Conjoint Analysis to Measure the Perceived Quality in Volume Rendering*

Joachim Giesen Klaus Mueller *Senior Member, IEEE* Eva Schuberth Lujin Wang
Peter Zolliker *Member, IEEE*

Abstract—Visualization algorithms can have a large number of parameters, making the space of possible rendering results rather high-dimensional. Only a systematic analysis of the perceived quality can truly reveal the optimal setting for each such parameter. However, an exhaustive search in which all possible parameter permutations are presented to each user within a study group would be infeasible to conduct. Additional complications may result from possible parameter co-dependencies. Here, we will introduce an efficient user study design and analysis strategy that is geared to cope with this problem. The user feedback is fast and easy to obtain and does not require exhaustive parameter testing. To enable such a framework we have modified a preference measuring methodology, conjoint analysis, that originated in psychology and is now also widely used in market research. We demonstrate our framework by a study that measures the perceived quality in volume rendering within the context of large parameter spaces.

Index Terms—Conjoint Analysis, Parameterized Algorithms, Volume Visualization

1 Introduction

The main purpose of visualization is to produce images that allow users to gain more insight into the illustrated data. This is a complex issue, depending on many factors of the visualization system, starting from human-computer interaction, to rendering speed, to rendering style and algorithm, and finally human perception and cognition. With the exception of the last component all of these factors have been designed by humans and many diverse technologies have emerged, and are still emerging, over the years. But in the end, human perception is the ultimate judge that determines which of these are the most effective. A popular focus of the field of visualization is the modeling and optimization via engineering and mathematics tools and frameworks, and often the designer/engineer him/herself judges the success of the method. Here, the easiest parameters to measure are rendering speed and memory consumption and others, which are all engineering quantities. However, in light of the importance of the last element in the chain, the human observer, a more recent focus has become to also conduct adequate user studies to measure the success of a proposed method. This practice is already common place in the field of human-computer interaction, and to a more limited extent also in information visualization, but less so in scientific and medical visualization. In essence, user studies are always considered burdensome since in many cases there are a large number of parameters and algorithmic alternatives, requiring many trials, that is, human subjects and experiments, to produce statistically significant results. This has been a major obstacle in assessing a method's success in terms of the human perceptive and cognitive system. The pressing question is: can we make this task easier by introducing a more methodical and organized approach. For this it pays to look at other fields, especially those driven by heavy monetary investments. One then finds that user studies play a major and dominant role in product marketing, where it is important to tune the various parameters of a product before it is being launched to market (or determine its launch at all). Clearly, these studies must be conducted as thoroughly as possible, in order to maximize the outcome, but they also must be conducted as efficiently as possible in order to minimize the time, burden put on the participants in the study and samples needed to explore the vast parameter space in a statistically significant manner. A technique called *conjoint analysis* [5] is the answer to all of these design goals, and this paper makes this technique accessible to visualization researchers and their specific domain setting. Visualization researchers are faced with the task that a large number of algorithms need to be compared. However, the number of algorithms is too large for a single user to compare/rank all of them in reasonable time (and with reasonable accuracy). Fortunately, in many visualization areas, such as volume visualization, the algorithms are not strictly arbitrary but to some extent related; that is, they are all different incarnations of one parameterized algorithm and are obtained by fixing the parameter values. A comparison of the algorithms then leads to a ranking of the algorithms/parameter settings. This is essentially the same problem that market researchers face when eliciting consumers' preferences on substitute goods that can be described in terms of attributes and attribute levels. Conjoint analysis, as introduced above, is a well established family of questionnaire based techniques to elicit consumer's preferences. It frees the evaluator from the daunting burden of presenting the effects of all attribute levels to all users for evaluation but nevertheless allows statistically significant results.

A main contribution of this paper is the development of our own conjoint analysis technique as an extension of Thurstone's method of comparative judgment [14]. Like most techniques it is based on some assumptions (model), but it has the advantage that all assumptions can be tested, which is not the case with off-the-shelf conjoint analysis software like Sawtooth' software [1]. The model assumptions allow us to derive robust preference estimates from sparse data, i.e., every user needs to 'explore' only a small fraction of the large parameter space.

We demonstrate our conjoint analysis technique in four related studies that fit two important visualization purposes: visual aesthetics and the conveyance of detail. In this pursuit, we can gain further insights. For example, we determine the relative importance of the algorithm's parameters and their levels. This is important information if one has to tradeoff perceived quality against other objectives like time or file size. Conjoint analysis allows us to quantify these tradeoffs. We can also study the effects of age, gender, culture, or color deficiencies on users' preferences.

Our analysis framework is timely in light of the various recent efforts to optimize viewpoints [8, 13], transfer functions [10], sampling intervals [15, 2], high-level appearance descriptors [12], illustrative

*Authors are listed alphabetically.

- *Joachim Giesen is with the MPI Informatik, E-mail: jgiesen@mpi-inf.mpg.de.*
- *Klaus Mueller and Lujin Wang are with Stony Brook University, E-mail: {mueller,lujin}@cs.sunysb.edu.*
- *Eva Schuberth is with ETH Zürich, E-mail: sceva@inf.ethz.ch.*
- *Peter Zolliker is with EMPA Dübendorf, E-mail: peter.zolliker@empa.ch.*

Manuscript received 31 March 2007; accepted 1 August 2007; posted online 27 October 2007.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

rendering parameters [3], perceived salience [9], and others. All of these use mostly mathematical, engineering, but also sometimes aesthetics and perception-motivated arguments to devise their methods. Controlled user studies need to eventually decide which strategy is most effective and relevant for the human observer, especially in conjunctive terms. Furthermore, these user studies can also be helpful to fine tune the parameters of these methods, which may also be task and domain dependent. In the following sections we first overview the theory of our framework in accessible terms, and then apply it within a typical multi-parameter volume rendering scenario to demonstrate the analysis path. While the results are in fact insightful in their own right, they should also and perhaps predominantly be valued as indicators for the analytical power of our framework as a more general user assessment tool and as a guideline on how to conduct and analyze a conjoint study in the context of visualization algorithms evaluation/comparison.

2 Choice based conjoint analysis

A class of items has a conjoint structure if it can be described by the Cartesian product $A_1 \times \ldots \times A_n$ of attribute sets A_i. The elements of the attribute sets are called attribute levels. An item a is then represented by a vector $(a_1, \ldots, a_n)$ with $a_i \in A_i$, i.e., by fixing the attribute levels. Conjoint analysis is a family of techniques for eliciting from a population of people their ranking (on some scale) of the elements in $A_1 \times \ldots \times A_n$, i.e., on the items. Conjoint analysis techniques can be distinguished by two (not independent) parameters: firstly the elicitation procedure, i.e., the way preference data are obtained from respondents, and secondly the way the elicited data are processed in order to derive a representation of individual or aggregated preference information (typically in form of a value or utility function).

