



Web-based vs. controlled environment psychophysics experiments

Silvia Zuffi¹, Paolo Scala¹, Carla Brambilla², Giordano Beretta
Digital Printing and Imaging Laboratory
HP Laboratories Palo Alto
HPL-2006-187
December 15, 2006*

Web-based
experiments,
legibility,
readability

A recent trend in psychophysics experiments related to image quality is to perform the experiments on the World Wide Web with a large number of observers instead of in a laboratory under controlled conditions. This method assumes that the large number of participants involved in a Web investigation “averages out” the parameters that the experiments would require to keep fixed in the same experiment performed, following a traditional approach, under controlled conditions. In this paper we present the results of two experiments we have conducted to assess the minimum value of color contrast to ensure readability. The first experiment was performed in a controlled environment, the second on the Web. The result emerging from the statistical data analysis is that the Web experiment yields the same conclusions as the experiment done in the laboratory.

* Internal Accession Date Only

¹ITC, Consiglio Nazionale delle Ricerche, Milano, Italy

²IMATI, Consiglio Nazionale delle Ricerche, Milano, Italy

Published in and presented at 19th Annual Symposium Electronic Imaging Science and Technology, 28 January – 1 February 2007, San Jose, CA, USA. This paper made available with permission of SPIE and IS&T. One print or electronic copy may be made for personal use only.

Approved for External Publication

© Copyright 2006 SPIE and IS&T

1. INTRODUCTION

A recent trend in psychophysics experiments related to image quality is to perform the experiments on the World Wide Web with a large number of observers instead of in a laboratory under controlled conditions [1]. The latter allows to exactly reproduce reference conditions such as illuminant, illuminance, illumination mode, and adaptation in general; the experimenter then changes a single parameter to establish a metric or objective quantity well correlated to a perceptual phenomenon. Web experiments lack the possibility of controlling viewing conditions, and, due to the different devices and browsers employed, parameters like font size, color, contrast, and brightness are uncontrolled. While in the experiments performed in the traditional manner the participants' behavior can be monitored, and tests not properly conducted deleted, in Web-based experiments it is necessary to design strategies for filtering the results in order to eliminate invalid observations, which would not necessarily show up as outliers.

Web based experiments are interesting for studies that require few data from many subjects in a heterogeneous population. In general, this approach is claimed to be useful for experiments that do not require a precise timing, color, contrast, size or resolution [2]. In considering Web based experiments for evaluating visual characteristics, like the quality of images, the assumption is that the large number of observations made possible by this medium "averages out" the effect of viewing conditions and the other experimental parameters that should be fixed. However it is always possible that the presence of systematic errors changes the estimated result. As an example, consider experiments devoted to the study of problems related to Web visual interfaces. These studies assume the sRGB color space [3] and, if performed in a controlled environment, refer to sRGB reference viewing conditions to set ambient parameters like illumination level and chromaticity of the light source. If the same experiments are performed on the Web, they could exhibit a systematic error due to the fact that most displays are shipped with the white point set to D92 or higher, while the sRGB standard requires D65.

In order to compare traditional and Web-based studies, we have considered a psychophysics experiment in controlled conditions presented in a sister conference last year, and performed the same experiment on the Web. The subject of the study was the legibility of colored text on colored background for Web page design.

In the design of Web pages the ease of reading textual information is crucial for the effectiveness of information delivery. Many studies have been performed in recent years to evaluate legibility and readability of colored information displays on emissive devices. In these studies the terms **legibility** and **readability** are often used interchangeably, or with different meanings. In general, *readability* is a property that allows the recognition of characters into words and sentences and is more related to many aspects of the page layout, like character spacing, line spacing, and margins than to specific features of the individual characters. *Legibility* is a property of the text, and refers to whether it is possible to read a document, regardless of how comfortable it could be [4]. When a text is of low legibility, its readability is also low, but, on the other hand, when a text is not very readable, it is still possible that it is legible. Legibility is mostly related to the visual ability of the subject, but the term is often used in a generic sense, and may refer also to readability.

