measures and increased medical costs. Linear regression techniques such as moderator and mediator analyses (see [16]) depend on (partial) correlations but results are biased if the compositional structure is ignored [17].

These shortcomings also affect the statistical power. Regarding the correlation ρ of two random variables the Pearson correlation r estimates ρ indicating the linear dependency of two variables. However, the compositional data space is by no means linear [18]. Thus, the Simplex affects the estimation of the true value of correlation ρ as well as the statistical power of correlation analysis [14]. As noted by [19, 20, 21] the problem of low statistical power ("underpowerment") and significances at the edge of non-significance must not be neglected in psychometric analyses. The isometric log-ratio (ilr) approach proposed in this article could help to overcome these problems. It increases the statistical power in psychometric data analyses and provides unbiased parameter estimates. Moreover, compositional data should not be evaluated using standard statistical procedures. Evaluation of the ilr-transformed data instead of the raw data is expedient [22, 23]. Finally, the results can be back-transformed by means of the inverse ilr transformation [24].

3. Materials and methods

In this section the basic ideas about scales and limits of quantification are presented. Without loss of generality each of the three scales (response, Likert and trait scale) can be transformed into a standard scale ranging from 0 to 100. Furthermore, the paper presents evidence of the compositional properties of all scales and defines the compositional data scale used. The article presents an algorithm for scale transformation and data preparation analyses and the influence of the parameter p reflecting the limit of quantification (LOQ).

3.1. Basic ideas about bipolar scales

For illustration of the following, consider the personality trait openness. It is plausible to assume limits of its multiple manifestations among the public. Naturally, there exists a lower bound (representing no openness to anything) and an upper bound (representing openness to everything). Clearly, openness shows the properties of a bipolar trait. For the moment, set the bounds of openness to L (lower bound) and U (upper bound). A single person's openness is represented by a specific value $\mu \in [L; U]$. The closer μ and U are, the larger the order of magnitude of the corresponding trait (e.g., openness). On the other hand, the closer μ and L are, the larger the order of magnitude of the opposite trait (non-openness). μ incorporates this complementary (or bipolar) information, i.e., $\Delta_1 = \mu - L$ (reflecting the order of magnitude of openness) and $\Delta_2 = U - \mu$ (reflecting the order of magnitude of non-openness). For an illustration see Figure 1.

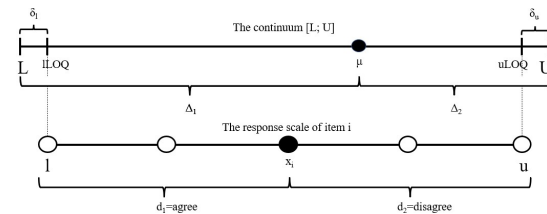


Figure 1. The complements Δ_1 and Δ_2 , both represent the order of magnitude of a trait. We have $\Delta_1 + \Delta_2 = U - L$.

Note that μ does not denote the population mean but the order of magnitude of the trait. A psychometric procedure (e.g., the Big-Five inventory, see [25]) is used to estimate μ .

Without loss of generality we can assume any real values of L and U as long as $L < U$ is satisfied. For example $\mu = 0.5$ is the midpoint of $[0; 1]$ whereas $\mu = 50$ represents the midpoint of $[0; 100]$. Both μ -values represent the same order of magnitude of the trait but on

different scales, so L and U can be chosen arbitrarily. Thus, $L = 0$ and $U = 100$ can be assumed.

3.2. Basic ideas about bipolar response scales

Consider a response scale (r_1, \dots, r_{k+1}) with $r_1 = l$ and $r_{k+1} = u$ (e.g., a discrete scale of 5 categories ranging from "strongly disagree (l)" to "strongly agree (u)"). Let $x \in \{r_1, \dots, r_{k+1}\}$ be an observed response value.

The closer x and u are, the larger an individual's agreement with the item assertion. On the other hand, the closer x and l are, the larger the disagreement. x incorporates this complementary (i.e., bipolar) information, $d_1 = x - l$ (reflecting the order agreement) and $d_2 = u - x$ (reflecting the order of disagreement). For an illustration see Figure 1.

Due to imperfect knowledge, uncertainty about situations and a complex environment [26, 27, 28] the extremes of both scales ($[L; U]$ and $[l; u]$) cannot be (numerically) identical. For example, someone who chose $x = u$ strongly agrees with the item statement, however, this does not imply "full agreement concerning all possible circumstances" because it is impossible for the individual to take into account all facts. Thus, the scale values should be adjusted such that $L < l$ and $u < U$. For an illustration refer to Figure 1. As a result, the scale $[L; U]$ is partly covered by $[l; u]$. A proportion $p = |[L; U] \setminus [l; u]| > 0$ remains where $|\cdot|$ denotes the cardinality of a set and p represents the proportion of $[L; U]$ that is not covered by the response scale.

3.3. Preparations concerning compositional data

The orders of agreement and disagreement with an item assertion (d_1 and d_2) as well as their counterparts on the trait scale (Δ_1 and Δ_2) indicate the compositional structure dissenting the assumption of an interval scale. Ignoring complementary information corrupts further evaluation of data. Even simple statistics such as the arithmetic mean \bar{x} of a set of item responses or Pearson correlations cannot be computed in a straightforward manner. Although the computation of sums, means and Pearson correlations has been propagated widely in psychometrics it leads to biased results when applied to compositional data [29, 18, 30, 31, 32, 33].

Let $lLOQ$ and $uLOQ$ be the lower and upper limits of quantification, respectively. Given a psychometric procedure (e.g., BFI-10), $lLOQ$ represents the lowest value practically measurable using the response scale, i.e., $L < lLOQ$. Accordingly, $uLOQ < U$ is defined. Define $\delta_l = [L; L + lLOQ]$ and $\delta_u = [U - uLOQ; U]$. δ_l and δ_u represent the lower and upper edge areas

of $[L; U]$, respectively, that the response scale cannot measure. For an illustration see Figure 1.