In recent years choice based conjoint analysis has become the most popular conjoint analysis technique. It got its name from the method employed for elicitation, namely, preferences are elicited in a sequence of choice tasks. In a choice task a small number of items (typically between two and four) is presented to a respondent who has to state which one out of these she/he prefers most. Choice tasks are popular in market research since they resemble real buying situations and thus tend to provide the most reliable information.

There are many different ways to analyze the data obtained from several respondents and several choice tasks each, but any analysis method defines a *scale* on which the items are compared. A scale assigns to each item a number. In conjoint analysis there are essentially two types of scales used: on *ordinal scales* the numbers assigned to the items are their ranks in a linear order. Note that the nominal difference between ranks has no meaning. On *interval scales* an item is preferred over another if it gets assigned a larger number. Differences of the assigned numbers have a meaning on interval scales, but these scales have no *natural zero*. Note that translating all scale values on an interval scale has no effect.

Another difference in analysis methods is whether they define a scale for each respondent, or just a scale for a population of respondents (aggregated scale). Our analysis method defines an interval scale for a population of respondents.

3 Data Collection

As mentioned several times our goal was to measure the perceived quality of a visualization algorithm for different parameter settings that we describe in Section 4. We chose choice based conjoint analysis as our elicitation procedure, where each choice task was a paired comparison between two renderings, i.e., between two parameter settings. Note that the cognitive burden increases with the number of items from which to choose. Higher cognitive burden should result in poorer data quality. We decided to use choice tasks with the least cognitive burden, namely paired comparisons.

For our study we used two data sets. The first data set FOOT is meant to cover the medical application domain, whereas the second data set ENGINE covers the engineering applications area. See Figure 2 for various images rendered for these two data sets. The ENGINE data size is $256 \times 256 \times 256$, and the FOOT data size is $154 \times 263 \times 222$. For the FOOT data set we had 2250 different parameter settings resulting in 2250 different renderings (images) and for the ENGINE data set we had 2700 different parameter settings.

Perceived quality itself can be measured along different directions. We made this more explicit by asking two different questions: *Which image do you like best?* and *Which image shows more detail?* We will later refer to the first question as AESTHETICS and the second as DETAIL. Note that the second question is more specific than the first, which is fairly general.

Each combination of data set and question is considered as a different study, i.e., we conducted the four different conjoint studies [ENGINE, AESTHETICS], [ENGINE, DETAIL], [FOOT, AESTHETICS] and [FOOT, DETAIL].

We elicited data for our studies from visitors at an exhibition that took place to celebrate the 25th anniversary of the computer science department at ETH Zürich. Our survey took place in a room at the exhibition that was darkened using light-impermeable black curtains. The room had six work places each having a computer with mouse and LCD screen with resolution 1280x1024 pixels. During the survey the light in the room was switched off.

786 visitors of the exhibition participated in our study. From the participants we collected the following data: age, gender and color deficient (yes or no). To test for color deficiencies we used the Ishihara test [7]. Every respondent took part in exactly two of our conjoint studies, i.e., a respondent always had to answer the same randomly chosen question on both data sets. That is, each respondent participated in one study for each data set FOOT and ENGINE, respectively. We did not conduct the two studies one after the other, but interleaved them: alternately a respondent was shown ENGINE image pairs and FOOT image pairs to choose from, altogether 20 pairs for each data set. The image pairs for the comparisons were determined as follows: the first image was drawn uniformly at random from the set of all images. The second image was then drawn uniformly at random from the set of images having for each parameter a different value than the first image. The images were presented side by side on the screen with a black stripe separating them. The background of the screen was set to black. All images had a resolution of 512×512 pixels. Respondents chose an image by clicking on it, after a click the next image pair was shown. Typically the respondents needed three to five minutes to complete the survey.

4 Rendering Algorithm

We have chosen a relatively standard volume visualization scenario to demonstrate our user study framework. Using GPU-accelerated ray casting rendering, the visualization of each object can be described in terms of the parameters COLORMAP, RENDERING, VIEWPOINT, RESOLUTION, STEP SIZE and BACKGROUND. The parameter COLORMAP has three levels which correspond to different color maps that are applied for transfer function design. For all transfer functions, the alpha channel has been set to always reveal most of the object's structures, in order to suppress 'occlusion' to act as an independent variable. The parameter RENDERING describes the applied rendering mode and has five levels: DVR (Direct Volume Rendering), DVRNS (Direct Volume Rendering with No Shading, just compositing), DVRGM (Direct Volume Rendering with Gradient Modulation to highlight surfaces), XRAY (Colored X-Ray) and MIP (Colored Maximum Intensity Projection). The parameter VIEWPOINT has six levels for the ENGINE and five levels for the FOOT data set. It describes the viewpoint under which the observer sees the object. Different viewpoints are chosen in such a way that most structures are always kept visible, again to prevent 'occlusion' from playing a significant role in the study. The parameter RESOLUTION describes the screen resolution used for rendering. We render at the resolution of the dataset and twice that. Note that in the end the image size was always 512×512 (the image rendered at reduced resolution, that is, at volume resolution, was scaled up with bilinear filtering). STEP SIZE is the ray traversal increment (measured in voxel size), which has three levels, 0.2, 0.5 and 1.0. Finally the parameter BACKGROUND describes the color of the background and has five levels: BLACK, WHITE, DARK

GREEN, DARK BLUE and YELLOW. Combining these parameters results in the 2700 ENGINE images and in the 2250 FOOT images.

5 DATA ANALYSIS: THEORY

In this section we describe a method to define an interval scale for a population of respondents from their choices in paired comparisons. Note that in general any 1 out-of k choice task provides information about $k-1$ paired comparisons. Our method extends Thurstone's method of comparative judgement [14], which does not assume a conjoint structure. In a nutshell, our method works as follows: At first we estimate scale values for all levels of a single attribute. To this end we interpret any paired comparison as a comparison of just the two levels of the given attribute that are present in this comparison, ignoring differences in the levels of all other attributes. We can apply this method to all attributes to obtain scale values for all their levels. Note that the scale values for levels of different attributes need not be comparable yet. To make them comparable we design a rescaling method. It builds on the fact that for the computation of the scale values for any attribute, always the same stated preferences are used, namely, the outcomes of all paired comparisons. Finally, the scale value of an algorithm, i.e., complete parameter set, is just the sum of the scale values of the parameter values.