Old studies for the evaluation of legibility were focused on printed text; experiments on monitors were conducted on achromatic devices, or color was simulated using filters [5]. Recently, the evaluation of legibility for Web page design has become of interest. This is due to many reasons, among which the large diffusion of the Web, also accessed by old people with, presumably, low visual acuity; the increase of device diversity, due to the application of Web technologies to small devices like mobile phones; and the "anarchy" in the generation of Web pages, due to a large population of unskilled Web designers, who easily publish their content without considering basic principles of Web accessibility and page design.

In the design of Web pages, colors should be selected as to guarantee legibility of the text for many displays and environments. Studies in the specific context of Web applications have assumed that color was specified within the restricted table of Web-safe colors, without managing color appearance for different devices. This approach, while interesting for color preference evaluations, cannot be exploited for quantitative assessments on legibility, due to the different appearance unmanaged color specifications generates on different devices. A good strategy to address device differences is to adopt a well-known standardized color space, like sRGB, where viewing conditions are also defined.

A very important issue in color selection for displaying information is to take into account color vision deficiencies (CVDs). A person affected by an anomalous color vision may have problems in discriminating hues, due to an anomalous sensitivity of the cones (anomalous trichromate) or to the overlap of two cone sensitivity peaks (dichromate). Monochromatic vision is very rare. A good strategy for legibility, which implicitly takes into account CVDs, is to ensure a sufficient lightness difference between foreground and background colors. Studies by Legge [5] and others have concluded that luminance

contrast is the basis of text readability [Knoblauch1991]. Equiluminant colors with high chromatic contrast can also generate legible text, but no advantages of color contrast were found for low-vision reading [7]. Since visual acuity is much better for lightness changes than for changes in hue and saturation, the lightness and contrast of the color against the background predominate when we try to resolve fine detail. Since luminance contrast does not affect CVD, for readable text a good strategy is to provide a significant lightness difference between the information and the background [8].

In a previous study [9], we evaluated the threshold in lightness difference for readability in a controlled environment conforming to the sRGB specifications. Statistical analysis on the data collected for that experiment suggested that a sufficient lightness difference is about 30 CIELAB units. We have repeated last year's experiment under tighter control and better experimental design. We then developed a Web based incarnation of this experiment, in order to estimate the lightness threshold under Web conditions. This allowed us to:

- Refine the lightness minimum difference value for readability by performing the experiment with an improved design of the color combination selection (see details in the next paragraph);
- Analyze if performing the experiment on the Web gives the same results.

THE EXPERIMENTS

We conducted experiments for the estimation of the suitable threshold for the lightness difference between text and background colors. We evaluated this difference by means of a readability test, under the hypothesis that the time necessary to perform a readability test is a measure of legibility for the color combination. There are many measurements that can be used as correlates of readability. These include reading speed, identification of wrong words, searching for pre-defined words or characters. The purpose of these different tasks is to try to isolate the many factors that influence readability, in order to correlate it to the parameter under investigation. The test we considered was a characters counting task. The user was asked to count the occurrences of a random character in a sequence of words composed by a random selection of characters. We decided to avoid the use of meaningful words to exclude any possibility of bias in the results due to the semantics of the text.

The average length of the words was set at the average length of words in the English version of the novel "Alice in Wonderland". Even if most of the participants to the test in the lab were Italians, we considered English as the reference language for the Web. The average length of the words was set to five characters. The number of characters to be found was random with average 30. The font was Arial with a dimension of 11 pixels, which corresponds to 8 points on a 96 dpi display. In general, Web design rules for accessibility require a larger font. We have verified, by exploring some popular Web sites, that it is very frequent that a font as small as 8 points on a 96 dpi display is used. As our aim is to derive legibility rules to be used on the Web, we decided to consider this worst case in terms of font size.

