Web-based psychometric evaluation of image quality

Iris Sprow, Zofia Barańczuk, Tobias Stamm and Peter Zolliker

Swiss Federal Laboratory for Materials Testing and Research (Empa), Dübendorf, Switzerland

ABSTRACT

The measurement of image quality requires the judgement by the human visual system. This paper describes a psycho-visual test technique that uses the internet as a test platform to identify image quality in a more time-effective manner, comparing the visual response data with the results from the same test in a lab-based environment and estimate the usefulness of the internet as a platform for scaling studies.

Keywords: image quality, web-based evaluation, psychophysics

1. INTRODUCTION

An important technique in evaluating image quality in colour reproductions is the evaluation by the human observer through psychophysical assessment. Visual evaluations have been carried out widely in the past for research involving colour reproduction judgements, where viewing conditions are usually based on the constantly revised ISO 36641 viewing condition standard which states guidelines for print-based psycho-visual analysis and later also for monitors.

Psycho-visual studies are traditionally carried out in standardized laboratory environments to ensure repeatability and comparability between studies. However, in-lab tests are time-consuming and a real effort has to be made to include enough observers to obtain statistically relevant results. Although a large number of data adds to the robustness of the results, it also introduces complexity and observer stress. The point is, that while psychophysical assessment results gain in stability as more images are being evaluated, gathering data through lab environment set-ups slows down this process. Collecting data where people can access images easily is desirable.

The context of this work is the need to find more time-effective, cost-saving methods for determining subjective image quality. As society is increasingly connected to the World Wide Web, one alternative is the use of the internet as a platform for a scaling study with easier access to visual response data but with the drawback of little or no control over monitor settings and viewing environments. The trend of assessing visual information online is addressed in the work by Zuffi et. al.¹ for example, as well as by Moroney,² who describes an ongoing web-based approach to name colors. Using the web for collecting data, one is able to collect much more data and therefore gains statistical power. Another recent activity in the area of online testing to the field of image quality is discussed by R. Rasmussen.³ These experiments all together report good accuracy and efficiency with the need of further testing to confirm these conclusions.

Here, we implement the same psychometric test for a calibrated laboratory environment and the non-calibrated internet setting in order to study the potential of web-based psychometric studies. As test study we choose a gamut mapping study, a typical task involving the measuring of image quality. Gamut mapping is the rendering from one device color space to another device color space. For an overview, see Morovic.⁴ Our motivation is to efficiently collect visual response data with the web as a study tool. Therefore, we embed an earlier conducted lab study, see⁵ to realize the current work. The basic task of the observer is to compare an original image to two images mapped to a smaller gamut using parameterized gamut mapping algorithms. The data analysis is performed using conjoint analysis, a common method in market research and recently applied to visualization and image quality studies.⁵⁻⁷

Due to the pair comparison design which includes the original image, changes between images stay relative. This is also a good type of set-up to test to which extent J. von Kries' adaption rule^{8,9} holds. It states that

Further author information: (Send correspondence to I. Sprow)

I. Sprow: E-mail: iris.sprow@empa.ch, Telephone: +41 44 823 4033

the human observer perceives relative changes in different viewing conditions equally, even though the viewing conditions might change.

The paper is organized as follows: In the next section we describe the necessary parts to realize the comparison of a psycho-visual survey between the laboratory and the web environment. Collection of data and the according observer groups are explained. The employed gamut mapping algorithms are described in this section as well. In the third section we show and discuss results on the comparison between web-based to laboratory tests. In the fourth section we show the advantage and potential of a large data collection method (using the web) for a conjoint study of parameterized gamut mapping algorithms.

2. EXPERIMENTAL SET-UP

The same visual study was carried out within the different environments. While the **laboratory** set-up was carried out in a controlled, closely to CIE viewing standard adjusted environment, the **web**-based test was carried out by observers on their own systems. Observers were asked to make their choice to the following instruction:

"Choose the best representation of the original. If you see no difference, click the original".

As a test study, we chose a gamut mapping task. Test images were computed to four different device gamuts and were viewed as simulations on the display. To determine the importance of the different gamut mapping parameters, we opted for a pair comparison method where 3 images were displayed on-screen: the original (sRGB) image on top and the two different rendered images on the bottom.

