Accounting for Diversity in Subjective Judgments

Evangelos Karapanos[1], Jean-Bernard Martens[1], Marc Hassenzahl[2]

Eindhoven University of Technology[1]
Department of Industrial Design
P.O. Box 513, 5600 MB
Eindhoven, The Netherlands
{E.Karapanos, J.B.O.S.Martens}@tue.nl

Folkwang University[2]
Department of Industrial Design
Universitätsstraße 12, 45117
Essen, Germany
marc.hassenzahl@folkwanghochschule.de

ABSTRACT

In this paper we argue against averaging as a common practice in the analysis of subjective attribute judgments, both across and within subjects. Previous work has raised awareness of the diversity between individuals' perceptions. In this paper it will furthermore become apparent that such diversity can also exist within a single individual, in the sense that different attribute judgments from a subject may reveal different, complementary, views. A Multi-Dimensional Scaling approach that accounts for the diverse views on a set of stimuli is proposed and its added value is illustrated using published data. We will illustrate that the averaging analysis provides insight to only $1/6^{th}$ of the total number of attributes in the example dataset. The proposed approach accounts for more than double the information obtained from the average model, and provides richer and semantically diverse views on the set of stimuli.

Author Keywords

Repertory Grid, Multi-Dimensional Scaling, quantitative methods, subjective judgments, user experience.

ACM Classification Keywords

H5.2. User Interfaces: Evaluation/methodology.

INTRODUCTION

Subjective measures for assessing the quality of interactive products has always been of interest to the field of Human-Computer Interaction (HCI) [11, 17]. However, with the recently increased interest in User Experience (e.g., [18]), personal attribute judgments are becoming more and more used.

A number of multivariate techniques such as Factor Analysis (FA), Multidimensional Scaling (MDS) [12] and Structural Equation Modelling (SEM) [33] are employed traditionally for exploring the relations between different obtained attributes. For instance, Schenkman and Jönsson [34]

analyzed users' judgments on a number of predefined attributes, such as beauty, comprehension, meaningfulness, for a set of websites, while Heidecker and Hassenzahl [21], in a similar context, elicited users' idiosyncratic attributes using the Repertory Grid Technique [13]. Both datasets were analyzed using MDS. In both papers averaged models were employed, without much consideration for the underlying remaining diversity.

Approaches such as the Repertory Grid typically emphasize the idiosyncratic nature of perception and evaluation of objects. In other words, individuals perceive interactive products, such as websites, through different, individual "templates". This in turn leads to a certain amount of diversity in obtained attributes and attribute ratings. Some people may use entirely different attributes to evaluate a website, whereas others may use the same attributes but apply them differently. An idiosyncratic approach embraces this diversity and treats it as valuable information.

However, there is also a problem inherent in this approach. As an analyst, one is confronted with as many idiosyncratic views as participants. Views may overlap or even contradict one another; it is in any way complicated to systematically explore this diversity. In practice, the consequence is either an idiosyncratic analysis with a "narrative" summarization [15] or the use of average models. Averaging, however, treats diversity among participants as error and thereby contradicts the basic idea of the underlying approaches.

This paper argues against averaging as a common practice in the interpersonal analysis of subjective judgments. More precisely, we suggest a quantitative, exploratory MDS procedure to identify homogeneous sub-models, thereby reducing the number of different views to be considered while gaining a deeper insight than an averaging approach by accounting for more and of greater semantic diversity in attributes. It will be demonstrated that even single subjects can handle more than one view on a set of stimuli. We will show that by averaging interesting views are overlooked due to majorization bias.

BACKGROUND

As the focus of HCI shifts from task performance towards the more experiential aspects of product use, such as issues of *trust* in online transactions [8], the increased *social connectedness* that awareness systems bring among family

members [29], and the hedonic aspects [16] of users' experience with interactive products, subjective perceptions and evaluations become a core interest.

The development and validation of questionnaires that operationalize new constructs has been the common practice, both in HCI and in many other fields. Using validated questionnaires, one can measure how two or more artifacts compare on a given quality dimension (e.g. trust), or explore the relations between different quality dimensions to advance the theoretical understanding of the field. See, for instance, the ongoing discussion on the role of aesthetics in the acceptance of interactive systems [19].

This traditional approach to the analysis of subjective judgments has however two limitations.

The first limitation lies in the way in which it is being applied in the field. The development of a questionnaire is often described as a three-step process: item (attribute) generation, scale development, and reliability assessment [22]. While the two latter stages aim at improving the discriminant and convergent validity of the questionnaire (i.e. that each individually scaled item correlates highly with the latent construct it attempts to measure, and weakly with all other constructs of the questionnaire), the first stage, namely item generation, is as crucial, as it contributes to the questionnaire's content validity (i.e. that the proposed constructs capture a complete view on the domain of interest).

We consider current approaches to item generation often to be rather superficial. Items are often generated purely on the basis of prior literature and brainstorming (most of the times involving only experts). At best, constructs may be driven by psychological theories; this however introduces other shortcomings, especially in new domains, as constructs that are not supported by theory, will evidently be neglected. In rare cases, where user studies are employed in the item generation process, they are mostly restricted within one research group, often involving a limited set of products and contexts of use. A by-product of this emphasis on the latter two stages of the questionnaire development process is a limited reporting on the exact procedure and intermediate results of item generation, thus making it difficult for researchers to expand and further validate an existing questionnaire. In other words, there is a natural force to undermine the qualitative part in the process of developing a new questionnaire. Instead, we argue, that item generation should be at the core of researching and reporting when first attempts to measure new constructs are being made.