5.1 Thurstone's method

A good introduction into Thurstone's work is given in [4]. Here we summarize it only briefly. Our goal is to define a meaningful scheme to assign scale values on an interval scale to n items that we label by $1,\dots,n$. Thurstone's intuition was that the relative frequency $F_{i\succ j}$ that i was preferred over j by the respondents is an indirect measure for the distance between i and j. He derived an interval scale from this intuition under the assumption that the scale values S_i of the items i are uncorrelated normally distributed random variables with expectations μ_i and variances $\sigma_i^2 \equiv \sigma^2$, i.e., all the variances are the same.

The idea is to assign to each item i the value μ_i, on an interval scale, which still has to be defined. To do so we need to estimate all the μ_i's from the paired comparison data that we have available. It turns out that it is easier to estimate the differences $\mu_i - \mu_j$. We use the latter and assign to each item i the value

$$\frac{1}{n}\sum_{j=1}^{n}(\mu_i-\mu_j) \;=\; \mu_i - \frac{1}{n}\sum_{j=1}^{n}\mu_j \;=:\; \mu_i - \bar{\mu}.$$

That is, we only shift the scale that assigns μ_i to item i by the value $\bar{\mu}$, i.e., as interval scales both scales are the same. Note that by the properties of normal distributions, the differences $S_i - S_j$ are normally distributed with expectations $\mu_i - \mu_j$ and variance $2\sigma^2$. This yields

$$P[S_i - S_j > 0] = \frac{1}{\sqrt{4\pi\sigma^2}}\int_0^\infty e^{-\frac{(x-(\mu_i-\mu_j))^2}{4\sigma^2}}dx = \Phi\left(\frac{\mu_i-\mu_j}{\sqrt{2}\sigma}\right),$$

where Φ is the cumulative distribution function

$$\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{y^2}{2}}dy$$

of the standard normal distribution. Hence,

$$\mu_i - \mu_j = \left(\sqrt{2}\sigma\right)\Phi^{-1}\big(P[S_i - S_j > 0]\big).$$

We can estimate $P[S_i - S_j > 0]$ by the observed quantity $F_{i\succ j}$, i.e., the relative frequency that item i was preferred over j and thus estimate $\mu_i - \mu_j$ by $\sqrt{2}\sigma\Phi^{-1}\big(F_{i\succ j}\big)$ and the scale value of item i by $s_i = \frac{\sqrt{2}\sigma}{n}\sum_{j\neq i}\Phi^{-1}\big(F_{i\succ j}\big)$. The choice of σ essentially fixes the scale, but the ratio of differences of scale values is not affected by the choice of σ, i.e., any fixed choice of σ would work. A natural but arbitrary choice is $\sigma = 1$.

5.2 Conjoint structure case

Recall that in conjoint analysis we assume that the items come from set A with conjoint structure, i.e., $A = A_1 \times \ldots \times A_n$. Thus we have choice data (from paired comparisons), where the elements in A were compared. The set A is typically fairly large and we do not have enough choice information to apply Thurstone's method directly. Instead we take a decompositional approach and first order the levels in each of the attributes $A_i, i = 1,\dots,n$ on interval scales. For that purpose, whenever in a choice task $(a_1,\dots,a_n) \in A$ was preferred over $(b_1,\dots,b_n) \in A$, we consider this as a_i was preferred over b_i, provided $a_i \neq b_i$. That is, we derive choice information on the attribute level from choice information on the item level. We then can apply Thurstone's method to the choice data on attribute level to get a scale for each attribute. When applying Thurstone's method we set all variances to 1, which fixes the scales for all the attributes but does not necessarily make these scales comparable. For any attribute A_i with levels $a_{i1},\dots,a_{ik_i}$ let $s_{i1},\dots,s_{ik_i}$ be the scale values that we get from applying Thurstone's method on the attribute level. The next step is to aggregate these scales. We make the following assumption that extends the distribution assumption needed for Thurstone's method for a single attribute:

Assumption. For any attribute A_i the scale values $S_{i1},\dots,S_{ik_i}$ are all normally distributed with variance σ_{i1}^2 and expectations drawn from another normal distribution which itself has expectation 0 and variance σ_{i2}^2. That is, we assume that the scale values for the levels of attribute A_i are drawn from a normal distribution N_i with expectation 0 and variance $\sigma_{i1}^2 + \sigma_{i2}^2$ (the convolution of the two normal distribution functions introduced before).

As when applying Thurstone's method the value σ_{i1}^2 is the same for all the S_{ij}, but not necessarily 1. Later it will be chosen such that the scales for all attributes become comparable, i.e., the scaled scale values $\sigma_{i1}s_{ij}$ will be comparable. Now remember that we derived the scales on the attribute level from paired comparisons on the item level. That is, all the distributions $N_i, i = 1,\dots,n$ should describe the distribution of scale values on the item level, i.e., the item scale values should follow the N_i distributions and all these distributions should be the same, i.e.,

$$\sigma_{i1}^2 + \sigma_{i2}^2 = \sigma_{j1}^2 + \sigma_{j2}^2 \equiv 1 \quad \text{for all attributes } A_i \text{ and } A_j,$$

here the value 1 is arbitrary (we just need to choose one fixed value). Note that if we knew the values σ_{i2}, then these equalities would determine the values for the σ_{i1} (that we kept variable so far) and by that make the scales for all the attributes comparable. We can estimate the σ_{i2} from the scaled observed scale values $\sigma_{i1}s_{ij}$ by taking the (biased) estimator of the standard deviation, i.e., by

$$\sqrt{\frac{\sigma_{i1}^2}{k_i-1}\sum_{j=1}^{k_i}s_{ij}^2}, \text{ remember that } \frac{1}{k_i}\sum_{j=1}^{k_i}s_{ij} = 0.$$

Using this estimate we can solve $\sigma_{i1}^2 + \sigma_{i2}^2 = 1$ for σ_{i1} to estimate σ_{i1} as

$$\sigma_{i1} = \frac{1}{\sqrt{1 + \frac{1}{k_i-1}\sum_{j=1}^{k_i}s_{ij}^2}}.$$

That is, in order to make the scales for the different attributes comparable we need to rescale the s_{ij} that we computed with Thurstone's method (with constant variance 1) by this estimate of σ_{i1}.

