An Unsupervised Learning Approach to Evaluate Questionnaire Data— What One Can Learn from Violations of Measurement Invariance

MAX HAHN-KLIMROTH (ID)

PAUL W. DIERKES (ID)

MATTHIAS W. KLEESPIES (ID)

*Author affiliations can be found in the back matter of this article

]u[ubiquity press

ABSTRACT

In several branches of the social sciences and humanities, surveys based on standardized questionnaires are a prominent research tool. While there are a variety of ways to analyze the data, some standard procedures have become established. When those surveys want to analyze differences in the answer patterns of different groups (e.g., countries, gender, age), these procedures can only be carried out in a meaningful way if there is measurement invariance; i.e., the measured construct has psychometric equivalence across groups. As recently raised as an open problem by Sauerwein et al. (2021), new evaluation methods that work in the absence of measurement invariance are needed.

This paper promotes an unsupervised learning-based approach to such research data by proposing a procedure that works in three phases: data preparation, clustering of questionnaires, and measuring similarity based on the obtained clustering and the properties of each group. We generate synthetic data in three data sets, which allows us to compare our approach with the PCA approach under measurement invariance and under violated measurement invariance. As a main result, we obtain that the approach provides a natural comparison between groups and a natural description of the response patterns of the groups. Moreover, it can be safely applied to a wide variety of data sets, even in the absence of measurement invariance. Finally, this approach allows us to translate (violations of) measurement invariance into a meaningful measure of similarity.

CORRESPONDING AUTHOR:

Max Hahn-Klimroth

Goethe-University Frankfurt, 13 Max-von-Laue St, Frankfurt 60438, Germany

hahnklim@mathematik.uni-frankfurt.de

1. INTRODUCTION

In several branches of the social sciences and humanities, a prominent research tool is to conduct surveys using standardized questionnaires. One reason for the prominence of questionnaire-based studies may be that they are inexpensive, relatively easy to administer, and if the responses are standardized, it is easy to compile the data. Measurement instruments in such questionnaires can either consist of single questions that measure separate variables, such as questions about preferences or daily activities, or they can consist of multiple questions that can be aggregated into a single value or index. In the latter case, it is common to say that all the items in the questionnaire measure the same *construct*. They are often used when measuring attitudes (Milfont and Duckitt 2010), connection to nature (Mayer and Frantz 2004) or environmental behavior (F. G. Kaiser 1998). A common research question in both cases is whether several groups differ in their preferences, attitudes, or other environmental psychological measures. In this way, differences and similarities between age groups (Liefländer et al. 2013), between genders (Kleespies and Dierkes 2020) or across different disciplines (Feucht, Dierkes, and Kleespies 2023) can be examined.

There are a variety of ways to analyze and evaluate questionnaire data. Over the years, some standard procedures have become established in social science research and have been used in countless studies. When a large data set consisting of different psychological variables and constructs has been collected, a structurally simplifying procedure, such as a factor analysis or principle component analysis (PCA), is usually carried out to simplify the interpretation of the results (Costello and Osborne 2005; Yong and Pearce 2013).

However, these procedures require that the data collected are also suitable for carrying out such an analysis. To verify the applicability of these methods, Bartlett's test of sphericity and the Kaiser, Meyer and Olkin criterion (KMO) are usually applied (Bartlett 1950; H. F. Kaiser 1970). The KMO criterion assesses the sampling adequacy for each variable of the model and for the entire model, and Bartlett's test of sphericity tests whether there are correlations between the single items. A factor analysis or PCA only makes sense if these criteria are met. As a general rule, it is assumed that a factor analysis or PCA can only be applied if the Bartlett's test finds significant deviations from the hypothesis of having no correlations and the KMO is above the value of 0.7 (Dziuban and Shirkey 1974).

In these structure-simplifying procedures, similar items are assigned to the same higher-level factors. In the analysis, the individual items of the higher-order factors can be summarized by calculating a mean value. If there is only a single factor, this is referred to as a unidimensional model. In order to confirm the internal consistency (the inter-relatedness between the test items) and validity of the individual factors or components, Cronbach's alpha is often calculated (Cronbach 1951). The mean values of the individual factors can then be used to carry out group comparisons using for example hypothesis tests.

When making comparisons between different groups, it is only possible to carry out these procedures in a meaningful way if there is measurement invariance, i.e., the measured construct shows psychometric equivalence between groups (Putnick and Bornstein 2016). For example, it is possible that the perception of a measurement instrument differs between different cultural groups and that the factor analysis therefore produces a different factor structure for each group. In this case, there is a lack of measurement invariance and a comparison of the different groups is not easily possible. Verifying measurement invariance is a complex and multi-stage process that involves different stages (Knickenberg et al. 2019). Currently, methods are still needed that allow research data to be analyzed despite the lack of measurement invariance, as recently posed as an open problem (Sauerwein and Theis 2021). Such methods of analysis could help to carry out cross cultural studies, of which more are needed especially in environmental psychology (Tam and Milfont 2020).

1.1. OUR CONTRIBUTION

In this contribution, we propose an unsupervised learning-based approach towards such research data. As already described, standard methods require either a similar data structure in all subgroups, or at least comparable pairwise correlations between the individual items across all groups. However, especially when comparing heterogeneous groups, this cannot

be guaranteed, so the application of the standard methods described is not appropriate. In addition, missing data pose a major challenge, as the standard approach requires missing data to be replaced by the mean of the questionnaire, regardless of any correlations or similarities between different items. As an alternative, missing values are often simply ignored by excluding cases, although this has a negative impact on the sample size. Finally, from a more statistical point of view, following the standard approach results in dealing with multiple groups comparisons. With an increasing number of groups, the number of pairwise comparisons also increases, which quickly becomes confusing and very difficult to interpret when there are many groups. In this case, applied error corrections can contribute to the result becoming inaccurate. Our approach will analyze a questionnaire data set in three steps; more precisely, we describe an algorithm that:

(1) prepares the questionnaire data,

(2) clusters the questionnaires according to their *response types*,

(3) measures the similarity between groups using the proportion of questionnaires of each response type in the group.

In the data preparation step, the algorithm takes care of missing values in the original data using k-nearest neighbor imputation, and prepares the data for the actual clustering step. The clustering step clusters the individual questionnaires, and the centroids of the clusters will be called *response types*, as they refer to the *typical questionnaire* in each cluster. Finally, the proportion of each response type per group provides a very natural measure of similarity between groups, and further statistical analyses that might explain group similarities or differences can be applied based on this quantity.

In this paper, we give examples of this method applied to synthetic data and compare the result with the classical methods when they can be applied. We also give examples where our approach can be easily applied, but standard methods fail. Of course, the unsupervised learning approach itself (the actual clustering) and the imputation approach (nearest neighbor imputation) are well known and extensively studied methods in the data science community. However, the main goal of this paper is to combine these methods and to promote this approach for the evaluation of questionnaire data to a wide range of researchers who evaluate questionnaire data in different fields.

2. IMPORTANT DEFINITIONS AND NOTATION

2.1. STUDIED DATASETS

Below, we describe three synthetic datasets that are used throughout the paper. The datasets were created using the NumPy package in the Python programming language, and for completeness, the generated data are provided in the Supplementary Material. The first and second datasets are based on questionnaires consisting of seven items, where each item takes an integer value in [1, 5], and the third dataset consists of questionnaires with only three such items.

Questionnaires are often used to measure a *construct* using multiple questions. In this case, the value of each item x_i in a questionnaire can be described as a noisy measurement of some base value $b \in [1, 5] \cap \mathbb{Z}$. In particular, a questionnaire like (4, 4, 4, 5, 4, 5, 4) is expected to be observed, while (1, 3, 5, 2, 4, 1, 5) is very unlikely to occur if the questionnaire is answered truthfully. The first data set, $\mathcal{D}_1$, models this case.

As already explained, in the case of $\mathcal{D}_1$, the standard PCA can easily be applied because the data in each group show a similar, one-dimensional component structure, and only the distributions of the different base values of the groups might differ. However, in the other data sets, we model the case that some items do not measure the same construct as the others, thus the measurement invariance is violated. Formally speaking, in those cases, we are given multiple base scores per questionnaire (each corresponding to a construct). If, in each group, the same items measure the same constructs, PCA can be conducted and will find multiple relevant components. Now, the groups can be compared on each of those components independently. However, if different items measure different constructs in different groups, the standard approaches can be applied, but it is not clear how the results can be interpreted (Schmitt and

Kuljanin 2008; Van De Schoot et al. 2015). However, our approach is still a valid method in this case. Data set $\mathcal{D}_2$ models a corresponding questionnaire survey.

