Increasing normal approximation in psychometric health care data analyses using a compositional data approach

[1,2]René Lehmann
[1]FOM University of Applied Science,
Essen, Germany
rene.lehmann@fom.de rene1.lehmann@ovgu.de

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

Abstract

Psychometric health care focuses on the development and improvement of psychotherapeutic measures. Adequate psychological profiling and advanced statistical evaluation are fundamental to assessing the efficacy and measure-associated benefits. Consider psychological constructs operationalized as means or sums of item response values of bipolar Likert scales. Using estimates of the effect size and statistical tests the relevance of a psychotherapeutic measure can be assessed, e.g., via the computation of (partial) correlations of different constructs. Many statistical procedures depend on approximate normal distribution, e.g., t-tests, linear regression and partial least squares path modeling. Increasing the degree of approximate normality of means and sums of item responses enhances the quality of statistical evaluations. Via simulation we provide evidence that applying the isometric log-ratio (ilr) transformation to bipolar Likert scales data prior to the computation of item response means or sums increases the degree of approximate normality. That is, a shift towards normality is observed enhancing the quality of subsequent statistical analyses. As a result, the quality of statistical evaluations enhances. The reliability of psychological diagnostics increases and the development of psychometric scales can be improved enhancing patient welfare. Moreover, reliability and significance affect grant funding in health economics.

1. Introduction

Reducing unnecessary pain of the patients and saving time and costs of a treatment are important goals for ethical and economic reasons. Thus, the choice of an appropriate psychotherapeutic measure is essential. Concerning psychometric patient profiling and therapeutic success the relevance and efficacy of a health measure can be associated with the statistical effect size. The effect size can be assessed via pre-post-comparisons (e.g., using t-tests and Cohen's d) and the degree of convergent and discriminant validity between similar and complementary constructs (e.g., using Pearson correlations and the significance test based on Student's t-distribution), see [1].

Bipolar Likert scales (see [2]) are widespread tools in psychological health care. They are used to quantify a test person's order of magnitude of psychological traits and constructs (e.g., locus of control, neuroticism, risk disposition).

Often, results are at the edge of significance and measures of effect size are small [3]. There is a need to improve statistical techniques in psychometric data analyses. Suboptimal data evaluation can cause false positive or negative medical border cases favouring misdiagnoses, corrupting treatment plans occasioning additional treatment costs and compromising patient welfare. Therefore, it is of fundamental interest to apply statistical procedures providing large statistical power [4].

In the following, if not otherwise stated, the term "normality test" is used synonymously with the term "test of the null-hypothesis that the considered data are approximately normally distributed".

Many statistical procedures applied to bipolar scales data assume approximately normally distributed means and sums of item responses (e.g., the correlation test based on Student's t-distribution, t-tests, linear regression, structural equation models, factor analysis). The better the statistical prerequisite of normality is fulfilled the larger is the statistical power of the procedure. Propositional logic indicates effect size causal concerning statistical power. Effect size can be considered the most important determinant of power [1]. Increasing the statistical power results in two aspects. First, the same effect size can be detected with a higher probability. Second, a smaller effect size can be detected with the same probability. That is, the efficacy of a therapeutic measure can be assessed more accurately affecting quality-adjusted life years (QALY) index values in health economics.

URI: https://hdl.handle.net/10125/106571
978-0-9981331-7-1

In practice, bipolar Likert scales data are often assumed to yield approximately interval scaled data [5, 6]. The assumption of approximate normality of means and sums of item response values is justified by central limit theorem (CLT) of statistics in its various versions (among them allowing for non i.i.d. random variables and other generalizations, see [7]) ([8]). [9] revealed the true compositional structure of bipolar scales data, the so-called Simplex. Applying statistical procedures designed for interval scaled or ratio scaled data to compositional data causes serious bias [10]. Using the isometric log-ratio (ilr) transformation, compositional item response data are transformed towards the real valued interval scale $\mathbb{R}$ where statistical standard procedures can be applied [11, 12]. Concerning the CLT, means and sums of non-transformed and of ilr transformed item responses, both are asymptotically normally distributed. The question arises whether or not the ilr approach increases the convergence towards normal distribution.

Via simulation we generate item response data with respect to different data generating processes (DGP) including floor and ceiling effects and psychometric parameters (e.g., the limit of quantification (LOQ)) [13]. Applying a set of normality tests (e.g., the Shapiro-Wilk test [14]), we analyze the ilr-transformed and non-transformed data. Finally, we compare the proportions of non-rejected null-hypothesis Δ in 5,000 simulation runs.

2. Literature review

The debate about the true nature of Likert type data obtained using discrete response scales continues. For example, [15] argued that rank data are approximately interval scaled while [16] and [17] point out numerous disadvantages of this idea. Recently, [18] and [6] argued that statistical procedures based on normal distribution can be applied to bipolar scales data because the introduced error is small.

In the following, if not otherwise stated, the term "a shift [..] towards normality" is used synonymously with the term "the convergence of the distribution of items response means towards a normally distributed random variable increases".

As noted by [19] compositional data structures in psychometric measure scales can be overseen, e.g., regarding Thurstonian scales. [9] revealed the Simplex structure of bipolar scales data.

Simplex data must not be evaluated using methods designed for interval data [20]. For example, the Pearson correlation r yields a biased estimate of the true correlation ρ if the compositional structure is ignored [12] affecting the results of linear regression techniques such as moderator and mediator analyses (see [21]). Mean values and standard deviations based on untransformed data yield biased psychometric standards (see [10]). Statistical bias affects the statistical power of significance tests.