Regarding p as the proportion of $[L; U]$ that is not covered by the response scale we have $p = |[L; U] \setminus [l; u]| = |\delta_l \cup \delta_u|$.

In the following, without loss of generality, assume $L = 0$. For details refer to 3.1. δ_l and δ_u strongly depend on the quality of the response scale used. The better a psychometric procedure is, the closer $lLOQ$ and $uLOQ$ will be to L and U , respectively. That is, levels of quantification vanish when $p \rightarrow 0$ and $\delta_l \cup \delta_u \rightarrow \emptyset$. However, this is only an asymptotic result. In practice, it is plausible to assume that the response scale covers a proportion of $(1 - p) \in (0, 1)$ of the trait scale $[0; U]$. Thus, we have $\delta_l \neq \emptyset \neq \delta_u$ and $\delta_l \cup [lLOQ; uLOQ] \cup \delta_u = [0; U]$ with $|\delta_l \cup \delta_u| = p \cdot U$ where $|\cdot|$ represents the cardinality of a set.

Thus far, p , $lLOQ$ and $uLOQ$ can be considered parameters that should be chosen carefully, e.g., via expert judgement. For example, consider $lLOQ = 0.025 \cdot U$ and $uLOQ = 0.975 \cdot U$. That is, at both ends of the trait scale $p/2$ (e.g., 2.5%) is not covered by the response scale. What follows is that $\delta_l = [0; (p/2)U]$ (e.g., $\delta_l = [0; 0.025U]$) and $\delta_u = [(1 - p/2)U; U]$ (e.g., $\delta_u = [0.975U; U]$). In section 3.6 a simulation is used to assess the effect of p on the statistical power of the well-known correlation test based on Student's t-distribution which can be used to assess criterion-related validity.

3.4. The compositional structure of the data revealed

In this subsection we derive the compositional structure of bipolar scale data. To achieve this purpose, response scales are shifted and levels of quantification are used. The details are presented algorithmically.

Let $l = r_1, \dots, r_{k+1} = u$ be a discrete response scale ranging from "strongly disagree (l)" to "strongly agree (u)" in k steps of width $sw = (u - l)/k$, i.e., $r_j = r_1 + (j - 1)sw$ with $j = 2, \dots, k + 1$. Setting $r_1 = 0$, $r_{k+1} = 100$ and $sw = 100/k$ with $r_j = r_1 + (j - 1)sw$ for $j = 2, \dots, k + 1$ the numeric assignments of the original scale are reassigned to match a scale ranging from "strongly disagree (0)" to "strongly agree (100)" in k steps of width $sw = 100/k$. However, no substantial change is imposed because the verbal attributes remain unchanged and the numeric assignments remain equidistant. Thus, we assume $L = 0$, $U = 100$, $l = 0$ and $u = 100$ without loss of generality.

The article presents the derivation of the final transformed response scale algorithmically. Additionally, it provides further insight by means of an

example. Consider a response scale $0 = r_1, \dots, r_{k+1} = 100$ consisting of k steps and step width $sw = 100/k$ (e.g., $r_1 = 0, r_2 = 25, r_3 = 50, r_4 = 75, r_5 = 100$ with $k = 4$ and $sw = 25$). Choose $p \in (0, 1)$ and set the limits of quantification to $lLOQ = \frac{p}{2} \cdot 100$ and $uLOQ = (1 - \frac{p}{2}) \cdot 100$ (e.g., $p = 0.05$, $lLOQ = 2.5$ and $uLOQ = 97.5$). The following algorithm returns a modified scale r_1^*, \dots, r_{k+1}^* which is called the "response scale*".

1. Define $r_1^* := lLOQ = p/2$ and $r_{k+1}^* := uLOQ = (1 - p/2)100$ (e.g., $r_1^* = 2.5$ and $r_{k+1}^* = 97.5$). The range of the modified scale is given by the difference $r_{k+1}^* - r_1^* = (1-p)100$ (e.g., $97.5 - 2.5 = 95$).
2. Recalculate the step width as $sw^* = (r_{k+1}^* - r_1^*)/k = (1 - p)100/k$ (e.g., $sw^* = 95/4 = 23.75$).
3. For $j = 1, \dots, k$ define $r_j^* := r_1^* + (j - 1)sw^*$ (e.g., $r_1^* = 2.5, r_2^* = 26.25, r_3^* = 50, r_4^* = 73.75, r_5^* = 97.5$).

In the following, we present the mathematical definition of compositional data and show that item responses on the response scale* match the definition. Let $x \in \mathbb{R}^D$ with $x = (x_1, \dots, x_D)^T$ where $()^T$ denotes the transpose. For any constant $\kappa \in \mathbb{R}$ define the compositional data space as $\mathcal{S} := \{x \mid \sum_{j=1}^D x_j = \kappa \text{ and } x_j > 0 \forall j = 1, \dots, D\}$.

Consider a test persons response x^* to a single item on the response scale* (i.e., $x^* \in \{r_1^*, \dots, r_{k+1}^*\}$). Due to bipolarity of the scale the complement $100 - x^*$ exists. While x^* quantifies the test person's order of magnitude of agreement with the item assertion $100 - x^*$ quantifies the order of magnitude of disagreement. Setting $x = (x^*, 100 - x^*)^T$ with $\kappa = 100$ and $D = 2$, x satisfies the definition of compositional data.