The second limitation is that of assuming homogeneity in the ways that different individuals perceive and evaluate products. Previous research has raised awareness of the diversity between individuals' perceptions [24]. Karapanos et al. [27] showed how diversity may exist at two different stages in the formation of an overall evaluative judgment. *Perceptual diversity* lies in the process of forming product quality perceptions (e.g. novel, easy to use) on the basis of product features. For instance, different individuals may infer different levels on a given quality of the same product, e.g. disagree on its novelty. *Evaluative diversity* lies in the process of forming overall evaluations of the product (e.g. good-bad) on the basis of product quality perceptions. For instance, different individuals may form different evaluative judgments even while having no disagreement on the perceived quality of the product, e.g. both might think of it as a novel and hard-to-use product, but they disagree on the relative importance of each quality. In extreme cases, individuals might even use entirely different attributes to evaluate a product, reflecting the qualities they consider important for the specific product being evaluated.

One might assume a certain hierarchical structure on the importance of different qualities, that is universal across different individuals, such as that proposed by Jordan [23] on the relative importance of functionality, ease-of-use and pleasure. While this might hold true to a certain extent, empirical findings have shown this hierarchy to be modulated by a number of contextual aspects such as the user's motivational orientation [20], and time of ownership [25, 28].

All in all, research suggests that individuals may disagree on the perceived quality (e.g. ease-of-use) of a given product, or may even infer the overall value of a product on a different basis. To some extent, one could even wonder whether rating a product on quality dimensions that are imposed by the experimenter is always a meaningful activity for the participant, for example when the participant does not consider a quality dimension as relevant for the specific product.

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. Many different interview approaches have been proposed in the fields of Constructivist and Economic Psychology. For instance, *Free Elicitation*, rooted in theories of Spreading Activation [c.f. 5], probes the participants with a stimulus and asks them to rapidly express words that come to mind. The *Repertory Grid Technique (RGT)*, rooted in Kelly's Theory of Personal Constructs [c.f. 13], provides three alternatives to the participants and asks them to define dimensions in which the three products are meaningfully differentiated. The *Multiple Sorting Procedure*, rooted in Facet Theory [c.f. 1], asks the participant to sort products in a number of piles, and only later on define a label for each pile. Comparing the different techniques is not the focus of this paper; see [2, 3, 35, 36] for more information on this. While this paper illustrates the analysis procedure using Repertory Grid data, it may also be well applied to data derived from any of the other attribute elicitation techniques.

The Repertory Grid Technique
The RGT is one of the oldest and most popular attribute elicitation techniques. It originates from Kelly's Personal

Construct Theory (PCT) which suggests that people form idiosyncratic interpretations of reality based on a number of dichotomous variables, referred to as personal constructs or attributes. A personal construct is a bi-polar similarity-difference judgment. For example, when we meet a new person we might form a construct *friendly-distant* to interpret her character. In this process we perform two judgments: one of similarity and one of dissimilarity. Both judgments are done in comparison to reference points: people that we regard as friendly or distant.

To elicit the idiosyncratic attributes of each individual, the RGT employs a technique called *Triading,* where the participant is presented with three products and is asked to *"think of a property or quality that makes two of the products alike and discriminates them from the third"* [10]. This can be repeated for all possible combinations of products and until no new attribute arise. The result is a list of attributes that the specific individual uses to differentiate among a set of products. The attributes may then be employed in rating scales, typically Semantic Differentials [31], and each participant rates the set of products on his own elicited attributes. Participants' ratings are subsequently analyzed with exploratory techniques such as Principal Components Analysis (PCA) and Multi-Dimensional Scaling (MDS).

With the recently increased interest in user experience (e.g., [18]), the RGT has become popular in the field of HCI. Hassenzahl and colleagues employed the RGT to evaluate the outcome of parallel design [13] and analyze the perceived character of websites [14]. Fallman [9] elicited users' experiences with mobile technology devices, while Boyd Davis and Carini [6] explored player's experience of fun in video games. Karapanos & Martens [24] explored the differences between designers' and users' perceptions on a set of user authentication techniques for multi-user printers, while Hertzum et al. [11] studied the differences between designers' and users' perceptions for three diverse cultural settings.

It, thus, becomes evident that an increasing number of researchers in HCI, emphasize the idiosyncratic nature of subjective judgments on the quality of interactive products. To our knowledge, however, all RGT approaches up to date have been employing averaging techniques for the quantitative analysis of personal attribute judgments [26]. We believe this to be due to a lack of more advanced techniques that can account for diversity in users' subjective judgments, eventually undermining the core motivation for the RGT and other personal attribute elicitation methods. In the remainder of the paper, we suggest a quantitative, exploratory MDS procedure to account for the diverse views that one or more individuals may have on a set of products. It will be demonstrated that even single subjects can handle more than one view on a set of stimuli. We will show that by averaging interesting views are overlooked due to majorization bias.

THE STUDY

The data for the present analysis was taken from Heidecker and Hassenzahl's [21] study of individuals' perceptions of eight university websites. The study was part of a larger project aiming at understanding how the Technical University of Darmstadt (TUD) is perceived in comparison to other regional competitors. Ten individuals, all students at TUD, participated in the study.

The eight university websites were presented to participants in the form of color A4 screenshots of the main page. Using the Repertory Grid Technique [13], a number of attributes on which the eight websites differ, were elicited from each participant . Participants were then asked to rate the websites on their own elicited attributes, using 7-point Semantic Differential scales. The resulting data set consisted of a total of 118 attributes (10 to 14 per participant) on which ratings for the eight different stimuli were elicited.

Table 1. Attribute categories and examples

Attribute category	Example
Layout	Graphical layout – Textual layout
	Colorful – Pale colors
	Professional - playful
University Image	Technical studies – Social studies
	Emphasis on achievement – Average univ.
	Refers to student life – Modern organization
Information Access	Fast access to information – time-intensive
	Legible – Tangled

MDS FOR THE ANALYSIS OF QUALITY JUDGMENTS

Multivariate techniques such as FA and MDS aim at modeling relations between stimuli (e.g. websites), attributes (e.g. "professional – unprofessional") and overall judgments (e.g. preference). More specifically, MDS looks for a K-dimensional configuration for the stimuli such that the coordinates of the stimuli in the configuration space along different axes can be monotonically related to the observed attribute ratings of the participants [30].