For example, [9] showed that the statistical power of the well-known correlation test based on Student's t-distribution increased when using means of ilr transformed instead of non-transformed item response data. One reason could be the ilr-induced unbiased parameter estimates. Furthermore, a shift towards normality could have contributed to the enhanced statistical power because normal distribution is located on the real-valued interval scale [8]. Simplex data, however, cannot be normally distributed [22]. Enhanced approximate normality affects many more statistical procedures in psychometric bipolar scales data analyses. That is, a shift towards normality increases the statistical power of all procedures based on approximate normality (e.g., linear regression, factor analysis, structural equation modeling, partial least squares path modeling, t-tests, analysis of variance, moderator analysis, mediator analysis) implying more accurate results and enhanced effect size [1] affecting grand funding.

Statistical bias implies reduced statistical power [1]. As noted by [3] the problem of low statistical power ("underpowerment") and significances at the edge of non-significance must not be neglected in psychometric analyses. Ignoring the Simplex causes biased psychometric profiling possibly leading to suboptimal psychotherapeutic measures and increased medical costs.

The ilr approach proposed in this article could help to overcome these problems. It transforms Simplex data towards the real valued interval scale. Evaluation of the ilr-transformed data instead of the raw data is expedient [23]. The results can be back-transformed by means of the inverse ilr transformation [24].

3. Materials and Methods

This section provides an overview of the different types of scales used in psychometrics. We briefly introduce the compositional data space (the Simplex) and related psychometric parameters (e.g., the LOQ). The simulation process is described including different DGP and other simulation parameters.

We differentiate between the trait scale (TS) of a personality trait (i.e., the continuum of all possible manifestations of a trait) and the Likert scale (LS, i.e., a set of statements/items represented by the sum or

mean value of their corresponding responses). The LS represents a model of the TS for estimating the order of magnitude of a personality trait (OMT) ([2]). The item response scale (RS) measures the order of magnitude of a person's agreement (OMA) or disagreement (OMD) towards a statement. Associating verbal responses of the RS (e.g., ranging from "not at all" to "very much") with numerical values (e.g., $1,\ldots,5$) is common practice [25]. In the following, if not otherwise stated, the term scale refers to a bipolar scale.

3.1. Bipolar constructs and psychometric scales

Psychometric scales provide estimates of individual values of psychological constructs. For example, think of the Big 5 trait openness. The items of a questionnaire (e.g., the BFI-10 inventory of [26]) 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 [27, 28] implying the existence of a LOQ [9]. For an illustration see Fig. 1.

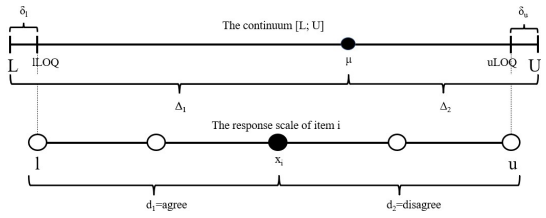


Figure 1. Illustration of the different types of scales used in psychometrics. The continuum [L; U] represents the TS. The lower scale represents the RS.

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 $i = 1,\ldots,I$ associated with a responses scale, 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 response scale are different from L and U reflecting the lower (lLOQ) and upper (uLOQ) limit of quantification. The edge area of the construct scale which is not covered by the items and their respective response scale are named δ_l and δ_u. Any response x_i towards an item assertion reflects the order of magnitude of agreement (d_1) and the order of magnitude of disagreement towards the item assertion (d_2). For example, let l=lLOQ=2.5, u=uLOQ=97.5, $x_i = 73.75$, $d_1 = 50$, $d_2 = 50$, $\delta_l = [0; 2.5)$ 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.

3.2. The compositional structure in brief

Concerning the TS [L; U], any limit values $L, U \in \mathbb{R}$ can be assumed as long as $L < U$ is satisfied, e.g., $L = 0$ and $U = 100$. For example $\mu_1 = 0.5$ is the midpoint of the TS $[L; U] = [0; 1]$ whereas $\mu_2 = 50$ represents the midpoint of the TS $[L; U] = [0; 100]$. μ_1 and μ_2 both represent the same order of magnitude of the trait but on different scales, so L and U can be chosen arbitrarily.

Without loss of generality consider a RS $r = \{r_1,\ldots,r_{k+1}\}$ with $r_1 = 1$, $r_{k+1} = k+1$, $k \in \mathbb{N}$, $r_{s+1} - r_s = 1\ \forall s \in 1,\ldots,k$ (e.g., the discrete scale $\{1, 2, 3, 4, 5\}$ of $k+1 = 5$ categories ranging from "not at all (1)" to "very much (5)").

Let $p \in (0; 1)$ quantify the LOQ. Symmetric values of lLOQ and uLOQ are assumed, that is, $lLOQ = 100 \cdot p/2$ and $uLOQ = 100(1 - p/2)$. Therefore, the edge areas are also symmetric with $|\delta_l| = |\delta_u| = p/2$.

Let $x' \in \{r_1,\ldots,r_{k+1}\}$ be an observed response value and let $p \in (0; 1)$ be the LOQ. The algorithm presented transforms any response value x' towards the TS $[0; 100]$ with due regard to p. In the following, if not otherwise stated, assume $L = 0$, $U = 100$, $r = \{1, 2,\ldots,k+1\}$ ($k \in \mathbb{N}$).