3.5. Compositional data and the (inverse) ilr transformation

Any compositional data point x depends on the Aitchison metric [15]. However, most standard statistical procedures (e.g., computation of arithmetic means, Pearson correlation, (multiple) linear regression) are based on the Euclidean metric. These methods cannot be applied to compositional data, therefore, data must be transformed before standard statistical procedures are utilized. The ilr transformation yields interval scaled data underlying the Euclidean metric [34]. By means of the ilr and the inverse ilr, data and statistical results (e.g., mean values) can easily be

(back-)transformed. The ilr transformation is defined as $ilr((x_1, \dots, x_D)^T) := (z_1, \dots, z_{D-1})^T$ with

$$z_s = \sqrt{\frac{s}{s+1}} \ln \frac{\sqrt[s]{\prod_{j=1}^s x_j}}{x_{s+1}}, s = 1, \dots, D-1 \quad (1)$$

In the present case of $D = 2$ the ilr reduces to $ilr((x^*, 100 - x^*)^T) = z_1$ with

$$z_1 = \sqrt{\frac{1}{2}} \ln \frac{x^*}{100 - x^*}. \quad (2)$$

Consider the example in subsection 3.4. For a presentation of the original response scale as well as its corresponding response scale* and the ilr-transformed response scale, see Table 1.

Table 1. Different representations of the response scale $r_1 = 1, \dots, r_5 = 5$ using $p = 0.05$.

response scale	response scale*	ilr response scale
1	2.5	-2.59
2	26.25	-0.73
3	50	0
4	73.75	0.73
5	97.5	2.59

Please note that the bounds of the ilr response scale depend on p . The smaller $p \in (0, 1)$ is, the larger the spread of the ilr response scale. For example consider the response scale $r = (r_1 = 1, \dots, r_5 = 5)$. Using the ilr approach with $p = 0.02$, the scale $(-3.25, -0.76, 0, 0.76, 3.25)$ is obtained while $p = 0.1$ yields $(-2.08, -0.69, 0, 0.69, 2.08)$.

Using the inverse ilr transformation we can back-transform any $z \in \mathbb{R}^{D-1}$ to an $x \in \mathcal{S}$ yielding the Simplex representation of the data. The inverse ilr is defined as follows. Let $z = (z_1, \dots, z_{D-1})^T \in \mathbb{R}^{D-1}$.

$$y_s := \sum_{j=s}^D \frac{z_j}{\sqrt{j(j+1)}} - \sqrt{\frac{s-1}{s}} z_{s-1}; z_0 := z_D := 0 \quad (3)$$

$$x_s := \kappa \cdot \frac{e^{y_s}}{e^{y_1} + \dots + e^{y_D}}, s = 1, \dots, D \quad (4)$$

Like ilr, the inverse ilr is simplified in the present case. The corresponding x^* is obtained by setting $z_0 := z_D := 0$ and $\kappa = 100$ with

$$x^* = 100 \cdot \frac{e^{y_1}}{e^{y_1} + e^{y_2}} \text{ with } y_1 = \frac{z_1}{\sqrt{2}} \text{ and } y_2 = -\sqrt{0.5}z_1. \quad (5)$$

Again, the complete compositional data point is given by $x = (x^*, 100 - x^*)^T$. Applying the inverse ilr transformation to the ilr response scale yields the response scale*. For example, we obtain $\text{invilr}(0.73)=73.75$ (see Table 1). Regarding the ilr-transformed scale (which is an interval scale) we can compute mean values, sums, and correlation coefficients and apply other statistical procedures [14, 22, 15, 17]. Thus, the idea is straight forward:

1. Apply ilr transformation to obtain interval-scaled data.
2. Analyse the ilr-transformed data using any appropriate statistical procedure (e.g., Pearson correlation coefficient, linear regression)
3. Interpret the results on the interval scale.
4. If necessary: use inverse ilr transformation to back-transform the results to the Simplex (e.g., mean values) and interpret.

The term "isometric" refers to the fact that the Aitchison distance d_A of two compositional data points $x, y \in \mathcal{S}$ is identical to the Euclidean distance d_E of their ilr transforms, i.e., $d_A(x, y) = d_E(\text{ilr}(x), \text{ilr}(y))$ [35]. For example, consider two test persons who answered 2 and 3 on the original scale $r_1 = 1, \dots, r_5 = 5$. Their answers correspond to $x^* = 26.25$ and $y^* = 50$ on the response scale* with $r_1^* = 2.5, r_2^* = 26.25, r_3^* = 50, r_4^* = 73.75, r_5^* = 97.5$ (see Table 1). Thus, we have $x = (26.25, 73.75)^T$ and $y = (50, 50)^T$ with $\text{ilr}(x) = -0.73$ and $\text{ilr}(y) = 0$. The Aitchison distance is given by $d_A(x, y) = d_E(\text{ilr}(x), \text{ilr}(y)) = \sqrt{(0 - (-0.73))^2} = 0.73$. Clearly, the Aitchison distance 0.73 differs from what we would obtain when applying the Euclidean distance (erroneously) to the raw data or to the response scale* data, that is $d_E(2, 3) = \sqrt{(2 - 3)^2} = 1$ or $d_E(26.25, 50) = \sqrt{(26.25 - 50)^2} = 23.75$, respectively.

3.6. Simulation study on correlations

Correlations are often used to assess (e.g., criterion-related) validity. However, two major questions arise regarding the mathematical fundamentals of compositional data. First, does p affect the statistical power of the classical correlation test of the null-hypothesis $H_0 : \rho = 0$ using Student's t-distribution (see [36]) where ρ denotes the true coefficient of correlation? Second, does the application of the ilr transformation affect the statistical power compared to classical correlation analysis of the untransformed data? To answer these questions we

conduct a simulation study and present the results in section 4.

Regarding Likert scales and common practice in measuring personality traits we calculate means or sums of item responses. The central limit theorem of statistics in its various versions (among them allowing for non i.i.d. random variables and other generalizations, see [37]) grants that means and sums are asymptotically normally distributed [38]. As means and sums differ by only the constant "1/number of observations" they are equivalent. Thus, in the simulation we focus on means.