Finally, it is possible that a questionnaire does not try to measure a specific construct, but each item corresponds to the opinion on a certain topic. These topics are usually related in studies, but in principle they could be completely unrelated. The standard approach of a factor analysis, or a PCA, cannot be used to compare groups on such datasets. However, with the proposed approach, similarity between groups can be measured quite naturally. The third dataset, $\mathcal{D}_3$, is presented as an example of this scenario.

2.1.1. Case 1: measurement invariance is given

The first data set, $\mathcal{D}_1$, describes the situation where all items in the questionnaire measure the same construct in each group, but different groups have different perceptions of the items. This means that measurement invariance is given along the groups. Four different groups are simulated and each group contains 1,000 questionnaires. The groups differ in their perceptions of the construct, or more formally, the choice of the base value b differs between the groups. Group 1 is supposed to contain mostly questionnaires with high item values, while Group 2 typically gives moderately high answers. Finally, in Group 3, each opinion is roughly equally distributed, and an average person in Group 4 has either a high or a relatively low base value.

Formally, for each group i, 1,000 entries $v \in ([1, 5] \cap \mathbb{Z})^7$ are first sampled independently from a probability law $\mathcal{L}_i$, and in a second step the individual entries are perturbed. To formally describe the laws from which we sample in the first step, we denote by $\delta_x = \delta_{(x,x,x,x,x,x,x)}$ the Dirac measure on $\underline{x} \in \mathbb{Z}^7$, hence

$$\mathbb{P}(\delta_x(\tau)) = \mathbf{1}\{\underline{x} = \tau\},$$

and let unif (1, 5) denote the uniform distribution on $\{1, 2, 3, 4, 5\}^7$. Then,

$$\mathcal{L}_1 = 0.25 \cdot \delta_5 + 0.55 \cdot \delta_4 + 0.15 \cdot \delta_3 + 0.05 \cdot \text{unif}(1, 5),$$
$$\mathcal{L}_2 = 0.25 \cdot \delta_4 + 0.45 \cdot \delta_3 + 0.25 \cdot \delta_2 + 0.05 \cdot \text{unif}(1, 5),$$
$$\mathcal{L}_3 = 0.19 \cdot \delta_5 + 0.19 \cdot \delta_4 + 0.19 \cdot \delta_3 + 0.19 \cdot \delta_2 + 0.19 \cdot \delta_1 + 0.05 \cdot \text{unif}(1, 5),$$
$$\mathcal{L}_4 = 0.3 \cdot \delta_5 + 0.175 \cdot \delta_4 + 0.175 \cdot \delta_2 + 0.3 \cdot \delta_1 + 0.05 \cdot \text{unif}(1, 5).$$

In the second step, the value of each element x is perturbed by the following noise function F, so that

$$F(x) = \max\left(1, \min\left(5, \lfloor x + \mathcal{N}(0, 0.66)\rceil\right)\right).$$

This means that an independent Gaussian noise with mean 0 and variance 0.66 is added to the value of each element and the result is rounded to the nearest integer. Also, values above 5 and below 1 are truncated to 5 and 1, respectively.

2.1.2. Case 2: violations of measurement invariance

The second data set, $\mathcal{D}_2$, describes the situation where the items that do not measure the same construct are not the same across groups, and thus measurement invariance is violated. This means that, either the standard approaches do not provide an interpretation of the results, or even may not be applied. However, our approach can still be used to analyze similarities and differences across groups.

To create $\mathcal{D}_2$, we extend the first data set $\mathcal{D}_1$ by three additional groups, each consisting of 1,000 independent samples. The typical questionnaire in the additional groups, Group 5, Group 6, and Group 7, is described as follows. In Group 5, the first six items measure the same construct and typically assign a high value to that construct, but item 7 is expected to be answered with a low score. Group 6 questionnaires are expected to measure the same construct with moderately small baseline scores on items 1, 2, 3, 5, 6, 7, but item 4 is expected to be atypically high. Finally, in Group 7, item 4 is expected to be large, item 7 is typically small, and the other items measure the same construct with either large or small base values.

Formally, we describe the probability laws from which we sample, as follows. Again, we denote by δ_v the Dirac measure on v and by unif (1, 5) the uniform distribution on $\{1, 2, 3, 4, 5\}$. Then,

$$\begin{aligned}
\mathcal{L}_5 =\ & 0.24 \cdot \delta_{(5,5,5,5,5,5,1)} + 0.24 \cdot \delta_{(5,5,5,5,5,5,2)} \\
& + 0.24 \cdot \delta_{(4,4,4,4,4,4,2)} + 0.24 \cdot \delta_{(4,4,4,4,4,4,1)} + 0.04 \cdot \mathrm{unif}(1,5), \\
\mathcal{L}_6 =\ & 0.32 \cdot \delta_{(1,1,1,4,1,1,1)} + 0.32 \cdot \delta_{(3,3,3,5,3,3,3)} \\
& + 0.32 \cdot \delta_{(2,2,2,5,2,2,2)} + 0.04 \cdot \mathrm{unif}(1,5), \\
\mathcal{L}_7 =\ & 0.12 \cdot \delta_{(5,5,5,5,5,5,2)} + 0.12 \cdot \delta_{(5,5,5,4,5,5,1)} + 0.12 \cdot \delta_{(4,4,4,4,4,4,1)} \\
& + 0.12 \cdot \delta_{(4,4,4,5,4,4,2)} + 0.12 \cdot \delta_{(2,2,2,4,2,2,1)} + 0.12 \cdot \delta_{(2,2,2,5,2,2,1)} \\
& + 0.12 \cdot \delta_{(1,1,1,5,1,1,1)} + 0.12 \cdot \delta_{(1,1,1,4,1,1,1)} + 0.04 \cdot \mathrm{unif}(1,5).
\end{aligned}$$

Having sampled 1,000 elements for each group independently, the same perturbation as before is applied, meaning that each value is perturbed by

$$F(x) = \max\left(1, \min\left(5, \lfloor x + \mathcal{N}(0, 0.66) \rceil\right)\right).$$

2.1.3. Items are unrelated and show differences between groups

The last example data set is called $\mathcal{D}_3$. The main idea is that each item asks about a different construct, which means that we expect the items to be weakly correlated. In this case, we simulate only three items per questionnaire, so that each questionnaire is represented as a point in $[1, 5]^3 \cap \mathbb{Z}^3$. The typical response to each item differs by group. Four groups are modeled: Group 8 is expected to have very different answers to each question, Group 9 is expected to have a high score for item 1, Group 10 is expected to have a low score for this item, and finally Group 11 is expected to have a relatively high score for item 2.

Formally, for Group i we sample 1,000 questionnaires from the corresponding probability law $\mathcal{L}_i$ and then perturb the individual entries, but the perturbation will be different from the previous cases. Analogously as before, let $\mathrm{unif}(1, 5)$ denote the uniform distribution on $\{1, 2, 3, 4, 5\}^3$ and let

$$\sigma_1 = (5,3,1), \quad \sigma_2 = (1,3,5), \quad \sigma_3 = (3,3,3), \quad \sigma_4 = (5,1,3), \quad \sigma_5 = (1,5,3), \quad \sigma_6 = (3,5,1).$$

Then define

$$\begin{aligned}
\mathcal{L}_8 &= 0.16 \cdot \delta_{\sigma_1} + 0.16 \cdot \delta_{\sigma_2} + 0.16 \cdot \delta_{\sigma_3} + 0.16 \cdot \delta_{\sigma_4} + 0.16 \cdot \delta_{\sigma_5} + 0.16 \cdot \delta_{\sigma_6} + 0.04 \cdot \mathrm{unif}(1,5), \\
\mathcal{L}_9 &= 0.4 \cdot \delta_{\sigma_1} + 0.02 \cdot \delta_{\sigma_2} + 0.12 \cdot \delta_{\sigma_3} + 0.4 \cdot \delta_{\sigma_4} + 0.02 \cdot \delta_{\sigma_5} + 0.02 \cdot \delta_{\sigma_6} + 0.02 \cdot \mathrm{unif}(1,5), \\
\mathcal{L}_{10} &= 0.06 \cdot \delta_{\sigma_1} + 0.4 \cdot \delta_{\sigma_2} + 0.04 \cdot \delta_{\sigma_3} + 0.02 \cdot \delta_{\sigma_4} + 0.4 \cdot \delta_{\sigma_5} + 0.06 \cdot \delta_{\sigma_6} + 0.02 \cdot \mathrm{unif}(1,5), \\
\mathcal{L}_{11} &= 0.07 \cdot \delta_{\sigma_1} + 0.07 \cdot \delta_{\sigma_2} + 0.07 \cdot \delta_{\sigma_3} + 0.07 \cdot \delta_{\sigma_4} + 0.35 \cdot \delta_{\sigma_5} + 0.35 \cdot \delta_{\sigma_6} + 0.02 \cdot \mathrm{unif}(1,5).
\end{aligned}$$

After the sample procedure, we perturb each item's value by the noise function G, where

$$G(x) = \max\left(1, \min\left(5, \lfloor x + \mathcal{N}(0,1) \rceil\right)\right).$$

Compared to $\mathcal{D}_1$ and $\mathcal{D}_2$, the perturbation is slightly larger to be able to observe a variety of different questionnaires representing that all items measure different constructs.