The participants in the experiment were asked to perform the test several times. Each time the colors of foreground and background were chosen at random but constrained to a given lightness difference, which was randomly selected among the values 20, 25, 30, and 35. In our previous experiment [9] the color selection was performed on the Munsell Atlas. Since the atlas is a discrete space where colors have a minimum lightness distance of about 10 CIELAB units, we made the atlas continuous, in order to consider differences equal to those requested. We used the following procedure: first, we selected randomly the hue and chroma values for foreground and background. Then, we looked for the couple of colors in the atlas with the corresponding hue and chroma values and with lightness distance closest to that requested. Finally, we modified the value of lightness for the foreground color in order to have exactly the desired distance, paying attention to keep the generated color inside the sRGB gamut. Figure 1 plots of some of the colors considered in the (a^* , b^*) plane of the CIELAB color space.

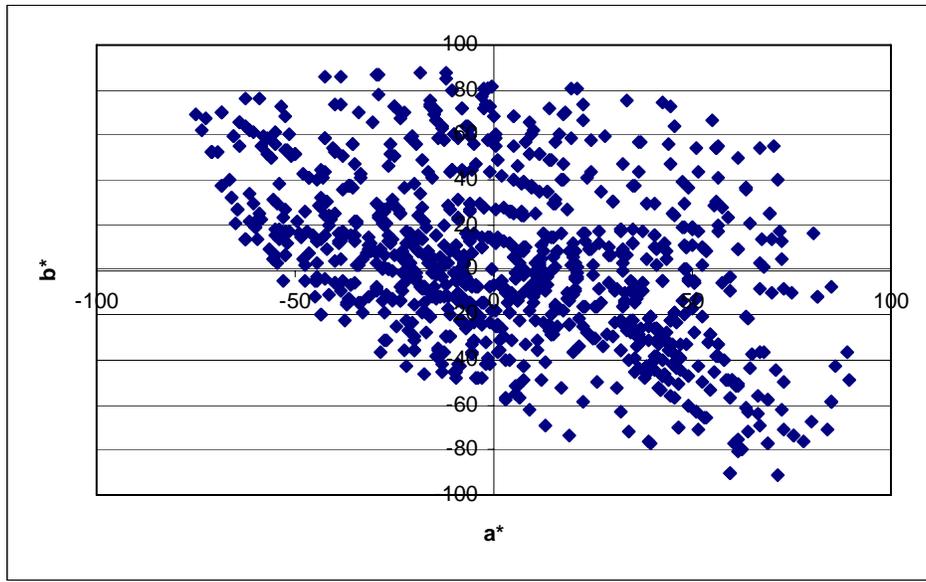


Figure 1. A subset of the colors considered in the experiment plotted in the (a^* , b^*) plane of the CIELAB color space.

We performed a traditional experiment in the lab under controlled conditions, and a Web experiment.

The **traditional experiment** under controlled conditions was conducted in a room without windows with a fixed illumination, which was regulated in order to correspond to the Reference Viewing Conditions of the sRGB Color Space [3]. The LCD monitor used for the experiment, an HP2335, was also calibrated to the sRGB specifications. In the following we will refer to this monitor as the Reference Display. The Reference Display screen size was 1920x1200, which corresponds to a resolution of 96 dpi. Each participant's color discrimination aptitude was assessed with the Farnsworth-Munsell 100 Hue Test.[®]

The **Web experiment** was conducted with the same parameters of the controlled experiment. We had however to consider that the lighting conditions of the Web users can be very different from sRGB specifications and the display characteristics vary considerably as monitors differ for type, dimension, screen size in terms of pixels, color depth, color gamut, white point, and brightness. To transfer the controlled experiment to the Web, we assume that all uncontrolled parameters average to the sRGB specification, and that the controlled parameters are as in the controlled experiment.

Unfortunately, on the Web the text font size cannot be controlled; to ensure that all Web users see the characters at the same physical dimension, we would need to know the physical dimension of the monitor and the screen dimension in pixels. While the second information is easy to acquire automatically, the user should specify the first. We have discovered that many users are not able to correctly answer the question "what is the dimension of your monitor in inches?" Due to this, we have specified a font size of 11 pixels, which corresponds to an 8-point dimension on a display having resolution of 96 dpi, like that of our Reference Display.