Imagine two hypothetical personality traits, T_1 and T_2 (e.g., T_1 =openness and T_2 =risk disposition). Let ζ_1 and ζ_2 be a test individual's order of magnitude of T_1 and T_2 in the ilr-transformed space. Let z_1 and z_2 be the means of the ilr-transformed item responses, that is, z_i estimates ζ_i ($i = 1, 2$). In this study, we simulated z_1 and z_2 assuming a bivariate normal distributions of a random vector $(Z_1, Z_2)^T$ with expectation $\mu \in \mathbb{R}^2$ and covariance matrix Σ . We apply a bivariate normal distribution using the `rmvnorm()` function of the R package `mvtnorm` (version 1.1-1) [39, 40]. For an illustration of the procedure, see Figure 2.

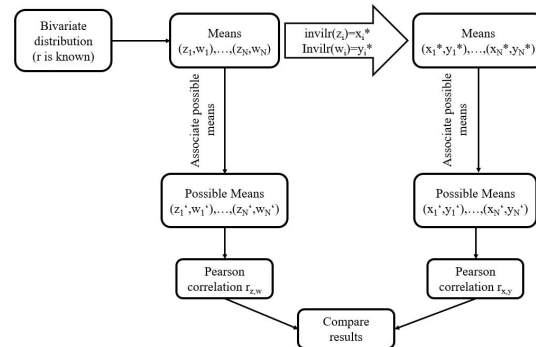


Figure 2. After simulating values using a bivariate normal distribution data are associated with their closest possible means (left-hand path). By means of the inverse ilr transformation the simulated values are transformed to the response scale* and associated with their closest possible means on the original response scale (right-hand path). $H_0 : \rho = 0$ is tested in both paths and the proportions of rejections of H_0 are obtained. Furthermore, the Pearson coefficients of correlation are evaluated in both paths, and the results are compared to the real value of correlation ρ of the specific simulation scenario.

Without loss of generality select $\mu = (0, 0)^T$ because the correlation ρ does not depend on the expectation of the distribution but on the covariance s_{12} and the variances s_{11} and s_{22} of Z_1 and Z_2 , i.e., $\rho = s_{12}/\sqrt{s_{11}s_{22}}$. During

the simulations, we apply different correlations $\rho \in \{-0.65, -0.60, -0.55, \dots, 0.55, 0.6, 0.65\} \setminus \{0\}$ by setting $s_{12} = \rho\sqrt{s_{11}s_{22}}$ in the covariance matrix Σ . We do not consider correlations $|\rho| > 0.65$ because the results indicate powers of 100% irrespective of analysing data of the original data scale or the ilr-transformed data scale.

To consider the influence of the variances s_{11} and s_{22} on the statistical power, we choose $s_{11}, s_{22} \in \{0.25, 0.5, 0.75, \dots, 1.75, 2\}$. Define $s = s_{11} + s_{22}$ as the total variance representing the overall dispersion. We estimate the statistical power of the correlation test of $H_0 : \rho = 0$ in a specific scenario by the proportion of rejected null-hypotheses in 1000 simulation runs.

Calculating means of a finite number of item responses yields a discrete set of possible means. For example, using two items and the ilr response scale $r_1 = -2.59, r_2 = -0.73, r_3 = 0, r_4 = 0.73, r_5 = 2.59$ (see Table 1) the set of possible means is given by $\{-2.59, -1.66, -1.30, -0.93, -0.73, -0.37, 0, 0.37, 0.73, 0.93, 1.30, 1.66, 2.59\}$. To obtain realistic values we replace any simulated mean z_i ($i = 1, 2$) with its nearest possible mean μ_i^{ilr} ($i = 1, 2$) according to the Euclidean metric. In the above example the nearest possible mean of $z = 0.82$ is given by $\mu^{ilr} = 0.93$. Note that the number of possible means depends on the number of responses $k + 1$ and the number of items $I \in \mathbb{N}$. Thus, during the simulation we used different numbers of items $I \in \{1, 4, 10, 30\}$ and numbers of responses $k + 1 \in \{5, 6, 10\}$.

We transform the simulated means to the response scale* via the inverse ilr transformation and replace them with their nearest possible mean. Although the Aitchison metric should be used on the response scale*, the Euclidean metric is used to obtain the nearest possible means. This approach is necessary because in common practice means and correlations are calculated without considering the compositional structure of the response data. The intention of the simulation is to show the effects of disregarding the compositional structure on the statistical analysis. Note that each possible mean of response scale* corresponds to a possible mean of the original response scale. Thus, any simulated mean z_i ($i = 1, 2$) could also be assigned to its nearest possible mean μ_i^{orig} ($i = 1, 2$) of the original response scale. For example, let $z = 0.82$ be a simulated mean. By using the inverse ilr transformation we obtain $\text{invilr}(0.82) = 76.13$. Consider the response scale $r_1 = 1, r_2 = 2, r_3 = 3, r_4 = 4, r_5 = 5$ and the response scale* with $r_1 = 2.5, r_2 = 26.25, r_3 = 50, r_4 = 73.75, r_5 = 97.5$ consisting of $I = 2$ items. The sets of possible means on both scales are $\{1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5\}$ and $\{2.5, 14.38, 26.25, 50,$

$61.88, 73.75, 85.63, 97.5\}$. The nearest possible mean of 76.13 on the response scale* is given by 73.75 which represents the mean $\mu^{orig} = 4$ on the initial response scale.

To analyse the influence of the LOQ on the statistical power we use different values of $p \in \{0.02, 0.04, \dots, 0.2\}$. Overall, the simulation incorporates a total number of 36 (#variance combinations) $\cdot 26$ (#correlations) $\cdot 3$ (#k) $\cdot 4$ (#I) $\cdot 10$ (#p) = 112,320 scenarios. Each scenario is simulated 1000 times with 200 simulated pairs of means (z_1, z_2) resulting in 112,320,000 simulation runs of 200 pairs of means.