Since now the scales of all the attributes are comparable we can get the scale value for an item just as the sum of the scale values of the attribute levels involved, i.e., the scale value of $(a_{1j_1},\dots,a_{nj_n}), a_{ij_i} \in A_i$ is given as $\sum_{i=1}^{n}\sigma_{i1}s_{ij_i}$. Note also that on comparable scales each value σ_{i2} can be interpreted as a measure of how important attribute A_i is (contributes larger values to the sum). But we have to be careful, additivity only holds when the attributes are independent. For example

a black foreground and a black background might independently contribute a lot to the perceived quality in visualization, but their combined contribution is negative.

5.3 Error analysis

Analytic error estimate. Let us briefly describe our error analysis. Our observed quantities are the relative frequencies $F_{i\succ j}$. We assume that any comparison of items i and j is an independent Bernoulli trial with success probability p (here "success" means that i is preferred over j). We want to estimate p by $F_{i\succ j}$. For Bernoulli trials, $F_{i\succ j}$ converges to p when the number of trials goes to infinity, but here we make only a finite number m_{ij} of comparisons which procures some error. This error can be estimated by the standard deviation

$$\sqrt{\frac{F_{i\succ j}\left(1-F_{i\succ j}\right)}{m_{ij}}}.$$

To compute errors of our scale values we use error propagation.

Resampling error estimates. We will also simulate errors by randomly dividing the respondents into two groups. For each group we can compute the scale values for all attribute levels (on comparable) scales as described above. So we get for each attribute level a scale value from each group. Averaging the absolute difference of these two scale values over several random groupings of the respondents provides us with an experimental error for the scale value of this attribute level.

Similarly we also compute experimental errors by randomly dividing the paired comparisons into two groups.

5.4 Testing the model

In our model of scale values we made two assumptions, one on the attribute level and one on the item level. The assumption on the attribute level is, that the scale values for all levels of a given attribute are uncorrelated and have the same variance, and the assumption on the item level is, that the scale value of an item is the sum of the scale values of its attribute levels. The latter is essentially the assumption that the attributes are preferentially independent.

Preferential independence (additivity). Here we want to describe how to test the second assumption of our model, i.e., the additivity (or preferential independence) assumption. Let A_1 and A_2 be two attributes and let $C = A_1 \times A_2$ be the new attribute that results from combining A_1 and A_2 and let $c_1,\ldots,c_k$ be its levels. We compute scale values for the levels of C in two different ways. First, for every level $c_i = (a_{i1}, a_{i2})$ with $a_{i1} \in A_1$ and $a_{i2} \in A_2$ we add up the comparable scale values for a_{i1} and a_{i2} that we compute as described before. Let $s_1,\ldots,s_k$ be the resulting scale values. Second, we apply Thurstone's method directly to the combined attribute C and make the resulting scale values comparable with the scales values of all levels of attributes different from A_1 and A_2. This results in scale values $s'_1,\ldots,s'_k$.

If additivity holds, then we expect that $s_i \approx s'_i$. Thus, our hypothesis is that $s_i = s'_i$ for all $1 \le i \le k$. As test statistic we use

$$\chi^2 = \sum_{i=1}^{k} \frac{(s_i - s'_i)^2}{\sigma_i^2 + \sigma_i'^2},$$

where σ_i and σ'_i are computed by error propagation from the errors of the observed frequencies. If the hypothesis is true then the test statistic χ^2 is approximately χ^2-distributed with $k-1$ degrees of freedom. The hypothesis is rejected at a significance level of α if $\chi^2 > \chi^2_{1-\alpha,k-1}$ where $\chi^2_{1-\alpha,k-1}$ is the $1-\alpha$ quantile of the χ^2-distribution with $k-1$ degrees of freedom.

Mosteller's test. On the attribute level we make the assumptions that the scale values are uncorrelated and normally distributed with equal variances. A test for this assumption was devised by Mosteller [11] and is also described in [4]. Here we only briefly review Mosteller's test, which boils down to test if our model can explain the observed frequencies $F_{i\succ j}$. To this end we compute

$$p_{ij} = \frac{1}{2}\int_0^\infty e^{-\frac{(x-(s_i-s_j))^2}{4}}\,dx = \Phi\left(\frac{s_i - s_j}{\sqrt{2}}\right),$$

where we use s_i and s_j as computed by Thurstone's method with $\sigma = 1$. Then we transform both $F_{i\succ j}$ and p_{ij} into angles θ_{ij} and θ'_{ij}, respectively, using the arcsine transformation given by

$$\theta_{ij} = \arcsin\left(2F_{i\succ j} - 1\right) \quad \text{and} \quad \theta'_{ij} = \arcsin\left(2p_{ij} - 1\right)$$

The arcsine transformation converts binomially distributed frequencies into asymptotically normally distributed variables with variance $1/m_{ij}$, where m_{ij} is the number of comparisons of level i with level j for the given attribute. Our hypothesis is that θ_{ij} is normally distributed with expectation θ'_{ij} and variance $1/m_{ij}$ for all $i < j$. As test statistic we use

$$\chi^2 = \sum_{i<j} m_{ij}(\theta_{ij} - \theta'_{ij})^2.$$

If the hypothesis is true then the test statistic χ^2 is approximately χ^2-distributed with $\binom{n-1}{2}$ degrees of freedom. Thus, at level α we have to compare our test statistic to the $1-\alpha$ quantile of the χ^2-distribution with $\binom{n-1}{2}$ degrees of freedom.

6 Data Analysis: Results

In this section we report on how we applied our data analysis method that we described in Section 5 to obtain meaningful scale values for our four conjoint studies. All subsequent results refer to respondents that are more than 10 years old[1] and have passed the Ishihara test for color blindness. Among all respondents fulfilling these two criteria 317 respondents participated in the two studies with test question DETAIL and 366 respondents participated in the other two studies with test question AESTHETICS.

In a first step we computed scale values using the method described in Section 5.2. These scale values need not be meaningful since model assumptions that underlie these computations might not be met in our studies. Hence we discuss in the following how to obtain meaningful scale values from the initially computed ones.

6.1 Testing preferential independence

As pointed out earlier, if the parameters are preferentially independent, then the scale values for different parameters are comparable and we can determine the scale value of an image (rendering for a specific choice of parameter values) by adding up the scale values for the parameter values used to render the image. The top ranked image that we get this way for the FOOT data set and AESTHETICS question does not look like a reasonable first choice, see Figure 1. The reason is not surprising: the parameters COLOR and BACKGROUND are not preferentially independent for this study.

We tested all pairs of parameters on interdependencies for all four studies using the additivity test described in Section 5.4. Table 1 summarizes the result of this test for all combinations of parameters.