Observers in the lab study judged 50 comparisons and web-participants 30 (or more) each. For the laboratory tests we used 24" Eizo CG 241W and 22" Eizo CG 220 LCD monitors calibrated to show sRGB correctly with a white point set to 6500 Kelvin. Images were displayed with size 400x600 pixels or 400x400 pixels on our 96 dpi monitors. Monitor flaps around the screen prevented flare. The monitor's background was set to a neutral gray.

For the internet based part we had to consider deviating viewing conditions and displays to those from the laboratory set-up (concerning brightness, size, resolution, white point and color gamut). Therefore, additional information was collected from the web study participants concerning their employed system (ambient illumination, display type and size, internet browser and operating system). To get a feel for the observers' perceived dynamic range and possible color shifts of their display, a few adjustment images had to be evaluated by the observer as well. Any sort of calibration, however, was avoided - the intention was to conduct the study on the actual system from online users. We used JPEG images with very low compression and a maximal width or height of 400 pixels which resulted in about 150 KB per image. The resulting test pages have been verified to be presentable on common operating systems, browsers and even with small laptop displays.

2.1 Observers

Three classes of distinct users participated in the study: about 70 people completed the laboratory controlled environment, about 700 people participated in the web-based test and 43 users completed both studies. The last group serves as a control group to get a reliable link between both test settings.

Test persons for the laboratory environment were recruited from participants of the 2008 Color Management Symposium in Munich who were mostly color experts and participated voluntarily. Each test person had passed the Ishihara test for color deficiency. We had 70 test persons participating in our study resulting in a total of 3500 comparisons.

Observers from the web study were contacted via the ECI mailing list, posting of the test on our work group's web page and soliciting the test to co-workers, friends and family. About 700 people took part in the web test resulting in a total of about 30.000 comparison.

The cross-link study observers (one observer takes part in both environments and is identified with an own user ID) was carried out by 43 students from the Swiss Federal Institute of Technology Zurich.

2.2 Data Sets

In summary, we had the following for data sets:

- (1) A test in a controlled environment close to laboratory conditions was carried out at Fogra's Colour Management Symposium in Munich. Further denoted as *symposium*.
- (2) Test in an uncontrolled environment over the internet. A questionnaire was used to get crude estimations of the observer's viewing environment. The participants each only judged 30 image pairs to avoid 'dropout', early cancelation of the test. However they were allowed to perform more comparisons if they wanted. Further denoted as *internet*.
- (3) Observers from the control group took this test, identified by user ID. Further denoted as *ctrl_lab*.
- (4) Observers from the same control group took the test on a system of their choice using the internet, defined by user ID. Further denoted as *ctrl_web*.

2.3 Implementation

The pair comparison test for the laboratory study was implemented as a C++ Windows application. The program displayed the computed images next to the original image in a pair comparison manner with a gray surround. Visual preference data was recorded as a text file. While the laboratory test was run with one user at a time and on the same systems, we had to consider a different scenario for the web implementation of the study.

Data assessment of the web-based test was realized with the use of XHTML(Extensible HyperText Markup Language) containing PHP (PHP: Hypertext Preprocessor) files which were connected to a MySQL (Standard Query Language) database. To present the images to the user, we implemented a web-based internet site: <http://www.empamedia.ethz.ch>. To cope with the many possible combinations of operating systems, display types, display sizes and browsers, we used XHTML encoding stripped down to the very essential parts necessary for the desired presentation of the images. Parallel access of more than one user at a time is guaranteed as well as a clear arrangement of data storage and handling. The data of a particular user is stored by assigning an ID to each user and then stored in a database.

Information collection of web observers' displays was realized with the use of a questionnaire on the one hand and adjustment images on the other hand. The web-questionnaire provided answers in form of a drop-down menu (see Table 1) from which observers chose the answer that best described their viewing environment. Adjustment images had to be judged visually by the user.

2.4 Questionnaire

Providing a web questionnaire before the test started, gave us the possibility to collect information about the participant's viewing environment. Contrary to the laboratory environment, displays, illuminations and browsers vary when observers take the pair comparison test online. Not each parameter can be determined with a questionnaire but we opted for information that would likely influence image quality preferences made on a display, keeping in mind to not overburden the observer with too many questions, see Table 1.