The experiment's home page is displayed in Figure 2. At the beginning, observers see a blank frame without characters; then, when they start the test by clicking the "Start" button, the character to be found and the random text appear. They have to scan the text and click the "Count(+1)" button each time they read the character. Finally, the observers stop the time counter by clicking on "Stop". Subsequently the observers are presented a second page where they are asked to provide information about their type of monitor (LCD or CRT) and a subjective judgment on their color combination liking.

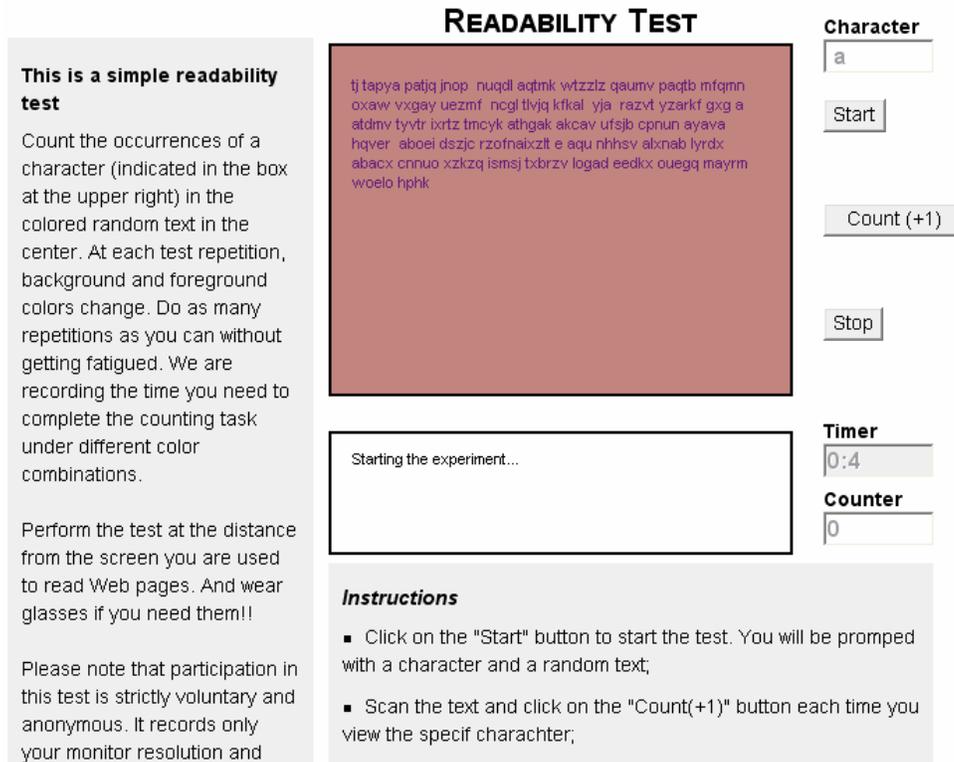


Figure 2. The Web experiment home page (<http://daedalus.itc.cnr.it/readability>)

At the end the observer opts whether to upload the data to the Web server or discard them. The uploaded data consist of a record with the time employed, the character to be found, the colors of the foreground/background combination, the screen size and color depth, the counter, and the judgment about liking.

To avoid outliers, we kept the experiment as simple as possible, and gave clear instructions to the Web observer. However, in the traditional experiment, the participant was assisted by an expert, and could ask instructions if in need of help. Records of the controlled experiments were in general errorless. This was not the case in the Web records. We then made the assumption that the average character counting performance of the participants to the traditional experiment can be considered a control group for detecting invalid observations on the Web. We recorded the number of counted characters and discarded Web records meeting the following criteria:

- records with a null counter;
- records with counts much greater than 30, which was the average number of characters to detect;
- records with counts lower than 15, which we considered an indication that the observer stopped reading before the end. We didn't set a limit equal to the number of characters to detect, because we determined that there could be some missing clicks on the counter, even when the observer detects the character;
- records with time greater than 80 sec, which we defined on the basis of the traditional test results as the maximum time employed to perform the test correctly.

