Error estimation of paired comparison tests for Thurstone's Case V

Peter Zolliker, Zofia Barańczuk, Laboratory for Media Technology; Swiss Federal Laboratory for Materials Testing and Research (EMPA), Duebendorf

Abstract

Pair comparison methods based on Case V of Thurstone's Law of Comparative Judgment are widely used to derive interval scales for perceptual image quality. A thorough treatment of the involved statistical errors is often neglected, even though this is the base for computing confidence intervals and other statistical tests. In this paper we show, that consequent error estimation through all steps of the data analysis provides a simple and reliable method to compute confidence intervals. Monte Carlo simulations are used to verify the results and to compare the proposed error estimation with other known methods.

Keywords: paired comparison, Thurstone's Law of Comparative Judgment, error estimation

Introduction

Psycho-visual tests are used for subjectively evaluating the quality of images. A typical example is the measurement of the quality of images transformed by gamut mapping algorithms. There are a variety of methods to carry out such tests. The aim of the tests is to compare images with respect to perceived quality. Often an interval scale for the different variants of an image is computed from these comparisons, where a scale value is a measure for the relative quality of a specific image variant. For one important application, namely gamut mapping, CIE guidelines for algorithms [1] provide specific experimental methods, viewing conditions, and reference algorithms. Three kinds of psycho-visual tests are recommended for evaluating the quality of gamut mapping algorithms: pair comparison, rank order, and category judgment. The most widely used method is pair comparison. It is also the easiest for the observers, especially if the differences between the images are small.

A pair of images from a set A is presented to an observer. He or she is then asked to choose the one that better fulfills instructions of the test. In the gamut mapping case, the instructions usually state that one should choose the more aesthetic image, or the image more similar to the original. In the latter case the original image is shown along with the transformed images.

A thorough treatment of the involved statistical errors is often neglected, even though this is the base for computing confidence intervals. Morovic [2] gives a simple formula to estimate confidence intervals. The formula depends only on the number of observations N per pair of stimuli but not on the number of stimuli n. Since then this method seems to have been used in many psycho-visual studies on gamut mapping. Only a few of them cite the used formula explicitly [3]. The CIE-guidelines [1] give a reference to Morovic's thesis concerning confidence intervals of paired comparison.

In a recent paper Montag [4] has investigated the problem of missing dependency on the number of stimuli n using Monte Carlo simulations. He derived an empirical formula showing approximate dependency of the estimated error with the square root of the product of N as well as and n. Newer gamut mapping studies have used this formula for error estimation [5]. Older standard books on psycho-visual scaling [6, 7] give more detailed description for error estimations, however their results are tied to their specific data analysis and a direct application to the standard evaluation of the Thurstone's Case V is not obvious.

In this paper we will give a direct derivation of error estimation for Thurstone's Case V. It is based on error propagation. We will use Monte Carlo simulations to compare the results with other commonly used methods and to find the region of applicability as a function of the number of observations N the number of stimuli n and scale value range.

Methodology

Thurstone's Law of Comparative Judgment

Thurstone's Law of Comparative Judgment is a method used for evaluating data obtained in a pair comparison test [8]. It falls into the class of discrete choice models. We consider only Case V, i.e. that the discriminal differences follow a Gaussian distribution of equal width and that there is no correlation between two stimuli. For our methodology we leave it open, whether a pair comparison test was made by one or several observers. We also ignore whether a single image or several images were used in the test. However we make the rather strong assumption, that all judgments are independent of a specific observer and image. Furthermore we assume forced choice for the pair comparison test to avoid treatment of tie choices.