Each simulation run generates two data sets: $ILR = \{(\mu_{1,i}^{ilr}, \mu_{2,i}^{ilr}) \mid i = 1, \dots, 200\}$ and $ORIG = \{(\mu_{1,i}^{orig}, \mu_{2,i}^{orig}) \mid i = 1, \dots, 200\}$. We apply the correlation test based on Student's t-distribution to test $H_0 : \rho = 0$ for the ILR and ORIG data sets. The two proportions of rejections of H_0 in 1000 runs represent the estimates of the statistical powers of the correlation test on both scales, the ilr scale and the original scale, that is, $Power^{ilr}$ and $Power^{orig}$. The difference $\Delta Power = Power^{ilr} - Power^{orig}$ indicates the superiority or inferiority of the ilr approach.

4. Results of the simulation study and conclusions

This section describes the results of the simulation study, which are summarized in Tables 2-4. Figures 3-4 provide graphical representations of the results. The `splinefun` function of the R statistic software package is used to derive the Figures 3-4 and Tables 2-4. For details refer to the `fmm` method of [41].

Below, we present the main results of the simulation with respect to different combinations of numbers of items $I \in \{1, 4, 10, 30\}$, numbers of responses $k + 1 \in \{5, 6, 10\}$ of the response scale, variances $s_{11}, s_{22} \in \{0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2\}$ (with total variance $s = s_{11} + s_{22}$) and values of $p \in \{0.02, 0.04, \dots, 0.2\}$ reflecting the LOQ. Please note that $\Delta Power > 0$ indicates superiority of the ilr approach.

1. The ilr approach tends to yield higher statistical power especially if the underlying correlation is close to 0, see Tables 2-4).
2. We observe a range of $\Delta Power \in (-0.03, 0.18)$, see Figure 5.
3. If $|\rho| > 0.35$ the ilr approach is neither superior nor inferior to the classical evaluation and $\Delta Power$ reduces to 0.

4. A moderate (or sometimes large) increase of statistical power of the ilr approach can be observed for the majority of sets of parameter combinations.
5. Figure 3 shows that $\Delta Power$ increases as p increases.
6. Concerning the total variance parameter s , $\Delta Power$ increases as s increases (see Figure 4).
7. The total variance s has a considerable effect on $\Delta Power$ (see Figure 4). By contrast, the parameter p is less influential (see Figure 3).

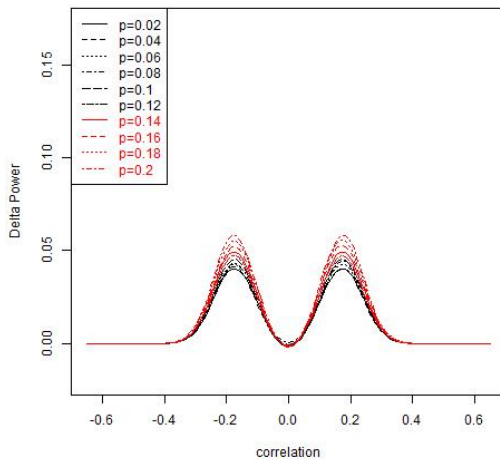


Figure 3. $\Delta Power$ of rejection of $H_0 : \rho = 0$ for different correlations ρ .

5. Application to real data

The Corona pandemic started in 2019. It is of great interest to identify influential variables affecting the (un-)willingness to receive a COVID 19 vaccination. Recently, [8] suggested several potential sources. They evaluated the orders of magnitude of the Big-Five personality traits [25], locus of control (including its three facets "chance", "powerful others" and "internal", see [42]), paranoia [43], altruism (including its three facets "identify", "care" and "help", see [44]), conspiracy beliefs [45], authoritarianism [46] and social dominance [47] of more than 2000 test persons in the UK and more than 1000 test persons in Ireland (IRL). Besides the psychological constructs [8] measure people's attitudes towards a vaccination. The three groups considered are as follows: people accepting a vaccination, people hesitant and people

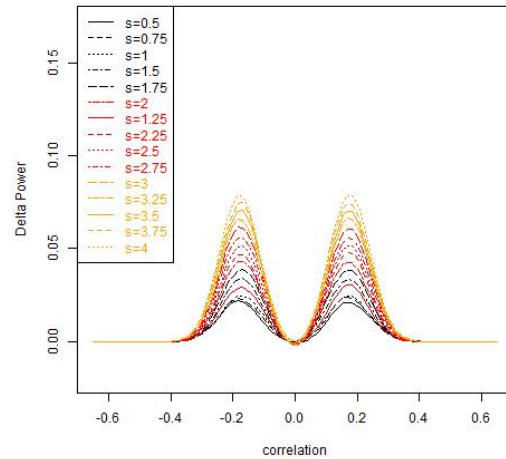


Figure 4. $\Delta Power$ of rejection of $H_0 : \rho = 0$ for different correlations ρ .

resistant towards a vaccination. Concerning the three groups [8] evaluates the orders of magnitudes of the psychological constructs using pairwise two sample t-tests of independent samples (acceptance vs. hesitance, acceptance vs. resistance, acceptance vs. pooled data of hesitance and resistance). We re-evaluate the data using the ilr approach. To date, an optimal selection of the parameters p , $lLOQ$, $uLOQ$ is unavailable. Therefore, we re-evaluate the data using different values $p \in \{0.02, 0.1, 0.2\}$ with $lLOQ = 100 \cdot (p/2)$ and $uLOQ = 100 \cdot (1 - p/2)$.