RESULTS

Twenty observers did the traditional experiment, producing a total of 664 observations. On the Web 35% of the recorded observations were classified as invalid. After discarding them, we obtained 546 observations. Statistics of the time employed for performing the experiments are reported in Table 1 and depicted in Figures 3 and 4.

Traditional experiment							
L* diff.	Min.	1 st Qu.	Median	Mean	3 rd Qu.	Max.	sd
20	28.00	40.00	46.00	48.37	56.25	74.00	10.3
25	24.00	39.00	46.50	48.47	57.00	78.00	12
30	25.00	39.00	44.00	45.73	51.00	75.00	10
35	22.00	39.75	45.00	46.30	53.25	73.00	11

a)

Web experiment							
L* diff.	Min.	1 st Qu.	Median	Mean	3 rd Qu.	Max.	sd
20	22.00	39.00	47.50	48.58	54.00	78.00	12.04
25	24.00	40.00	47.50	48.02	57.00	76.00	12.12
30	24.00	36.00	43.50	45.85	55.25	80.00	12.44
35	20.00	38.00	44.50	45.96	52.75	76.00	10.65

b)

Table 1. Statistics of the time employed to perform the character counting task for four lightness differences between foreground and background (L* diff) for the traditional experiment (a) and for the Web experiment (b). Times are in seconds.

The times employed to perform the counting task with different lightness differences were compared using the Wilcoxon rank sum test (or Mann-Whitney test), a nonparametric test for comparing locations of two different populations [9]. The nonparametric approach was chosen because of lack of normality in the data. We made comparisons between the lightness differences of 20 and 25, 25 and 30, 30 and 35. We performed a one-side test, assuming as alternative hypothesis that the median time employed with the lower difference is bigger. The results of the test are shown in Table 2.

Traditional experiment			
Comparison	Null hypothesis	Alternative hypothesis	p value
20 vs 25	no difference in median time	median(20) > median(25)	>>0.10
25 vs 30	no difference in median time	median(25) > median(30)	0.03
30 vs 35	no difference in median time	median(30)>median(35)	>>0.10

Web experiment			
Comparison	Null hypothesis	Alternative hypothesis	p value
20 vs 25	no difference in median time	median(20) > median(25)	>>0.10
25 vs 30	no difference in median time	median(25) > median(30)	0.05

30 vs 35	no difference in median time	median(30)>median(35)	>>0.10
----------	------------------------------	-----------------------	--------

Table 2. Results of the Wilcoxon rank sum test for the traditional experiment (a) and for the Web experiment (b).