Based on the outcome of the additivity test we decided to combine the parameters RENDERING and STEPSIZE into a single new parameter RENDERING-STEPSIZE for all four studies. For the FOOT data set and both questions we also combined the parameters COLORMAP and BACKGROUND into the new parameter COLORMAP-BACKGROUND[2].

[1] We found no significant differences between respondents younger and older than 17, respectively. See also Section 7.

[2] Note that though the top ranked image for the study [FOOT, DETAIL] looks reasonable, see Figure 1, it turns out that we have to combine the two color parameters also for this study.

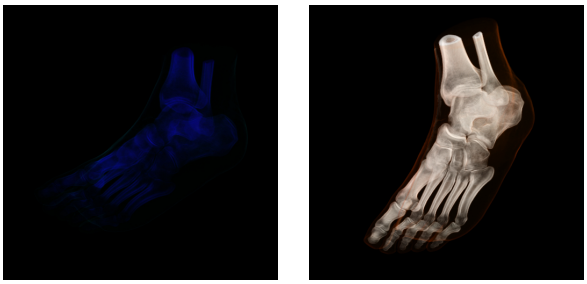

Fig. 1. The images with the highest scale values for the studies [FOOT,AESTHETICS] (left) and [FOOT, DETAILS] (right) before taking care of parameter dependencies.

FOOT		C	R1	V	R2	S	B
Colormap	C	*	4				1
Rendering	R1	2	*	5		3	2
Viewpoint	V			*			
Resolution	R2				*		
StepSize	S		4			*	6
Background	B	1	3				*
ENGINE		C	R1	V	R2	S	B
Colormap	C	*					
Rendering	R1		*			1	
Viewpoint	V			*			
Resolution	R2				*		
StepSize	S		1			*	
Background	B		2				*

Table 1. Test for pairwise preferential independence of the parameters with significance level $\alpha = 0.01$. The shown numbers denote the rank order of relevance (only for significant dependencies), i.e., smaller values indicate more relevant dependencies. The values below the diagonal are for the AESTHETICS question and the values above the diagonal are for the DETAILS question.

That is, we compute new scale values for the combined parameters and use them to replace the scale values for the original parameters. This already gives our final scale values that we summarize in Table 3 and Figure 2. The figure shows the best ten and the worst ten renderings for each of the four studies.

6.2 Mosteller's test

We also tested our model assumptions on the parameter (attribute) level using Mosteller's test, see Section 5.4, on all parameters (including the combined ones). With a few exceptions all parameters passed the test at the $\alpha = 0.01$ significance level. All exceptions concerned the RENDERING parameter. Possible reasons are unequal variances of the distributions of the scale values for different levels, inappropriateness of a one-dimensional scale or an underestimation of the error.

To further investigate the last point, underestimation of the error, we compared the computed sample size error with the two experimental errors described in Section 5.3. All computed theoretical sample size errors are within 15% of the experimental errors, except for the parameter RENDERING which shows an underestimation of up to 40%. This finding also puts the results on the preferential independence tests involving the RENDERING parameter into a new perspective. Some of the detected interdependencies in Table 1 are not significant anymore if the error estimates for RENDERING are adjusted.

7 RESULTS

In this section we discuss the results obtained for our four visualization case studies.

Relative importance of parameters. As we pointed out at the end of Section 5.2 the standard deviation σ_{i2} for attribute A_i can be interpreted as the relative importance of attribute A_i. In our setting the attributes are the parameters of the visualization algorithm. Using the estimated standard deviation we get the rank ordering of the parameters as shown in Table 2. From these results it is safe to conclude that overall the rendering mode (combined parameter RENDERING-STEPSIZE) is the most important parameter. The importance of this parameter is relatively higher for the DETAIL than for the AESTHETICS question. A second important parameter is the color scheme used (or the background), although this finding is not as pronounced. The viewpoint is somewhat important (mostly for the FOOT), while the resolution is somewhat important for the ENGINE. The other parameters are relatively unimportant, at least at the levels we have measured.

Most preferred levels. The results of Tables 2 and 3 as well as Figure 2 reveal a good deal of useful information. We observe that the algorithms XRAY and MIP are not considered useful by our respondents (but note that these were non-expert viewers – doctors can see a lot more in those renderings). The DVRGM algorithm performs (slightly) better than DVR, which performs better than DVRNS. This ranking shows that the more structure enhancement, the better.

There is also a clear preference for achromatic backgrounds. Only blue is also found to be somewhat useful, possibly because blue is a monocular depth cue in that colors very far away shift to the blue spectrum, or because of the background shade of blue and the object. Highly saturated backgrounds are generally disliked. Interestingly, there are also differences between the two achromatic backgrounds: a black background is considered more aesthetic, whereas white seems to show detail better. This is particularly true for the ENGINE which is overall a more complex dataset. It is most likely also an object that is less familiar to the respondents. Therefore they require more detail; higher resolution is also more important (than for the less complex FOOT).

For the ENGINE, the color map applied does not seem to matter as much, but for the DETAIL question, the FOOT (bone) is strongly preferred to be seen in a color resembling that of bright bone (skin grey). This indicates that for object inspection, viewers like to see objects in colors that are most natural and at the same time bright (when such a color is generally agreed on), but for objects less defined in that respect the color choice is a matter of taste (as is the case for the ENGINE), as long as they are bright and define contrast well. In the AESTHETICS category viewers still preferred a natural color (for the FOOT), but the brightness condition was no longer so important (by definition of the task criterion).

An interesting observation can also be made with respect to the viewpoint. A common feature is that viewers prefer to see objects at oblique angles, which generally gives objects a more three-dimensional appearance and also reveals more features (such views are also used for product advertisements). But the engine was in general preferred to be situated as standing on a surface — the views where the engine was rotated at an arbitrary angle (and appeared as it were flying towards the viewer) were rated low. On the other hand, the foot was acceptable at most orientations. We believe that the 'flying' engine was deemed unrealistic, and perhaps even dangerous and therefore unappealing, while a foot is seen commonly at general orientation in real life (just not as a bone).

Dependency on the respondent. We observed that the experimental error, see Section 5.3, was larger when dividing respondents into different sets than when dividing choice tasks into different sets. This indicates that although the respondents answered only 20 choice tasks for each data set, we can already detect a dependency on the individual's preferences, i.e., preferences are not homogeneous over the population.