Given a set A of n stimuli, e.g. gamut mapping algorithms, and choice data of the form $i \succeq j$ with $i, j \in A$. We know the frequency f_{ij} (F-matrix) of stimulus i being preferred over stimulus j (number of times i is preferred over j). We consider the proportion q_{ij} (Q-matrix) of stimulus i being preferred over stimulus j defined by

$$q_{ij} = \frac{f_{ij} + \delta}{f_{ij} + f_{ji} + 2\delta} \tag{1}$$

as an indirect measure for the distance of the "qualities" (named scale values) v_i of i and v_j of j, respectively. We introduced the bias correction δ in order to eliminate numerical problems for pairs of items, which have zero entries in the frequency matrix. Except where stated we used in this paper $\delta = 0.2$. For a discussion of different bias correction formulae see also Engeldrum [9, chapter 9.4].

Discrete choice models build on the assumption that the observed choices are outcomes of random trials: confronted with the two options $i, j \in A$ an observer assigns quality values $u_i = v_i + \varepsilon_i$ and $u_j = v_j + \varepsilon_j$, respectively, to the stimuli, where the error terms ε_i and ε_j are drawn independently from the same distribution. The observer then prefers the stimulus with larger quality value. Hence the probability p_{ij} that i is preferred over j is given as

$$\begin{aligned} p_{ij} &= Pr[u_i \geq u_j] \\ &= Pr[v_i + \varepsilon_i \geq v_j + \varepsilon_j] = Pr[v_i - v_j \geq \varepsilon_j - \varepsilon_i]. \end{aligned}$$

In Thurstone's model [8] the error terms ε_i are drawn from a normal distribution $N(0,\sigma^2)$. Thurstone's Case V model assumes that the variances for all stimuli are equal. The difference $\varepsilon_j - \varepsilon_i$ is also normally distributed with expectation 0 and variance $2\sigma^2$ and thus

$$\begin{aligned} p_{ij} &= Pr[u_i \geq u_j] = Pr[v_i - v_j \geq \varepsilon_j - \varepsilon_i] \\ &= \Phi\left(\frac{v_i - v_j}{\sqrt{2}\sigma}\right), \end{aligned}$$

where Φ is the cumulative distribution function of the standard normal distribution

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

This is equivalent to

$$v_i - v_j = \sqrt{2}\sigma\Phi^{-1}(p_{ij}). \tag{2}$$

Let us notice, that q_{ij} is an empirical approximation of p_{ij}. Using the proportion q_{ij} of i being preferred over j we set

$$z_{ij} = \Phi^{-1}(q_{ij}) \quad [Z-\text{matrix}]. \tag{3}$$

Note, that the Z-matrix is antisymmetric, thus $z_{ij} = -z_{ji}$. Mosteller [10] has already mentioned that the least squares solution of the system of equations for the scale values v_i can be determined by averaging the columns

$$v_i = \frac{1}{n}\sqrt{2}\sigma \sum_j z_{ij}. \tag{4}$$

In order to fix the arbitrary offset of the scale values, the sum of all scale values v_i is assumed to be zero.

Error estimation.

In order to gauge the statistical significance of differences between scale values as well as for statistical tests a good error estimation is needed. In this paper consider the estimated standard deviation of a value as the estimated error of that value. These estimated errors serve as a basis for the calculation of confidence intervals and other statistical tests such as χ^2-tests. Typical applications are model verification using Mosteller's Test or the testing whether scale values from different data sets (e.g. expert versus general observers) are statistically indistinguishable. There are a few known methods for estimating errors for the Thurstone Case V, which will be described in the following.

Morovic's error estimation. Morovic [2, chapter 4] gives the following formula to estimate the 95 per cent confidence interval:

$$CI_S = 1.96\frac{\sigma}{\sqrt{N}}. \tag{5}$$

With $\sigma = 1/\sqrt{2}$ we can compute the underlying estimated standard deviation for the scale values

$$E_m = \sqrt{\frac{1}{2N}} \tag{6}$$

Montag's error estimation. Montag [4] has published an empirical formula for the estimated standard deviation of scale values based on Monte Carlo simulations.

$$E_e = b_1(n-b_2)^{b_3}(N-b_4)^{b_5} \tag{7}$$

with $b_1 = 1.76$, $b_2 = -3.08$, $b_3 = -0.613$, $b_4 = 2.55$ and $b_5 = -0.491$. It shows the expected approximate dependency of the estimated error with the square root of the product of N and n ($b_3 \approx -0.5$ and $b_5 \approx -0.5$).