Different questions were asked in a drop-down menu. Answers are collected and stored directly into the connected MySQL database. Observers had the possibility to not provide answers and take the test regardless. Almost all users have opted to provide answers to the questionnaire however. After submission, a user state is generated, identifying the user by an unique ID and the IP-address to save results avoid data manipulation.

Table 1. Questionnaire

Questions	Drop-down
Gender	Male, Female
Operating System	Windows, Mac, Linux, Different
Browser	Internet Explorer (IE), Firefox, Safari, Opera, Different
Display Type	LCD, Laptop, Beamer, Led, CRT
Screen Size	Up to 16", up to 20", up to 25", Bigger
Surrounding Illumination	Room/Ambient, Office, Window, Living Room, Outdoor, Basement, Different illumination
Possible Color Deficiency	No Deficiency, All Colors, Red-Green, Blue-Yellow

2.5 Adjustment Images

Adjustment images were created and used in this experiment to collect information about the user's display. Information about the user's display is gathered according to the observer's visual preference of how an adjustment image, i.e. gray-scales looked best to the user on his/her monitor. No adjustments to the displays themselves were made, only adjustments which reflect the observer's visual preferences of the images were recorded.

The goal of this study was *not* to calibrate the user's monitors but rather to find out, to which extent the settings throughout the employed systems affect observer's preference choices.

One has to keep in mind that observer stress should be kept low to avoid an early dropout from the user and therefore, only limited gathering of user display information is possible. Still, some information of the user display was necessary to better understand current display settings that are employed in the web environment. Considering this restraint, the thought was to collect information only about two parameters which are important for display quality: the perceived brightness range and a possible hue shift of the observer's device.

2.5.1 White Point, Black Point

Two adjustment images to collect brightness information, from the visually perceived black point and perceived white point were created and displayed to the observer, see Figure 1. These gray-scale images show a brightness range of 5-255, defined in *RGB* values with a change of 5 *RGB* steps between each patch. The assumption is, that if the displays are set to a good contrast, the user should be able to differentiate between the lightest patch or darkest patch in the grayscale, compared to the background white/black, respectively. The user is prompted to select the darkest/brightest patch which is distinguishable from the white/black background. This choice is recorded to the database with an assigned user ID.

Figure 1. Gray patches to determine perceived white/black point

No statement about the gamma curve was made from these adjustment images but rather about the perceived brightness range. The gamma-curve has an impact on the contrast. The hypothesis for these gray-scale adjustment images is, that *if* the γ =standard, then differences between darkest and brightest patches compared to the background should be visually noticeable for the user.

2.5.2 Hue Shift

Repetitions of the hue shift image determined the visually perceived hue shift of the display. This image consists of 7 neutral patches. The middle patch denotes the correct neutral patch (128, 128, 128) while the surrounding ones are neutral shifted patches. The first repetition shows the correct neutral patch in the middle while the surrounding patches show hue shifts of 12 steps in *RGB* away from the middle patch. The patch which is selected as neutral by the observer is taken into the second repetition and again, placed into the middle (according to which patch was chosen).

The surrounding hue shifts get smaller in *RGB* steps as the cycle goes on: 12 steps in the first cycle, 8 steps in the next and 4 steps in the last cycle. Only if the observer chooses the same patch to be neutral as in his neutral choice from the preceding cycle, the next cycle with less hue variations is displayed. When the user selected his perceived neutral option to be the same in the three cycles, will this patch be recorded as *perceived neutral* and the hue shift adjustment is over. Figure 2 supports this idea visually:

Figure 2. Repetitions: left: 16 step changes in *RGB*, middle: 8 steps, right: 4 steps

2.6 Test Images

For the laboratory test (*symposium*) the images were presented with a resolution between 400×400 pixels and 400×600 pixels on the 96 dpi screen which resulted in 8.5×12.5 cm to 8.5×8.5 cm sized images on screen. The images of the internet test (*internet*, *ctrl_web* and *ctrl_lab*) had a resolution of a maximal width or height of 400 pixels. Its size on the display varied depending on the user's monitor size and resolution.

The image set included the obligatory "Ski" image that is specified by the CIE guidelines. To average out a judgment bias due to image content, a wide range of scenes, including 97 different images, was used in the experiment. In addition to a set of ISO test images, they were taken from royalty free libraries as well as from private stock.