We also analyzed preferential differences between different subgroups male vs. female and young vs. old, respectively) of our population respondents [3]:

[3] We also collected preference data from 37 persons showing color deficiencies, but the sample size was not sufficient to detect significant differences to

		AESTHETICS		DETAIL	
FOOT	1.	RENDERING-STEPSIZE	(0.31)	RENDERING-STEPSIZE	(0.52)
	2.	COLORMAP-BACKGROUND	(0.3)	COLORMAP-BACKGROUND	(0.35)
	3.	VIEWPOINT	(0.14)	VIEWPOINT	(0.12)
	4.	RESOLUTION	(0.05)	RESOLUTION	(0.08)
ENGINE	1.	RENDERING-STEPSIZE	(0.56)	RENDERING-STEPSIZE	(0.77)
	2.	BACKGROUND	(0.19)	RESOLUTION	(0.09)
	3.	RESOLUTION	(0.12)	VIEWPOINT	(0.08)
	4.	VIEWPOINT	(0.09)	BACKGROUND	(0.05)
	5.	COLORMAP	(0.05)	COLORMAP	(0.01)

Table 2. Rank order of the parameters used in our four studies. the rank order is derived from estimated variances (shown in brackets).

We only found significant differences between male and female respondents for the COLORMAP parameter in the [FOOT,AESTHETICS] study: female respondents mostly prefer BLUECYAN (scale value: 0.07(3))[4], which is also liked by the male respondents (0.07(2)) but not as much as SKINGRAY (0.99(2)), which is the least preferred color of the females (-0.04(3)). Magenta is least preferred by the males (-0.12(2)), whereas females (-0.03(3)) prefer it over SKINGRAY.

In general we found no significant differences between the two age classes 17 years or younger (teenagers) and older than 17 years (adults). We only found two exceptions concerning the AESTHETICS question. For adults the preferences within the RENDERING parameter are more pronounced than for teenagers, though the rank order of the individual levels is the same. On the other hand teenagers tend to have more pronounced preferences concerning the background color, again with basically the same order on the individual levels as for the adults.

Altogether these findings have interesting consequences if one wants to personalize visualization systems: it seems hard to do so based on socio-demographic data (as age and gender) only.

Dependency on the data set. Preferences obtained for the FOOT differ significantly from preferences for the ENGINE dataset. This difference is most pronounced for the combined parameter RENDERING-STEPSIZE[5], which is much more important for the ENGINE dataset for both questions.

Dependency on the question. The observed preferences in the DETAIL studies are significantly different from the preferences in the AESTHETICS studies. The question about detail separates the preferences for different parameter values better. This means that there is more mutual consent in the test population about detail. We believe this is due to the fact that the question about detail is more specific, and less subject to personal taste. The question about details separates the preferences on the ENGINE data set into two distinct preference classes (DVRXX against XRAY/MIP). This separation does not show in the [ENGINE,AESTHETICS] study.

Parameter interdependence. As discussed earlier, our additivity test shows that the independence assumption is not fulfilled for the parameters COLORMAP and BACKGROUND for the FOOT data set. This finding seems very reasonable since similar object and background color certainly should have a negative impact on the perceived image appearance. Furthermore details are better visible if the contrast between foreground and background color is high.

The additivity test also shows that the parameters RENDERING and STEPSIZE are not independent. The observed interdependency is less intuitive than the one between COLORMAP and BACKGROUND, but can be explained also. The scale values for the combined parameter show that the changes in STEPSIZE do not induce the same magnitude of change for the scale values of the different RENDERING levels. In particular for XRAY and MIP levels the changes in STEPSIZE seem to have no or only marginal influence. This can be due to the fact that MIP and XRAY algorithms lack coherency in structure and are mostly used for quick survey modalities, but not for careful diagnosis. Our study indicates that the visual system cannot detect all errors or even inconsistencies, and thus viewers do not become aware of possible errors,

8 DISCUSSION

We too first steps to demonstrate that conjoint analysis can be a useful and efficient tool to gauge influences of a rich set of rendering parameters on human perception in visualization tasks. We believe that the data analysis technique that we have developed here can even be used to analyze data gathered in the first phase of the 'human-in-the-loop' method of House, Bair and Ware [6]. Note that our analysis method only needs paired comparisons between renderings that even can be obtained from measurement of how well a test person performs a task on different renderings.

We have tested the framework within a familiar visualization environment, a parameterized volume renderer, where we have taken great care to reduce the effects of competing adverse parameters, such as image size and occlusion, without reducing the effects of the relevant tested parameters, such as color schemes and rendering precision and algorithm. In this process we verified a few known results, such as the effect of rendering fidelity, but we also teased out some lesser-known but important results, such as preferred object orientations, color schemes, and the relationship of step size and rendering modality. Another interesting finding is that our conjoint analysis method can help to resolve tradeoff decisions. In particular for the DVRGM algorithm it is not necessary to go down to step size 0.2—step size 0.5 even gives perceptually better results. That is, it is often not worthwhile to spend the extra computing time required by smaller step size (time-quality tradeoff). A second tradeoff concerns perceived quality and file size, which is to a large extent determined by the resolution. Our methods allow us to quantify this tradeoff, i.e., to answer the question of how much quality gets sacrificed when the file size (resolution) decreases.

With our careful error analysis we obtained insights beyond gauging of preferences by scale values: we were able to conclude from the computed experimental errors that preferences depend on the individual, which in itself is not so surprising, but we also found that one cannot predict an individual's preferences from the socio-demographic data available to us (age and gender).

In future work we want to investigate limitations of the applicability of conjoint analysis to visualization. Possible concerns are: the large number of respondents needed (though the burden on each respondent is low); need for more systematic ways to estimate the number of required respondents; important parameters may over-shadow the results for not so important ones (rendering statements about the latter dubious); restrictiveness of the distribution assumptions; influence of framing effects or the surrounding in general (we conducted our study in a controlled environment and tried to control for framing effects by alternating questions from two studies).

the rest of the population.

[4]Numbers in parenthesis show the estimated standard deviation in units of the last shown digit.

[5]The parameters VIEWPOINT and COLOR can not be compared directly for the two datasets, because different colors and viewpoints were used as parameter levels.

