# Interpreting the Diversity in Subjective Judgments

**Jean-Bernard Martens**
Eindhoven University of Technology
Department of Industrial Design
P.O. Box 513
Eindhoven, the Netherlands 5600 MB
j.b.o.s.martens@tue.nl

## ABSTRACT

In a CHI paper from 10 years ago, entitled "Accounting for Diversity in Subjective Judgments", an interesting dichotomy was reported between, on the one side, the increased use of idiosyncratic constructs when judging the user experience of diverse products and, on the other hand, the statistical methods available to analyze such data. The paper more specifically proposed a method to extract diverse perspectives (called views) from experimental data. The current paper provides three improvements of this existing method by: 1) showing that a little-known approach for clustering attributes, called VARCLUS, can be applied and extended to provide a more optimal algorithm, 2) showing how the VARCLUS method can be applied to perform both within- and across-subject analysis, and 3) providing access to the VARCLUS method by incorporating it in ILLMO, a user-friendly and freely available program for interactive statistics.

## CCS CONCEPTS

• **Human-centered computing → User studies**; *Empirical studies in HCI*;

## KEYWORDS

User Experience, Diversity, Repertory Grid, Quantitative Methods, Clustering

**ACM Reference format:**
Jean-Bernard Martens. 2019. Interpreting the Diversity in Subjective Judgments. In *Proceedings of CHI Conference on Human Factors in Computing Systems Proceedings, Glasgow, Scotland UK, May 4–9, 2019 (CHI 2019),* 11 pages.
https://doi.org/10.1145/3290605.3300449

*CHI 2019, May 4–9, 2019, Glasgow, Scotland UK*
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-5970-2/19/05…$15.00
https://doi.org/10.1145/3290605.3300449

## 1 INTRODUCTION

In the field of human-computer interaction, we take it for granted that human judgements, and not technical features, decide the (potential) value of the products that we design. The products that we design, such as smart phone or website applications, are becoming increasingly complex so that only a fraction of their features may appeal to a specific user or in a specific context of use. This results in users developing an idiosyncratic perspective on the (expected) user experience when faced with a new product.

In the field of user-experience research it has long been recognized that standardized, and therefore generically formulated, questionnaires may fail to fully capture such personal perspectives. In the paper "Accounting for Diversity in Subjective Judgments" [7], the authors therefore formulated the increasingly popular point of view that "An alternative approach to posing predefined questionnaires to participants lies in a combination of *structured interviewing*, that aims at eliciting the attributes that are personally meaningful for each individual, with a subsequent *rating process* performed on the attributes that were elicited during the interview". Mixing personalized attributes with attributes from standardized questionnaires is another popular approach. Note that the interpretation of standardized questions may also vary across subjects, so it can even be advisable to treat the answers to such questions as personal attributes.

A very popular technique for structured interviewing is the Repertory Grid Technique (RGT) [3, 15], in which subjects are presented with triplets of products and their task is to come up with criteria for distinguishing two of the products from the third one. Subsequent laddering can be used to dissect such criteria into more basic personal attributes and to establish how such criteria contribute to the end-user judgement on the products.

While the above approach is by now widely accepted and practiced in diverse application areas in the field of user-experience research [4, 8, 12, 13], it also creates a problem, as discussed in more depth in [7]. Most often we are interested in generalizing the observations made by individual participants in a study, but also not up to the point where all individual perspectives are averaged out. This implies that, in the analysis of such data, we need some flexibility in controlling the level of generalization or, if viewed from the opposite perspective, the level up to which individual perspectives should be reflected.

The analysis method in [7] is a proposal for how to derive a number of distinct perspectives from ratings of personal attributes. Rather than constructing a single multi-dimensional model from the experimental data, a number of distinct two-dimensional (2D) models, called *views*, are derived. The proposed method is however quite rigid and difficult to execute, which is why the current paper presents a new method that is not only more optimal, but also more flexible and more easily accessible and easy to use. The latter aspect has been accomplished by integrating the new method in a freely available program for doing interactive statistics [10].

## 2 INTRODUCTION TO 2D VIEWS

Before demonstrating how multiple 2D views can be derived from an extensive experimental data set, we start with a fairly simple example illustrating the use of a single 2D view in product design. We want the reader to be aware of the fact that clustering (of which the VARCLUS method discussed in this paper is one example) is an exploratory data analysis technique, aimed at giving users insights into their data, and that there are no widely accepted objective criteria for comparing clustering algorithms (which explains why there are some many variations). In the end it comes down to the insights that users can extract from the visualisations (such as the 2D views) being generated.

A student (Daphne Menheere) at our department of Industrial Design collected user impressions about 8 commercially available bracelets, six of which were (physical) activity bracelets (see http://www.daphnemenheere.nl/M12/pdf for more details). She first conducted structured interviewing using the Repertory Grid Technique, which resulted in nine attributes that users proposed as viable ways to compare the bracelets. These attributes were subsequently used to create a questionnaire in which subjects needed to rate (on a 7-point scale) their agreement with statements such as: this bracelet can be considered as {Jewelry, Feminine, Masculine, Unisex, Sportive, Tough, Luxurious, Graceful, Easy to Combine}. Three points on the 7-point scale were labeled as completely disagree (1), neutral (4) and completely agree (7).

The responses from 5 subjects for all 8 bracelets and 9 attributes were used by the VARCLUS algorithm, discussed in the next section, to derive the 2D view shown in the left graph in Figure 1. This view provides insight into the existing design space. Especially noteworthy is the negative correlation between the semantically distinct attributes *A2:Feminine* and *A5:Sportive*, which reveals a premise that seems to be implicit in the design space of the products that were investigated. Assuming that sportive and feminine are opposite attributes is obviously nonsense in general but nevertheless seems to hold true within the set of commercially available bracelets. This led the student to develop a bracelet that scored high on both of these attributes (and that would hence not fit very well in the 2D view that inspired it in the first place). In this sense, we could state that 2D views are not only intended at visualizing the current situation, but also at broadening our view towards new opportunities. Combining features of existing products, which corresponds to imagining products that are in between existing products in the 2D view, can be considered as evolutionary design, while breaking some relationships (or dependencies) that seem to be implicitly present in the view on existing products, can be characterized as more revolutionary design.

The case where all empirical data can be represented into a single 2D view is rather unusual. In cases where not all attributes can be adequately represented, existing clustering algorithms adopt a multi-dimensional space as model. The problem with such multi-dimensional spaces is of course that they are hard to visualize and hence do not serve well their primary purpose, which is to help explore relationships in the data. Some form of interaction is often provided in order to allow the user to generate (a succession of) 2D images that correspond to distinctive perspective views on the more-dimensional space. The strategy adopted here is not to rely on such user interaction to find interesting perspectives, but instead to aim directly at generating 2D views that offer complementary perspectives (technically speaking, a number of such 2D views could always be combined into a multi-dimensional model).