Figure 1 illustrates a 2D MDS configuration with two stimuli and two attributes. The relative positions of the stimuli on a given attribute axis reflect subjects' ratings for the stimuli on this attribute. For instance, website j can be perceived as being both more *legible* and *colorful* than website i.

An important motivation for MDS is the *principle of homogeneity of perception* which states that attribute judgments from different participants are related and thus can be represented in a common configuration space [12, 30]. This view, although it holds in perceptual judgments, has recently been challenged in more cognitive judgments where the quality dimensions of interactive products are assessed [24].

This paper will highlight the trade-offs of an averaging approach by demonstrating that only 1/6th of the attribute judgments (18 out of 118) are taken into account when the
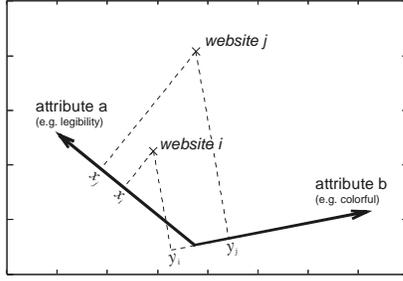
Figure 1. A two-dimensional MDS configuration of two websites using ratings from two attributes. Website *j* is perceived as more *legible* and *colorful* than website *i*.

analysis is restricted to an average view from all participants. It will further challenge this view by providing evidence for the fact that even single individuals can handle several different views in parallel when assessing a set of products.

AN MDS APPROACH TO ACCOUNT FOR DIVERSITY
The starting point of the proposed approach is that of *identifying the different views* that each participant has on the set of products. In this step, an average model is attempted for each participant. However, attributes that are not adequately predicted by the average model (see Table 2) are removed and used in deriving a second model, i.e. a second view for the specific participant (Figure 2 illustrates two diverse views derived for one subject).

Once the diverse views of all individuals have been identified, t*he similarity among them is assessed* and views are clustered into groups of increased homogeneity.

A final set of diverse configurations is formed by grouping the similar views, which are then used to model the attributes from all participants.

Identifying the different views
In identifying the different views that an individual might hold, one tries to model the individual's perceptions in one or more *non-trivial K-dimensional* models, each explaining *adequately* a part of his/her attribute judgments. The maximum dimensionality *K* is limited by the number of degrees of freedom in the data, but may also be set a priori by the data analyst. For the example data set considered below the dimensionality was fixed to *K*=2 so that different visualizations can be easily presented on paper. Note that models of degree higher than 2 need multiple 2D views to be assessed anyhow. However, in this latter case, the views are different 2D projections of a shared multi-dimensional configuration. The 2D views that we will present in this paper, on the other hand, can be independent.

A *two-step procedure* is proposed to establish whether zero, one or two models with dimension *K*=2 can adequately model the attribute scores of a single observer. In the first step, all attributes of a participant are modeled together, as is common practice in MDS (average model). However, only the attributes that satisfy a particular goodness-of-fit

criterion are considered to be adequately modeled. These attributes are analyzed to form the first model, i.e. the individual's most dominant view on the set of products.

In the second step, the attributes that displayed the least fit to the average model are grouped and used to attempt a second model. By selecting the least-fit attributes, instead of all remaining attributes, we promoted the diversity between the two models. The same goodness-of-fit criteria are applied for the second model to select the attributes that are retained.

Table 2. Goodness of fit Criteria. Attributes that are adequately predicted are employed in model 1. A second model is attempted only on attributes that display the least fit, to ensure diversity between the two models.

	R^2	R_k
1. Adequate fit	$R^2 > .5$	$R_k > 6$
2. Average fit (Excluded)		$4 < R_k < 6$
3. Least fit (attempt 2nd model)		$R_k < 4$

Defining goodness-of-fit criteria
We suggest a combined goodness of fit criterion. First, for an adequately predicted attribute, a substantial amount of its variance should be accounted for by the model. This proportion of explained variance is the R^2 statistic (i.e., the squared multiple correlation coefficient). A threshold $R^2 > 0.5$ was set, implying that only attributes are retained for which at least 50% of their variance is accounted for by the model. A limitation of this criterion is that it is insensitive to the range of the ratings for the different stimuli on a given attribute. An attribute might make no meaningful differentiation between stimuli (e.g. if all stimuli are rated as 4 or 5 on a 7-point scale) but can nevertheless be well-predicted by a model. To account for this limitation, we combine it with a second criterion.

This second criterion is a modification of a measure originally proposed by Draper and Smith [7; p244]. It is the ratio of the maximum range of the predicted scores for attribute *k* divided by the standard deviation σ_k of the estimation error [30] in the attribute scores (1a).

$$R_k = \frac{\hat{A}_{k,max} - \hat{A}_{k,min}}{\sigma_k} (1a) \qquad \frac{1}{n^2}\sum_{i,j}\frac{[\widehat{A}_{ki} - \hat{A}_{kj}]^2}{\sigma_k^2} (1b)$$