2.2. WARD'S CLUSTERING METHOD

The clustering obtained in this paper is due to performing a standard agglomerative clustering with Ward's minimum variance criterion as the objective function (Ward 1963). When this method is applied to a data set of size *n*, in the first step of the clustering algorithm, all *n* data points form their own cluster. Now, in each step, the two clusters whose merging minimizes the total within-cluster distance are merged, and the cluster center of the new cluster is computed as the point minimizing the sum of squares distance to all points in the cluster. More precisely, in each step, the clustering algorithm must find the pair of clusters that leads to the minimum increase in the total within-cluster variance after merging (Cormack 1971; Ward 1963). Formally, we choose those clusters *A* and *B* with centers c_A and c_B that minimize

$$\Delta(A, B) = \frac{|A||B|}{|A| + |B|} \| c_A - c_B \|^2 .$$

This clustering approach is quite popular because it usually produces compact and comparably sized clusters (Szmrecsanyi 2012).

However, the algorithm requires a stopping criterion. This can either be calculated automatically, e.g., when the increase of the within-cluster variance exceeds a certain threshold, or when a certain number of clusters is reached. We follow the second approach, where the user determines the number of clusters obtained by the method. This choice can be guided by clustering indices.

2.3. DETERMINING THE NUMBER OF CLUSTERS

A fairly intuitive, however, quite recent, approach to determine a suitable number of clusters during agglomerative clustering, is based on the so-called *gap statistic* (Mohajer, Englmeier, and Schmid 2010; Tibshirani, Walther, and Hastie 2001). It compares the cluster compactness of a given clustering with a null reference distribution of the data, which is data with no (obvious) clustering. The number of clusters suggested by the method is the value for which cluster compactness on the original data is significantly smaller than the cluster compactness on the reference data (Mohajer, Englmeier, and Schmid 2010; Tibshirani, Walther, and Hastie 2001). Hence, we are looking for a (local) maximum in a scree plot which plots the number of clusters against the gap value. Intuitively, this corresponds to 'unnaturally large gaps' in a corresponding dendrogram. A dendrogram, which represents a tree, illustrates the arrangement of clusters produced by an agglomerative clustering process. The leaves of the tree are the individual data points, and whenever two clusters are merged, an edge is used to visualize the merging. The corresponding height, the distance from the leaves, is equal to the 'distance' of the cluster centroids at that moment. The clusters are induced by a horizontal line, so the tree is cut into a forest by removing all lines above this line. The height of such a natural horizontal line should correspond to the existence of a large level gap in the dendrogram (Tibshirani, Walther, and Hastie 2001).

Of course, there are several other indices which can be used to measure the goodness of a clustering, like the Calinski-Harabasz index (Calinski and Harabasz 1974; Liu et al. 2013) or the Silhouette coefficient (Rousseeuw 1987). However, while the gap statistic can be used with any clustering algorithm, the latter indices are known to prefer convex clusters over non-convex clusters, even if a non-convex variant might intuitively reflect the better clustering (Liu et al. 2013; Rousseeuw 1987), in particular if an underlying community structure is supposed to exist. Due to this reason, we decided to use the gap statistic in this contribution.

3. EVALUATION ALGORITHM FOR QUESTIONNAIRE DATA

In this section, we present the proposed evaluation method in detail. As previously described, the algorithm used to evaluate the questionnaire data runs in three phases. The first phase is the data preparation phase and consists of the following steps.

Phase 1 Data preparation.

Data: Original data matrix A

Result: Prepared data matrix D

Fill missing values by k-nearest neighbor imputation and call the resulting matrix A'

Balance group sizes by upsampling rows of smaller groups

Perturb each item value independently by a defined noise function

---

The first step of the data preparation phase is called data imputation. More specifically, there may be questionnaires in which some items were not answered. Such missing data must be imputed (or the questionnaire ignored). One method of imputing data that takes into account the correlations between items is k-nearest neighbor imputation (Cover and Hart 1967; Troyanskaya et al. 2001). It samples the k most similar questionnaires given the existing items, and fills in the missing items with the average score of the items in the sample. More formally, suppose that the questionnaire data is expressed by $v \in A \subset \mathbb{Z}^d$ and $M \subset \{1,...,d\}$ are the indices of the items with no answers. Then, we define $W = \{x_{\overline{M}} \in \mathbb{Z}^{d-|M|} \mid x \in A\}$ and find the k closest points $X_k(v) \subset W$ to $v_{\overline{M}}$. Finally, set $v_\ell = \text{avg}_{x \in X_k(v)}(x_\ell)$ for $\ell \in M$.

The next step, the data balancing step, is required to obtain a meaningful clustering of the questionnaires. As we assume that the distribution of questionnaires might vary between

groups, and the sample sizes between the groups might also vary, we need to make sure that the questionnaires appearing (only) in groups with a comparably small sample size are not irrelevant during the clustering step. A standard approach to guarantee this in (supervised) learning tasks is to balance the data by oversampling *minority* groups and/or downsampling *majority* groups (Chawla et al. 2002; Menardi and Torelli 2012). In our case, we propose to oversample groups with smaller sample sizes until all groups contain equally many questionnaires, as the actual evaluation will only be with respect to original data and not the synthetically oversampled data (see Phase 2).

The last data preparation step is to perturb each item's value slightly with independent additive Gaussian noise with mean zero and standard deviation 0.1. The main purpose is that the clustering (e.g., the cluster *centers*) becomes much more stable towards adding or removing single data points if the raw data is augmented. This is a well-known principle, not only in clustering, but in various machine learning tasks in which the models generalize much better if the training data is augmented by random noise (Belinkov and Bisk 2018; Ding et al. 2007; Min et al. 2020; Zhang, Kinoshita, and Kiya 2020). Moreover, the data matrix becomes full rank as there are no duplicate rows anymore (with high probability), which increases the numerical stability of the computation.

Data: Prepared data matrix D

Result: Cluster centers R (*response types*) and the groups' *fingerprints* per group F

Obtain l cluster centers R using the data D by agglomerative clustering with Ward's linkage

For each row of A, calculate the closest cluster center of V and assign the cluster to the corresponding questionnaire.

Calculate the fingerprints types per group F

Phase 2 Questionnaire clustering.

The second phase, the clustering phase, starts with clustering the oversampled and augmented questionnaires given by D. As a clustering algorithm, we propose to perform a standard agglomerative cluster analysis with Ward's minimum variance method as the objective function (see Section 2.2). The number of clusters l is determined visually using the gap statistic (Tibshirani, Walther, and Hastie 2001) based on a scree plot as well as a dendrogram. As explained earlier, the number of clusters is expected to be at a local maximum in the scree plot. The main idea is described in Section 2.3.

Once the clusters are obtained, we compute the corresponding cluster centers (geometrically speaking, the centroid of each cluster) and call these points *response types*. While mathematically the response types are really just cluster centroids, the name should reflect the fact that we expect a *typical questionnaire* in the cluster to follow that response type. We call the set of all response types R and fix an arbitrary order. Let the ordered response types be $r_1, \ldots, r_l$.