Each image had 1536 possible mapping combinations. In a pre-selection, the pairs were randomly selected. About 30.000 pairs were used in the study. The test images are exactly the same than used in Baranczuk.⁵ The five addressed parameters described in the following arise from a parameterized gamut mapping scenario. The parameters include mapping settings, as well as destination gamut size and shape. The basic gamut mapping algorithm is quite simple, it maps any color point in the source gamut along a line segment connecting the color point and a focal point into the destination gamut. Note that we always used sRGB as source gamut, i.e. we did not consider the source gamut as a parameter. Gamut mapping parameters with according levels were:

Compression. The compression parameter describes how the mapping along the line segment is done. We studied four different strategies: linear compression (lin), clipping (clip) and two sigmoidal compression algorithms (sig1 and sig2).⁵

Detail enhancement. We use the detail enhancement method described by Zolliker et.al.¹⁰ with different levels of enhancement, namely 0.5, 1, and 1.5, respectively. The levels of the detail enhancement parameter are E1, E2, E3, and N which means that no detail enhancement was applied.

Gamut size. To gauge the importance of the destination gamut we also tested a parameter that describes the size of the destination gamut. We tested four different destination gamuts. The smallest was ISONewspaper, the largest ISOCOated. The two remaining gamuts were created from ISONewspaper and ISOCOated gamut, as weighted average with weights $\frac{2}{3}$ and $\frac{1}{3}$, and $\frac{1}{3}$ and $\frac{2}{3}$, respectively. We refer to the levels of the gamut parameter as sml, med1, med2, and lrg, respectively.

Gamut shifts. This parameter describes a shift of the destination gamut. We considered six levels with shifts along the gray axis by vectors (5,0,0) and (-5,0,0) and shifted in the color plane by (0,3,0), (0,-1.5,1.5 $\sqrt{3}$), (0,-1.5,-1.5 $\sqrt{3}$). We refer to the levels of this parameter in the order that we have described them as G-D, G-D+, G-D-, G-a+, G-a-b+, and G-a-b-, respectively.

Working color space and Gamut rotations. In the last parameter we test two features: One is to use two different working color spaces CIE-LAB¹¹ and IPT.¹² CIE-LAB is known to preserve hue not accurately, especially in the blue region, whereas IPT was designed to preserve hue. The other is to add specific rotations to the destination gamut. Thus the first two levels of this parameter describe the working color space used for the gamut mapping algorithm (called LAB, IPT). Two additional levels describe rotations of the destination gamut in the IPT in IPT space by -0.1, 0, +0.1 radians. We refer to the levels of this parameter as G-LAB,G-IPT-, G-IPT, G-IPT+.

3. RESULTS

Thurstone's method. In this paper we investigate user preferences in several scenarios. A well-established method for this purpose in the context of paired-comparison tests is Thurstone's method of Comparative Judgement, see Engeldrum¹³ for details and notations not explicitly declared here. In principle, we would like to apply Thurstone's method directly but the huge number of involved comparisons for our total of 1536 mapping combinations makes this impractical. However, after a concentration of the captured data we can apply Thurstone's method to a single parameter. The influence of all other parameters are implicitly accounted for in the Gaussian distribution describing the discriminial differences. As far as needed for this paper we shortly recapitulate the basics from Baranczuk.⁵

First of all, every image in the test corresponds to a unique parameter vector $a = (a_1, \dots, a_5)$ where the i -th parameter has the discrete levels $k_{i,1}, k_{i,2}, \dots, k_{i,r_i}$, $4 \leq r_i \leq 6$. For each image only a subset of all possible parameter variants are present in the data set. Following the Case V Solution of Thurstone's method we collect a *frequency matrix* F_i , $1 \leq i \leq 5$, for every parameter. Whenever an image with parameter vector $a = (\dots, k_{i,j}, \dots)$ was preferred over an image with parameter vector $b = (\dots, k_{i,h}, \dots)$ then the entry $F_i[k_{i,j}][k_{i,h}]$ was increased by 1. Please note that one comparison contributes to all matrices F_i , $1 \leq i \leq 5$. Comparisons with a tie as result are ignored. Then we derive *proportion matrices* $P_1, \dots, P_5$ from $F_1, \dots, F_5$ and finally compute a scaling value $S_i(k_{i,j})$, $1 \leq i \leq 5$, $1 \leq j \leq r_i$, for every level of every parameter level. In Figure 3 we show the scale values for the four considered data sets (Internet, Symposium, Ctrl-Web, Ctrl-Lab), depicted side by side.