## 3 MULTIPLE 2D VIEWS

### Experimental Data

In order to facilitate comparison to the paper of Karapanos et al. [7], we adopt the same experimental data that they used to illustrate their method. The data concerned emerged from an RGT experiment [5] with 10 students of the Technical University of Darmstadt (TUD) who rated the homepage of 8 German universities on personal constructs. The RGT experiment produced 118 personal attributes (10 to 14 per subject). The 10 subjects rated the 8 websites on a separate 7-point Likert scale for each of the attributes that they proposed.

In subsequent analyses performed in this paper the eight university websites will be referenced as: (s1) Darmstadt, (s2) Mannheim, (s3) Heidelberg, (s4) Mainz, (s5) Frankfurt, (s6) Aachen, (s7) Karlsruhe and (s8) Munchen.

### Original Method

The method proposed in [7] consists of a number of steps. In the *first step*, a 2D view is constructed from the attributes

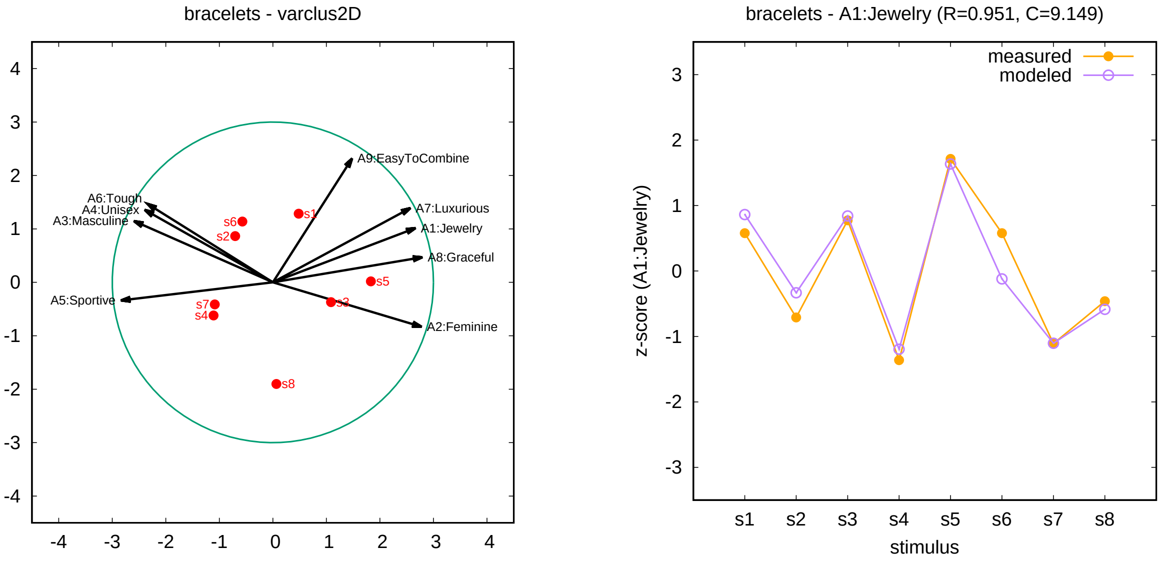


**Figure 1: The 2D view on the left is derived from the 9 scaled attribute judgements on 8 (activity) bracelets. The length of the arrows are proportional to the fraction of explained variance, where the circle corresponds to a fraction of 1 (perfect correlation). The graph on the right shows the correlation between the measured values on the first attribute and the predictions according to the 2D view, which are the coordinates of the 8 bracelets (s1-s8) in the direction of the *A1:Jewelry* vector.**

of each subject using an established method for dimension reduction such as factor analysis or principal component analysis (PCA) [2].

In the *second step*, the goodness-of-fit between the original attribute scores $A_{ki}$ for atribute $k$, where $i$ identifies the distinct websites being judged, and the predictions $\hat{A}_{ki}$ according to the 2D view, is established (in a similar way as shown in the right graph of Figure 1) using a combination of two effect sizes. The first effect size for attribute $k$ is $R_k^2$, the *squared correlation coefficient* (or fraction of explained variance). The second effect size is defined as

$$C_k = \frac{\hat{A}_{k,max} - \hat{A}_{k,min}}{\sigma_k}, \qquad (1)$$

where $\sigma_k$ is the estimated standard error. This second effect size is equal to *Cohen's d* [2] between the two most extreme predicted scores, and has been proposed earlier [1] as a goodness-of-fit measure. Three levels of goodness of fit are distinguished based on these two effect sizes:

(1) **good fit:** $R_k^2 > 0.5$ and $C_k > 6$
(2) **adequate fit:** $R_k^2 > 0.5$ and $C_k > 4$
(3) **poor fit**: $R_k^2 \leq 0.5$ or $C_k \leq 4$

In the *third step* of the method, the attributes with poor fit are used to construct one or more additional 2D views.

In the *fourth step*, only 2D views that contain at least two attributes with good fit are retained, so that the resulting number of 2D views can vary across subjects. In the example case there were 2 subjects with 0 views, 3 subjects with 1 view and 5 subjects with 2 views. These 13 views provided good predictions for 62 (out of the 118) attributes.

In the *fifth step*, a distance measure was introduced to compare pairs of 2D views. It was used by a hierarchical clustering algorithm to reduce the number of views. For the example data, it was concluded that the 13 views could be clustered into 3 clusters, and the 2D views in the same cluster were averaged to end up with the final three 2D views.

In the *sixth step*, the goodness-of-fit of all attributes in the clustered views was established. In the example, the 3 clustered views provided a good fit for 38 of the 118 attributes, which was substantially more than in case of a single (averaged) view which only provided a good fit for 18 attributes.

## VARCLUS method

The original method is interesting in terms of the key proposition that it makes, i.e., that multiple views can provide a good fit to substantially more attributes than a single averaged view. The reason why more views are better is explained in the original paper. If the number of views is too low, the effect is NOT that the approximation/modelling of all observed attributes is reduced, but that the resulting views are mostly determined by a subset of the attributes, and that the remaining ones are effectively discarded.

The clustering method in [7] also adopts a number of interesting principles that we subscribe to. The first principle is that the outcome of a statistical analysis should preferably be presented (graphically) as low-dimensional (more specifically, 2D) spaces or views. Users can be expected to have a hard time interpreting spaces with a dimension higher than 2, such as the ones that arise from more established clustering methods (such as factor analysis). The second principle is that any single 2D view can only provide a good fit for a limited subset of the observed attributes, so that attributes will inevitably have to be divided over different views. This latter principle will be supported quantitatively later on (see Figures 3 and 4).