1. Choose $p \in (0; 1)$. Set $lLOQ = 100 \cdot p/2$ and $uLOQ = 100 \cdot (1 - p/2)$ (e.g., $p = 0.05$, lLOQ=2.5, uLOQ=97.5)

2. Define the $range := uLOQ - lLOQ$ and the step width $sw := range/k$ (e.g., $range = 97.5 - 2.5 = 95$ and $sw = 95/4 = 23.75$).

3. Let the observed response value be $x' = r_s \in \{r_1,\ldots,r_{k+1}\}$ with $s \in \{1,\ldots,k+1\}$ (e.g., $x' = 3$ corresponds to $s = 3$).

4. Calculate the response value $x^* = lLOQ + sw \cdot (s - 1)$ (e.g., $x' = 3$ and $x^* = 2.5 + 23.75 \cdot (3 - 1) = 50$).

For example, the algorithm transforms the RS $r = \{1, 2, 3, 4, 5\}$ towards the RS* $r^* = \{2.5, 26.25, 50, 73.75, 97.5\}$ ($p = 0.05$).

Please note that the bounds of r^* depend on p. $x^* \in (lLOQ; uLOQ)$ reflects the transformed order of magnitude of agreement (OMA) towards the item assertion. Any OMA value implies an order of magnitude of disagreement (OMD) towards the item assertion as a complement, say $100 - x^*$. Define $x = (x_1, x_2)^T \in \mathbb{R}^2$ with $x_1 := x^*$, $x_2 := 100 - x^*$, $x_1, x_2 > 0$ and $x_1 + x_2 = 100$ where $()^T$ denotes the transpose.

Generally, the compositional data space (the Simplex) is defined as $\mathcal{S} := \{x = (x_1, \ldots, x_D)^T \in \mathbb{R}^D \mid \sum_{i=1}^{D} x_i = \kappa \in \mathbb{R},\ x_i > 0\ \forall\ i = 1, \ldots, D\}$. With $D = 2$ and $\kappa = 100$ x fulfills the definition of compositional data [29, 30, 22, 20, 9]. Fig. 2 illustrates the Simplex of bipolar scales data.

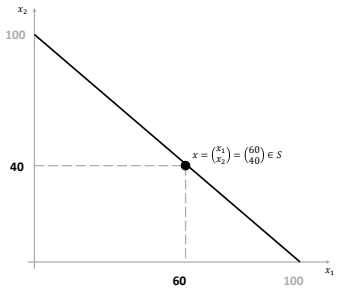


Figure 2. The black line illustrates the Simplex of bipolar scales data. x_1 (x_2) represents the OMA (OMD) towards the item assertion, respectively. The exemplary point $x = (60, 40)^T$ illustrates an OMA of 60 and an OMD of 40.

3.3. Ilr and inverse ilr transformation

Any compositional data point x depends on the Aitchison metric. However, most standard statistical procedures (e.g., computation of arithmetic means, Pearson correlation, (multiple) linear regression, t-tests) are based on the Euclidean metric. The ilr transformation yields interval scaled data underlying the Euclidean metric [10]. 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 $ilr(x) = ilr((x_1, \ldots, x_D)^T) := (z_1, \ldots, z_{D-1})^T$ with

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

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

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

For example, the ilr transform of the RS $r^* = \{2.5, 26.25, 50, 73.75, 97.5\}$ denotes $ilr((2.5, 97.5)^T) = -2.59$, $ilr((26.25, 73.75)^T) = -0.73$, $ilr((50, 50)^T) = 0$, $ilr((73.75, 26.25)^T) = 0.73$ and $ilr((97.5, 2.5)^T) = 2.59$.

Please note that the bounds of the ilr RS depend on p because the bounds of r^* depend on p. The smaller $p \in (0, 1)$ is, the closer are the bounds of r^* to 0 and 100, respectively. Therefore, $\lim_{p \to 0} \frac{r_1^*}{r_{k+1}^*} = 0$, $\lim_{p \to 0} \frac{r_{k+1}^*}{r_1^*} = \infty$ and $\lim_{p \to 0} \ln \frac{r_1^*}{r_{k+1}^*} = -\infty$, $\lim_{p \to 0} \ln \frac{r_{k+1}^*}{r_1^*} = \infty$, i.e., the spread of the ilr RS increases as $p \to 0$.

The inverse ilr is used to 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 $z = (z_1, \ldots, z_{D-1})^T \in \mathbb{R}^{D-1}$.

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

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

Like the 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

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

Again, the complete compositional data point is given by $x = (x^*, 100 - x^*)^T$. Applying the inverse ilr transformation to the ilr RS yields the RS r^*, e.g., invilr(0.73)=73.75 in the above example.

Please note that the ilr transformation differs from the logistic transformation only by the scaling factor $\sqrt{0.5}$. Both transformations consider $\ln \frac{x^*}{100 - x^*}$ in order to obtain interval scaled data. That is, mathematically they are practically identical if $D = 2$.

The idea of data evaluation is straight forward:

1. Apply the ilr transformation to obtain interval-scaled data.

2. Analyse the ilr transformed data using any appropriate statistical procedure (e.g., Shapiro-Wilk test, t-test, linear regression etc.)

3. Interpret the results on the interval scale.

4. If necessary: use the inverse ilr transformation to back-transform the results to the Simplex (e.g., apply the invilr to the arithmetic mean of ilr transformed data) and interpret.

3.4. The simulation

Table 1. Probability densities of the DGP used for simulation. $K = k + 1$ represents the number of responses of the RS. S=symmetric, LS=leptocurtic symmetric, U=U-shaped, EU=extremely U-shaped, PS=positively skewed and EPS=extremely positively skewed.