Comparison of data sets. The observation is that differences between the four scenarios are small, see Figure 3, which leads us to the central question of this section: *Are these small observed differences statistically significant?* To answer this question we will apply a χ^2 -test to the entries of the proportions matrices $P_1, \dots, P_5$. On the assumption that any comparison between the images corresponding to $(\dots, k_{i,j}, \dots)$ and $(\dots, k_{i,h}, \dots)$ can be considered as an independent Bernoulli trial the entries in the proportion matrices are binomial distributed. Next, we apply the arcsin-transformation to the proportions which is also the basis of Mosteller's test¹⁴ for the preconditions of Thurstone Case V. The arcsin-transformation changes the proportions into normal deviates of known variance, a precondition to set up a χ^2 -test. As we compare data sets with different numbers of observers, we have to adjust the variance. The difference of normal distributions with variances $\frac{1}{j}$ and $\frac{1}{j'}$ respectively is a

Figure 3. Scale values for all parameter levels: Laboratory test and Internet test. The error bars show estimated standard errors.

normal distribution with variance $\frac{1}{j} + \frac{1}{j'}$. Finally, we get our sample for the χ^2 -test of goodness of fit as

$$\sum_{i=1}^5 \sum_{j < l} \left(\frac{J_{jl} \cdot J'_{jl}}{J_{jl} + J'_{jl}} \right) \cdot (\arcsin(2p_{jl} - 1) - \arcsin(2p'_{jl} - 1))^2$$

With a χ^2 test we can now check the hypothesis that the results of the studies cannot be differentiated on the basis of our data. The results are summarized in Table 2.

$\chi^2_{\alpha=0.1} = 47.2, \chi^2_{\alpha=0.01} = 58.6$ degree of freedom 36	χ^2
<i>ctrl_lab</i> - <i>ctrl_web</i>	48.2
<i>ctrl_lab</i> - <i>symposium</i>	32.5
<i>ctrl_lab</i> - <i>internet</i>	26.0
<i>ctrl_web</i> - <i>symposium</i>	51.4
<i>ctrl_web</i> - <i>internet</i>	32.4
<i>symposium</i> - <i>internet</i>	49.4

Table 2. χ^2 -test for comparison of test sets.

The comparison of visual preference data between the controlled laboratory and uncontrolled web scenario shows no significant differences as can be seen visually in the results when using Thurstone's method in Figure 3 and in the according χ^2 -test in Table 2 where all χ^2 values are smaller then the limit for $\chi^2 = 0.01$.

Homogeneity of web data. The influence of observers' individual web setups was tested against several factors based on the internet questionnaire: Screen Size, Display Type, Illumination, Gender, Operating System and Browser. The small number of color deficient people does not allow to draw a conclusion concerning a difference between color deficient vision and normal vision. To be on the safe side we did not include data of observers known to have a color deficiency. Likewise only data from observers with a effective brightness range (White-Black) of at least 200 were included in the evaluations.

χ^2 tests were carried out for the factors determined from the questionnaire. Only categories with data from at least 50 observers were used in comparisons. Display sizes were grouped into a small and a large category with boundary set at 16". The results of the χ^2 -Comparison Tests are summarized in Table 3.

$\chi^2_{\alpha=0.1} = 47.2$, $\chi^2_{\alpha=0.01} = 58.6$ degree of freedom: 36	χ^2		χ^2		χ^2
Browser		Display Type		Illumination	
Firefox - IE	39.8	LCD - CRT	54.8	Office - Window	38.5
Firefox - Safari	39.2	LCD - Laptop	47.6	Living - Window	34.4
Safari - IE	42.7	CRT - Laptop	68.6	Living - Office	36.5
Screen Size		Operating System		Gender	
Large - Small	47.1	Mac - Windows	51.0	Male -Female	35.0

Table 3. χ^2 -Comparison Test for categories of web setup parameters

In general, no significant difference could be found between categories of the investigated setup parameters. Only the display type shows a significant deviation for the CRT versus Laptop comparison. An inspection of the resulting scale values for the different display types shows that the CRT data generally yield higher scale values than laptop data, and the scale data for LCD displays lie in between. The relative scale values however are very similar, thus we still can include data of all display types in our data evaluation. Laptops generally have lower quality displays, thus it is not surprising to see such a tendency. It is rather remarkable, that the differences among display type is not larger.