A combined criterion thus takes into account both the accounted variance in the attribute as well as the range of the scores for the different stimuli (i.e. the attribute's strength). The obvious limitation of the second measure is its sensitivity to outlier scores. However, in single-stimulus scales such as the semantic differential scales, these outlier scores may actually be very valuable, since they point at the stimuli that most strongly influence the existence of the attribute scale in the first place. When using more sensitive scales such as paired comparisons [24], one might consider adopting the modified measure (1b) that averages across differences in predictions. Draper and Smith [7] proposed a minimum ratio value of four, meaning that any attribute predic-
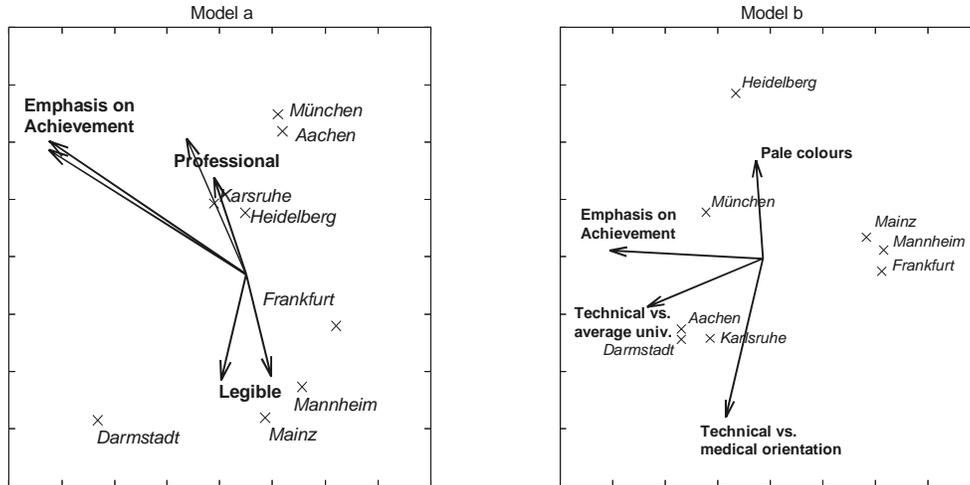
Figure 2. Two diverse views for one subject

tor with a ratio below four hardly makes any distinction between the stimuli and is pretty useless. Predictors with a ratio value above ten are considered to be excellent. We decided to use an *acceptable ratio* of six for the data analysis reported in Table 1.

We are aware of the fact that the criteria that we introduce for assigning attributes to models may come across as somewhat arbitrary. The main objective of this paper is to illustrate that multiple views can provide richer modeling of heterogeneous data than a single (averaged) view. It is hence not particularly crucial at this stage whether or not our strategy for partitioning the attributes is optimal. Finding more optimal ways of partitioning attributes is an issue that can be addressed in more depth after the usefulness of having multiple views is firmly established. A more optimal partitioning strategy will only help to strengthen our claim.

Two diverse views for one subject
Table 3 illustrates the analysis process on the attribute judgments of a single subject. A first (average) model was attempted on all attributes of the subject. Attributes (2,4,6,8,12,13; in bold) were adequately predicted by the average model, using the two criteria that were discussed before, i.e. $R^2 > .5$ & $R_k > 6$. Model 1 was then derived by optimizing the average model only for the attributes that were adequately predicted by the average model.

Note that the R^2 and R_k values are identical (at least in this exact decimal point) for Model 1 and the average model. This implies that when removing the attributes that are not adequately predicted (even with these arbitrary criteria), the 2D configuration space (which is represented in the model parameters) displays virtually no change. In other words, the attributes that were removed (according to the arbitrary criteria) had no contribution to the configuration space. Thus, the information contained in these attributes is not modeled when attempting an averaging analysis and therefore it is lost.

Out of the all the attributes that were not adequately predicted, attributes (1,5,7,9,10,11; in italics) displayed the least fit by model 1, i.e., $R_k < 4$. These were used to derive a second model. Out of them, only attributes (5,7,9,10) turned out to be adequately predicted by model 2, using the same goodness of fit criteria as used in model 1.

Figure 2 illustrates the different insights that the two diverse views bring. One can note that the two views highlight semantically different attributes. Each attribute is visualized as an arrow, i.e. a dimension, on which the relative positions of the websites can be compared. The length of each arrow highlight's the strength of the attribute, reflecting the variance in the attributes ratings for the different stimuli; on some attributes all websites might be rated as 4 or 5 on a 7-point scale, while others might make strong differentiations between sites, i.e. across the whole range of the scale.

The first view provides overall three different insights. First, that the universities of Frankfurt, Manheim and Mainz

Table 3. Goodness of fit statistics for the two diverse models of subject one. Attributes (2,4,6,8,12,13) were adequately predicted $(R^2 > .5$ & $R_k > 6)$ by model 1. Attributes (1,5,7,9,10,11) displayed the least fit $(R_k < 4)$ and were used to derive a second model. Attributes (5,7,9,10) were adequately predicted by model 2.

No	Attribute Variance (σ^2)	Avg. Model R^2	Avg. Model R_k	Model 1 R^2	Model 1 R_k	Model 2 R^2	Model 2 R_k
1	2.6	.47	2.2	*.47*	*2.2*	.36	3.3
2	3.8	.89	7.3	**.89**	**7.3**		
3	0.6	.73	4.1	.73	4.1	.56	2.6
4	1.9	.98	18.6	**.98**	**18.6**		
5	3.7	.49	2.3	.49	*2.3*	**.95**	**13.7**
6	2.2	.99	40.5	**.99**	**40.5**		
7	1.7	.48	2.4	*.48*	*2.4*	**.99**	**39.6**
8	6.3	.93	9	**.93**	**9**		
9	4.1	.63	4.8	*.63*	*4.8*	**.99**	**40.1**
10	4.5	.26	2.1	*.26*	*2.1*	**.61**	**6.5**
11	3.9	.08	0.9	*.08*	*0.9*		
12	1.9	.88	6.8	**.88**	**6.8**	.48	2.8
13	5.6	.99	50.4	**.99**	**50.4**	.85	5.5

5

Subj.	Total	View a	View b	Remain
1	13	5	4	4
2	13	6	4	3
3	14	5	-	9
4	10	5	4	1
5	12	8	-	4
6	11	6	4	1
7	13	-	-	13
8	11	-	-	11
9	11	4	-	7
10	10	4	3	3

are perceived as putting less *emphasis on achievement,* as compared to the remaining five universities. This may be induced by the websites but may also reflect prior beliefs of the individual. Second, the websites of the universities of München, Aachen, Karlsruhe and Heidelberg have a more *professional layout* as opposed to the remaining four which have a more *playful* one. Last, the subjects perceive this same group of websites as legible as opposed to the remaining four in the upper part of the figure that are perceived as having no clear structure.