		K=4			
S	LS	U	EU	PS	EPS
0.175	0.1	0.35	0.4	0.25	0.5
0.325	0.4	0.15	0.1	0.45	0.25
0.325	0.4	0.15	0.1	0.2	0.15
0.175	0.1	0.35	0.4	0.1	0.1
		K=5			
S	LS	U	EU	PS	EPS
0.075	0.075	0.35	0.375	0.15	0.5
0.225	0.15	0.125	0.1	0.4	0.25
0.4	0.55	0.05	0.05	0.25	0.125
0.225	0.15	0.125	0.1	0.125	0.075
0.075	0.075	0.35	0.375	0.075	0.05
		K=6			
S	LS	U	EU	PS	EPS
0.1	0.05	0.325	0.375	0.15	0.1
0.15	0.1	0.125	0.1	0.4	0.55
0.25	0.35	0.05	0.025	0.2	0.125
0.25	0.35	0.05	0.025	0.125	0.1
0.15	0.1	0.125	0.1	0.075	0.075
0.1	0.05	0.325	0.375	0.05	0.05

We apply six different types of DGP to simulate item response data. They can be described as "symmetric" (S), "leptocurtic symmetric" (LS), "U-shaped" (U), "extremely U-shaped" (EU), "positively skewed" (PS) and "extremely positively skewed" (EPS). The PS and EPS DGP refer to floor effects [13]. For details please refer to Table 1. Inverting the labels of the RS (e.g. from $1,\ldots,5$ towards $5,\ldots,1$) transforms a positively skewed DGP into a negatively skewed DGP. Therefore, negatively skewed DGP representing ceiling effects are redundant.

In the following, we introduce relevant simulation parameters. According to [9] the number of responses $K = k + 1 \in \{4, 5, 6\}$ of the RS could affect the results. The number of items varies among real questionnaires. Choosing different numbers of items $I \in \{1, 2, 5, 10, 25, 50, 100\}$ we cover a broad range of real questionnaires. As statistical accurateness depends on the sample size we select different numbers of test persons $N \in \{25, 50, 100, 200\}$. The scale specific LOQ value affects the results of the ilr transformation. Selecting $p \in \{0.02, 0.1, 0.2\}$ we provide different values of the lower and upper LOQ, say $lLOQ = 100 \cdot p/2$ and $uLOQ = 100 \cdot (1 - p/2)$. We apply the inversion method to generate item response data. Let F be the cumulative distribution function (CDF) of a DGP of Table 1. Consider the generalized inverse CDF, say $F^{-1}(u) = \inf\{x|\ F(x) \geq u\}$ with $u \in [0;1]$. Let U be a continuous and uniformly distributed random variable with CDF

$$F_U(u) = \begin{cases} 0, & u < 0 \\ u, & u \in [0;1] \\ 1, & u > 1 \end{cases} \tag{6}$$

Using the R function `runif` random values $u \in [0;1]$ are generated [31]. Then, $F^{-1}(u)$ yields randomly generated item responses. The simulation is applied $B = 5000$ times to each combination of parameters (i.e., p, I, N, K) yielding 252 scenarios and an overall number of 1,260,000 simulation runs.

Generally, two real situations are possible: the construct considered is of low complexity (say LC, e.g., think of risk disposition concerning financial investments) or the construct is highly complex (HC, e.g., think of risk disposition in general). The larger the complexity of a construct the larger the number of items necessary for its operationalization. That is, a LC construct can be operationalized using a small number of items while HC constructs demand large numbers of items possibly representing various facets (i.e., sub-scales) of the construct.

Assume positively coded item responses such that a large OMA value implies a large OMT value. Responses to negatively formulated items can be recoded by reversion of the RS (e.g., $\{1, 2, 3, 4, 5\}$ recodes to $\{5, 4, 3, 2, 1\}$). The item responses of a LC construct can be assumed to result from a single DGP while HC constructs consisting of numerous facets imply different DGP. This paper considers the case of a HC construct, i.e., the DGP differs among items.

To provide a broad set of possible DGP combinations among items we randomly select the DGP with replacement using the `sample` function of the R `base` package (version 3.6.2). For example, consider a five-item scale and the random DGP set $\{S,U,S,LS,U\}$ providing a symmetric DGP (item 1), a U-shaped DGP (item 2), a symmetric DGP (item 3), a leptocurtic-symmetric DGP (item 4) and a U-shaped DGP (item 5). Using the DGP set N times we simulate a set of item responses of N test persons.

3.5. Evaluation procedure

Using the stats R package (version 3.6.2) we apply the well-known Shapiro-Wilk test (SW) [14]. Additionally, the Anderson-Darling test (AD) [32], the Lilliefors modification of the Kolmogorov-Smirnoff test (LF) [33], the Cramer-von-Mises test (CM) [33] and the Shapiro-Francia test (SF) [34] are applied using the `nortest` R package (version 1.0-4).

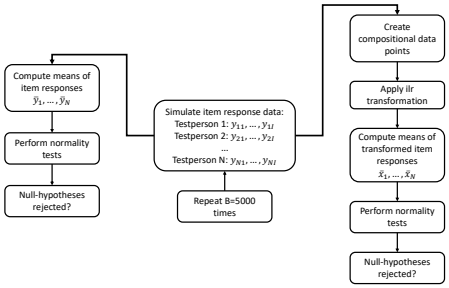


Figure 3. Visualization of the simulation process. The number of test persons N and the number of items I varies, see 3.4. The proportions of non-rejected null-hypotheses in both paths are compared.