Analytic error estimation. Here we derive an analytic error estimate for Thurstone's method. The basic approach is to estimate the error in the image choice process and then propagating the error through the data evaluation steps: This process of choosing one image from the pair of images can be modeled as a Bernoulli trial with success probability p_{ij}. The standard deviation for p_{ij} equals to the standard deviation for a Bernoulli variable in N trials

$$\sigma_{p_{ij}} = \sqrt{\frac{p_{ij}(1-p_{ij})}{N}} \tag{8}$$

As we approximate p_{ij} by the empirical value q_{ij} the estimated error $E_{q_{ij}}$ for the proportion q_{ij} in equation (1) can be written as

$$E_{q_{ij}} = \sigma_{q_{ij}} = \sqrt{\frac{q_{ij}(1-q_{ij})}{f_{ij}+f_{ji}+2\delta}}. \tag{9}$$

To compute the errors of the entries z_{ij} in the Z matrix, we propagate the error using equation (3)

$$E_{z_{ij}} = E_{q_{ij}}\frac{d}{dq_{ij}}\Phi^{-1}(q_{ij}). \tag{10}$$

Using equation (4) the errors of the scale values v_i are computed as

$$E_{v_i} = \frac{1}{n}\sqrt{2}\sigma\sqrt{\sum_{b\in A; a\neq b} E^2_{z_{ij}}}. \tag{11}$$

Approximation of errors. An approximate error estimate can be derived for Thurstone's Case V if the probabilities p_{ij} are not far from $1/2$. Then their standard deviation is

$$E_{q_{ij}} \approx const = \sqrt{\frac{1}{4N}} \tag{12}$$

and the error of the Z-matrix elements z_{ij} can be approximated as

$$E_z \approx \sqrt{\frac{1}{4N}}\frac{d}{dq}\Phi^{-1}(q) \tag{13}$$

for $q = 0.5$ yielding

$$E_z \approx \sqrt{\frac{\pi}{2N}} \tag{14}$$

independent off i and j. Assuming $\sigma = 1/\sqrt{2}$ the error for the scale values v_i is approximatively

$$E_v \approx \frac{1}{n}\sqrt{\frac{\pi(n-1)}{N}} \tag{15}$$

independent on i. This formula shows also the expected approximate dependency of the estimated error with the square root of the product of N and n if n is not too small.

Experimental error estimation. Experimental error estimation [11] is an approach complementing above methods. It is based on a minimum of assumptions. It samples the error by dividing the choice data randomly into two groups. For both groups scale values are individually computed and errors are estimated from the differences of the values obtained from both groups. This process is repeated several times and the results are averaged to increase the accuracy of the error estimation. If all model assumption are fulfilled this error estimation should