There is an existing clustering algorithm that has been developed with very similar principles in mind. Unfortunately, it is limited to deriving one-dimensional (1D) spaces or views from observed attributes. Moreover, it is not available in popular statistical packages (such as SPSS). The clustering algorithm in question is sometimes referred to as *Oblique Principal Component Cluster Analysis* (OPCCA). It is also known by the name VARCLUS under which it is implemented in the statistical package SAS [14]. Interestingly enough, the name of the original inventor and implementor of the VARCLUS algorithm seems to have been lost, and details about the method are actually quite hard to find.

The implementation that we will use is based on the description in an unpublished paper [6] available through ResearchGate. Specific about VARCLUS is that, unlike other clustering methods, it does not attempt to minimize the number of clusters, but instead primarily tries to create clusters with high internal consistency. This may for instance result in VARCLUS producing three clusters with high internal consistency, where other clustering algorithms such as VARIMAX (or PROMAX) would produce two dimensions with smaller internal consistency. A popular indicator for the consistency of a number of attributes that are assigned to a single 1D cluster is *Cronbach's alpha* [2].

The VARCLUS algorithm is applied on z-scores rather than on the original attribute scores, i.e., the linearly transformed scores that are used as input into the algorithm are normalized to have zero mean and unit variance. The effect of this normalization is that all attributes receive an equal weight (of 1) and are a priori considered to be equally important.

The clustering starts with all attributes in a single cluster and uses PCA to decide whether or not this single cluster should be split into two sub-clusters. The PCA determines so-called *eigenvalues* which are delivered in decreasing order. VARCLUS decides to split a cluster if the second eigenvalue $\lambda_2$ is larger than a threshold value $\lambda_2(thr)$. This threshold value is the main parameter of the VARCLUS algorithm as it allows to control the level of generalization. More specifically, increasing the value of this threshold parameter will reduce the number of clusters up to the point where all attributes are allocated to a single cluster.

The complicating part of VARCLUS, which we will not attempt to explain in detail, is that the original attributes are iteratively re-assigned to optimize internal consistency every time an additional cluster is introduced. It is this iterative procedure that ensures that the algorithm is more optimal than the procedure proposed in [7] which contains only a single iteration (in which the attributes with a poor fit are removed from the initial cluster to define a new cluster). The sketched procedure is applied recursively until the second eigenvalue for all clusters is below the threshold. The details of the VARCLUS (or OPCCA) method are described in the unpublished paper cited above.

While the VARCLUS algorithm possesses several characteristics that we aspire for in a clustering algorithm, it is restricted to clustering into 1D views. It was therefore necessary to devise an extension towards 2D views. The corresponding clustering algorithm, which will be denoted by VARCLUS2, again uses PCA to determine whether or not a cluster of attributes should be split into sub-clusters. While VARCLUS uses PCA up to order 2, VARCLUS2 uses PCA up to order 4, resulting in 4 eigenvalues per cluster, denoted by $\lambda_i$, for $i = 1, \ldots, 4$, in decreasing order. These eigenvalues are combined pairwise into

$$\tilde{\lambda}_i = \sqrt{\lambda_{2i-1}^2 + \lambda_{2i}^2}, i = 1, 2 \tag{2}$$

and a cluster is split into two sub-clusters if the second value $\tilde{\lambda}_2$ exceeds the specified threshold value of $\sqrt{2} \cdot \tilde{\lambda}_2(thr)$. The iterative re-assignment of attributes every time an additional cluster is introduced, done in order to optimize internal consistency, also needed to be generalized (using a concept called communality that is an extension of the well-known concept of correlation for 1D spaces). The detailed textual and mathematical descriptions of both VARCLUS and VARCLUS2 can be found in [11].

## 4 SINGLE-LEVEL APPLICATION OF VARCLUS

### Within-subject analysis

In order to allow for a direct comparison with the within-subject analysis (i.e., the first four steps) of the original method, we performed a similar analysis with VARCLUS2 (with parameter value $\tilde{\lambda}_2(thr) = 1$), the results of which are summarized in Table 1.

In total, 12 views were derived which allowed for good fit of 58 attributes (and adequate fit for another 31 attributes), which is comparable to the 62 attributes with good fit in the 13 views generated by the original method (see Table 4 in [7]). There are however some obvious differences between the outcomes of both methods. The first is that VARCLUS2 provides at least 1 view for each subject (while there were

**Table 1: Number of attributes modeled by the (up to two) 2D views, for all 10 participants (VARCLUS2 with $\tilde{\lambda}_2(thr) = 1$). The number of attributes with good and adequate fit are specified by the number outside and inside the brackets, respectively.**

| Participant | Total | View a | View b | Remain |
|---|---|---|---|---|
| 1 | 13 | 4 (7) | | 9 (2) |
| 2 | 13 | 6 (5) | | 7 (2) |
| 3 | 14 | 4 (3) | 7 (0) | 3 (0) |
| 4 | 10 | 9 (0) | | 1 (1) |
| 5 | 12 | 9 (0) | | 3 (3) |
| 6 | 11 | 6 (4) | | 5 (1) |
| 7 | 13 | 4 (4) | 3 (2) | 6 (0) |
| 8 | 11 | 5 (3) | | 6 (3) |
| 9 | 10 | 4 (5) | | 6 (1) |
| 10 | 11 | 7 (0) | | 4 (4) |

no views for 2 subjects in the original method), and only requires a second view for 2 of the subjects (3 and 7), while in the original method 2 views were required for 5 subjects (1, 2, 4, 6 and 10). The reason that the VARCLUS algorithm is able to come up with one 2D view per subject, where the original algorithm did not, is that it performs an exhaustive search over a large number of possible groupings of attributes in order to find the ones that result in views where multiple attributes have a good fit. Manually trying all these alternatives is simply not practically feasible. In the original method, groupings were established in an interactive way by the user, inspecting the outputs of existing clustering methods and adopting the step-by-step method explained above, so that only few alternatives can actually be tried.

Having to discard attributes (or subjects) that are a priori of interest (e.g., based on their description) because the analysis method is not able to handle them obviously constitutes a problem. So continuing to generate views until at least all attributes (or subjects) of interest have a good fit makes sense. Using the parameter of VARCLUS2 it is easy to increase the number of attributes that have a good fit, especially for those subjects where a substantial number of the attributes are only adequately or poorly fit. For instance, by lowering the VARCLUS2 parameter to $\tilde{\lambda}_2(thr) = 0.8$, we obtain the results in Table 2. The number of views has increased from 12 to 17, but so have the number of attributes that have a good fit (increases from 58 to 88). As there are fewer attributes left that do not have a good fit, the number of attributes with an adequate fit decreases to 24 (so that 112 of the 118 attributes now have a good or adequate fit, versus 89 in case of VARCLUS2 with $\tilde{\lambda}_2(thr) = 1$).