The evaluation consists of two paths illustrated in Fig. 3. The normality tests apply to the means of non-transformed (NT) and ilr transformed item responses. We consider the proportions of non-rejected null-hypotheses ($PNR \in [0;1]$) in $B = 5000$ simulation runs. The difference $\Delta = PNR^{ilr} - PNR^{NT}$ indicates the shift towards normality when using the ilr approach.

4. Results and pracical implications

Overall, the results of the simulations indicate a shift towards normality when using the ilr approach. The order of magnitude of the shift seems to depend on the parameters I,K,N.

In practice, the LOQ is unknown and the choice of p depends on expert judgement [9]. Fig. 4 suggests that $p \in \{0.02, 0.1, 0.2\}$ hardly affects the values of Δ. Therefore, in the following we consider the arithmetic means of the three Δ values referring to the different values of p.

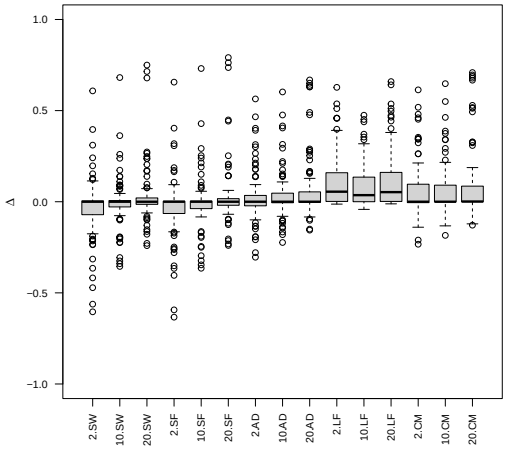


Figure 4. The normality tests used are SW=Shapiro-Wilk, SF=Shapiro-Francia, AD=Anderson-Darling, LF=Lilliefors, CM=Cramer-von-Mises.

The three boxplots referring to one specific normality test indicate the minor influence of p on Δ. That is, different values of p hardly affect Δ. However, Δ differs among the normality test used. Let $\tilde{\Delta}$ be the median Δ value and $\Delta_{0.25}$ and $\Delta_{0.75}$ be the 0.25 and the 0.75 quantile, respectively. In the following, the term Δ_{all} refers to all of the three values simultaneously. A value of $\Delta_{all} > 0$ indicates a shift towards normality, that is, the ilr approach is superior to traditional data analysis.

Let $Nortest \in \{SW,SF,AD,LF,CM\}$ be a normality test. Figure 5a indicates $\Delta_{all} \approx 0$ if $I = 1$. The number of items I also represents the number of summands. A value of $I = 1$ implies that no item response means can be computed and the CLT does not apply to either ilr transformed or non-transformed data resulting in missing approximate normal distribution in both cases. Thus, Δ_{all} must be near 0. On the other hand, if $I \geq 30$ means are approximately normally distributed [35]. Concerning $I = 50$ $\Delta_{all} \approx 0$ suggests that the normal approximation of the ilr approach and traditional analysis are similar. Increasing the number of summands enhances the CLT-postulated normal approximation [7]. 5a suggests $\Delta_{all} > 0$ if $I = 100$, that is, superiority of the ilr approach. Concerning $I \in \{2,5\}$ the convergence of item response means towards normality is small because of the small number of

summands. However, 5a suggests a larger convergence if the ilr approach is used. On the other hand, if $I \in \{10, 25\}$ means of non-transformed item response data show a larger convergence towards a normal distribution (SW, SF), a slightly larger convergence (AD, CM) and a smaller convergence (LF). That is, if $I \in \{10, 25\}$ the results are indifferent depending on the sensitivity of the normality test used. Regarding psychometric scales consisting of few items ($I \in \{2, 5\}$) or large numbers of items $I = 100$ the ilr approach seems most beneficial.

Fig. 5b suggests an increased convergence towards a normal distribution as K increases (compare AD, LF and CM). Additionally, K hardly affects Δ_{all} if $Nortest \in \{SW, SF\}$. The results of AD, LF and CM indicate the superiority of the ilr approach because $\Delta_{all} > 0$ while SW and SF indicate equal results, say $\Delta_{all} \approx 0$.

As the number of responses $K = k + 1$ increases Δ_{all} decreases (see Fig. 5b). While $\Delta_all > 0$ indicates the superiority of the ilr approach $\tilde{\Delta}$ indicates that it is most superior when using a short RS. Overall, the simulations provided evidence that the shift towards normality increases as K decreases and the overall shift towards normality seems substantial irrespective of the value of K.