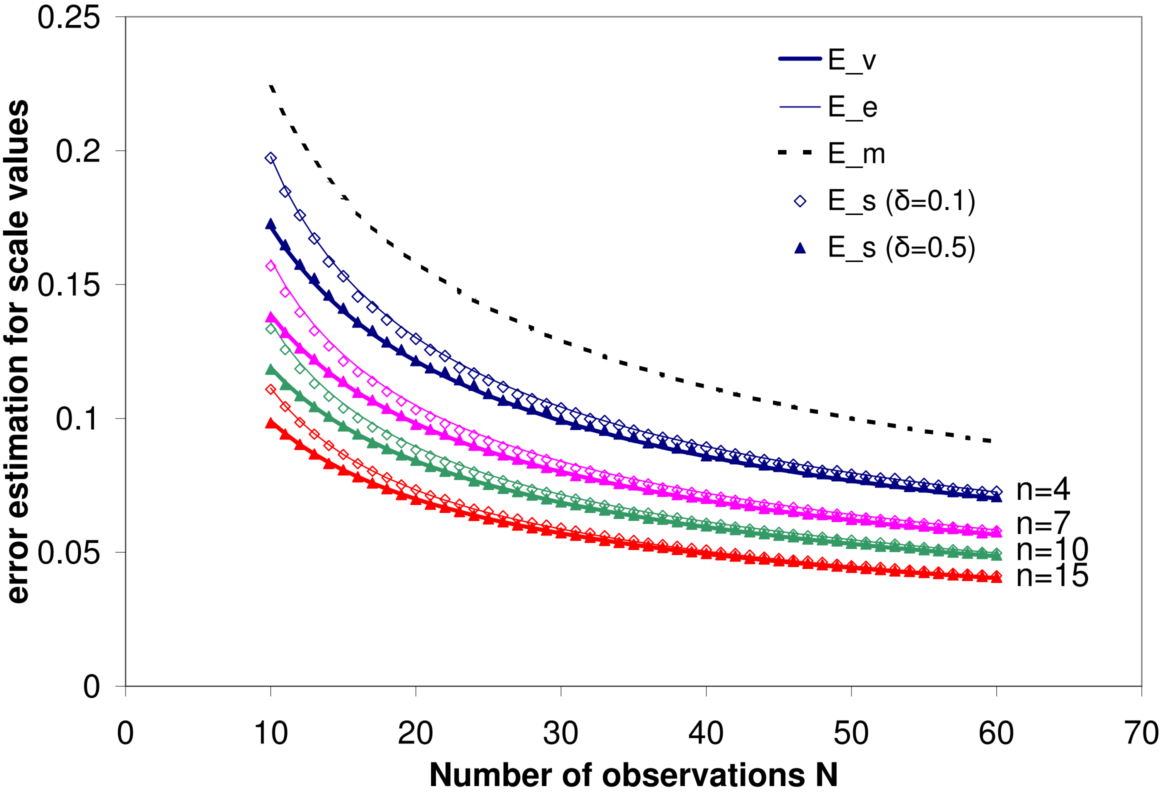

Figure 1. *Simulated scale value errors E_S compared to error error estimates E_m (eq. 6), E_e (eq. 7) and E_v (eq 15) for different number of stimuli. Triangles and circles are for simulated errors and lines are for estimated errors.*

within the statistics deliver the same error as the one obtained from analytic error estimation.

A special option of this method is to test the heterogeneity among the observers (or among the individual images). In this case individual observers (or individual images) are randomly divided into two groups and errors are computed from the average difference of the scale values between the two groups. Here error estimation using such biased samplings allows to test whether the choices depend on individual observers (or images).

Relation of error propagation with Mosteller's test.

Based on the error estimation $E_{q_{ij}}$ the error of the arcsine transformed scale values $\theta_{ij} = \arcsin(2q_{ij} - 1)$ can be derived using error propagation

$$E_{\theta ij} = E_{q_{ij}} \frac{d}{dq_{ij}} \arcsin(2q_{ij} - 1)) \tag{16}$$

Using equation (9) and the derivative of arcsin the errors of the θ_{ij} values simplify to

$$E_{\theta ij} = \sqrt{\frac{q_{ij}(1-q_{ij})}{N}} \sqrt{\frac{1}{q_{ij}(1-q_{ij})}} = \frac{1}{\sqrt{N}} \tag{17}$$

This result confirms the usefulness of the arcsine transformation in Mosteller's χ^2 test. With this transformation the variance of the θ-values is independent of the proportion q_{ij} and depends only on the number of observations N.[1] Thus estimation of errors and confidence intervals based on the arcsine transformation can be linked to the error estimation of the scale values by error propagation.

Simulation

We used Monte Carlo simulation in order to compare the different error estimations and to investigate their validity as a function of the number of observation N, the number of stimuli n and the scale value range. For all simulations we assumed a psychological continuous scale that conforms to Thurstone's Case V, i.e., that the discriminal differences follow a Gaussian distribution of equal width and that no correlation exist between two stimuli i and j. Furthermore we assumed no correlation neither in the responses of an individual observer nor in responses for an individual image. Thus ideal conditions are assumed for the simulated experiments.