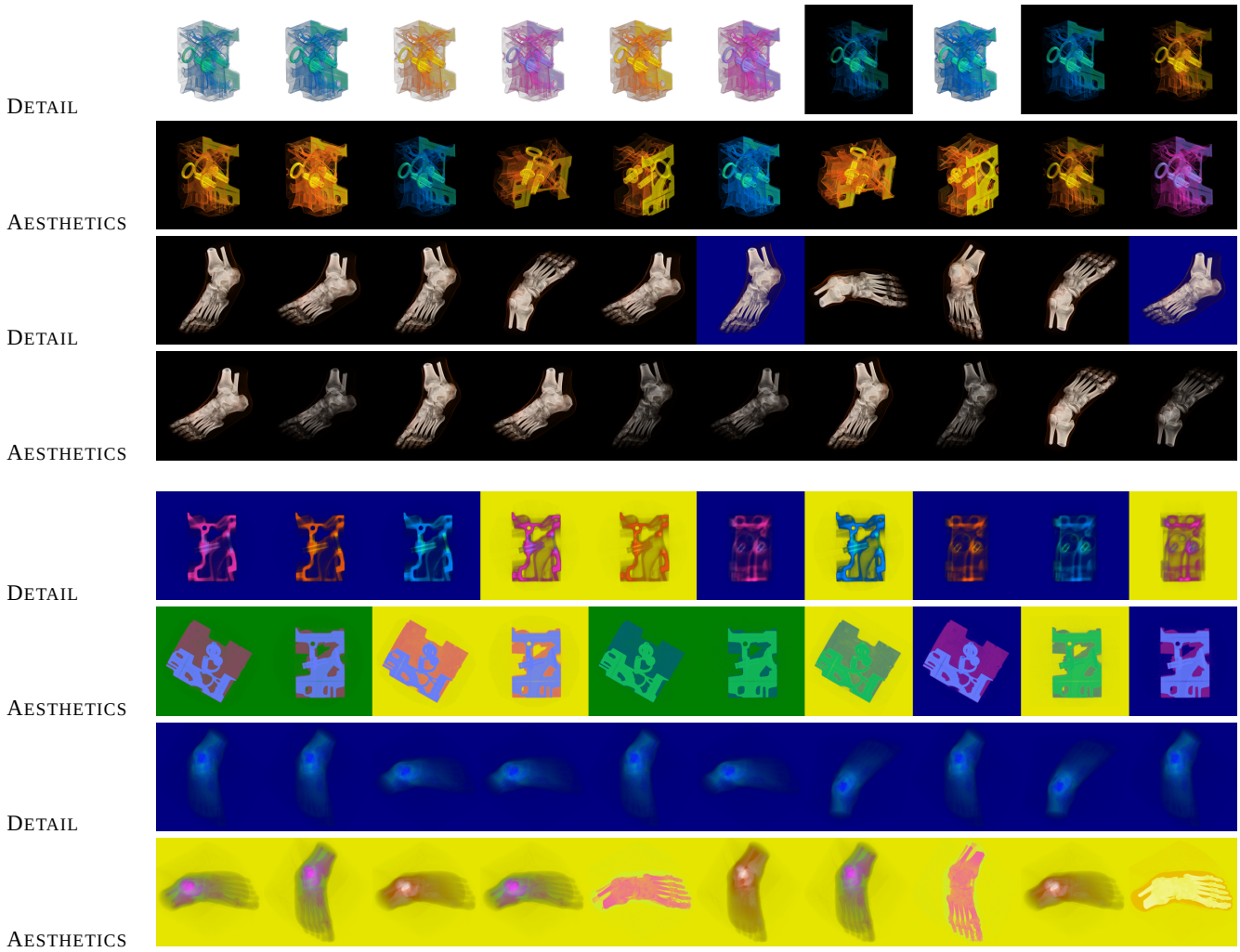

Fig. 2. On top: Best ten renderings (ranking decreasing from left to right). On bottom: Worst ten renderings (ranking increasing from left to right) for our four conjoint studies.

Our vision is to create a (web based) user study analysis suite that can be used by researchers to conduct and analyze multi-parameter user studies. Conjoint analysis should be an integral component of such a suite.

ACKNOWLEDGEMENTS

Joachim Giesen and Eva Schuberth are partially supported by the Swiss National Science Foundation in the project "Robust Algorithms for Conjoint Analysis". Klaus Mueller and Lujin Wang are partially supported by NSF CAREER grant ACI-0093157 and NIH grant 5R21EB004099-02.

REFERENCES

[1] `www.sawtoothsoftware.com`.

[2] S. Bergner, T. Möller, D. Weiskopf, and D. J. Muraki. A spectral analysis of function composition and its implications for sampling in direct volume visualization. *IEEE Transactions Visualization and Computer Graphics*, 12(5):1353–1360, 2006.

[3] S. Bruckner, S. Grimm, A. Kanitsar, and M. Gröller. Illustrative context-preserving exploration of volume data. *IEEE Transactions Visualization and Computer Graphics*, 12(6):1559–1569, 2006.

[4] P. G. Engeldrum. *Psychometric Scaling, A Toolkit for Imaging Systems Development*. Imcotek Press, Winchester MA, USA, 2000.

[5] A. Gustafsson, A. Herrmann, and F. Huber. Conjoint analysis as an instrument of market research practice. *Conjoint Measurement. Methods and Applications. A. Gustafsson, A. Herrmann and F. Huber (editors)*, pages 5–45, 2000.

[6] D. H. House, A. Bair, and C. Ware. An approach to the perceptual optimization of complex visualizations. *IEEE Trans. Vis. Comput. Graph.*, 12(4):509–521, 2006.

[7] S. Ishihara. *Ishihara's tests for colour blindness*. Isshinkai, 1962.

[8] G. Ji and H. Shen. Dynamic view selection for time-varying volumes. *IEEE Transactions Visualization and Computer Graphics*, 12(5):1109–1116, 2006.

[9] Y. Kim and A. Varshney. Saliency-guided enhancement for volume visualization. *IEEE Transactions Visualization and Computer Graphics*, 12(5):925–932, 2006.

[10] J. Kniss, S. Premoze, M. Ikits, A. Lefohn, C. Hansen, and E. Praun. Gaussian transfer functions for multi-field volume visualization. *Proceedings of IEEE Visualization 2003*, pages 497–504, 2003.

[11] F. Mosteller. Remarks on the method of paired comparisons. *Psychometrika*, 16:207, 1951.

[12] P. Shanbhag, P. Rheingans, and M. desJardins. Temporal visualization of planning polygons for efficient partitioning of geo-spatial data. *Proceedings of the 2005 IEEE Symposium on Information Visualization*, page 28, 2005.

[13] S. Takahashi, I. Fujishiro, Y. Takeshima, and T. Nishita. A feature-driven approach to locating optimal viewpoints for volume visualization. *Proceedings of IEEE Visualization 2005*, pages 495–502, 2005.

[14] L. L. Thurstone. A law of comparative judgement. *Psychological Review*, 34:273–286, 1927.

[15] H.-C. Wong, H. Qu, U.-H. Wong, Z. Tang, and K. Mueller. A perceptual framework for comparisons of direct volume rendered images. *IEEE Pacific-Rim Symposium on Image and Video Technology*, 2006.