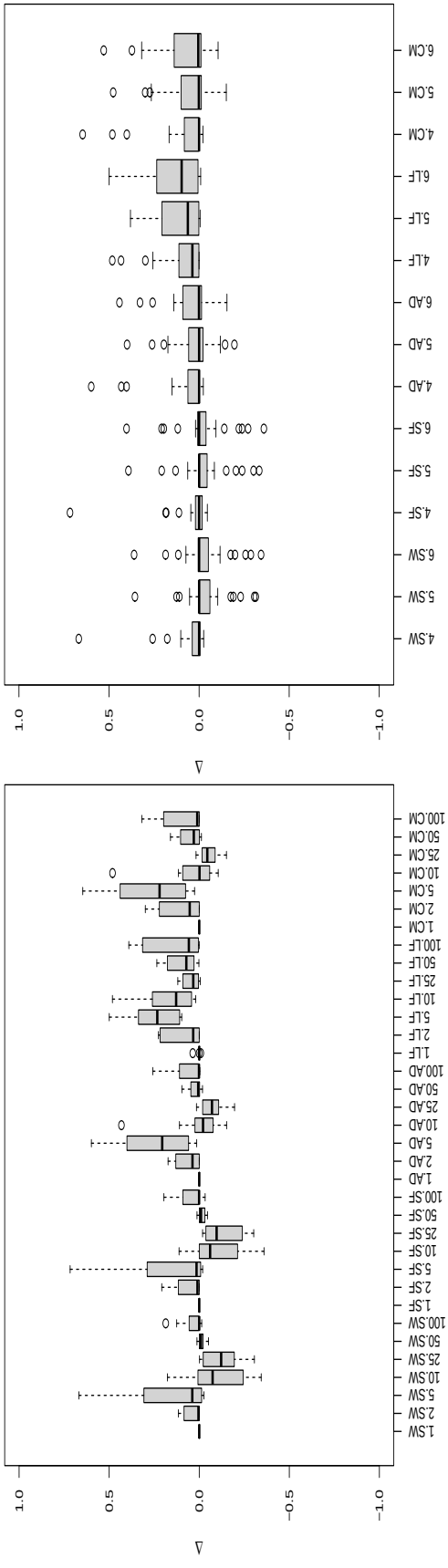
If $Nortest \in \{SW, SF\}$ Fig. 6a suggests $\Delta_{all} \approx 0$ for all numbers of test persons $N \in \{25, 50, 100, 200\}$ indicating no shift towards or away from normality. However, we have $\Delta_{all} > 0$ if $Norstest = AD$ ($N \in \{50, 100\}$) and $Nortest \in \{LF, CM\}$ ($N \in \{25, 50, 100, 200\}$) indicating a shift towards normality. That is, the ilr approach seems to increase the convergence of the distribution of item response means towards a normal distribution or not affect it all. Either way, a shift away from normality seems implausible.

Fig. 6b indicates $\Delta_{all} > 0$ if $Nortest \in \{AD, LF, CM\}$ and $\Delta_{all} \approx 0$ if $Nortest \in \{SW, SF\}$. That is, there is evidence that the ilr approach causes a shift towards normality or no shift at all. A shift away from normality seems implausible.

From a practical point of view, the ilr approach should generally be applied to bipolar psychometric scales data. In times of automation and high computational power there is no excuse to resign the benefits.

5. Practical implications in health economics

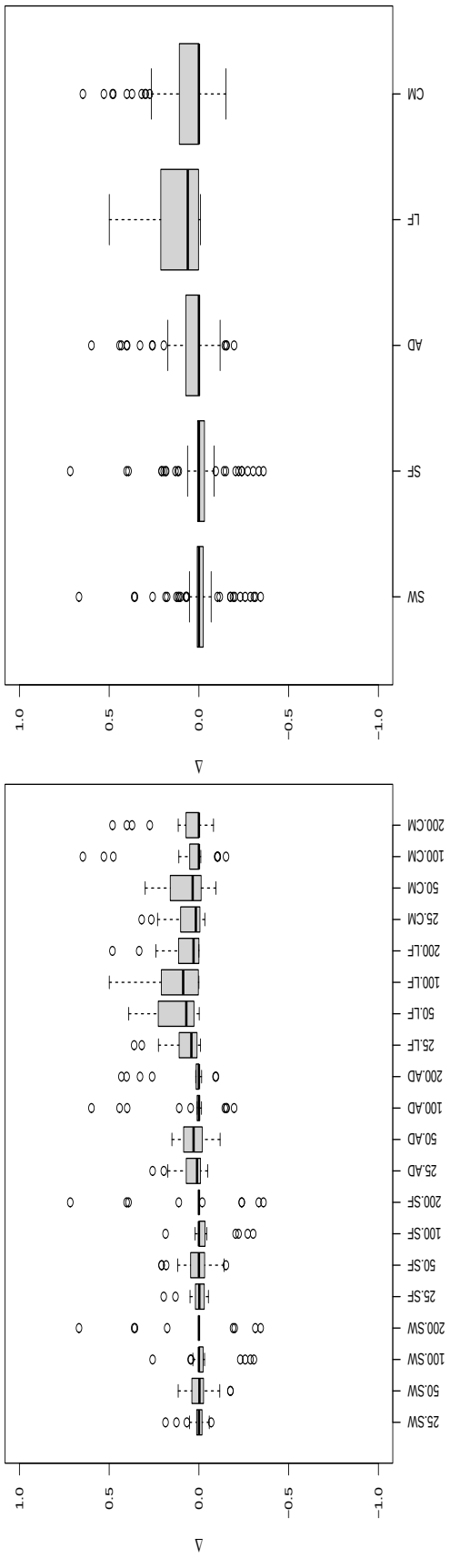
Measuring therapeutic success in health psychology focuses on pre-post-comparisons and associations of constructs. Obvious measures are Cohen's d, d_z and f as well as Pearson correlations [1]. The ilr approach



(a) The influence of I on Δ.

(b) The influence of $K = k + 1$ on Δ

Figure 5. The influence of the parameters I and $K = k + 1$ on Δ. $I \in \{1, 2, 5, 10, 25, 50, 100\}$, $K = k + 1 \in \{4, 5, 6\}$, the normality tests used are SW=Shapiro-Wilk, SF=Shapiro-Francia, AD=Anderson-Darling, LF=Lilliefors, CM=Cramer-von-Mises.