**Table 2: Number of attributes modeled by the (up to two) views for all 10 participants (VARCLUS2 with $\tilde{\lambda}_2(thr) = 0.8$). The number of attributes with good and adequate fit are specified by the number outside and inside the brackets, respectively.**

| Participant | Total | View a | View b | Remain |
|---|---|---|---|---|
| 1 | 13 | 6 (1) | 5 (0) | 2 (1) |
| 2 | 13 | 7 (1) | 4 (1) | 2 (0) |
| 3 | 14 | 4 (3) | 7 (0) | 3 (0) |
| 4 | 10 | 9 (0) | | 1 (1) |
| 5 | 12 | 9 (0) | | 3 (3) |
| 6 | 11 | 6 (4) | | 5 (1) |
| 7 | 13 | 4 (4) | 3 (2) | 6 (0) |
| 8 | 11 | 5 (2) | 3 (1) | 3 (0) |
| 9 | 10 | 4 (1) | 5 (0) | 1 (0) |
| 10 | 11 | 7 (0) | 0 (4) | 4 (0) |

The example hence supports our claim that the VARCLUS2 algorithm constitutes an improvement in both performance and flexibility when compared to the method proposed in [7]. It guarantees that all attributes are assigned to one 2D view only, and uses an iterative method to minimize the number of views required. The flexibility is created by controlling the parameter $\tilde{\lambda}_2(thr)$ of the VARCLUS2 algorithm in such a way that at least the attributes of interest are represented with a good or (at least) adequate fit.

As an example of the generated output, we show the two views derived for subject 3 in Figure 2. The left view has 4 attributes with a good fit (represented by the bold arrows), and 3 attributes with an adequate fit (represented by the regular arrows). The right view has 7 attributes with good fit. The lengths of all vectors are proportional to the fraction of explained variance (or squared correlation coefficient), where the circle corresponds to a perfect correlation of 1. This correlation coefficient is determined by a linear regression between the measured attribute values and *the orthogonal projections* of the points representing the stimuli (i.e., the 8 websites indicated by $s1$ to $s8$) *onto the vector that represents this attribute.* These projected values (see Figure 1 in [7]) are hence *the model predictions* for the measured attribute.

The low-contrast dashed arrows in a view correspond to the attributes that are not assigned to this view (and which are represented better in the other view). The fact that the lengths of many of these vectors are much smaller than 1 indicates that the view does indeed not account for the judgements expressed in these attributes, and that a second complimentary view is required to model them.

The left view in Figure 2 contains attributes that relate to the positioning or image of the university (such as A13:

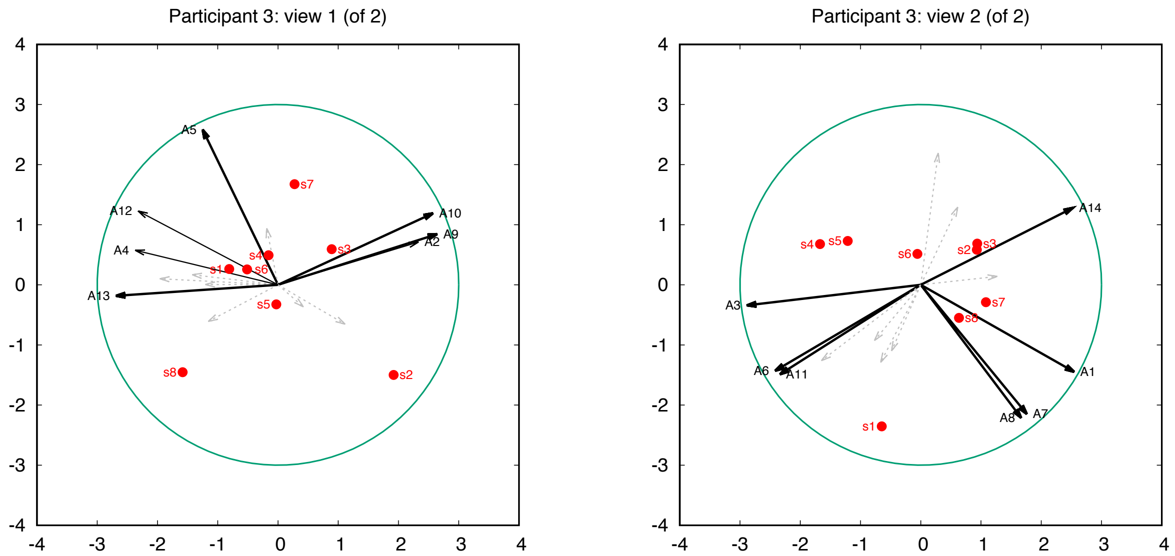


**Figure 2: Two 2D views derived from the 14 scaled attributes of participant 3. The length of the arrows are proportional to the fraction of explained variance, where the circle corresponds to a fraction of 1 (perfect correlation). Attributes with good fit are represented by bold arrows, while attributes with adequate fit are represented by regular arrows. The low-contrast dashed arrows correspond to the attributes that are represented in the alternative view.**

externally oriented, A4: open to new directions, and A12: not driven by fixed values or tradition), ways of providing information (such as A10: clear layout, A9: sober layout, and A2: concrete) and A5: care invested in website. In the right view of Figure 2, a very noticeable feature is that the own university (s1) is very much distinguished from the other universities. Especially the aspects of A3: spontaneity, A6: broad offer and A11: possibility to explore, on the one hand, and A7: focus on technology and A8: focus on objectivity, on the other hand, positively distinguish the own university in the perception of this student. Properties of the website that can also be distinguished in this view are A1: good overview and A14: focus on own university. The fact that most of the other universities cluster together in this second view could be real, but might also point at a lack of detailed knowledge of this student about the educational program and didactical methods offered at other universities. In summary, where the left view helps to capture the student's perception of the university image and way of communication on the website, the right view provides insight into his understanding of the nature of the program being offered.

## Across-subject analysis

Rather than creating individual 2D views, and subsequently clustering these views to reduce their number, as was done in the last two steps of the original method to construct a joint interpretation of all collected attributes, we propose a more straightforward method which is to analyze all 118 attributes simultaneously using VARCLUS2. The default value of $\tilde{\lambda}_2(thr) = 1$ results in 8 views that provide a good fit for 90 attributes and an adequate fit for 24 attributes (in total, this provides a fit for 114 of the 118 attributes).

By stepwise increasing the threshold value $\tilde{\lambda}_2(thr)$ from 1.0 to 3.5, the number of views can be decreased all the way down to 1, as illustrated in Table 3 and Figure 3. A comparison with the original method can be made by looking at the case where 3 views are produced; with the original method 38 attributes had good fit, while with the VARCLUS2 method a slightly larger number of 41 attributes have good fit.