	Data Set:	Engine		Data Set:	Foot	
Parameter	Param. Value	Question: Aesthetics	Details	Param. Value	Question: Aesthetics	Details
Colormap	MagentaBlue	-0.061(17)	-0.006(18)	SkinGrey	0.039(17)	0.146(18)
	RedYellow	0.065(17)	-0.001(18)	BlueCyan	0.070(17)	-0.079(18)
	BlueGreen	-0.004(17)	0.007(18)	Magenta	-0.109(17)	-0.067(18)
Background	Black	0.378(26)	0.049(28)	Black	0.419(27)	0.246(28)
	White	-0.034(26)	0.078(28)	White	-0.063(26)	0.047(27)
	Green	-0.162(26)	-0.018(28)	Green	-0.105(26)	-0.097(28)
	Blue	-0.063(26)	-0.062(28)	Blue	0.007(26)	-0.103(28)
	Yellow	-0.120(26)	-0.046(28)	Yellow	-0.258(26)	-0.093(28)
Rendering	DVR	0.514(25)	0.719(28)	DVR	0.095(26)	0.361(27)
	DVRNS	0.353(24)	0.530(25)	DVRNS	-0.001(25)	0.058(26)
	DVRGM	0.385(24)	0.629(26)	DVRGM	0.484(26)	0.561(27)
	XRAY	-0.305(23)	-1.005(31)	XRAY	-0.308(26)	-0.758(28)
	MIP	-0.947(28)	-0.872(28)	MIP	-0.270(25)	-0.223(26)
StepSize	0.5	0.028(17)	0.026(18)	0.5	0.035(17)	0.065(18)
	0.2	0.051(17)	0.066(18)	0.2	0.061(17)	0.038(18)
	1.0	-0.078(17)	-0.093(18)	1.0	-0.096(17)	-0.103(18)
Viewpoint	side-front	0.132(30)	0.118(32)	side-60	0.126(26)	0.174(28)
	side-back	0.052(30)	0.071(33)	top-90	-0.158(26)	-0.118(28)
	side-top	0.060(30)	0.027(32)	top-0	-0.133(26)	-0.151(28)
	side-down	-0.120(30)	-0.041(32)	side-30	0.208(27)	0.098(28)
	front	-0.007(30)	-0.073(32)	top-135	-0.044(26)	-0.003(28)
	side	-0.117(29)	-0.101(32)			
Resolution	high	0.115(10)	0.091(11)	high	0.045(10)	0.080(11)
	low	-0.115(10)	-0.091(11)	low	-0.045(10)	-0.080(11)
Rendering -StepSize	DVR, 0.5	0.60(5)	0.81(7)	DVR, 0.5	0.00(5)	0.17(5)
	DVR, 0.2	0.51(5)	0.86(6)	DVR, 0.2	0.29(5)	0.64(5)
	DVR, 1.0	0.41(5)	0.49(5)	DVR, 1.0	0.02(5)	0.26(5)
	DVRNS, 0.5	0.18(4)	0.41(5)	DVRNS, 0.5	0.08(5)	0.17(5)
	DVRNS, 0.2	0.44(4)	0.64(5)	DVRNS, 0.2	-0.01(5)	-0.03(5)
	DVRNS, 1.0	0.40(4)	0.49(5)	DVRNS, 1.0	-0.09(5)	0.02(5)
	DVRGM, 0.5	0.63(5)	0.85(6)	DVRGM, 0.5	0.67(5)	1.07(6)
	DVRGM, 0.2	0.48(5)	0.71(5)	DVRGM, 0.2	0.60(5)	0.64(5)
	DVRGM, 1.0	0.07(4)	0.32(4)	DVRGM, 1.0	0.16(5)	0.04(5)
	XRAY, 0.5	-0.29(4)	-0.95(5)	XRAY, 0.5	-0.31(5)	-0.80(5)
	XRAY, 0.2	-0.32(4)	-1.00(6)	XRAY, 0.2	-0.36(5)	-0.77(5)
	XRAY, 1.0	-0.29(4)	-1.05(8)	XRAY, 1.0	-0.25(5)	-0.73(6)
	MIP, 0.5	-0.89(5)	-0.89(5)	MIP, 0.5	-0.24(5)	-0.26(5)
	MIP, 0.2	-0.93(5)	-0.86(5)	MIP, 0.2	-0.26(5)	-0.23(5)
	MIP, 1.0	-1.03(6)	-0.83(6)	MIP, 1.0	-0.31(5)	-0.19(5)
Colormap -Background	MagBlu-BBlk	0.29(5)	-0.05(5)	SkinGray-BBlk	0.73(5)	0.74(6)
	MagBlu-BWht	-0.12(5)	0.05(5)	SkinGray, Wht	-0.29(5)	-0.30(5)
	MagBlu-BGrn	-0.22(5)	0.07(5)	SkinGray, Grn	-0.11(5)	0.11(5)
	MagBlu-BBlu	-0.10(5)	-0.11(5)	SkinGray, Blu	0.24(5)	0.47(5)
	MagBlu-BYel	-0.17(5)	0.01(5)	SkinGray, Yel	-0.40(5)	-0.24(5)
	RedYel-BBlk	0.44(5)	0.20(5)	BluCya, Blk	0.30(5)	-0.13(5)
	RedYel-BWht	-0.08(5)	0.11(5)	BluCya, Wht	0.26(5)	0.36(5)
	RedYel-BGrn	-0.06(5)	-0.13(5)	BluCya, Grn	0.02(5)	-0.11(5)
	RedYel-BBlu	0.07(5)	-0.01(5)	BluCya, Blu	-0.26(5)	-0.75(6)
	RedYel-BYel	-0.04(5)	-0.17(5)	BluCya, Yel	0.04(5)	0.17(5)
	BluGrn-BBlk	0.40(5)	-0.00(5)	Mag, Blk	0.20(5)	0.16(5)
	BluGrn-BWht	0.10(5)	0.06(5)	Mag, Wht	-0.14(5)	0.08(5)
	BluGrn-BGrn	-0.20(5)	0.01(5)	Mag, Grn	-0.19(5)	-0.29(5)
	BluGrn-BBlu	-0.17(5)	-0.06(5)	Mag, Blu	0.03(5)	-0.07(5)
	BluGrn-BYel	-0.15(5)	0.02(5)	Mag, Yel	-0.42(5)	-0.19(5)

Table 3. Scale values for all parameter levels (also combined ones) of the four conjoint studies (Engine studies on the right and Foot studies on the left). Numbers in parenthesis show the estimated standard deviation in units of the last shown digit.