The second view partly provides overlapping information (emphasis on achievement), but also gives three new insights. First, the website of the University of Heidelberg is differentiated from all others by having a less colorful layout. Second, the Universities of Darmstadt, Aachen and Karlsruhe are differentiated as universities that provide primarily technical studies, as opposed to the universities of Mainz, Mannheim and Frankfurt that are referred to as universities of average quality, and third, as opposed to the university to Heidelberg that is perceived as a university offering primarily medical studies.

Note that an attribute may range from being purely descriptive, i.e. referring to specific features (e.g. allow search), to having an evaluative tone, e.g. referring to the perceived quality of the product (e.g. easy to use) or the product's overall appeal (e.g. good). This enables the researcher to gain a better understanding of the inferences individuals make as they form evaluative judgments of products.

The resulting views
Table 1 summarizes the results of the above analysis for all ten participants. For two of the ten participants (7, 8), no substantial agreement between their attribute judgments is observed, i.e., no satisfactory MDS-model can be derived. This implies that they either have as many different views as their attribute judgments, or more likely, that their ratings are too noisy to be analyzed in a meaningful way. For another three participants (3,5,9) only one satisfactory model is determined, which accounts for roughly half of their attributes (17 of 37). The remaining five participants (1,2,4,6,10) have two different, complementary models, i.e.,

the number of attributes in the first model (26) is comparable to the number of attributes in the second model (19). This shows that diversity is prevalent. Half of the participants even hold two different views, explaining subgroups of attributes.

All together, 13 different views emerged from the ten individuals. These views may partly overlap, which motivated us to group similar views and identify the major diverse of this user group.

Assessing the similarity between different views
In grouping the diverse views one has to derive a distance measure that reflects the degree of dissimilarity between configurations. Each configuration can be regarded as a NxK matrix, where N is the number of stimuli and K the number of dimensions of the configuration space. The distance between configurations X_n and X_m can be calculated using the MATFIT procedure, developed by Ramsay [32]. MATFIT seeks for a transformation matrix M that minimizes the distance measure:

$$d^2 = trace[(X_nM - X_m)^t(X_nM - X_m)]$$

An arbitrary KxK transformation matrix M was applied. The procedure was repeated with the matrices in reverse order as a means to calculating both distances: with X_n as independent and X_m as dependent, and vice versa. The resulting distances were visualized in three dimensions using the program XGms [30]. A hierarchical (minimum variance) clustering algorithm was applied to the 3-D configuration (a cluster is denoted by the lines connecting the different views). Figure 3 represents a 2-D perspective on the 3-D configuration of individual models. Note that the distances in this 2D perspective do not necessarily reflect the true distances in 3D, which is why one should rely on the lines that visualize the clusters (clustering was performed in 3D). Participant 7 and 8 are excluded, because no individual model could be fitted. In case of two fitting models per participant (1,2,4,6,10) the first model is denoted as a, the second as b.
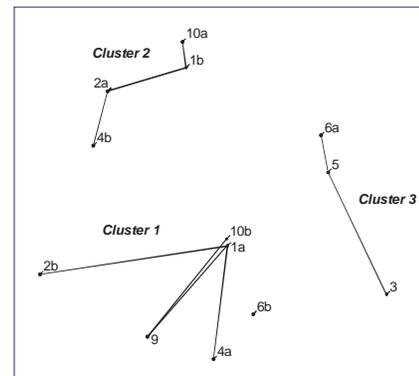

Figure 3. A 2-D perspective of the 3-D visualization of distances between individual's different views.

Three clusters of models emerged. Cluster 1 summarizing 6 of the 13 single models (1a, 2b, 4a, 6b, 9, 10b), cluster 2 summarizing 4 models (1b, 2a, 4b, 10a) and cluster 3 summarizing the remaining 3 models (6a, 5, 3). The complementary models (a & b) for these five participants appear to be quite dissimilar as illustrated in figure 3 by the fact that they belong to different clusters. These clusters represent homogenous views, which can subsequently be mapped out.

Grouping the homogeneous views
In the last phase we establish a final set of configurations that represent the major diverse views across all subjects and all attributes, on the set of stimuli. Views that belong in the same cluster are analyzed together and a shared MDS configuration is sought. Attributes that are not adequately predicted by the model are eliminated with the same criteria as in phase 1. The resulting 'averaged' views are then used for modeling the attributes from all participants. Attributes are allowed to exist in more than one configuration if they are adequately explained by all of them. When attributes in the same semantic category are not significantly different (which can be deduced from the fact that they have overlapping confidence ellipses in the K-dimensional configuration space), they are grouped. Attributes that cannot be grouped (have no replicates) are eliminated since no evidence exists that they contain reliable information.

How do the diverse views compare to the average view?
We will address this question in three ways. Firstly, we will illustrate that the average model predicts less than half of the attributes predicted by the three diverse models together (attributes that are adequately explained by more than one model are only counted once for the model that explains them best). Secondly, we will illustrate that, for the attributes that are predicted by the three diverse models, these models provide a better fit than the average model, as demonstrated by the amount of explained variance in the attribute data and the values of the well established Akaike Information Criterion (AIC) for model selection. Thirdly, by exploring the resulting views, we will illustrate that the diverse models, not only account for more attributes and with a better fit, but that they also result in semantically richer insights, i.e., introduce more semantically different attributes.