Simulation for small scale values

The first series of simulation experiments was set up to compare the analytic error estimation with the estimation and simulation given in Montag's [4] article. We used the following number of stimuli $n = [4, 7, 10, 15]$. The number N of observations per pair of stimuli was in the range of $[10...60]$. We used stimuli with small scale value differences compared to the width of the distribution of the discriminal process. Thus the n scale values were assumed to be uniformly distributed in the range $[-0.25... + 0.25]$. Each experiment was repeated 10'000 times for each combination of parameters. The simulated error E_s of the scale values was calculated using the standard deviation of the scale values from the experiments. Furthermore it was veri-

[1] Note, that if we want to find a transformation which has the property of equalizing the standard deviation of all probability values p_{ij} based on eq. (8) we end up with an arcsine function.

fied, that the average scale values extracted from the simulation compare to the scale value of the model with an accuracy well within the error estimation. Three different values for the bias correction were used in the simulation ($\delta = 0.1$, 0.2 and 0.5).

In Fig. 1 the experimental error E_s is compared to the error estimates E_m (eq. 6), E_e (eq. 7) and E_v (eq. 15). The error E_m generally overestimates the simulated error E_s. The overestimation increases with the number n of stimuli. The error estimates E_e and E_v are in good agreement with the simulated error for all investigated combinations of n and N and compare well with the results given by Montag [4]. The differences between E_e and E_v get smaller with higher N. For small N the differences are in the variability range of different bias correction values δ. Simulated errors using a bias correction $\delta = 0.1$ follows Montag's error estimation E_e, where a use of $\delta = 0.5$ is in good agreement with E_v.

Error simulation for individual scale values

The estimated error E_{v_i} in eq. (11) depends on the scale value. This dependency gets important as soon as the percentages q_{ij} differ substantially from 0.5. In a second series of simulation experiments this dependency has been investigated. We used the following parameters: The number of stimuli was $n = [4, 8, 16]$, the n scale values were assumed to be uniformly distributed in the range $[-1.0 ... +1.0]$. The number of observations was fixed to a rather large number $N = 200$ to avoid a significant influence of bias corrections. The result is shown in Fig. 2. Each experiment was repeated 10'000 times for each combination of parameters.

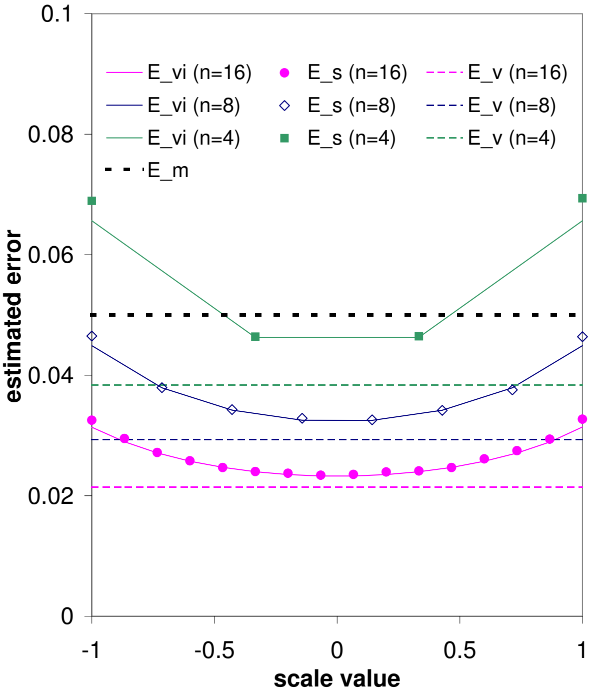

Figure 2. Simulated scale value errors as a function of z-scale compared to error estimations for different number of stimuli. squares and circles are for simulated errors, full lines for estimated errors E_{v_i}, dashed lines for E_m and E_v.