[8] obtain 31 (UK) and 13 (IRL) significant results out of the 2×45 t-tests of independent samples applied to the data. Using the ilr approach, we obtain four (UK) and six (IRL) additional significances while losing one significance (UK) and one significance (IRL). Ignoring the Simplex causes serious bias. Thus, the two significances lost can be interpreted as statistical type-I errors revealed by using the ilr transformation. The ilr approach yields a total of 34 (UK) and 18 (IRL) significances. The proportion of significant results increases by 6.67% (UK) and 11.11% (IRL). The number of additional significances (4+6) overcomes the minor losses of significances (1+1). Nine out of the ten additional significances are observed regardless of the value of p . One additional significance of the IRL data set is observed only if $p \in \{0.1, 0.2\}$. Overall, it seems that the value of p hardly affects the results of the ilr approach in practice. However, this is an object of upcoming research. The results indicate that more significant results can be expected when using the ilr approach.

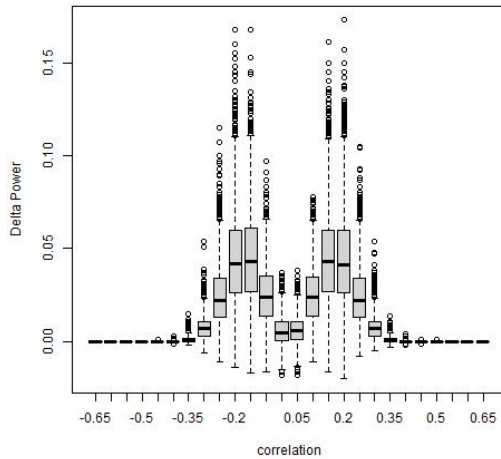


Figure 5. Δ Power of rejection of $H_0: \rho = 0$ for different correlations ρ .

6. Discussion, limitations and practical implications

Overall, the ilr approach proved to be more powerful than the classical approach. Scenarios with Δ Power > 0 were observed more often than scenarios with Δ Power < 0 . Moreover, the mean gain in statistical power outweighs the potential minor losses. Concerning correlations $|\rho| \leq 0.2$, an additional power of up to 18% is possible.

Partial correlations of the variables are estimated for multiple linear regression models or structural equation models [48]. The larger the number of relevant variables is, the closer the partial correlations will be to 0 because a variable's proportional contribution to the explained variance is reduced by additional variables. Practically relevant models provide a large number of variables resulting in (partial) correlations of $0.05 < |\rho| < 0.2$ where the gain in statistical power using the ilr approach can reach 18%. Thus, the ilr approach could help to overcome the problem of low statistical power and significance at the edge of non-significance [19, 20, 21]. Moreover, the ilr approach contributes to the problem of ethics in medical research. It is possible to decrease the sample size (i.e., the number of test individuals) while maintaining at least the same statistical power as in classical data analysis reducing ethical issues [49] and economic effort. Furthermore, psychometric profiling can be improved because minor (but relevant) coherences of traits or between-groups effects can be revealed more efficient.

The results of the ilr transformation depend slightly

on the value of p . In practice, proper selection of p is based on expert judgement because it refers to the quality of the psychometric scale. The better the psychometric procedure is, the closer p will be to 0. Thus, p depends on the scale used. Assuming a high quality of the measure scale $p = 0.02$ could be reasonable. Figures 3-4 show the minor influence of p on Δ Power. A small value of p reduces the improvement in statistical power induced by the ilr approach compared to large values of p . However, the reduction seems to be negligible. Appropriate selection of p is a qualitative rather than a technical task. Further research is needed to determine appropriate and scale specific values of p .

Performing repeated data evaluations using different values of p could be expedient if an appropriate selection of p seems impossible. Alternatively, p could be chosen at random within reasonable bounds (e.g., using a uniform distribution on a predefined interval, e.g., (0;0.2)). Another idea is to apply different proportions, e.g., $p_l > 0$ and $p_u > 0$ at both ends of the response scale* where $p_l \neq p_u$ indicates differences in the order of magnitude of the limits of quantification at the scale ends, i.e., $|\delta_l| \neq |\delta_u|$. Overall, the selection of p will be considered in upcoming research.

A limiting factor of the simulation is the finite number of scenarios. Thus far, the results appear to be plausible and generalizable towards common values of I , $k + 1$, s_{ii} and symmetric limits of quantification (i.e., $|\delta_l| = |\delta_u|$). However, further research on the influences of non-symmetric limits of quantification and non-symmetric or heavy-tailed data generating processes on Δ Power is necessary.

Table 2. Summary of Δ Power over all values of p and s . ρ denotes the true correlation.

ρ	Δ Power
± 0.35	0.001
± 0.3	0.008
± 0.25	0.025
± 0.2	0.044
± 0.15	0.045
± 0.1	0.025
± 0.05	0.006

Table 3. Summary of Δ Power for different values of p over all values of s . ρ denotes the true correlation.

ρ	p				
	0.02	0.04	0.06	0.08	0.1
± 0.05	0.005	0.005	0.006	0.006	0.006
± 0.1	0.022	0.022	0.023	0.022	0.023
± 0.15	0.038	0.038	0.040	0.041	0.043
± 0.2	0.038	0.038	0.039	0.041	0.042
± 0.25	0.021	0.021	0.021	0.022	0.023
± 0.3	0.007	0.007	0.007	0.007	0.007
± 0.35	0.001	0.001	0.001	0.001	0.001

ρ	p				
	0.12	0.14	0.16	0.18	0.2
± 0.05	0.006	0.006	0.007	0.007	0.007
± 0.1	0.025	0.026	0.028	0.029	0.031
± 0.15	0.045	0.046	0.050	0.053	0.054
± 0.2	0.044	0.046	0.049	0.052	0.055
± 0.25	0.024	0.026	0.028	0.029	0.031
± 0.3	0.008	0.008	0.009	0.009	0.010
± 0.35	0.001	0.001	0.001	0.002	0.002

References

[1] D. R. Johnson and C. R. Creech, "Ordinal measures in multiple indicator models: a simulation study of categorization error," *American Sociological Review*, vol. 48, no. 3, pp. 398–407, 1983.