Now we *forget* the augmented data D and return to using the original (but imputed) data matrix A'. For each group i, we compute the group's *fingerprint* f_i as a point in the l-dimensional standard simplex as follows. For each $j \in \{1,\ldots,\ell\}$, we let f_i be the proportion of questionnaires in group i in the cluster corresponding to response type j.

Data: Response types R and fingerprints F

Result: Group similarity

Interpretation of the response types R content-wise

Measure the similarity of the fingerprints F

(optional) use additional data and methods to analyze the groups' fingerprints

---

Phase 3 Measuring group similarity.

The last phase of the proposed method combines the explorative data-driven approach with the actual content interpretation. First, the response types can be interpreted as a typical response to a questionnaire in that cluster. The fingerprints thus reflect the distribution of people following a certain response type in the different groups. The more similar two fingerprints are between two groups, the more similar people answered the questionnaires, which is a natural measure of similarity between groups.

In Section 2.3, it was already explained that a dendrogram (a tree representation of a clustering algorithm) can be used to determine how many natural or robust clusters exist. It also yields a very intuitive description of similarity between data points, as those points whose clusters merge *earlier* are more similar. Such a notion of similarity is also standard outside of data science; for example, in ecology and evolution such dendrograms are known as phylogenetic trees, and show the evolutionary relationships among species (Letunic and Bork 2006).

As an optional last step, the fingerprints in combination with the response types may be explained by group specific properties. This step is not related to cluster analysis, nor is it part of the proposed method, but for completeness we present it here. For example, suppose the groups are different countries, and the response types are easy to interpret: response type 1 might reflect a high interest in conservation, response type 2 might reflect a high interest in conservation in principle, but some parts of preservation are irrelevant to the people, etc. Thus, fingerprints with a high value in response type 1 reflect countries where the majority of people are highly interested in conservation, and fingerprints with a high value in response type 2 reflect countries where people are also interested in conservation, but certain aspects are irrelevant. These results could be explained by indices that describe countries, such as wealth indices or a country's forest cover. A simple but powerful way to test such hypotheses is to measure the rank correlation between the marginal of the fingerprint representing a particular response type and the corresponding index.

4. RESULTS

4.1. STRUCTURE OF THE CONSTRUCT EQUALS IN ALL GROUPS

4.1.1. Factor analysis or PCA with follow-up testing

In the first phase, the applicability of PCA was assessed for each of the four groups using Bartlett's test and the Kaiser-Meyer-Olkin (KMO) criterion. Given the significance of the Bartlett's test ($p < 0.001$) and a KMO criterion of over 0.700 in all groups, PCA was considered appropriate. The PCA in all groups showed that the 7 items could be combined into one higher-order component according to the Kaiser criterion (F. G. Kaiser 1998). The calculation of the Cronbach's alpha for this component showed a high internal consistency and reliability between the items for all four groups ($\alpha > 0.700$). After determining the component, the mean values of the items were calculated for each data point. To determine differences between the groups, those mean values were compared using a hypothesis test between groups. We applied a Kruskal-Wallis test followed by the Dunn-Bonferroni post-hoc test. The level of significance was adjusted by the Bonferroni correction. The results demonstrated pairwise significant differences between all groups, with the exception of Groups 2–3 ($p = 0.564$) and 3–4 ($p = 1.00$).

4.1.2. Our approach

First, the data preparation described in Phase 1 was applied to the data. Second, we needed to determine an appropriate number of response types. Figure 1 contains the gap values of the gap statistic as well as the corresponding dendrogram. Clearly, this indicated that 5 response types should be used. Since the underlying data was artificially generated, we knew that we should expect about 5 response types in $\mathcal{D}_1$, as all items in most questionnaires measured the same construct, but the expected value was different. However, due to random fluctuations, we could have observed more response types. For completeness, we give an example of the stability of our approach with respect to more response types in section 4.4. The response types reflect, as one would expect, the questionnaires (1, ..., 1) to (5, ..., 5), with some slight noise.

The next step is to compute the fingerprints of the 4 groups and express the similarity of the groups (see Figure 2). The four groups are visually different. Group 1 is concentrated on response types corresponding to large uniform responses, Group 2 is concentrated on medium large answer patterns, Group 3 contains roughly equally many questionnaires of any response type, while in Group 4, most questionnaires contain either quite small or quite large answers. When compared to the model used to generate the data, it is immediately apparent that this is a very good reconstruction of the actual data which is also easy to interpret content-wise.
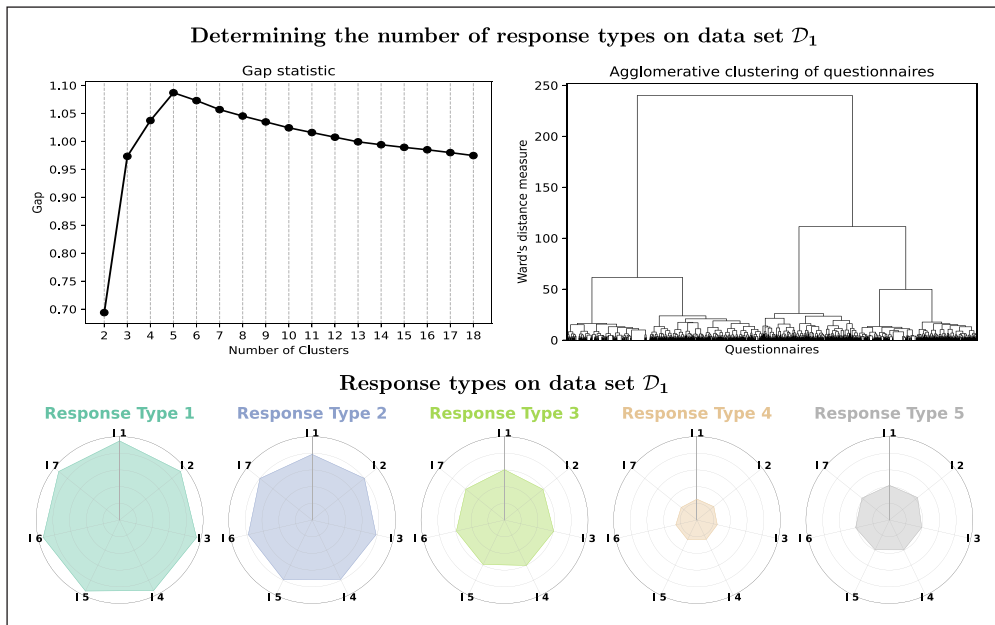
4.2. STRUCTURES OF THE CONSTRUCT DIFFERS BETWEEN GROUPS

4.2.1. Factor analysis or PCA with follow-up testing

In the case of the data set $\mathcal{D}_2$, the standard approach using factor analysis or PCA with subsequent follow-up testing encounters problems. The results of the Bartlett's test and KMO criterion indicated that a PCA can be performed for Groups 5, 6, and 7 ($p < 0.001$; KMO > 0.700). However, several problems arise from the results of the PCAs. In Group 5, item 7 has a much lower loading compared to the other items, while in Group 6, item 4 has a noticeably lower loading. In Group 7, according to the Kaiser criterion, an additional second component is formed by items 4 and 7. As the results of the PCA differ considerably from one another, it is not possible to simply calculate a mean value to compare the groups, as in the previous data set $\mathcal{D}_1$. A common approach in such cases would be to delete the items with low factor landings or cross-landings from the data set (Yong and Pearce 2013). However, this is likely to result in information being lost, and in some cases is not possible due to the low number of items. Especially if measurement invariance is not given, it is not possible to test the constructs between the groups in a meaningful way (Putnick and Bornstein 2016).

4.2.2. Our approach

Again, Figure 3 gives an overview of the indices that determine the number of response types as well as the similarity of the groups. The local maximum in the gap statistic is at 10 clusters, and those 10 clusters are highly visible in the corresponding dendrogram. As in the previous case, this fits well with the model used to generate questionnaires. We expect up to five *symmetric* clusters in which all items have roughly the same value, as well as clusters in which the typical item is small to moderate but item 4 is large (response types 1, 2 & 5), and finally clusters in which the typical item is large but item 7 is small (response types 7 & 8), see Figure 3.