We can conclude, that collecting data using the internet does not require to have much more observers than with controlled conditions to get equally significant results. Our test results show how well von Kries adaptation rule holds for viewing conditions as different as the variety of typical web user environments. However, one has to keep in mind that most differences between images in our test were rather large and easily detectable even for non-experts. Also, test images were always compared to the original image shown on the same display. Would we test mainly small differences between the test images such a test might reveal significant differences for web results compared to lab results as well as between experts and non-experts.

4. APPLICATION OF CONJOINT ANALYSIS TO DATA

In this part we summarize the results from the conjoint analysis on combining all four data sets (called *combined*) compared to the results of the *symposium* data set. For details on the underlying evaluation method see also the papers by Baranczuk et.al.⁵ and Giesen et.al.⁶ In the previous section of this paper, scale values for every parameter were calculated using Thurstone's method (see Figure 3). This is also the first step in our conjoint analysis method. In order to make different parameters comparable, we have to re-scale the scale values $s_i(k_{i,1}), \dots, s_i(k_{i,r_i})$ in order to use them as part-worths in a conjoint structure.⁵ The re-scale factor σ_1 is calculated as follows:

$$\sigma_1 = \frac{1}{\sqrt{1 + \frac{\sum_{j=1}^{n_j} \sum_{l=1}^{n_l} \Delta S_{jl}^2 F_{j>l}}{2 \sum_{j=1}^{n_j} \sum_{l=1}^{n_l} F_{j>l}}}}$$

where ΔS_{jl} are the elements of the z-Score-Matrix and $F_{j>l}$ the frequency of preferring stimulus j to stimulus l . The normalized quality values of the parameter levels are our part-worths that we assume to contribute linearly to the quality of a stimulus, i.e., the quality value of a stimulus $(a_1, \dots, a_5) \in A$ is $\sum_{i=1}^5 \sigma_{i1} s_i(k_{i,j})$ which is the

sum of the part-worths of the parameter levels present in the stimulus. The resulting part worth values are shown in Figure 4.

Figure 4. Part-worth values from the conjoint analysis based on the *combined* data set. The error bars show estimated standard errors.

In order to investigate the validity of our model we tested the following three assumptions which underlie our computations and error estimations:

- (1) The scale values for all levels of a given parameter are uncorrelated, normally distributed and have the same variance. This is tested with the Mosteller test.
- (2) The judgment of all pair comparisons are uncorrelated with respect to individual observers and individual images. To test this, we compare theoretical error with experimental errors.
- (3) The quality value for a parameterized stimulus—in our case an image—can be obtained by summing up the part-worths of the parameter levels present in the stimulus. This essentially means that the part-worths for the different parameters are uncorrelated. Here we use the Linearity Test.

(1) **Mosteller’s test.** A description of Mosteller’s test can be found in Engeldrums book.¹³ Results are presented in Table 4. For the *symposium* data set all test data are well within the limit. However for the

Mosteller	χ^2 <i>symposium</i>	χ^2 <i>combined</i>	$\chi^2, \alpha = 0.1$	$\chi^2, \alpha = 0.01$
Gamut Rotation	2.4	2.3	6.3	11.2
Gamut Shift	11.2	9.2	16.0	23.0
Compression	7.7	20.7	6.3	11.2
Detail Enhancement	2.8	0.3	6.3	11.2

Table 4. Mosteller’s test for parameters compared to χ^2 with significance level $\alpha = 0.1$ and $\alpha = 0.01$ Note: performing a Mosteller test for the gamut size was not possible due to the fact, that only partial data was collected, omitting large differences in gamut size.

combined data the test fails for the 'Compression' parameter. Closer inspection of the matrix entries show, that the entries comparing the extreme positions 'linear' and 'clip' is underestimated by the model. This suggests, that a distribution with a wider tail may be more appropriate than a normal distribution.