[2] G. M. Sullivan and A. R. Artino, "Analyzing and interpreting data from likerttypescales," *Journal of Graduate Medical Education*, vol. 5, pp. 541–542, 2013.

[3] G. Norman, "Likert scales, levels of measurement and the laws of statistics," *Advances in Health Sciences Education*, vol. 15, pp. 625–632, 2010.

[4] S. Jamieson, "Likert scales: How to (ab)use them," *Medical Education*, vol. 38, pp. 1217–1218, 2004.

[5] J. Carifio and R. J. Perla, "Ten common misunderstandings, misconceptions, persistent myths and urban legends about likert scales and likert response formats and their antidotes," *Journal of Social Sciences*, vol. 3, pp. 106–116, 2007.

[6] L. Carifio and R. Perla, "Resolving the 50 year debate around using and misusing likert scales," *Medical Education*, vol. 42, pp. 1150–1152, 2008.

[7] G. Pennycook, Z. Epstein, M. Mosleh, A. A. Arechar, D. Eckles, and D. G. Rand, "Shifting attention to accuracy can reduce misinformation online," *Nature*, vol. 592, pp. 590–595, mar 2021.

[8] J. Murphy, F. Vallières, R. P. Bentall, M. Shevlin, O. McBride, T. K. Hartman, R. McKay, K. Bennett, L. Mason, J. Gibson-Miller, L. Levita, A. P. Martinez, T. V. A. Stocks, T. Karatzias, and P. Hyland, "Psychological characteristics associated with covid-19 vaccine hesitancy and resistance in ireland and the united kingdom," *Nature Communications*, vol. 12, no. 29, 2021.

[9] R. Likert, "A technique for the measurement of attitudes," *Archives of Psychology*, vol. 22, no. 140, pp. 5–55, 1932.

Table 4. Summary of Δ Power for different values of s over all values of p . ρ denotes the true correlation.

ρ	s				
	0.5	0.75	1	1.25	1.5
0.05	0.002	0.003	0.003	0.004	0.005
0.1	0.010	0.012	0.013	0.015	0.018
0.15	0.020	0.023	0.023	0.029	0.031
0.2	0.019	0.021	0.023	0.028	0.031
0.25	0.012	0.012	0.012	0.015	0.017
0.3	0.005	0.004	0.004	0.005	0.005
0.35	0.001	0.001	0.001	0.001	0.001

ρ	s				
	1.75	2	2.25	2.5	2.75
0.05	0.005	0.005	0.006	0.006	0.007
0.1	0.022	0.023	0.026	0.028	0.028
0.15	0.036	0.041	0.045	0.048	0.052
0.2	0.036	0.039	0.045	0.048	0.052
0.25	0.020	0.022	0.024	0.026	0.029
0.3	0.006	0.006	0.008	0.008	0.009
0.35	0.001	0.001	0.001	0.001	0.001

ρ	s				
	3	3.25	3.5	3.75	4
0.05	0.007	0.008	0.010	0.010	0.009
0.1	0.031	0.034	0.035	0.039	0.041
0.15	0.057	0.062	0.066	0.070	0.073
0.2	0.057	0.062	0.065	0.069	0.075
0.25	0.033	0.035	0.037	0.040	0.043
0.3	0.010	0.011	0.012	0.013	0.013
0.35	0.002	0.002	0.002	0.002	0.003

[10] M. Rhemtulla, P. E. Brosseau-Liard, and V. Savalei, "When can categorical variables be treated as continuous? a comparison of robust continuous and categorical sem estimation methods under suboptimal conditions," *Psychological Methods*, vol. 17, no. 3, pp. 354–373, 2012.

[11] U. D. Reips and F. Funke, "Interval-level measurement with visual analogue scales in internet-based research: Vas generator," *Behavior Research Methods*, vol. 40, pp. 699–704, 2008.

[12] J. Aitchison, *The statistical analysis of compositional data*. Chapman and Hall, 1986.

[13] A. Brown, "Thurstonian scaling of compositional questionnaire data," *Multivariate Behavioral Research*, vol. 51, no. 2-3, pp. 345–356, 2016.

[14] P. Filzmoser and K. Hron, "Correlation analysis for compositional data," *Mathematical Geosciences*, vol. 41, pp. 905–919, 2009.

[15] P. Filzmoser, K. Hron, and C. Reimann, "Univariate statistical analysis of environmental (compositional) data: Problems and possibilities," *Science of the Total Environment*, vol. 407, pp. 6100–6108, 2009.

[16] T. Loeys, W. Talloen, L. Goubert, B. Moerkerke, and S. Vansteelandt, "Assessing moderated mediation in linear models requires fewer confounding assumptions than assessing mediation," *British Journal of*