The simulated error E_s is increasing with higher absolute scale values. This increase is nicely reproduced by the error estimation E_{v_i} given in eq. (11). The error approximation E_v can be regarded as a lower limit for the simulated error. The same is true for E_e which, for this high number N, is basically identical to E_v.

Accuracy of error estimation

A third series of simulation experiments has been performed to test the accuracy of the four error estimations E_m, E_e, E_v and E_{v_i} as a function of N, n and the scale value range. For this purpose we investigate the relative accuracy in percentage of the error compared to the simulated error. Within one specific experiment the maximum relative deviation of an estimated error from the simulated error was taken as a measure for the accuracy of an estimated error. 40 different scale value ranges were used up to $[-2.0 ... 2.0]$, N was in the range $[2..100]$ in steps of 2 and the number of stimuli n was $[3, 4, 5, 8, 12, 16]$. For each experiment scale values for n stimuli were selected as follows: n values x_i were randomly chosen in the range $[-1.0 ... +1.0]$. Then these values were scaled such that the difference between the minimum and maximum scale corresponded to the target scale value range. For each combination of N, n and scale value ranges the experiment was repeated at least 2000 times and simulated errors E_s, average scale value errors E_{v_i}, approximate errors E_m and E_e were calculated. In Fig. 3 we show the results for 3, 5 and 8 stimuli.

The error estimation E_m is accurate in a range of 20% only for $n = 3$. It is not accurate for larger number of stimuli. This confirms, that the error estimation E_m should, if at all, be used only for a small number of stimuli, i.e. smaller than five. The error estimations E_e and E_v have regions with high accuracy for all numbers of stimuli, but only for small and moderate scale values up to about 1.0. For these scale value ranges, the error estimation for all scale values are approximately equal and E_e (as well as E_v) give a simple, quick and accurate error estimation. The best error estimation is given by E_{v_i}. The accuracy is better than 10% for all number of stimuli n and number of observations N up to a limiting scale value range. The limiting scale value range basically scales with the square root of N. The limit is reached when one or more expected f_{ij} in the frequency matrix are close to or smaller than one. Interestingly the accuracy of the error estimations E_e, E_{v_i} and E_v depend only marginally depend on the number of stimuli n.

Discussion

The simple error estimation given by Morovic [2] generally overestimates the error and is approximately correct only for small number of stimuli ($n \approx 3...4$). It is not suitable as a general error estimation method.

For many cases the error approximation E_v as well as the empirical error estimation E_e given by Montag [4] are sufficiently accurate. The advantage of deriving the error approximation analytically (as for E_v and E_{v_i}) is, that an adaption to other discriminal distributions such as the logistic distribution of Bradley-Terry [12] is straightforward. The inverse cumulative distribution function Φ^{-1} and its derivative in equations (3), (10) and (14) have to be replaced by the appropriate functions. Furthermore note that an empirical formula is always restricted to the parameter range used in the fitting process. It is questionable, whether the formula given by Montag can be extended to $n < 4$ or to large N. It will even fail for the (trivial) case of $n = 2$, because the parameter b_4 is larger then 2.

From Fig. 3 we can define a validity range of error estimation if we assume some limit for the accuracy. It is reasonable to assume that error estimation has to be accurate to 10%. Then the approximative errors E_m (eq. 7) and E_v (eq. 15) are good approximations for all large enough number N of observation, as long as the range of the scale values is small enough. The upper limit for the scale value range for large N is just above 1.0. The validity range of the error estimation E_{v_i} extends to much higher scale value range. The limiting factor of this error estimation is the expected number of judgments for the least probable entry in the frequency matrix f_{ij}. This is due to the fact that the entries in the Z-matrix column z_{ij} are averaged and the error of the entry with the highest error also has the highest contribution to the error of the scale value v_i. A weighted linear regression method such as described by Bock and Jones [6, chapter 6] could give more accurate scale values with smaller estimated errors in the case of a large scale value range. For these cases the proposed error estimation E_{v_i} has reached its limit.