Next, the fingerprints of the groups are computed and the similarity of the group is expressed in Figure 4. Again, the group fingerprints reflect the actual groups very well. Groups 1 to 4 are described similarly to the previous case; they are still concentrated on those response types which express uniform answers of different height. Moreover, the 'new' groups are also well
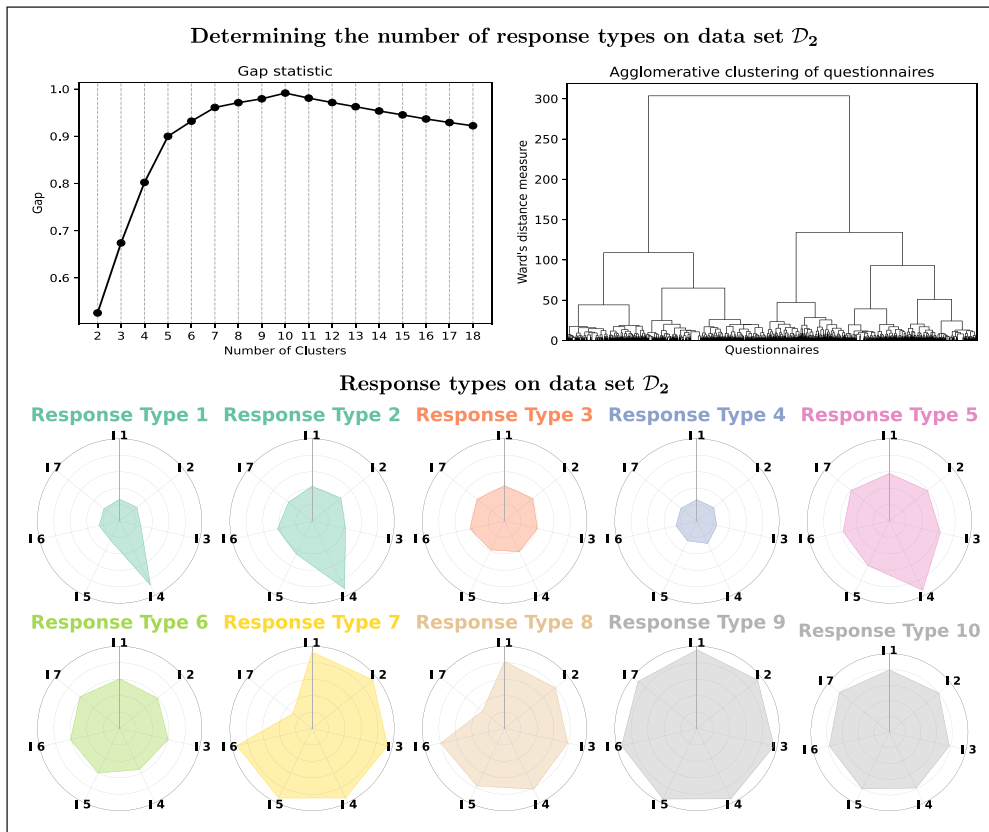
Figure 3 Graphical representation of the gap statistic as well as the dendrogram corresponding to the goodness of the clustering of the questionnaires in data set $\mathcal{D}_2$. Moreover, the corresponding response types are shown as a spider plot.
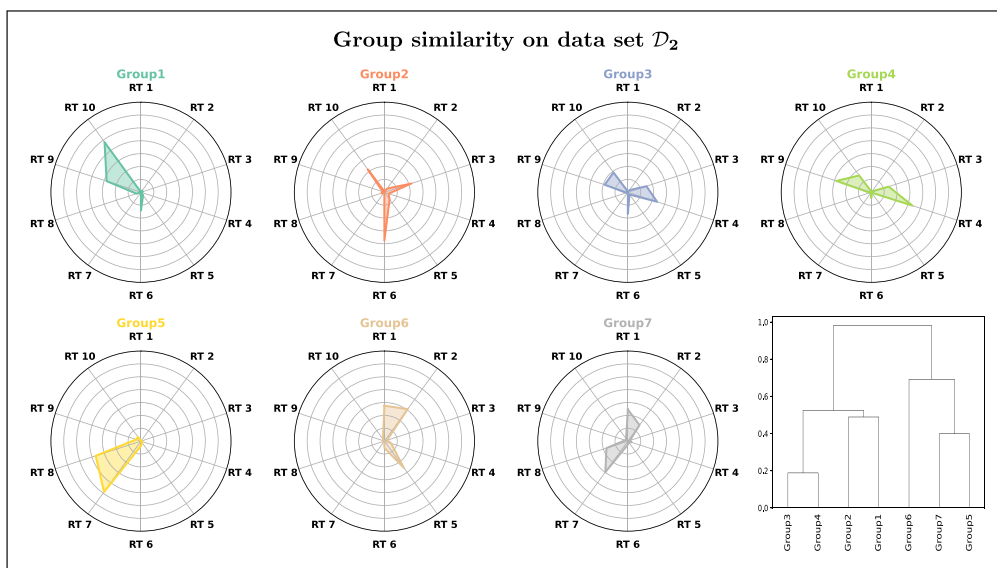

Figure 4 The fingerprints of the different groups regarding the response types as spider plots. The radial *y*-axes are scaled to (0, 0.7). Moreover, the group similarity on data set $\mathcal{D}_2$ is shown as a dendrogram.

described, in particular the distribution on the fingerprints yields the following interpretations: Group 5 has many high responses and item 7 is artificially small. Group 6 has small to medium responses but item 4 is large. Finally, in Group 7 we observe a large proportion of questionnaires in which all answers are small but item 4 is large (response types 1 & 2), but also questionnaires with high answers in which item 7 is comparatively small (response types 7 & 8). We also observe that the previously more similar groups (Group 3 and Group 4) are measured as similar again, and Group 1 and Group 2 are still more similar than either of them is to Group 3 or Group 4. The measure of similarity is thus stable with respect to adding data of more groups.

4.3. ITEMS ARE UNRELATED AND SHOW DIFFERENCES BETWEEN GROUPS

4.3.1. Factor analysis or PCA with follow-up testing

When items show only a low correlation with each other, for example, because they measure completely different constructs, a PCA or factor analysis is not applicable. For $\mathcal{D}_3$, the KMO criterion for all four groups indicates that the conditions for applying a PCA are not met (KMO

< 0.4*).* Therefore, the standard approach described above is not applicable for this data set. If a PCA is nevertheless conducted, it produces results that cannot be meaningfully interpreted. The items form two components for each of the groups, which cannot be separated from each other due to high cross-loadings. In addition, the components differ between the four groups. Standard rotation methods (such as varimax) do not improve the result. A meaningful evaluation, a group comparison or a follow-up test, is therefore not possible using this method.

4.3.2. Our approach

As before, we plot the response types in Figure 5. The gap statistic suggests the use of 5 response types, and in the dendrogram, one would choose 5–6 response types. Again, this fits well with the actual data generation, which is based on noisy instances of 6 types. The response types correspond to noisy measurements of five of the six ground-truth values $\sigma_1, \ldots, \sigma_6$, but $\sigma_3 = (3, 3, 3)$ does not appear as a response type. This might well be due to the relatively large noise applied to each coordinate, such that a typical sample from σ_3 will have different entries.
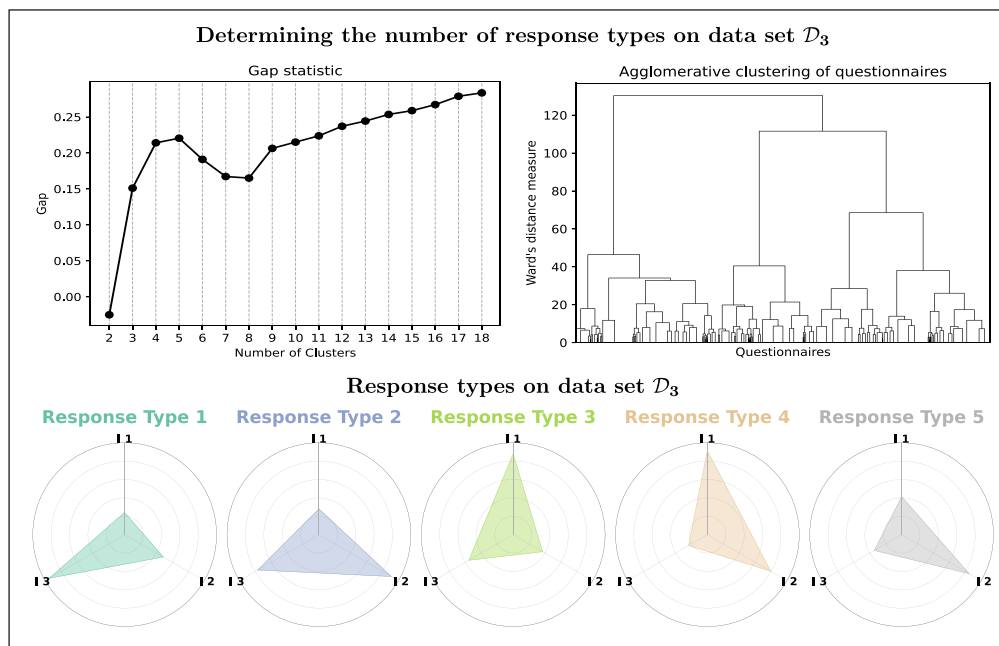
Regarding the interpretation of the group's fingerprints (see Figure 6), we observe that Group 8 contains questionnaires of each response type, Group 9 is mostly concentrated on response types with a high score for item 1, Group 10 is concentrated along those response types in which item 1 is small, and finally, Group 11 has large entries in item 2. This reflects the actual data model very well.
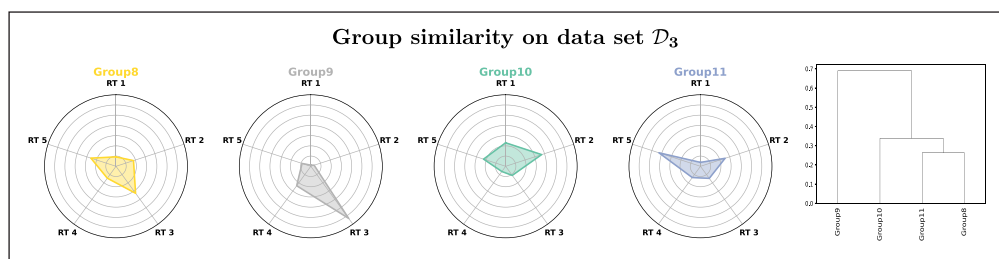
4.4. ROBUSTNESS TOWARDS THE NUMBER OF CLUSTERS