- Mathematical and Statistical Psychology*, vol. 69, pp. 352–374, oct 2016.
- [17] V. Pawlowsky-Glahn and J. J. Egozcue, “Blu estimators and compositional data,” *Mathematical Geology*, vol. 34, pp. 259–274, 2002.
- [18] J. Aitchison, *The statistical Analysis of Compositional Data*. Blackburn Press, reprint of 1986 containing additional material ed., 2003.
- [19] U. Simonsohn, L. D. Nelson, and J. P. Simmons, “P-curve: A key to the file-drawer,” *Journal of Experimental Psychology: General*, vol. 143, no. 2, pp. 534–547, 2014.
- [20] U. Simonsohn, L. D. Nelson, and J. P. Simmons, “p-curve and effect size: Correcting for publication bias using only significant results,” *Perspectives on Psychological Science*, vol. 9, no. 6, pp. 666–681, 2014.
- [21] U. Simonsohn, “Small telescopes: Detectability and the evaluation of replication results,” *Psychological Science*, vol. 26, no. 5, pp. 559–569, 2015.
- [22] K. Hron, M. Templ, and P. Filzmoser, “Imputation of missing values for compositional data using classical and robust methods,” *Computational Statistics and Data Analysis*, vol. 54, no. 12, pp. 3095–3107, 2010.
- [23] K. Varmuza and P. Filzmoser, *Introduction to Multivariate Statistical Analysis in Chemometrics*. CRC Press, Boca Raton, 2009.
- [24] R. Lehmann, “A new approach for assessing the state of environment using isometric log-ratio transformation and outlier detection for computation of mean pccdd/f patterns in biota,” *Environmental Monitoring and Assessment*, vol. 187, no. 1, p. 4149, 2014.
- [25] B. Rammstedt and O. P. John, “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*, vol. 41, pp. 203–212, 2007.
- [26] A. Romano, C. Mosso, and U. Merlone, “The role of incomplete information and others’ choice in reducing traffic: A pilot study,” *Frontiers in Psychology*, vol. 7, p. 135, 2016.
- [27] W. H. Loke, “The effects of framing and incomplete information on judgments,” *Journal of Economic Psychology*, vol. 10, no. 3, pp. 329–341, 1989.
- [28] J. James and G. Wood, “The effects of incomplete information on the formation of attitudes toward behavioral alternatives,” *Journal of Personality and Social Psychology*, vol. 54, no. 4, pp. 580–591, 1988.
- [29] C. Barceló, V. Pawlowsky, and E. Grunsky, “Some aspects of transformations of compositional data and the identification of outliers,” *Mathematical Geology*, vol. 28, no. 4, pp. 501–518, 1996.
- [30] J. Aitchison, *A Concise Guide to Compositional Data Analysis*. Department of Statistics University of Glasgow, 2003.
- [31] J. Aitchison and J. J. Egozcue, “Compositional data analysis: Where are we and where should we be heading?,” *Mathematical Geology*, vol. 37, pp. 829–850, 2005.
- [32] P. Filzmoser and K. Varmuza, *Introduction to Multivariate Statistical Analysis in Chemometrics*. Vienna, Austria: CRC Press, 2009.
- [33] J. Aitchison, G. Mateu-Figueras, and K. Ng, “Characterization of distributional forms for compositional data and associated distributional tests,” *Mathematical Geology*, vol. 35, pp. 667–680, 2003.
- [34] P. Filzmoser, R. G. Garrett, and C. Reimann, “Multivariate outlier detection in exploration geochemistry,” *Computational Geosciences*, vol. 31, pp. 579–587, 2005.
- [35] J. J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal, “Isometric logratio transformations for compositional data analysis,” *Mathematical Geology*, vol. 35, pp. 279–300, 2003.
- [36] H. E. Soper, A. W. Young, B. M. Cave, A. Lee, and K. Pearson, “On the distribution of the correlation coefficient in small samples. appendix ii to the papers of “student” and r.a. fisher. a co-operative study.,” *Biometrika*, vol. 11, no. 4, pp. 328–413, 1917.
- [37] H. Fischer, *A History of the Central Limit Theorem*. Springer, 2011.
- [38] J. Davidson, *Econometric Theory*. Blackwell Publishing, 2001.
- [39] R Core Team, *R: A Language and Environment for Statistical Computing*, 2020.
- [40] A. Genz, F. Bretz, T. Miwa, X. Mi, F. Leisch, F. Scheipl, and T. Hothorn, *mvtnorm: Multivariate Normal and t Distributions*, 2020. R package version 1.1-1.
- [41] G. E. Forsythe, M. A. Malcolm, and C. B. Moler, *Computer Methods for Mathematical Computations*. Wiley, 1977.
- [42] S. G. Sapp and W. J. Harrod, “Reliability and validity of a brief version of levenson’s locus of control scale,” *Psychological Reports*, vol. 72, pp. 539–550, 1993.
- [43] S. Melo, R. Corcoran, N. Shryane, and R. P. Bentall, “The persecution and deservedness scale,” *Psychology and Psychotherapy: Theory, Research and Practice*, vol. 82, no. 3, pp. 247–260, 2009.
- [44] S. McFarland, M. Webb, and D. Brown, “All humanity is my ingroup: a measure and studies of identification with all humanity,” *Journal of Personality and Social Psychology*, vol. 103, no. 5, pp. 830–853, 2012.
- [45] R. Imhoff and M. Bruder, “Speaking (un) truth to power: conspiracy mentality as a generalised political attitude,” *European Journal of Personality*, vol. 28, no. 1, pp. 25–43, 2014.
- [46] B. Bizumic and J. Duckitt, “Investigating right wing authoritarianism with a very short authoritarianism scale,” *Journal of Social and Political Psychology*, vol. 6, no. 1, pp. 129–150, 2018.
- [47] A. K. Ho, J. Sidanius, N. Kteily, J. Sheehy-Skeffington, F. Pratto, K. E. Henkel, R. Foels, and A. L. Stewart, “The nature of social dominance orientation: Theorizing and measuring preferences for intergroup inequality using the new sdo7 scale,” *Journal of Personality and Social Psychology*, vol. 109, no. 6, pp. 1003–28, 2015.
- [48] G. Marrelec and H. Benali, “A theoretical investigation of the relationship between structural equation modeling and partial correlation in functional mri effective connectivity,” *Computational Intelligence and Neuroscience*, vol. 2009:369341, 2009.
- [49] A. Avasthi, A. Ghosh, S. Sarkar, and S. Grover, “Ethics in medical research: General principles with special reference to psychiatry research,” *Indian Journal of Psychiatry*, vol. 55, no. 1, p. 86, 2013.