(2) Error estimation. In addition to the theoretical error estimations E_{th} we will 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 single comparisons provides us with an experimental error $E_{exp,Single}$ for the scale values. Similarly we also can compute experimental errors by randomly dividing the images into two groups (image based error $E_{exp,Img}$) or by randomly dividing the paired comparisons into two groups (observer based error $E_{exp,Obs}$). In the following Table 5 we compare these error estimations for the *symposium* and the *combined* data sets.

	Avg. Error <i>symposium</i>	Relative Error <i>symposium</i>	Avg. Error <i>combined</i>	Relative Error <i>combined</i>
E_{th}	0.0479	1	0.0150	1
$E_{exp,Single}$	0.0482	1.005	0.0148	0.99
$E_{exp,Obs}$	0.0511	1.06	0.0158	1.05
$E_{exp,Img}$	0.0542	1.13	0.0250	1.67

Table 5. Average error estimations for theoretical and experimintal errors for the *symposium* and *combined* data sets. The relative errors are with respect to the corresponding theoretical error.

The error estimation based on the *symposium* data set are quite similar. Only the image based error is slightly higher. For the *combined* data set however, the image based error is clearly larger than the other error estimations. The observer based error is slightly larger. On this evidence it is worth investigating the difference of individual image preferences, rather than the difference among individual observers.

(3) Linearity Test. We tested the linear model assumption by testing preferential independence of parameters. Let A_1 and A_2 be two parameters, let $C = A_1 \times A_2$ be their Cartesian product, and let $c_1, \dots, c_k$ be the levels of C . We compute part-worths for $c_1, \dots, c_k$ in two ways. First, we use part-worths for the single parameters and add the scale values for a_{i1} and a_{i2} , getting results $s_1, \dots, s_k$. Second, we apply Thurstone's method directly to the combined parameter C and re-scale the scale values as described above. We denote the results as $s'_1, \dots, s'_k$. If the parameters are additive (preferentially independent), then we should get $s_i \approx s'_i$. We test a null hypothesis that $s_i = s'_i$ for all $i = 0, \dots, k$ by using a χ^2 test with the following test statistic:

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

where σ_i and σ_i' are error estimations computed by linear regression from observed frequencies. The null hypothesis postulates that the test statistic is approximately χ^2 distributed with $k - 1$ levels of freedom.

The correlations among all possible attribute pairs are shown in Table 6. For most parameter pairs the linearity test supports the assumption, that part worths add in a linear manner. We have a few exceptions marked in bold: For the *symposium* data set only the combination 'Gamut Size - Gamut Shift' shows a small deviation from linearity. The *combined* data set shows a pronounced deviation from linearity for 'Gamut Size - Gamut Shift' and 'Compression - Details' and small deviation from linearity for the combination 'Gamut Size - Compression'. Of special interest are the results on 'Compression - Details'. The type of compression used for the gamut mapping likely had an influence on the quality gain achievable using detail enhancement.

From the linearity test we conclude that combining 'Compression - Details' and 'GamutSize - GamutRotation' should lead to a better quality scale in particular if enough data is available, such as for the *combined* data set. This increases the number of free parameters from 17 to 41 which is justifiable in view of the large data set.

	χ^2 <i>symposium</i>	χ^2 <i>combined</i>	$\chi^2_{\alpha=0.1}$	$\chi^2_{\alpha=0.01}$
Gamut Size - Compression	16.3	50.9	22.3	30.6
Gamut Size - Details	14.8	26.6	22.3	30.6
Gamut Size - Gamut Shift	50.3	83.9	32.0	41.6
Gamut Size - Gamut Rotation	14.1	11.9	22.3	30.6
Compression - Details	9.4	110.5	22.3	30.6
Compression - Gamut Shift	24.0	21.5	32.0	41.6
Compression - Gamut Rotation	9.3	18.5	22.3	30.6
Details - Gamut Shift	6.7	14.7	32.0	41.6
Details - Gamut Rotation	4.1	10.6	22.3	30.6
Gamut Shift - Gamut Rotation	15.6	16.7	32.0	41.6

Table 6. Linearity Test for every parameter pair for the *symposium* and *combined* data set. Figures in bold show χ^2 values above significance level of $\alpha = 0.01$

5. CONCLUSION

This study reveals close agreement between observers' preferences from the lab and the web. Surprisingly, no significant difference of preference scales between test conditions was found among the importance of the different parameters, such as compression type, local contrast preservation and gamut size. Also the wide range of monitor and viewing conditions used in the internet study had only little influence on the scaling results. For this type of study the adaptation of the human visual system was a key factor, well in line with J. von Kries' adaption rule.