Note that the largest increase in number of attributes with a good fit actually occurs when moving from 3 to 4 views, which could be used to argue for 4 instead of 3 views. An alternative way of displaying the same information is shown in Figure 4, where the average number of attributes being fitted per view is plotted. This figure reveals that at least 4 views are needed to ensure that the number of attributes with good fit exceeds the number of attributes with adequate fit. We could argue for an even larger number of views, as the total number of attributes with good or adequate fit only starts to saturate when the number of views exceeds 6.

Having a large number of views is not necessarily a problem, as long as each view provides a good or adequate fit for a substantial number of attributes. The views can indeed be seen as alternative ways to position a product of interest

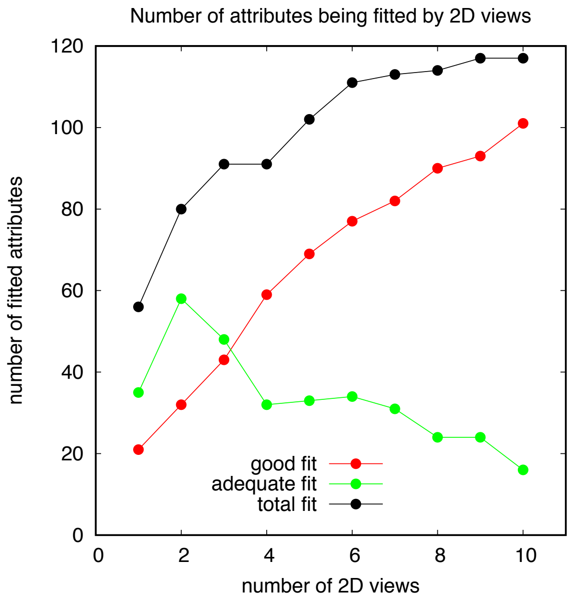


**Figure 3: Number of attributes with good and adequate fit (and their sum) as a function of the number of 2D views used to model all attributes.**

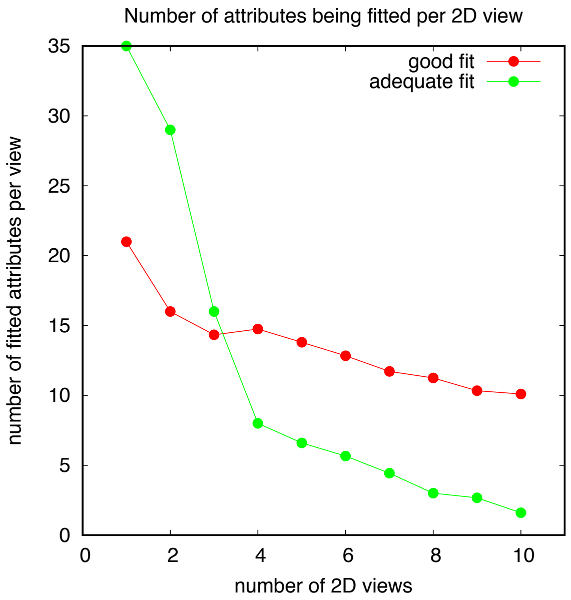


**Figure 4: Average number of attributes with good and adequate fit per 2D view as a function of the number of 2D views used to model all attributes.**

**Table 3: By changing the threshold value $\tilde{\lambda}_2(thr)$ of VARCLUS2 method, the number of 2D views used for modeling all attributes can be varied. For each solution, the table also provides the number of attributes that have good fit, the number of attributes that have adequate fit, and their sum.**

| $\tilde{\lambda}_2(thr)$ | Number Views | Good Fit | Adequate Fit | Total |
|---|---|---|---|---|
| 0.8 | 10 | 101 | 16 | 117 |
| 0.9 | 9 | 93 | 24 | 117 |
| 1.0 | 8 | 90 | 24 | 114 |
| 1.1 | 7 | 82 | 31 | 113 |
| 1.3 | 6 | 77 | 34 | 111 |
| 1.4 | 5 | 69 | 33 | 102 |
| 1.5 | 4 | 59 | 32 | 91 |
| 2.0 | 3 | 43 | 48 | 91 |
| 2.5 | 2 | 32 | 48 | 80 |
| 3.5 | 1 | 21 | 34 | 56 |

with respect to possible competing products, and to identify distinct strengths and weaknesses of such products. A 2D view in which a number of attributes have a good fit creates an opportunity for interpreting these attributes simultaneously, rather than individually, and to get insight into mutual dependencies and conflicts. Keep in mind that views are intended to support human (design) thinking and reflection rather than to create economic models from data.

For instance, the study that we use as example was executed to better understand the position of Darmstadt university (s1) with respect to other German universities. Any of the views can potentially trigger a discussion about opportunities for improving the current position. Take for instance view 5, which is reproduced in Figure 5. The choice for this view came about by picking an attribute of interest (A5) and selecting the view to which this attribute was assigned. The view reveals that s1 scores well on attributes A64, A69 and A70 (which all indicate a good Access to Information), which is obviously a desired feature, but that other universities, especially s6, s7 and s8, perform a lot better on attributes such as A5 (modern layout), A100 (professional appearance) and A9 (future-oriented), which could for instance be interpreted as a need for a more modern appearance of the website.

Note that views can also contain relationships that are seemingly strange (or even absurd). For instance, in the view of Figure 5, A5 (modern layout) is opposite to A67 (easy connection to the university), which are obviously not related attributes in general. We should however keep in mind that the views are derived from ratings of only the 8 university websites contained in the study, so that this questionable relationship is actually a property of the design space being considered. Adding more (and more diverse) websites

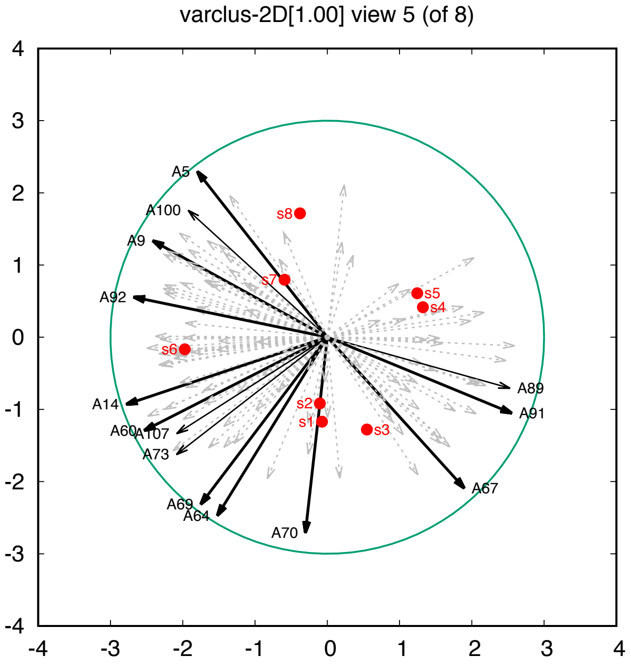


**Figure 5: Attributes with good and adequate fit in one of the 8 generated 2D views (view 5) that arise from the joint analysis of all collected attributes.**

into the comparison is likely to reduce the correlation between such semantically different concepts. As indicated before, imagining how the website of interest (s1) could be redesigned so that it scores well on both, seemingly conflicting, attributes could be an interesting design assignment.