Surprisingly enough, the average model could only predict 1/6th of all the attributes from the ten participants, i.e. 18 out of the 118 attributes. This means, that when deriving an average configuration to understand how individuals distinguish between these websites, only 1/6th of the attributes are taken into account. This is illustrated by the high correlation between the two resulting configurations (R=.99), the one derived using all 118 attributes and the one derived using only the 18 attributes that are well predicted. Thus, the consequence of averaging is that we account only for 1/6th of the information available. The three diverse models predict 12, 10, and 16 attributes respectively (attributes predicted by more than one model were excluded from the

ones that displayed the least fit). Thus, by accounting for diversity, even with our clearly sub-optimal procedure, we account for more than double the number of attributes than in the case of the average model.

Table 5 illustrates the goodness of fit of the average and the three diverse models for the 38 in total attributes resulting from models 1 to 3. As expected, a significant increase in the accounted variance (R^2) of the attribute data is observed as we the move from the average to the specific (i.e. diverse) model. But, does this increase in the goodness of fit of the model outweigh the increase in model complexity, i.e. going from one 2D to three 2D models? One of the most widely used criteria for model selection is the Akaike Information Criterion (AIC) [4] which is a function of the log likelihood value reflecting the goodness of fit of the model and the M degrees of freedom in the model reflecting its complexity:

$$AIC_c = -2\log\left(L(\hat{\theta})\right) + 2M\frac{n}{n-M-1}$$

Table 5. Goodness of fit of the average and the three diverse models for the 38 in total attributes resulting from models 1 to 3.

|	Attribute	R^2			
No	Variance (σ^2)	Average	Model 1	Model 2	Model 3
1	2.0	.36	.99		
2	4.3	.86	.99		
3	1.7	.91	.95		
4	4.6	.91	.94		
5	1.7	.32	.91		
6	2.7	.76	.90		
7	4.2	.91	.90		
8	6.6	.70	.87		
9	1.7	.21	.87		
10	6.1	.83	.86		
11	4.2	.75	.85		
12	3.0	.68	.69		
1	2.2	.39		1.0	
2	5.6	.78		1.0	
3	1.9	.70		.98	
4	6.3	.82		.94	
5	4.7	.68		.90	
6	2.8	.54		.90	
7	1.8	.89		.90	
8	3.8	.68		.89	
9	2.3	.76		.89	
10	1.9	.54		.88	
1	4.8	.75			1.0
2	6.0	1.0			1.0
3	7.0	.67			.95
4	1.7	.91			.94
5	7.0	.93			.94
6	2.6	.99			.93
7	4.3	.89			.93
8	1.6	.83			.92
9	3.7	.91			.91
10	5.9	.91			.91
11	2.6	.88			.90
12	7.4	.80			.90
13	6.1	.83			.90
14	1.7	.80			.87
15	4.7	.59			.85
16	2.7	.57			.83
		$AIC_{avg} = 1480$		$AIC_{123} = 1126$	$\Delta = 354$

7
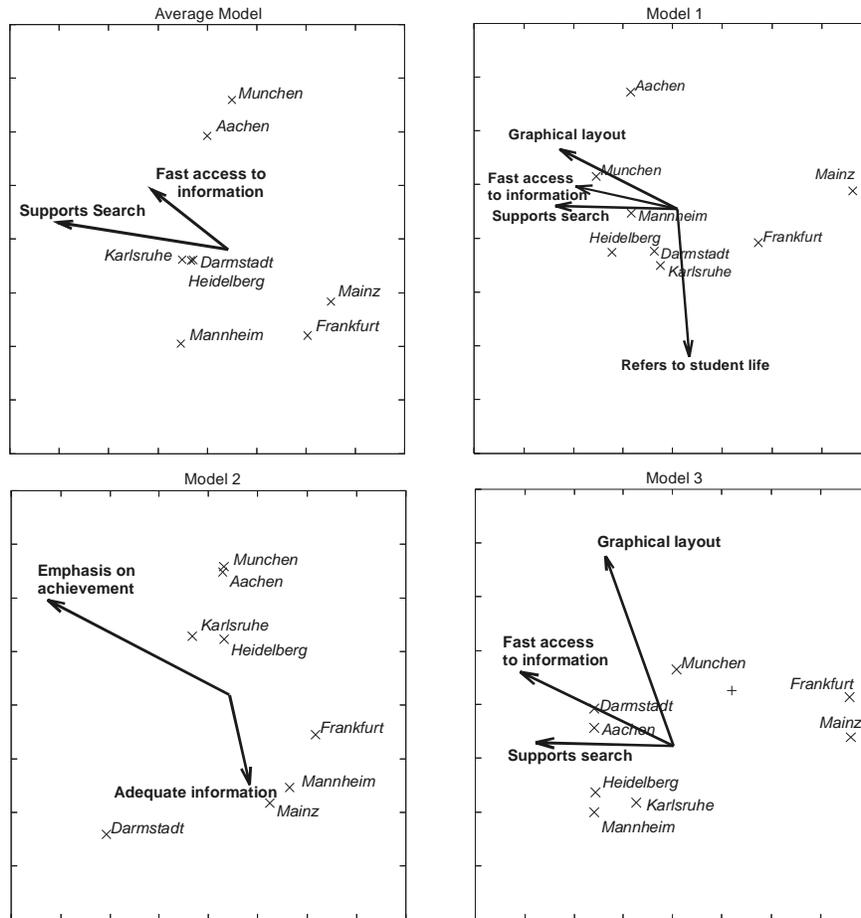
Figure 4. The average model and the three separate sub-models.

Burnham and Anderson [4] proposed a set of heuristics when comparing the AIC values of two models. Δ_i reflects the difference between the simpler (i.e., average) model and the more complicated one (i.e., consisting of three sub-models). $\Delta_i \leq 2$ provides significant evidence for the simpler model, $4 \leq \Delta_i \leq 7$ provides weak support for the heterogeneous model, while $\Delta_i \geq 10$ provides strong support for the heterogeneous model. In our case $\Delta_i = 354 >> 10$, providing significant evidence that the diverse models, despite the increase in the model complexity, perform better than the average model.