In this section, we briefly show that a slight over-estimation of the number of clusters does not change the similarity between groups significantly. In Figure 7, we present the dendrogram describing the groups' similarity on data set $\mathcal{D}_1$ for the optimal choice of 5 response types, as well as for 6, 7, and 8 response types. As can be easily observed, the similarity between groups does not change significantly.
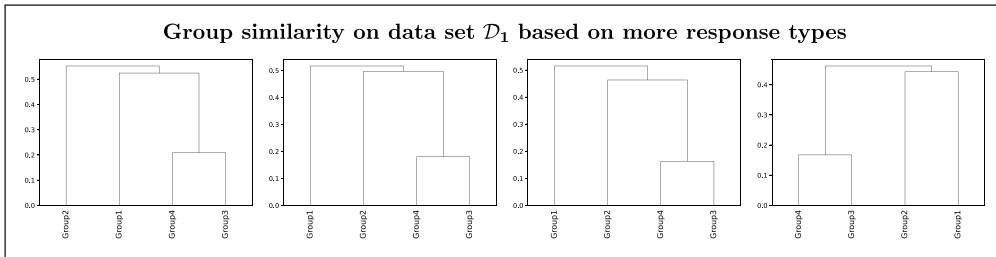
Group similarity on data set $\mathcal{D}_1$ based on more response types

While the similarity itself does not change, it is important to notice that over-estimation of the number of response types clearly has its drawbacks. The main challenge might arise, as the cluster centroids (the typical questionnaire per cluster) are no longer well separated and potentially harder to explain content-wise. Recall from Figure 1 that the five response types were very easy to interpret: they referred to the typical sheets 1...1–5...5 up to some noise. However, if eight response types are formed, they are not that easy to describe (see Figure 8). For example, we observe that response types 4 & 5 do clearly emerge from the previous response type 3 (see Figure 1). They contain questionnaires in which the answers are around the typical answer 3, but the cluster centroids are are slightly 'deformed' rather than being roughly uniform in all coordinates. This is not desirable, because, obviously, even if all items measure the same construct and the participant answers with care, the questionnaires (4, 3, 3, 3, 3, 2, 3) (in the cluster of response type 4) and (3, 3, 3, 4, 3, 3, 3) (in the cluster of response type 5) are highly likely to be observed and should, intuitively, both correspond to the same response type.
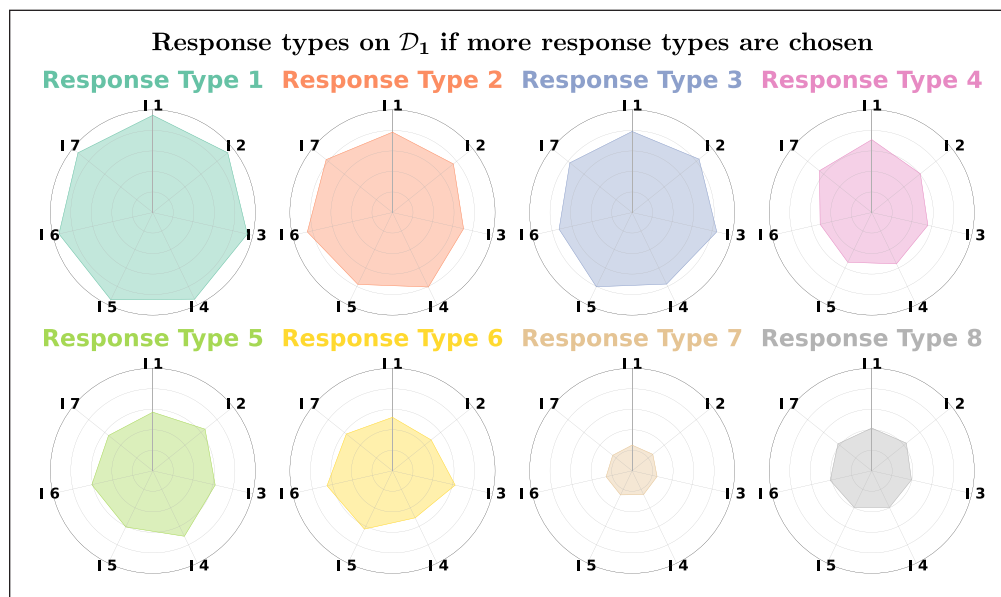

Response types on $\mathcal{D}_1$ if more response types are chosen

To summarize, it is unproblematic to over-estimate the number of response types with regard to the similarity of the groups, but the response types might become harder to interpret.

5. DISCUSSION & CONCLUSION

5.1. ON THE DATA PREPARATION STEP

All the steps used in the data preparation step are well known to the data science community, but in the described application, namely the evaluation of questionnaire data, these steps are unlikely to be found. The first step was to impute missing data by *k*-nearest neighbor imputation. Normally, such missing data are filled by simply taking the average (either of the row, or of the column) (Troyanskaya et al. 2001; Weaver and Maxwell 2014), or simply ignored (Husson and Josse 2013), but this either does not take into account the dependencies between different items, or it reduces the sample size. Therefore, especially since our method is also applicable to questionnaires measuring different constructs with different items, we need a better imputation technique. Since nearest neighbor imputation is well studied and often used for missing data in a variety of data science applications (Troyanskaya et al. 2001), we believe that it should be used in this case as well.

Next, the group samples are balanced by simple oversampling. This is necessary when the sample sizes of the different groups are different. For example, suppose one group has a completely different typical response than the other groups, but the number of questionnaires in that group is small. This tiny fraction of data points won't significantly affect the clustering metrics, and therefore the response type is unlikely to appear as a cluster centroid. This effect does not occur when the group samples are comparably large. While oversampling is a standard method in supervised learning, one usually has to be very careful not to overfit a model to a few examples (Chawla et al. 2002; Menardi and Torelli 2012). Note that this overfitting effect does not have a serious impact on the proposed method. First, we use the oversampled data set only to identify appropriate response types, while the actual evaluation (e.g., measuring similarity based on the fingerprints) is based only on the actual (not oversampled) data. Second, the obtained clustering is not intended to be applied to unseen data, but only to describe a data set. We emphasize that an alternative to oversampling would be weighted clustering, but linkage-based clustering algorithms as applied here are known to be incompatible with this approach (Ackerman et al. 2021). However, this choice results in a much higher computational cost, and in applications it may well be that non-linkage-based clustering algorithms could also be used. Furthermore, while oversampling allows for different sample sizes in the different groups, it is very important to note that the sample in each group is representative and valid for the question being studied.