## 5 MULTI-LEVEL APPLICATION OF VARCLUS

In the previous analysis, we implicitly assumed that all attributes generated by the participants reflect a worthwhile perspective and should be accounted for and be visualized (as much as possible). The fact that VARCLUS2 works on z-scores even implies that all attributes receive equal weight in this analysis. We could however also argue that it is unlikely that individuals develop 10-14 independent attributes to assess a limited number of stimuli (in casu, 8 websites). Maybe a participant actually distinguishes a much smaller number of internal constructs or discriminative processes (say 2-3) but externalizes them using a number of different verbalizations (i.e., uses several attributes to describe the construct). According to such a perspective, it makes sense to first establish these personal constructs, and to subsequently create 2D views in order to visualize mutual dependencies between them. Deriving such hidden constructs from ratings of attributes is what the original VARCLUS algorithm was developed for in the first place (e.g, the VARCLUS algorithm has been used to cluster the outcomes on different tasks in an IQ test). In order to establish a consistent terminology we will refer to such personal constructs as 1D views.

In Table 4 we summarize the results of a VARCLUS analysis with $\lambda_2(thr) = 1$ for each of the participants . The number of 1D views per subject derived from this analysis varies from 2 to 4. The total number of attributes with a good fit is 56, while the number of attributes with an adequate fit is 34. This leaves 28 attributes that are not adequately represented by any of the 28 1D views. Obviously, in case some of the attributes with poor fit are considered of interest, we can always increase the number of attributes that have a good or reasonable fit, at the cost of increasing the number of 1D views, by decreasing the value of the VARCLUS parameter $\lambda_2(thr)$.

The next step could be to remove or split some 1D views, such as the ones that have low internal consistency (such a 2b, 7d and 10c), as the evidence for a single underlying construct is weak in this case (in order not to overly complicate matters, we will not actually do so in our current example). Note that removing 1D views that are supported by few, or even a single, attribute may not always be advisable. For instance, in the example case, there are two instances of 1D views with only a single attribute, i.e., 4c and 8b. In both cases it turns out that these views represent unique perspectives, i.e., aspects that are only remarked upon by individual participants but that nevertheless could be highly relevant, i.e., 4c is about how lively the student scene is, while 8b is about the international orientation of the university.

In order to provide some evidence for the fact that the different 1D views indeed reflect different perspectives, we used the semantic classification of the attributes in 13 categories that the authors of the original paper [5] have established. They defined 10 main categories of attributes (including: website layout, focus on students, specific university image, access to information, clear target group, commercial orientation, etc.), but subdivided the layout category into 4 subcategories (layout of the menus, information and images and the use of color). We established that for the 97 attributes that belonged to the two most prominent 1D views (with the largest membership per subject), 30 were assigned to shared categories (i.e., both 1D views contained attributes that were in this category) while 67 were assigned to unique categories. This is equivalent to only 31% of these attributes ending up in both 1D views (with 95% confidence interval of [22%,41.2%]), supporting the argument that the different 1D views do indeed reflect complementary perspectives within individual participants.

The 1D views that are identified by the above procedure can subsequently be used by VARCLUS2 to create 2D views

**Table 4: Number of attributes modeled by the (up to 4) 1D views for all 10 participants (VARCLUS with $\lambda_2(thr) = 1$). For each view we specify the number of attributes assigned to this 1D view, followed by the consistency measure Cronbach's $\alpha$ between brackets. The number outside of the brackets after the semicolon is the number of attributes with good fit, while the number inside the brackets is the number of attributes with adequate fit.**

| Participant | Number of Attributes | View a | View b | View c | View d | Total | Remain |
|---|---|---|---|---|---|---|---|
| 1 | 13 | 5 (0.901): 2 (2) | 5 (0.901): 0 (1) | 3 (0.971): 3 (0) | | 5 (3) | 5 |
| 2 | 13 | 6 (0.950): 2 (3) | 2 (0.718): 2 (0) | 5 (0.957): 4 (0) | | 8 (3) | 2 |
| 3 | 14 | 6 (0.911): 3 (2) | 5 (0.891): 4 (0) | 3 (0.917): 3 (0) | | 10 (2) | 2 |
| 4 | 10 | 6 (0.935): 3 (3) | 3 (0.974): 3 (0) | 1 (1.000): 1 (0) | | 7 (3) | 0 |
| 5 | 12 | 7 (0.958): 3 (1) | 5 (0.867): 2 (1) | | | 5 (2) | 5 |
| 6 | 11 | 9 (0.947): 3 (2) | 2 (0.930): 2 (0) | | | 5 (2) | 4 |
| 7 | 13 | 5 (0.925): 1 (4) | 3 (0.907): 1 (2) | 3 (0.920): 2 (1) | 2 (0.780): 0 (2) | 4 (9) | 0 |
| 8 | 11 | 5 (0.898): 1 (2) | 1 (1.000): 1 (0) | 5 (0.927): 3 (0) | | 5 (2) | 4 |
| 9 | 10 | 6 (0.942): 2 (3) | 4 (0.830): 1 (1) | | | 3 (4) | 3 |
| 10 | 11 | 3 (0.879): 2 (1) | 4 (0.876): 2 (2) | | | 4 (4) | 3 |