The question arises, to what extend scaling studies need standardized viewing conditions. We have shown that web-based studies have a good potential to be used for psychometric scaling. This study accounts for one specific task using pair comparison relative to an original image shown on the same device. The use of a web-based test for tasks using other scaling methods such as absolute pair comparisons, rank order or magnitude estimation still needs to be investigated. The combination of a web-test with a laboratory test as used in this study allows to have access to a large number of observers via the internet, but still being able to confirm the results based on the evaluation of test data collected with well known viewing conditions.

First results from the conjoint analysis show that the large data set allows to gain more insight on the image quality aspects governing the psycho-visual tests. In particular interactions between the relevant gamut mapping parameters can be studied and evidence shows that the available data has the potential to study the influence of individual images on the scale values.

ACKNOWLEDGMENTS

Part of this research was supported by the Hasler Stiftung Proj. # 2200. We thank Phil Green (London School of Communication), Klaus Simon (Empa) and Joachim Giesen (University of Jena) for valuable discussions and ideas for the realization of this project. Furthermore we would like to thank everyone participating in our study.

REFERENCES

- [1] Zuffi, S. and Brambilla, C. and Eschbach, R. and Rizzi, A., "Controlled and Uncontrolled Viewing Condition in the Evaluation of Prints," *Color Imaging XIII: Processing, Hardcopy, and Applications* **6807**, 680714–1–7, SPIE/IS&T (2008).
- [2] Moroney, N., "Unconstrained web-based color naming experiment," *Color Imaging: Device Independent Color, Color Harcopy, and Graphic Arts VIII* **5008**, 36–46 (2003).
- [3] Rasmussen, D., "Online image quality surveys based on response time," in [*Image Quality and system performance*], 68080K–1–12, SPIE/IS&T (2008).
- [4] Morovic, J., [*Colour Gamut Mapping*], WileyBlackwell, ISBN 0470030321 (2008).
- [5] Baranczuk, Z., Sprow, I., Zolliker, P., and Giesen, J., "Conjoint analysis of parameterized gamut mapping algorithms," *16th Color Imaging Conference*, to be published, IS&T/SID, Portland, OR (2008).

- [6] Giesen, J., Mueller, K., Schuberth, E., Wang, L., and Zolliker, P., "Conjoint analysis to measure the perceived quality in volume rendering," *IEEE Transactions on Visualization and Computer Graphics* **13**, 1664–1671 (November 2007).
- [7] Wang, L., G. J., McDonnell K. T., Zolliker, P., and Mueller, K., "Color design for illustrative visualization," *IEEE Transactions on Visualization and Computer Graphics* (November 2008).
- [8] von Kries, J., "Chromatic Adaptation," *Festschrift der Albrecht-Ludwig-Universitaet* (1902).
- [9] Worthey, J. A. and Brill, M. H., "Heuristic Analysis of von Kries color constancy," *J. Opt. Soc. Am. A* **10**, 1709–1712 (1986).
- [10] Zolliker, P. and Simon, K., "Retaining local image information in gamut mapping algorithms," *IEEE Transactions on Image Processing* **16**, 664–672 (March 2007).
- [11] Central Bureau of the CIE, Vienna, *CIE Publication 116: Industrial color difference evaluation* (1995).
- [12] Ebner, F. and Fairchild, M. D., "Development and Testing of a Color Space (IPT) with Improved Hue Uniformity.," *Proceedings of 6th IS&T/SID Color Imaging Conference, Scottsdale AZ*, 8–13 (1998).
- [13] Engeldrum, P. G., [*Psychometric Scaling, A Toolkit for Imaging Systems Development*], Imcotek Press, Winchester MA, USA (2000).
- [14] Mosteller, F., "Remarks on the method of paired comparisons: III. A test of significance for paired comparisons when equal standard deviations and equal correlations are assumed," *Psychometrika* **16**, 203 (1951).