Finally, the data augmentation step adds some noise to the individual questionnaire items. Thus, the questionnaire items are no longer integers, but floating-point numbers. This makes the data matrix more likely to be of full rank, which is crucial for numerical stability, and the resulting clustering more stable against removal of individual data points. Moreover, it is well known that almost all machine learning models generalize much better when the training data is augmented with random noise (Belinkov and Bisk 2018; Ding et al. 2007; Min et al. 2020; Zhang, Kinoshita, and Kiya 2020). A further idea behind the augmentation is that it turns the naive over-sampling of step two into a SMOTE-similar over-sampling as the single over-sampled data points are subjected to noise and are no simple duplicates of the original data (Chawla et al. 2002).

5.2. COMPARISON TO FACTOR ANALYSIS AND PCA

First, we observe that certain patterns in the response types correspond to the applicability of the factor analysis approach. For example, it is highly unlikely that a PCA or factor analysis can be performed meaningfully if the centroids of the clustering of the questionnaire data, i.e., the response types, look *different*. In other words, if all response types have a similar shape (as in the set $\mathcal{D}_1$), or response types 1, 2, and 5 in the data set $\mathcal{D}_2$, then the principal components between different groups are expected to be the same.

On data set $\mathcal{D}_1$, our approach as well as the PCA approach could be safely applied. It is not very surprising that the hypothesis test following the PCA did not find a significant difference in the group median between Group 3 and Group 4 ($p = 1.00$) if this result is compared to the similarity dendrogram in which those groups are also close to each other. However, we observe a striking difference in the 'similarity' of Group 2 and Group 3 between the approaches. Our approach suggests that Group 2 and Group 3 are far apart from each other, while the hypothesis test following the standard PCA approach found no significant differences between these groups ($p = 0.564$). From a mathematical point of view, this is easy to explain: The hypothesis test is on the average response per group (hence a group mean or median, depending on the actual choice of statistical test) and tests the hypothesis that this value is the same between two groups. For Group 2 & 3, this value is expected to be the same (namely 3, based on the data generation). However, Group 2 contains mostly moderately answered sheets (expected values between 2 and 4), while Group 3 contains mostly extreme sheets (values either 1–2 or 4–5). This difference cannot be observed with a mean or median test. We emphasize that this is not a flaw of either method. If the hypothesis test result is interpreted correctly, no false conclusions should be drawn. However, in applications, such a test result is often misleading because it is interpreted as 'there are no group differences,' which is clearly wrong, because it does not take into account the lack of homogeneity of responses in certain groups. Be that as it may, our approach manages to naturally distinguish between groups with homogeneous and heterogeneous responses, and thus may not lead to false conclusions as easily.

Additionally, our approach gives a very natural and clean way to describe what a *typical answer* is supposed to mean. When a group's fingerprint is narrowly focused on one response type, that response type is close to the usual group average. In most cases, however, a fingerprint will have non-vanishing weight on more than one response type. In this case, the fingerprint yields several typical response patterns, and the weight indicates what proportion of the questionnaires belong to each typical pattern. Note that if $f \in [0,1]^\ell$ is the fingerprint of a group and $r_1, \ldots, r_\ell \in \mathbb{R}^d$ are the response types, the convex combination $\Sigma f_i r_i$ can be seen as the natural group mean. Thus, depending on the absolute size of each f_i, the fingerprint also directly gives a measure of the variability in the group. For example, if one $f_i = 1$, then only questionnaires of that response type belong to the group, while $f_i \sim l^{-1}$ for all i means a uniform distribution across all response types, i.e., a large variation. A natural measure of the heterogeneity in a group could therefore be the normalized entropy,

$$H(f) = -\frac{1}{\log \ell} \sum_{i=1}^{\ell} f_i \log(f_i)$$

of its fingerprint f, a quantity that is a very natural measure of variation in a variety of applications from physical systems to information theory (Wehrl 1978), and we propose to use the entropy of fingerprints to measure the heterogeneity of a group. A key property of the normalized entropy is that it takes values between 0 (all questionnaires belong to the same response type)[1] and 1 (there are equally many questionnaires of each response type), which quantifies homogeneity of a group quite naturally.

5.3. ADVANTAGES AND LIMITATIONS OF THE PROPOSED APPROACH

Probably the most limiting factor of our approach is its purely descriptive nature, i.e., it does not provide a typical measure of significance with respect to group differences. In addition, the number of required response types is usually hard to determine. While the gap statistic yields an algorithmic approach by taking the number of response types at a local maximum in the corresponding scree plot, it is not clear that such a unique or obvious maximum exists. Of course, visual inspection of the dendrogram can support the choice, but overall, this freedom of choice might reduce the inter-observer reliability.

On the positive side, unlike factor analysis or PCA, the proposed method can be applied to any questionnaire data set where differences between multiple groups are to be compared. The method comes with a very natural dimensionality reduction from individual questionnaires to fingerprints of groups that have a very natural interpretation, given reasonable response types. Moreover, the similarity between groups can be described in a straightforward way.

5.4. OVERCOMING SOME OF THE CHALLENGES

While the result of the method is indeed descriptive, the obtained similarity between groups can be seen as a quantification of how similar different groups are. When an agglomerative clustering algorithm is applied to the fingerprints, visual inspection of the dendrogram or the formal gap statistic (if the number of groups is large) can be used to identify *clusters* of fingerprints (Mohajer, Englmeier, and Schmid 2010; Tibshirani, Walther, and Hastie 2001). A difference between two groups exists if and only if these groups do not belong to the same cluster. Moreover, as given as an example, the proportion of each response type in a group can be compared to additional describing factors, e.g., country indices via standard methods. As long as the response types can be well interpreted, such analyses can be used to explain the observed effects.

Similarly, while the number of response types used during the algorithm has an impact on the results, we saw in 4.4 that the similarity between the fingerprints of groups is robust against choosing too many response types. The only effect was observed with respect to the interpretation of the response types (or, the groups' fingerprints on those response types). The most extreme case is a full clustering, where all possible questionnaire types form a singleton cluster (this corresponds to n^d response types, with d being the number of items and n being the possible number of scores). While the similarity can, in principle, still be measured based

---

1   As usual, we assume that $0 \log(0) = 0$.

on the n^d-dimensional fingerprints, an interpretation of the results becomes out of reach. However, due to the gap statistic backed up by visual inspection of the dendrogram, it is easy to determine a number of response types that is at least close to the potentially optimal choice. Here, we emphasize that the property of being a *local* maximum of the gap value is important. As can be observed in Figure 5, a local maximum appears around 5 response types. However, for more than 10 response types, the gap value increases again and becomes even larger as the local maximum. As the data set was generated based on noisy instances of 6 base types σ_1, ..., σ_6, it is not to be expected that considerably many answer patterns are observed regularly, such that the later response types, based on much more than 10 clusters, are likely to be artificial and cannot be well interpreted.

5.5. CONCLUSION

We presented a method to quantify the similarity between different groups based on questionnaire studies, and how it might be possible to explain group differences. The method does, in contrast to the standard approaches, not require measurement invariance, but it can even use variance in the measurements to better distinguish between the groups. The approach is easy to apply and relies on very well-known data scientific concepts and yields a natural interpretation of the results. Moreover, we observed that even in situations in which standard factor analyses could be conducted, following simply the standard approach, did not detect all occurring group differences if they were with respect to group homogeneity or heterogeneity rather than based on the average answer. Overall, we believe that the proposed approach may help a variety of applicants to analyze their complex data sets.

FUNDING INFORMATION

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR AFFILIATIONS

Max Hahn-Klimroth orcid.org/0000-0002-3995-419X
Goethe-University Frankfurt, Frankfurt, Germany

Paul W. Dierkes orcid.org/0000-0002-6046-6406
Goethe-University Frankfurt, Frankfurt, Germany

Matthias W. Kleespies orcid.org/0000-0002-8413-879X
Goethe-University Frankfurt, Frankfurt, Germany

REFERENCES

Ackerman, M, et al. 2021. Weighted clustering: Towards solving the user's dilemma. *Pattern Recognition*, 120: 108152. ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2021.108152

Bartlett, MS. 1950. Tests of significance in factor analysis. *Br. J. Stat. Psychol.*, 3(2): 77–85. DOI: https://doi.org/10.1111/j.2044-8317.1950.tb00285.x

Belinkov, Y and **Bisk, Y.** 2018. Synthetic and natural noise both break neural machine translation. In: *6th International Conference on Learning Representations, ICLR 2018*, Vancouver, BC, Canada, April 30–May 3, 2018, Conference Track Proceedings. OpenReview.net.