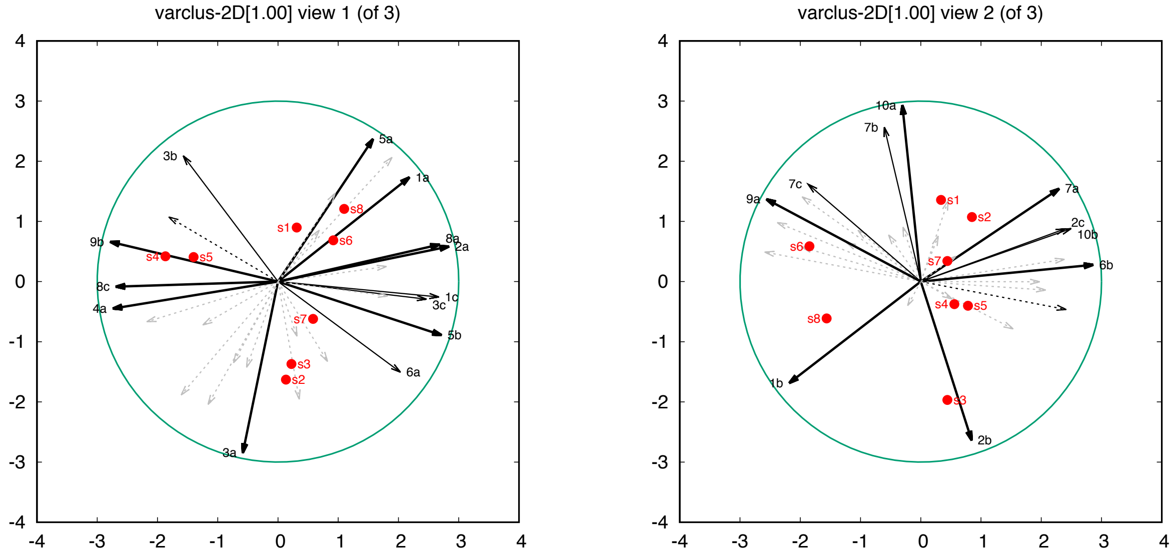
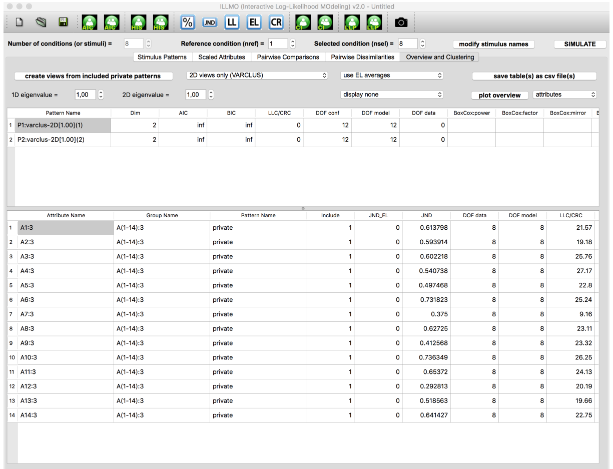


**Figure 6: Two of the three 2D views derived from the 28 personal constructs (or 1D views). The length of the arrows are proportional to the fraction of explained variance, where the circle corresponds to a fraction of 1. Personal constructs with good fit are represented by bold arrows, while those with adequate fit are represented by regular arrows. Personal constructs with inadequate fit are represented by dashed arrows. The low-contrast dashed arrows correspond to the personal constructs that are represented in one of the alternative views.**

that can provide insight into the relationships between these 1D views (or personal constructs). In case of the 28 1D views, applying VARCLUS2 with a parameter value of $\tilde{\lambda}_2(thr) = 1$ results in three 2D views. The two 2D views that provide a good fit for the majority (25 out of 28) of the 1D views are illustrated in Figure 4 (the third 2D view contains only three personal constructs and is not shown). The left view contains 14 personal constructs, 9 with good fit, 4 with adequate fit and 1 with inadequate fit. The right view contains 11 personal constructs, 6 with good fit, 4 with adequate fit and 1 with inadequate fit. It is worthwhile to mention that the three 2D views established using multi-level VARCLUS are actually quite different from the three 2D views produced by the original method.

As indicated before, one application of 2D views is to identify personal constructs that can be jointly interpreted. For



Number of conditions (or stimuli) = 8
Reference condition (nref) = 1
Selected condition (nsel) = 8

2D views only (VARCLUS)
use EL averages
display none
attributes

1D eigenvalue = 1,00
2D eigenvalue = 1,00

| | Pattern Name | Dim | AIC | BIC | LLC/CRC | DOF conf | DOF model | DOF data | BoxCox:power | BoxCox:factor | BoxCox:mirror | E |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | P1:varclus-2D[1.00](1) | 2 | inf | inf | 0 | 12 | 12 | 0 | | | | |
| 2 | P2:varclus-2D[1.00](2) | 2 | inf | inf | 0 | 12 | 12 | 0 | | | | |

| | Attribute Name | Group Name | Pattern Name | Include | JND_EL | JND | DOF data | DOF model | LLC/CRC |
|---|---|---|---|---|---|---|---|---|---|
| 1 | A1:3 | A(1-14):3 | private | 1 | 0 | 0.613798 | 8 | 8 | 21.57 |
| 2 | A2:3 | A(1-14):3 | private | 1 | 0 | 0.593914 | 8 | 8 | 19.18 |
| 3 | A3:3 | A(1-14):3 | private | 1 | 0 | 0.602218 | 8 | 8 | 25.76 |
| 4 | A4:3 | A(1-14):3 | private | 1 | 0 | 0.540738 | 8 | 8 | 27.17 |
| 5 | A5:3 | A(1-14):3 | private | 1 | 0 | 0.497468 | 8 | 8 | 22.8 |
| 6 | A6:3 | A(1-14):3 | private | 1 | 0 | 0.731823 | 8 | 8 | 25.24 |
| 7 | A7:3 | A(1-14):3 | private | 1 | 0 | 0.375 | 8 | 8 | 9.16 |
| 8 | A8:3 | A(1-14):3 | private | 1 | 0 | 0.62725 | 8 | 8 | 23.11 |
| 9 | A9:3 | A(1-14):3 | private | 1 | 0 | 0.412568 | 8 | 8 | 23.32 |
| 10 | A10:3 | A(1-14):3 | private | 1 | 0 | 0.736349 | 8 | 8 | 26.25 |
| 11 | A11:3 | A(1-14):3 | private | 1 | 0 | 0.65372 | 8 | 8 | 24.13 |
| 12 | A12:3 | A(1-14):3 | private | 1 | 0 | 0.292813 | 8 | 8 | 20.19 |
| 13 | A13:3 | A(1-14):3 | private | 1 | 0 | 0.518563 | 8 | 8 | 19.66 |
| 14 | A14:3 | A(1-14):3 | private | 1 | 0 | 0.641427 | 8 | 8 | 22.75 |

**Figure 7: Interface in the ILLMO program that provides access to the VARCLUS method. The spreadsheet at the bottom shows that the 13 attributes of subject 3 have been entered into the ILLMO program. VARCLUS2 has been executed with the default parameter value (2D eigenvalue) of $\tilde{\lambda}_2(thr) = 1$, as is evidenced by the spreadsheet at the top, which shows that two 2D views (shown in Figure 2) have been generated and stored as patterns.**

constructs that demonstrate a good fit in a 2D view, participants seem to at least agree on how the products under study (in casu, the 8 websites) relate to each other. For instance, in the left view in Figure 5 there appear to be clusters of websites such as (s2,s3), (s4,s5) and (s1,s6,s8) that behave similarly, and the personal constructs can help to identify the distinguishing characteristics of such clusters. For instance, personal constructs 4a, 8c and 9b, as well as 5b and 6a ,can help to identify in which sense websites (s4,s5) distinguish themselves, either positively or negatively, from the other websites, while other personal constructs such as 3a can help to identify aspects in which these same websites are not extreme at all.