Figure 4 illustrates the insights gained by the average and the three diverse models that are derived from the views corresponding to clusters 1 through 3. A significant overlap exists between models 1 and 3 (five common attributes), while model 2 provides a completely different view.

The average model, although it accounts for more attributes than each of the diverse models, fails to predict semantically similar attributes. Thus, replicate attributes (i.e. attributes pointing towards the same direction with overlapping confidence ellipses) exist only for two attribute categories, namely *"Fast access to information"* and *"Supports search"*. The websites of the university of München and Aachen are differentiated from the remaining ones as web-

sites that provide fast access to information, while the second attribute differentiates mainly the websites of the universities of Mainz and Frankfurt as the ones that do not support searching.

These two attributes are present also in two of the three diverse models, model 1 and model 3. Model 1 further differentiates the websites of Aachen and München as having a *"graphical layout"*, the website of the university of Aachen is mainly differentiated from all others as a website that does not *"refer to student life"*. On the contrary, model 2 provides a different insight. It reveals that the websites of the Universities of Mannheim, Frankfurt and Mainz put *"less emphasis on achievement"*. The set of websites can also be split in two groups based on the amount of information that they provide to the user.

CONCLUSION

Employing predefined questionnaires has been the common practice in the study of subjective judgments in HCI. In this paper we highlighted a limitation inherent in this approach, that of assuming homogeneity across individuals perceptions. Individuals may disagree on the perceived quality of a given product, or may even infer the overall value of a product on a different basis. Thus, to a certain extent, rating a product on pre-defined attributes may not always be a

meaningful activity for the subject, for example when the subject does not consider the quality attribute as relevant for the specific product. Relevant attributes may on the other hand be missing from the list of pre-defined attributes.

Approaches such as the Repertory Grid Technique typically emphasize the idiosyncratic nature of perception and evaluation of stimuli. Individuals rate stimuli only on attributes that were elicited when they were asked to qualitatively differentiate between stimuli. However, there is also a problem inherent in this approach. As an analyst, one is confronted with as many idiosyncratic views as participants. Views may overlap or even contradict one another; it is in any way complicated to systematically explore this diversity. In practice, the consequence is either an idiosyncratic analysis with a "narrative" summarization or the use of average models. Averaging, however, treats diversity among participants as error and thereby contradicts the basic idea of the underlying approaches.

In this paper we argued against averaging in the analysis of personal attribute judgments. We illustrated that when using averaging only 1/6th of the attributes in our study, i.e. 18 out of 118, were taken into account. A new MDS procedure that can better account for diversity in judgments was developed and its added value was illustrated through the reanalysis of published data. The analysis resulted in three diverse views on the data which were directly compared to the average view that is the common practice in RGT studies. The diverse models were found a) to account for more than double of the attributes accounted for by the average model, b) to provide a better model fit even for the attributes that were adequately predicted by the average model, and c) to result in semantically richer insights, since the diverse models can account for more semantically different attributes.

We further illustrated that diversity exists not only across different individuals, but also within a single individual, in the sense that different attribute judgments of a subject may reveal different, complementary, views. At any point in time individuals can have different, seemingly conflicting views. For instance, individuals may regard one car as beautiful, but at the same time expensive. Individuals' overall evaluations of the car might thus be modulated by contextual aspects such as their motivational orientation (e.g. whether they just saw it on a newspaper on a Sunday morning or they are in the process of purchasing it). Thus, being able to understand individuals' conflicting views is crucial for understanding how individuals infer the overall value of a product.

The proposed approach is a first step towards more exploratory procedures in the analysis of subjective judgments. It is thus not free of limitations. Firstly, the procedure that is currently used to assign attributes to different models is based on heuristics and not on an explicit optimization criterion. Developing a more structured (optimal) approach is clearly one of our objectives for the future. Secondly, the

analysis that was reported in this paper was purely descriptive, in the sense that it aimed at identifying the most dominant quality attributes on which users differentiate this set of products. Once a set of latent attributes is however established, one could explore the relations among them and establish potential theoretical models of the product domain. In this sense, one could perform exploratory path analysis which can be of significant value when limited theory exists in the field.

With this paper, we strongly advocate the view that the analysis of quality judgments of interactive products should not stop on a group level, but must be extended to the relations between the attribute judgments *within* an individual. The Repertory Grid combined with the suggested technique to analyze the resulting quantitative data is an important step towards the adequate account of homogeneity and especially diversity in individual quality judgments.

REFERENCES
1. Al-Azzawi, A., Frohlich, D., and Wilson, M., Beauty constructs for MP3 players. CoDesign, 2007. **3**(1 supp 1): p. 59 - 74.

2. Bech-Larsen, T. and Nielsen, N.A., A comparison of five elicitation techniques for elicitation of attributes of low involvement products. Journal of Economic Psychology, 1999. **20**: p. 315-341.

3. Breivik, E. and Supphellen, M., Elicitation of product attributes in an evaluation context: A comparison of three elicitation techniques. Journal of Economic Psychology, 2003. **24**: p. 77-98.

4. Burnham, K.P. and Anderson, D.R., Multimodel Inference: Understanding AIC and BIC in Model Selection. Sociological Methods & Research, 2004. **33**(2): p. 261.

5. Collins, A.M. and Loftus, E.F., A spreading-activation theory of semantic processing. Psychological Review, 1975. **82**(6): p. 407-428.

6. Davis, S.B. and Carini, C., Constructing a Player-Centred Definition of Fun for Video Games Design. HCI 2004 Conference, 2004: p. 117-132.