(a) The influence of N on Δ.

(b) The influence of the normality test on Δ.

Figure 6. The influence of the parameter N and the DGP on Δ. $N \in \{25, 50, 100, 200\}$. The DGP used are S=symmetric, LS=leptocurtic symmetric, U=U-shaped, EU=extremely U-shaped, PS=positively skewed, EPS=extremely positively skewed. The normality tests used are SW=Shapiro-Wilk, SF=Shapiro-Francia, AD=Anderson-Darling, LF=Lilliefors, CM=Cramer-von-Mises.

should be used to assess therapeutic success because it yields unbiased estimates of means, standard deviations and correlations [10]. Moreover, the ilr-induced shift towards normality increases the statistical power of procedures based on approximate normal distribution, e.g. t-test and correlation tests [36]. For example, [9] showed that the ilr approach increased the number of significances obtained when applying two sample t-tests to real data. Moreover, increasing statistical power via ilr transformation refers to an increase of the effect size in the ilr transformed data space, i.e., the effect size seems to be larger on the interval scale than on the traditional data scale [1]. As a result, the QALY index value associated with a therapeutic intervention increases affecting grant funding and incentives.

6. Discussion

The minor influence of p is congruent with the results of [9]. On the one hand, no Likert scale can cover all aspects of a construct. On the other hand, the scales used in practice have been validated, that is, the LOQ should be relatively small. Therefore, assuming $0 < p \leq 0.2$ seems reasonable. However, as p hardly affects the results the mid-value $p = 0.1$ seems to be a promising choice in practice.

The normality tests considered are not equally sensitive to distortion from a normal distribution. That is, the statistical powers differ depending on the properties of the DGP used [37]. Therefore, the influence of the normality test on Δ is by no means surprising. The different sensitivities of the tests, the randomness of the DGP sets and the different shapes of the individual DGP imply that no normality test can be considered superior to the other tests. That is, the results of all normality tests are relevant to the assessment of the ilr approach. Generally, we found evidence of a shift towards normality. However, the order of magnitude of the shift depends on different parameters, e.g., N and I.

Concerning I the results are surprising. Choosing $I \in \{10, 25\}$ seems to cause a minor/moderate shift away from normality while $I \in \{2, 5, 50, 100\}$ seems to shift the distribution of item response means towards a normal distribution. The above mentioned shift towards (away from) normality is suggested by all (not all) the normality tests applied. It seems that the values of I close to 30 with $I < 30$ represent values at the edge according to [35]. The ilr approach works satisfying if the number of items is small ($I < 10$) or large ($I \leq 50$). Concerning $I \in \{10, 25\}$ the results are inconclusive.

The larger the number of responses $K = k+1$ of a RS the more detailed are the information obtained. That is, the measurement becomes more precise [38] Increasing

the quality of the item response data also increases the quality of their arithmetic means. The results show that increasing K also increases PNR^{ilr} and PNR^{NT}. That is, both PNR values increase. It seems that the increase of PNR^{ilr} towards 1 is larger than the increase of PNR^{NT}, see Fig. 5b. Possibly, the randomness of the DGP sets affects the ilr approach less than the traditional approach supporting the impression of an ilr-induced shift towards normality.

The CLT postulates approximate normality of the sampling distribution of arithmetic means [39]. As stated by [35], the histogram of a set of arithmetic means approaches a bell-shape as the sample size (i.e, the number of arithmetic means N) increases. That is, the degree of approximate normality of a sample of arithmetic means depends on the number of means. Therefore, PNR^{ilr} and PNR^{NT} increase as N increases and, thus, $\Delta \to 0$. Consequently, the results presented in Fig. 6a are in accordance with statistical theory. Moreover, $\Delta > 0$ for small N indicates the superiority of the ilr approach.

7. Limitations

The simulations consider a set of six DGP while in practice the number of DGP is unlimited. Accounting for every nuance of symmetry, skewness or U-shape is impossible during simulation. However, the DGP considered cover a broad range of types of distributions including floor and ceiling effects and different numbers of responses $K = k + 1$.

We used the difference $\Delta = PNR^{ilr} - PNR^{NT}$ to indicate the shift towards normality when using the ilr approach. However, the choice is somewhat questionable. Reconsider the central limit theorem postulating an approximate normal distribution of the arithmetic mean. Thus, the null-hypothesis of the normality tests cannot be correct but it can be approximately correct. Actually, it is impossible to quantify the extent to which the null-hypothesis is approximately correct. The reason is disillusioning but true: one cannot define a certain degree of correctness or incorrectness, not even a convergence rate towards normality. That is, the statement "a decreased proportion of rejected null-hypotheses implies an increased normal approximation" is rather qualitative than quantitative.

Concerning the number of test persons the results seem generalizable towards $25 \leq N \leq 200$. Extrapolations towards $N > 200$ should be treated with caution. It is well-known that the sensitivity and specificity of normality tests depends on the sample size [37].

In practice the LOQ p is a scale specific unknown parameter. However, the results of the simulation and of [9] imply a marginal influence of the value of p. Future research should focus on an optimal choice of p concerning sensitivity and specificity of statistical tests. One of the questions to be answered is whether p should be chosen identically for all items of a scale or item specific. Should $lLOQ$ and $uLOQ$ be symmetric? What about chosing $p \in (0;1)$ at random or heuristcally according to [40].