In all our simulations we assumed the ideal Case V of Thurstone's Law of Comparative Judgment. Note that besides accurate computation of scale values also the estimation of their error is valid only if the underlying assumptions are valid. For example, if Mosteller's test fails the error estimation must also be questioned. Furthermore, correlations within individual observers' choices or within individual images have to be tested for example by using experimental error estimation [11]. If enough data is available for individual observers, a comparison of intra-observer errors with inter-observer error estimations [2] can also be used to verify whether such correlations have to be taken into account. The latter method can also be adapted to compare intra-image errors with inter-image errors.

Conclusion

We have shown, that analytic error estimation using error propagation gives good results for data evaluation of psycho-visual data using Thurstone Case V. The effort for this error computation is as small as the computation of the scale values themselves. This error estimation method should replace previous methods because it can be applied for a much larger range of psycho-visual scales. The simulation was based on an ideal Thurstone Case V model. Future simulations including correlations among individual users or individual images could give further insight to improve error estimation.

Acknowledgments

Part of this research was supported by the Hasler Stiftung Proj. # 2200.

References

[1] Central Bureau of the CIE, Vienna. *CIE Publication 156: Guidelines for the Evaluation of Gamut Mapping Algorithms*, 2004.

[2] J. Morovic. *To Develop a Universal Gamut Mapping Algorithm*. PhD thesis, University of Derby, UK, 1998.

[3] J. Y. Hardeberg, E. Bando, and M. Pedersen. Evaluating colour image difference metrics for gamut mapping images. *Coloration Technology*, 124(4):243–253, July 2008.

[4] E. D. Montag. Empirical formula for creating error bars for the method of paired comparison. *Journal of Electronic Imaging*, 15(1):010502 1–3, 2006.

[5] F. Dugay, I. Farup, and J. Y. Hardeberg. Perceptual evaluation of color gamut mapping algorithms. *Color Research and Application*, 33(6):470–476, 2008.

[6] R. D Bock and L. V. Jones. *The measurement and Prediction of Judgment and Choice*. Holden-Day, San Francisco, 1968.

[7] H. A. David. *The method of paired comparison*. Hafner Press, New York, 1969.

[8] L Thurstone. A law of comparative judgement. In *Psychological Review*, pages 273–286, 1927.

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

[10] F. Mosteller. Remarks on the method of paired comparisons: I. the least squares solution assuming equal standard deviations and equal correlations. *Psychometrika*, 16:3, 1951.

[11] P. Zolliker, Z. Baranczuk, Sprow I., and Giesen J. Conjoint analysis used for the evaluation of parameterized gamut mapping algorithms. *IEEE Transactions in Image Proceeding*, 19(3):758–769, March 2010.

[12] R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs, I. the method of paired comparisons. *Biometrika*, 39(3-4):324–345, 1952.

Author Biography

Peter Zolliker received a degree in physics from the ETH Zürich, and the Ph.D. degree in crystallography from the University of Geneva, Switzerland, in 1987. After two post-doc years at the Brookhaven National Laboratory he was a member of the R&D team at Gretag Imaging, working on image analysis, image quality, setup, and color management procedures for analog and digital printers. In 2003, he joined the EMPA, where his research is focused on digital imaging and psychophysics.

Zofia Barańczuk received her bachelor's degree in computer science and master's degree in mathematics from the Warsaw University. She works now on her doctorate and is engaged in psycho-visual tests and gamut-mapping.