7. Draper, N.R. and Smith, H., Applied Regression Analysis. 1998, New Yoek: John Wiley and Sons Inc.

8. Egger, F.N., " Trust me, I'm an online vendor": towards a model of trust for e-commerce system design. Conference on Human Factors in Computing Systems, 2000: p. 101-102.

9. Fällman, D., In Romance with the Materials of Mobile Interaction: A Phenomenological Approach to the Design of Mobile Information Technology. 2003: Univ.

10. Fransella, F., Bell, R., and Bannister, D., A Manual for Repertory Grid Technique 2003: Wiley.

11. Frøkjær, E., Hertzum, M., and Hornbæk, K., Measuring usability: are effectiveness, efficiency, and satisfaction really correlated?, in *Proceedings of the SIGCHI conference on Human factors in computing systems*. 2000, ACM Press: The Hague, The Netherlands.

12. Green, P.E., Carmone Jr., F.J., and Smith, S.M., Multidimensional Scaling, Concepts and Applications. 1989, Boston, MA: Allyn & Bacon.

13. Hassenzahl, M. and Wessler, R., Capturing design space from a user perspective: The Repertory Grid Technique revisited. International Journal of Human-Computer Interaction, 2000. **12**(3-4): p. 441-459.

14. Hassenzahl, M. and Trautmann, T., Analysis of web sites with the repertory grid technique, in *CHI '01* 2001, ACM Press: Seattle, Washington.

15. Hassenzahl, M., Character Grid: a Simple Repertory Grid Technique for Web Site Analysis and Evaluation, in *Human Factors and Web Development. Lawrence Erlbaum*, J. Ratner, Editor. 2002: Mahwah, NJ. p. 183-206.

16. Hassenzahl, M., The interplay of beauty, goodness, and usability in interactive products. Human-Computer Interaction, 2004. **19**(4): p. 319-349.

17. Hassenzahl, M. and Sandweg, N., From mental effort to perceived usability: transforming experiences into summary assessments, in *CHI '04 extended abstracts on Human factors in computing systems*. 2004, ACM Press: Vienna, Austria.

18. Hassenzahl, M. and Tractinsky, N., User experience - a research agenda. Behaviour & Information Technology, 2006. **25**(2): p. 91-97.

19. Hassenzahl, M., Aesthetics in interactive products: correlates and consequences of beauty, in *Product Experience, Elsevier, Amsterdam*, H.N.J. Schifferstein and P. Hekkert, Editors. 2007.

20. Hassenzahl, M. and Ullrich, D., To do or not to do: Differences in user experience and retrospective judgments depending on the presence or absence of instrumental goals. Interacting with Computers, 2007. **19**: p. 429-437.

21. Heidecker, S. and Hassenzahl, M., Eine gruppenspezifische Repertory Grid Analyse der wahrgenommenen Attraktivität von Universitätswebsites, in *Mensch & Computer 2007: Konferenz für interative und kooperative Medien*, T.Gross, Editor. 2007: Oldenbourg, Munich. p. 129-138.

22. Hinkin, T.R., A Review of Scale Development Practices in the Study of Organizations. Journal of Management, 1995. **21**(5): p. 967.

23. Jordan, P.W., Designing Pleasurable Products: An Introduction to New Human Factors. 2000, London: Taylor & Francis.

24. Karapanos, E. and Martens, J.-B., Characterizing the Diversity in Users' Perceptions, in *Human-Computer Interaction – INTERACT 2007*. 2007, Springer. p. 515-518.

25. Karapanos, E., Hassenzahl, M., and Martens, J.-B., User experience over time, in *CHI '08 extended abstracts on Human factors in computing systems*. 2008, ACM: Florence, Italy.

26. Karapanos, E. and Martens, J.-B., The quantitative side of the Repertory Grid Technique: some concerns, in *in the proceedings of the workshop Now Let's Do It in Practice: User Experience Evaluation Methods in Product Development, Human factors in computing systems, CHI'08*. 2008: Florence.

27. Karapanos, E., Wensveen, S.A.G., Friederichs, B.H.W., and Martens, J.-B., Do knobs have character? Exploring diversity in users' inferences in *CHI'08 extended abstracts on Human factors in computing systems*. 2008, ACM Press: Florence.

28. Karapanos, E., Zimmerman, J., Forlizzi, J., and Martens, J.-B., User Experience Over Time: An initial framework, in *Proceedings of the Twenty-Seventh Annual SIGCHI Conference on Human Factors in Computing Systems - CHI '09*. 2009, ACM: Boston.

29. Markopoulos, P., Romero, N., van Baren, J., Ijsselsteijn, W., de Ruyter, B., and Farshchian, B. Keeping in touch with the family: home and away with the ASTRA awareness system. 2004: ACM Press New York, NY, USA.

30. Martens, J.-B., Image technology design: A perceptual approach. 2003, Boston: Kluwer Academic Publisher.

31. Osgood, C.E., Suci, G., and Tannenbaum, P., The measurement of meaning. 1957, Urbana, IL: University of Illinois Press.

32. Ramsay, J.O., MATFIT: A Fortran Subroutine for Comparing Two Matriced in a Subspace. Psychometrika, 1990. **55**(3): p. 551-553.

33. Sarstedt, M., A review of recent approaches for capturing heterogeneity in partial least squares path modelling. Journal of Modelling in Management, 2008. **3**(2): p. 140-161.

34. Schenkman, B.N. and Jönsson, F.U., Aesthetics and preferences of Web pages. Behaviour & Information Technology, 2000. **19**: p. 367–377.

35. Steenkamp, J.-B.E.M. and Van Trijp, H.C.M., Attribute elicitation in Marketing Research: A comparison of Three Procedures. Marketing Letters, 1997. **8:2**: p. 153-165.

36. van Kleef, E., van Trijp, H.C.M., and Luning, P., Consumer research in the early stages of new product development: a critical review of methods and techniques. Food Quality and Preference, 2005. **16**(3): p. 181-201.