Generalization from $K \in \{4,5,6\}$ towards $K > 6$ and from $I \in \{1,2,5,10,25,50,100\}$ towards $I > 100$ should be treated with caution. Further research is needed to assess the effects of the ilr transformation on Δ. To date, the effects of larger values of K and I on $\Delta \to 0$ are unclear, compare Fig. 5b-5a.

The results of the simulations are generalizable towards sums of item response values, because sums and arithmetic means only differ by the constant $1/I$. The CLT postulates approximate normality for both, sums and arithmetic means of random variables.

The randomness of the DGP set seems to affect the results of the ilr approach. In practice, often highly complex constructs consist of several homogeneous facets or sub-scales (see, e.g., [26]). Considering one specific sub-scale a single DGP could be reasonable for all items. That is, further research should focus on the effects of different single DGP on Δ.

References

[1] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*. Routledge, may 2013.

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

[3] 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.

[4] K. S. Button, J. P. A. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. J. Robinson, and M. R. Munafò, "Power failure: why small sample size undermines the reliability of neuroscience," *Nature Reviews Neuroscience*, vol. 14, pp. 365–376, apr 2013.

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

[7] H. Fischer, *A History of the Central Limit Theorem*. Springer, 2011.

[8] J. Davidson, *Econometric Theory*. Blackwell Publishing, 2001.

[9] R. Lehmann and B. Vogt, "Reconsidering bipolar scales data as compositional data improves psychometric healthcare data analytics," in *Proceedings of the 56th Hawaii International Conference on System Sciences*, Jan. 2023.

[10] 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.

[11] J. Aitchison, *The statistical Analysis of Compositional Data*. Blackburn Press, reprint of 1986 containing additional material ed., 2003.

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

[13] T. A. DeWees, G. L. Mazza, M. A. Golafshar, and A. C. Dueck, "Investigation into the effects of using normal distribution theory methodology for likert scale patient-reported outcome data from varying underlying distributions including floor/ceiling effects," *Value in Health*, vol. 23, pp. 625–631, may 2020.

[14] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrike*, vol. 52, pp. 591–611, 1965.

[15] S. Labovitz, "The assignment of numbers to rank order categories," *American Sociological Review*, vol. 35, no. 3, pp. 515–524, 1970.

[16] L. S. Mayer, "Comment on "the assignment of numbers to rank order categories"," *American Sociological Review*, vol. 35, no. 5, pp. 916–917, 1970.

[17] R. M. O'Brien, "Using rank category variables to represent continuous variables: Defects of common practice," *Social Forces*, vol. 59, no. 4, pp. 1149–1162, 1981.

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

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

[20] 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.

[21] 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.

[22] 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.

[23] K. Hron, M. Templ, and P. Filzmoser, "Imputation of missing values for compositional data using classical and robust methods," *Computional Statistics and Data Analysis*, vol. 54, no. 12, pp. 3095–3107, 2010.

[24] R. Lehmann, "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*, vol. 187, no. 1, p. 4149, 2014.

[25] 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 withcovid-19 vaccine hesitancy and resistancein ireland and the united kingdom," *Nature Communications*, vol. 12, no. 29, 2021.

[26] 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.

[27] 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.

[28] W. H. Loke, "The effects of framing and incomplete information on judgments," *Journal of Economic Psychology*, vol. 10, no. 3, pp. 329–341, 1989.

[29] J. Aitchison, G. Mateu-Figueras, and K. W. Ng, "Characterization of distributional forms for compositional data and associated distributional tests," *Mathematical Geology*, vol. 35, pp. 667–680, 2003.

[30] J. Aitchison, *A Concise Guide to Compositional Data Analysis*. Department of Statistics University of Glasgow, 2003.

[31] R Core Team, *R: A Language and Environment for Statistical Computing*, 2020.

[32] T. W. Anderson and D. A. Darling, "A test of goodness-of-fit," *Journal of the American Statistical Association*, vol. 49, pp. 765–769, 1954.

[33] H. C. Thode, *Testing For Normality*. CRC Press, jan 2002.

[34] P. Royston, "A pocket-calculator algorithm for the shapiro-francia test for non-normality: An application to medicine," *Statistics in Medicine*, vol. 12, pp. 181–184, jan 1993.

[35] S. G. Kwak and J. H. Kim, "Central limit theorem: the cornerstone of modern statistics," *Korean Journal of Anesthesiology*, vol. 70, no. 2, p. 144, 2017.

[36] R. Lehmann and B. Vogt, "Increasing the power of two-sample t-tests in health psychology using a compositional data approach," in *Brain Informatics* (F. Liu, Y. Zhang, H. Kuai, E. P. Stephen, and H. Wang, eds.), Springer Cham (in press), 2023.

[37] K. L. Boedec, "Sensitivity and specificity of normality tests and consequences on reference interval accuracy at small sample size: a computer-simulation study," *Veterinary Clinical Pathology*, vol. 45, pp. 648–656, aug 2016.

[38] C. C. Preston and A. M. Colman, "Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences," *Acta Psychologica*, vol. 104, pp. 1–15, mar 2000.

[39] C. Schröder and S. Yitzhaki, "Reasonable sample sizes for convergence to normality," *Communications in Statistics - Simulation and Computation*, vol. 46, pp. 7074–7087, apr 2017.

[40] S. Lubbe, P. Filzmoser, and M. Templ, "Comparison of zero replacement strategies for compositional data with large numbers of zeros," *Chemometrics and Intelligent Laboratory Systems*, vol. 210, p. 104248, mar 2021.