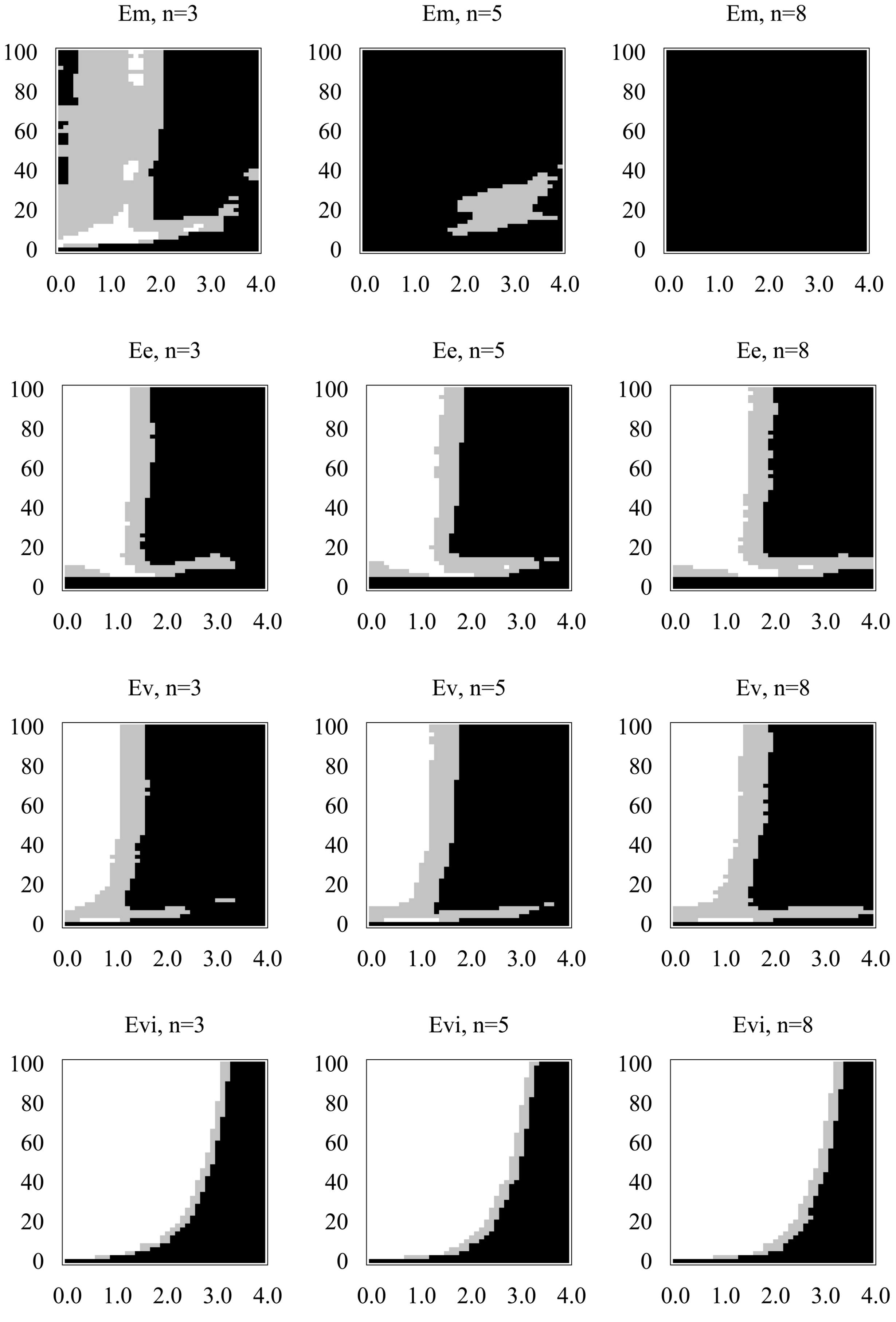

Figure 3. *Accuracy map of error estimation methods E_m (top), E_e (upper middle), E_v (lower middle) and E_{v_i} (bottom) for number of stimuli $n = 3$ (left column), $n = 5$ (middle column) and $n = 8$ (right column). Number N of observations is on vertical axis and scale value range on horizontal axis. Regions with an error estimation accuracy better than 10% are shown in white, those between 10% and 20% in gray and accuracies worse than 20% in black.*