It is potentially interesting to use such 2D views as input into a discussion with the original participants in the RGT experiment, something which is obviously no longer possible in the example case. Especially if the participants are provided with feedback about which of their original attributes relate to their diverse personal constructs, such a discussion might lead to a jointly shared interpretation of different directions in the space. This might in turn inspire design ideas in the way discussed before. What would be required to move one of the websites in a specific direction (evolutionary design)? Why do some personal constructs which are semantically quite different point in the same or opposite direction, and how could this coupling possibly be broken by a new (revolutionary) design?

## 6 SOFTWARE ACCESS TO THE VARCLUS METHOD

Presenting a method such as VARCLUS, and its extension to VARCLUS2, would hardly be useful if there was not an easy and user-friendly way to get access to the method. In order to allow for easy and free access, we have integrated the method in a freely available program for interactive statistics called ILLMO [9, 10]. The program (for both Mac OS and Windows) is available for download through the website http://illmoproject.wordpress.com, and data can be entered into ILLMO using standard means such as CSV (comma-separated values) files. Amongst others, the website contains references to an iBook [11] with more detailed information on the program and to a diversity of instruction videos (e.g., the videos on cluster analysis provide an introduction to the VARCLUS method).

In Figure 7, we show the interface of the ILLMO program that provides access to the VARCLUS method. Next to the button "create views from included private patterns" (in the upper left corner), there are 3 possible options to choose from, "1D views only" (VARCLUS), "2D views only" (VARCLUS2) and "1D followed by 2D views" (multi-level VARCLUS). The values "1D eigenvalue" and "2D eigenvalue" correspond to the parameters $\lambda_2(thr)$ and $\tilde{\lambda}_2(thr)$ of VARCLUS and VARCLUS2, respectively. Depending on the setting, a separate window with additional information in graphical format will be generated; this can be the views themselves, but also a comparison between original attribute values and model predictions (such as in the right graph of Figure 1). The program also generates gnuplot files (see http://http://gnuplot.sourceforge.net), which were for instance used to generative the graphs in this paper.

## 7 CONCLUSIONS

This paper has presented a way to systematically derive multiple views from ratings of products on individually defined attributes. Such views are alternative ways to look at the design space of the rated products. According to some views, products may behave very similarly, while in other views the same products may be seen as quite different. The attributes that are assigned to a view help to interpret the view and can inspire ideas for modified or new products. By integrating the method into a freely available program, we contribute to making it more accessible and easy to use, so that experimenting with it is much more feasible than with the original method proposed in [7].

## REFERENCES

[1] Norman Richard Draper and Harry Smith. 1966. *Applied regression analysis.* Wiley, New York [u.a.].

[2] Andy Field. 2013. *Discovering Statistics Using IBM SPSS Statistics* (4th ed.). Sage Publications Ltd., Los Angeles.

[3] Marc Hassenzahl and Tibor Trautmann. 2001. Analysis of Web Sites with the Repertory Grid Technique. In *CHI '01 Extended Abstracts on Human Factors in Computing Systems (CHI EA '01)*. ACM, New York, NY, USA, 167–168. https://doi.org/10.1145/634067.634169

[4] Sarah Hayes, Trevor Hogan, and Kieran Delaney. 2017. Exploring the Materials of TUIs: A Multi-Method Approach. In *Proceedings of the 2017 ACM Conference Companion Publication on Designing Interactive Systems (DIS '17 Companion)*. ACM, New York, NY, USA, 55–60. https://doi.org/10.1145/3064857.3079119

[5] Stephanie Heidecker and Marc Hassenzahl. 2007. Eine gruppenspezifische Repertory Grid Analyse der wahrgenommenen Attraktivität von Universitätswebsites. In *Mensch & Computer 2007: Interaktion im Plural - Konferenz für interaktive und kooperative Medien, Bauhaus-Universität Weimar, Weimar, Germany, September 2-5, 2007*. 129–138. https://dl.gi.de/20.500.12116/7246

[6] Ruud Hendrickx and Ton van Schaik. 2003. A Note on Oblique Principal Component Cluster Analysis. (Feb. 2003). https://www.researchgate.net/publication/242249665_A_Note_on_Oblique_Principal_Component_Cluster_Analysis

[7] Evangelos Karapanos, Jean-Bernard Martens, and Marc Hassenzahl. 2009. Accounting for Diversity in Subjective Judgments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. ACM, New York, NY, USA, 639–648. https://doi.org/10.1145/1518701.1518801

[8] Lucas Layman, Carolyn Seaman, Davide Falessi, and Madeline Diep. 2015. Ask the Engineers: Exploring Repertory Grids and Personal Constructs for Software Data Analysis. In *Proceedings of the Eighth International Workshop on Cooperative and Human Aspects of Software Engineering (CHASE '15)*. IEEE Press, Piscataway, NJ, USA, 81–84. http://dl.acm.org/citation.cfm?id=2819321.2819336

[9] Jean-Bernard Martens. 2012. Statistics from an HCI Perspective: Illmo - Interactive Log Likelihood Modeling. In *Proceedings of the International Working Conference on Advanced Visual Interfaces (AVI '12)*. ACM, New York, NY, USA, 382–385. https://doi.org/10.1145/2254556.2254629

[10] Jean-Bernard Martens. 2014. Interactive Statistics with Illmo. *ACM Trans. Interact. Intell. Syst.* 4, 1, Article 4 (April 2014), 28 pages. https://doi.org/10.1145/2509108

[11] Jean-Bernard Martens. 2017. Insights into Experimental Data: Interactive Statistics with the ILLMO Program. (Feb. 2017). https://itunes.apple.com/us/book/insights-in-experimental-data/id1210325588?mt=13

[12] Iain McGregor and Phil Turner. 2012. Soundscapes and Repertory Grids: Comparing Listeners' and a Designer's Experiences. In *Proceedings of the 30th European Conference on Cognitive Ergonomics (ECCE '12)*. ACM, New York, NY, USA, 131–137. https://doi.org/10.1145/2448136.2448164

[13] Mati Mõttus, Evangelos Karapanos, David Lamas, and Gilbert Cockton. 2016. Understanding Aesthetics of Interaction: A Repertory Grid Study. In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction (NordiCHI '16)*. ACM, New York, NY, USA, Article 120, 6 pages. https://doi.org/10.1145/2971485.2996755

[14] CORPORATE SAS Institute Inc. Staff (Ed.). 1988. *SAS-STAT User's Guide: Release 6.03 Edition.* SAS Institute Inc., Cary, NC, USA.

[15] Doménique van Gennip, Elise van den Hoven, and Panos Markopoulos. 2016. The Phenomenology of Remembered Experience: A Repertoire for Design. In *Proceedings of the European Conference on Cognitive Ergonomics (ECCE '16)*. ACM, New York, NY, USA, Article 11, 8 pages. https://doi.org/10.1145/2970930.2970942