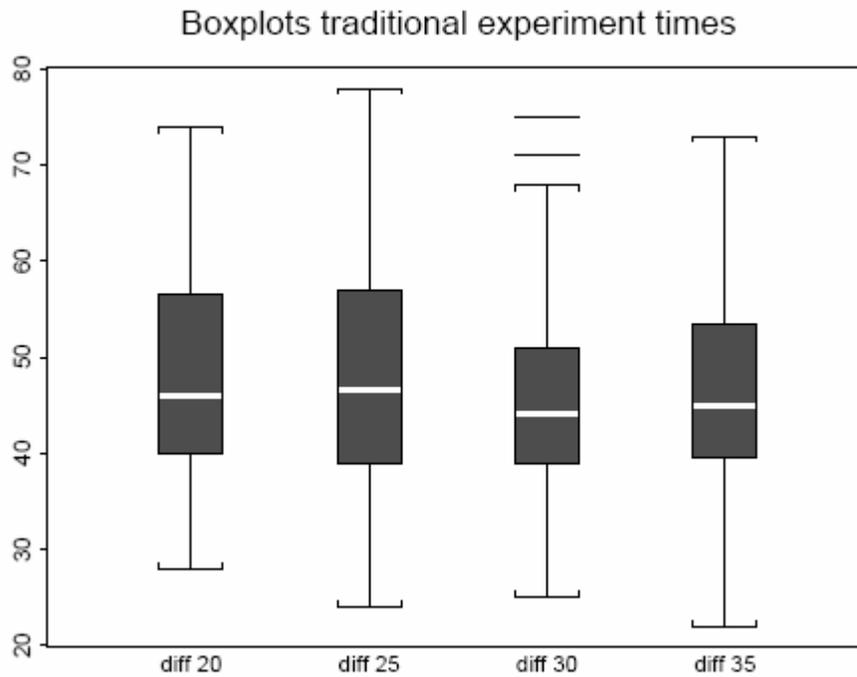


Figure 3. Traditional experiment. Box-plots of the times employed to perform the character counting task with the four lightness differences of 20, 25, 30, and 35. Boxes span the distances between the 25% and the 75% quartiles surrounding the medians (white lines). Whiskers are drawn to span the full data set with the exclusion of the points beyond 3/2 the interquartile range from the edge of the box.

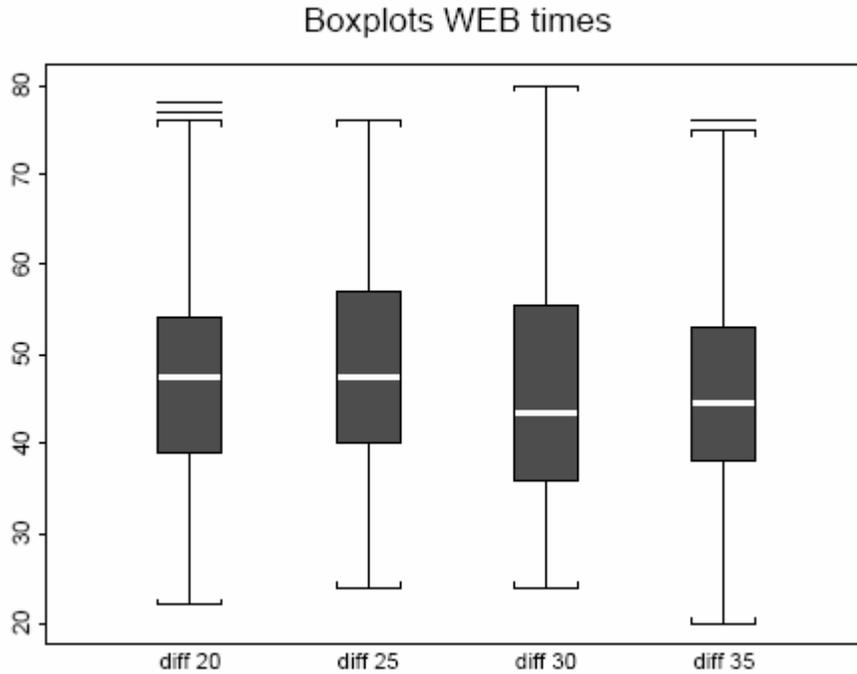


Figure 4. Web experiment. Box-plots of the times employed to perform the character counting task with the four lightness differences of 20, 25, 30, and 35. Boxes span the distances between the 25% and the 75% quartiles surrounding the medians (white lines). Whiskers are drawn to span the full data set with the exclusion of the points beyond 3/2 the interquartile range from the edge of the box.

The statistical analysis indicates that the traditional experiment and the Web experiment both suggest that the threshold for readability in the difference between the lightness of text color and background is between 25 and 30 CIELAB units.

The Wicoxon rank sum test was also used to compare the times obtained in the two experiments for each of the lightness differences taken into account. The results are shown in Table 3. They suggest that Web experiments related to visual tasks may yield the same results as experiments conducted in a laboratory under controlled conditions, although we are aware that other experiments and a more in-depth statistical analysis are needed to confirm this conclusion.

COMPARISON WEB VS. TRADITIONAL EXPERIMENT			
	Null hypothesis	Alternative hypothesis	p value
20	no difference in median time	median(Web) \neq median(trad.)	>>0.10
25	no difference in median time	median(Web) \neq median(trad.)	>>0.10
30	no difference in median time	median(Web) \neq median(trad.)	>>0.10
35	no difference in median time	median(Web) \neq median(trad.)	>>0.10

Table 3. Results of the Wilcoxon rank sum test for the comparison between Web and traditional experiments.