Calinski, T and **Harabasz, J.** 1974. A dendrite method for cluster analysis. *Communications in Statistics – Theory and Methods*, 3(1): 1–27. DOI: https://doi.org/10.1080/03610927408827101

Chawla, NV, et al. 2002. SMOTE: Synthetic Minority over-Sampling Technique. *J. Artif. Int. Res.*, 16(1): 321–357. ISSN: 1076-9757. DOI: https://doi.org/10.1613/jair.953

Cormack, RM. 1971. A review of classification. *Journal of the Royal Statistical Society. Series A (General)*, 134(3): 321. ISSN: 0035-9238. DOI: https://doi.org/10.2307/2344237

Costello, AB and **Osborne, J.** 2005. *Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis.* University of Massachusetts Amherst. DOI: https://doi.org/10.7275/JYJ1-4868

Cover, T and **Hart, P.** 1967. Nearest neighbor pattern classification. In: *IEEE Transactions on Information Theory*, 13(1): 21–27. DOI: https://doi.org/10.1109/TIT.1967.1053964

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

Ding, Y, et al. 2007 Robust clustering in high dimensional data using statistical depths. In: *BMC Bioinformatics*, 8(S7). DOI: https://doi.org/10.1186/1471-2105-8-S7-S8

Dziuban, CD and **Shirkey, EC.** 1974. When is a correlation matrix appropriate for factor analysis? Some decision rules. *Psychol. Bull.*, 81(6): 358–361. DOI: https://doi.org/10.1037/h0036316

Feucht, V, Dierkes, PW and **Kleespies, MW.** 2023. The different values of nature: a comparison between university students' perceptions of nature's instrumental, intrinsic and relational values. *Sustainability Science*, 18(5): 2391–2403. ISSN: 1862-4057. DOI: https://doi.org/10.1007/s11625-023-01371-8

Husson, F and **Josse, J.** 2013. Handling missing values in multiple factor analysis. *Food Quality and Preference,* 30(2): 77–85. ISSN: 0950-3293. DOI: https://doi.org/10.1016/j.foodqual.2013.04.013

Kaiser, FG. 1998. A general measure of ecological behavior. *Journal of Applied Social Psychology*, 28(5): 395–422. ISSN: 1559-1816. DOI: https://doi.org/10.1111/j.1559-1816.1998.tb01712.x

Kaiser, HF. 1970. A second generation little jiffy. *Psychometrika*, 35(4): 401–415. ISSN: 1860-0980. DOI: https://doi.org/10.1007/BF02291817

Kleespies, MW and **Dierkes, PW.** 2020. Impact of biological education and gender on students' connection to nature and relational values. *PLOS ONE*, 15(11): e0242004. ISSN: 1932-6203. DOI: https://doi.org/10.1371/journal.pone.0242004

Knickenberg, M, et al. 2019. Assessing dimensions of inclusion from students' perspective – measurement invariance across students with learning disabilities in different educational settings. *European Journal of Special Needs Education*, 35(3): 287–302. ISSN: 1469-591X. DOI: https://doi.org/10.1080/08856257.2019.1646958

Letunic, I and **Bork, P.** 2006. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, 23(1): 127–128. ISSN: 1367-4803. DOI: https://doi.org/10.1093/bioinformatics/btl529

Liefländer, AK, et al. 2013. Promoting connectedness with nature through environmental education. *Environmental Education Research*, 19(3): 370–384. ISSN: 1469-5871. DOI: https://doi.org/10.1080/13504622.2012.697545

Liu, Y, et al. 2013. Understanding and enhancement of internal clustering validation measures. *IEEE Trans. Cybern.*, 43(3): 982–994. DOI: https://doi.org/10.1109/TSMCB.2012.2220543

Mayer, F and **Frantz, CM.** 2004. The connectedness to nature scale: A measure of individuals' feeling in community with nature. *Journal of Environmental Psychology*, 24(4): 503–515. ISSN: 0272-4944. DOI: https://doi.org/10.1016/j.jenvp.2004.10.001

Menardi, G and **Torelli, N.** 2012. Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery*, 28(1): 92–122. ISSN: 1573-756X. DOI: https://doi.org/10.1007/s10618-012-0295-5

Milfont, TL and **Duckitt, J.** 2010. The environmental attitudes inventory: A valid and reliable measure to assess the structure of environmental attitudes. *Journal of Environmental Psychology*, 30(1): 80–94. ISSN: 0272-4944. DOI: https://doi.org/10.1016/j.jenvp.2009.09.001

Min, J, et al. 2020. Syntactic data augmentation increases robustness to inference heuristics. In: Jurafsky, D, Chai, J, Schluter, N and Tetreault, JR (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020,* Online, July 5–10, 2020, Association for Computational Linguistics, 2339–2352. DOI: https://doi.org/10.18653/v1/2020.acl-main.212

Mohajer, M, Englmeier, K-H and **Schmid, VJ.** 2010. A comparison of Gap statistic definitions with and without logarithm function. *LMU Department of Statistics: Technical Reports*, 96. DOI: https://doi.org/10.5282/ubm/epub.11920

Putnick, DL and **Bornstein, MH.** 2016. Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Dev. Rev.*, 41: 71–90. DOI: https://doi.org/10.1016/j.dr.2016.06.004

Rousseeuw, PJ. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20: 53–65. ISSN: 0377-0427. DOI: https://doi.org/10.1016/0377-0427(87)90125-7

Sauerwein, M and **Theis, D.** 2021. New ways of dealing with lacking measurement invariance. In: *Accountability and Educational Improvement.* Cham: Springer International Publishing. pp. 63–82. DOI: https://doi.org/10.1007/978-3-030-69345-9_5

Schmitt, N and **Kuljanin, G.** 2008. Measurement invariance: review of practice and implications. *Hum. Resour. Manag. Rev.*, 18(4): 210–222. DOI: https://doi.org/10.1016/j.hrmr.2008.03.003

Szmrecsanyi, B. 2012. *Studies in English language: Grammatical variation in British English dialects: A study in corpus-based dialectometry.* Cambridge, England: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511763380

Tam, K-P and **Milfont, TL.** 2020. Towards cross-cultural environmental psychology: A state-ofthe-art review and recommendations. *Journal of Environmental Psychology*, 71: 101474. ISSN: 0272-4944. DOI: https://doi.org/10.1016/j.jenvp.2020.101474

Tibshirani, R, Walther, G and **Hastie, T.** 2001. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 63(2): 411–423. ISSN: 1467-9868. DOI: https://doi.org/10.1111/1467-9868.00293

Troyanskaya, O, et al. 2001. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6): 520–525. DOI: https://doi.org/10.1093/bioinformatics/17.6.520

Van De Schoot, R, et al. 2015. Editorial: measurement invariance. *Front. Psychol.*, 6: 1064. DOI: https://doi.org/10.3389/fpsyg.2015.01064

Ward, JH. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301): 236–244. ISSN: 1537-274X. DOI: https://doi.org/10.1080/01621459.1963.10500845

Weaver, B and **Maxwell, H.** 2014. Exploratory factor analysis and reliability analysis with missing data: A simple method for SPSS users. *The Quantitative Methods for Psychology*, 10(2): 143–152. ISSN: 2292-1354. DOI: https://doi.org/10.20982/tqmp.10.2.p143

Wehrl, A. 1978. General properties of entropy. *Reviews of Modern Physics*, 50(2): 221–260. ISSN: 0034-6861. DOI: https://doi.org/10.1103/RevModPhys.50.221

Yong, AG and **Pearce, S.** 2013. A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in Quantitative Methods for Psychology*, 9(2): 79–94. ISSN: 1913-4126. DOI: https://doi.org/10.20982/tqmp.09.2.p079

Zhang, W, Kinoshita, Y and **Kiya, H.** 2020. Image-enhancement-based data augmentation for improving deep learning in image classification problem. In: *IEEE International Conference on Consumer Electronics – Taiwan, ICCE-TW 2020,* Taoyuan, Taiwan, September 28–30, 2020. IEEE, pp. 1–2. DOI: https://doi.org/10.1109/ICCE-Taiwan49838.2020.9258292