CONCLUSIONS

In this paper we presented the results of two studies we have conducted to compare traditional and Web-based experiments. The objective of the studies was the assessment of the minimum value of lightness difference between text and background to ensure the ease of reading of textual information. The readability test consisted in a character counting task, and the time of performing the task was considered a measure of readability. The first experiment was conducted in a controlled environment designed to match the sRGB specification. The second was performed on the Web. We have compared the times obtained to complete the test in the two different cases.

The Web experiment and the experiment conducted in the laboratory yielded the same results in terms of readability. No significant difference in readability was detected either for color combinations having a lightness difference of 20 and 25 CIELAB units, or for color combinations having a lightness difference of 30 and 35 CIELAB units. On the contrary, for both the experiments a significant difference was detected for color combinations having a lightness difference of 25 and 30 CIELAB units. These results confirm that a suitable threshold in lightness difference for readability could be about 27 CIELAB units, as stated in previous studies [8][9].

When we compared the times employed to perform the counting task in the two experiments separately for each lightness difference, we did not detect any significant difference. Together with the results just mentioned, this suggests that Web experiments could be a good strategy to perform studies related to visual tasks and that they could very well complement, if not replace, experiments conducted in laboratory under controlled conditions. However, as we already mentioned above, more experiments and an in-depth statistical analysis of the results of the Web experiments are needed to confirm this conclusion.

Of particular note in the Web experiment is that few people calibrate their monitors, and most monitors ship with a white point above D65. However, this did not produce a systematic error in the observations, suggesting that, for general use, monitor calibration may not be critical because the user sufficiently adapts to its device. Of course, in critical usages calibration is still paramount.

ACKNOWLEDGEMENTS

The authors thank Arnaud Pierson and Martina Trucco of HP University Relations for the equipment for the controlled experiment. We thank Mariangela di Paolo for managing the controlled experiment. Finally, we thank all our patient observers in the laboratory and on the Web.

REFERENCES

1. M. H. Birnbaum, "Human research and data collection via the Internet," *Annual Review of Psychology*, 55, 803-832, 2004.
2. H. van Veen, H. Bülthoff, G. Givaty, "Psychophysical experiments on the internet," *Proceedings of the 2nd Tübinger Conference of Perception*, H.H. Bülthoff, M. Fahle, K.R. Gegenfurtner, H.A. Mallot. Knirsch, Kirchentellinsfurt (Eds.), 1999.
3. M. Nielsen, M. Stokes, "The Creation of the sRGB ICC Profile," *Proceedings of IS&T Sixth Color Imaging Conference: Color Science. Systems and Applications*, 1998.
4. G. McKinstry, "Some typesetting conventions," *TUGboat* 9(3), Nov. 1988.
5. G.E. Legge, and G.S. Rubin, "Psychophysics of reading: IV. Wavelength effects in normal and low vision," *Journal of the Optical Society of America A*, 3, pp. 40-51, 1986.
6. K. Knoblauch, A. Arditi, and J. Szlyk, "Effects of chromatic and luminance contrast on reading," *Journal of the Optical Society of America A*, 8(2), pp. 428-39, 1991.
7. G.E. Legge, D.H. Parish, A. Luebker and L.H. Wurm. "Psychophysics of reading: XI. Comparing color contrast and luminance contrast," *Journal of the Optical Society of America, A*, 7, pp. 2002-2010, 1990.
8. G. Beretta, "Color Aspects of Variable Data Proofing," *Proceedings of IS&T/SPIE Internet Imaging VI*, San José, Jan 2005.
9. S. Zuffi. G. Beretta, C. Brambilla, "A Color Selection Tool for the Readability of Textual Information on Web pages," *Proceedings of IS&T/SPIE Internet Imaging VII*, San José, Jan. 2006.
10. M. Hollander, D. Wolfe, *Nonparametric Statistical Methods*, 2nd Edition, Wiley, New York, 1999.