Comparison of Naïve and Expert Observers in the Assessment of Small Color Differences between Textile Samples

Lina Cárdenas, David Hinks, Renzo Shamey, Rolf Kuehni, Warren Jasper and Melih Gunay
Polymer and Color Chemistry Program, North Carolina State University; Raleigh, NC

Abstract

Determination of the role of observer experience is of potentially critical concern regarding the development of accurate color difference formulae. As part of a larger multi-variable experiment investigating the minimum inter- and intra-observer variability possible among a statistically significant set of observers, a pilot study has been conducted to compare the performance of 25 naïve vs. 25 expert visual assessors for a set of 31 pairs of colored textile samples using a controlled psychophysical grayscale method.

No evidence of a training effect among the naïve observers was found using this method following three repeat assessments by each observer. However, a statistically significant difference between the judgments made by naïve and expert observers was found, demonstrating that observer experience is an important consideration in the development of visual datasets. The intra-observer variability among the naïve observers was equivalent to that of the expert assessors.

Introduction

An objective color difference formula accurately representing average perceptual assessments of observers is a desirable tool for color quality control of textile materials, and is arguably critical for effective electronic communication of colorimetric data for color management in a product supply chain. Existing formulae are based on several different sets of perceptual data that have been established under various experimental conditions, using samples representing a diverse range of substrates and different groups of observers. In the textile industry the CMC (2:1) color difference formula is used as standard [1-2]. Recently, however, the International Commission on Illumination (CIE) recommended the CIEDE2000 formula [3]. Luo et al. reported accuracy of prediction for several formulae against average data from a large visual data set that combined four separate experimental data sets. The large data set was used in the development of the CIEDE2000 formula. Using the PF/3 performance method, a value of 67.4 for the CIEDE2000 formula vs. 62.1 for CMC (2:1) was reported [4].

While the new formula produced an improvement for the combined data set the results remain unsatisfactory. Four subsequent independent field tests of CIEDE2000 vs. CMC (2:1) based on textile samples resulted in a similar level of accuracy for the two formulae [5-8]. No data can be found in the literature that provides a definitive answer to the disparity between the theoretical performance and the field-tested performance. However, our hypothesis is that the differences are mainly due to large inter-observer variability, differences in visual assessment protocols and insufficient fit of the formulas to mean observer data.

The work reported here is part of a larger study funded by the U.S. Department of Commerce through the National Textile Center with a primary goal of determining if, and to what degree, a significant improvement in accuracy of color difference formulae is possible. There are many variables that affect the degree of accuracy. Our project is focused on identifying and minimizing the variables in visual assessment of small color difference of textile materials and establishing the optimum level of intra- and inter-observer variability. These data will be used to determine the maximum performance of any color difference model.

Once the best experimental conditions (for textile samples) are established highly controlled replication experiments will be performed under identical conditions in different regions of the world (US, Europe and Asia).

The specific issue addressed in the pilot study reported here is: Do naïve and expert observers differ in terms of intra- and inter-observer variability in perceptual color difference assessments? (A naïve observer in this case is defined as a color-normal observer with no prior knowledge of commercial pass/fail color difference assessment, experts are defined as color normal observers whose employment involves, or has involved, commercial shade matching in the textile industry).

Experimental

Samples

For the purposes of the current experiment nine sets each consisting of a standard with six samples, dyed with disperse dyes on unbrightened plain weave spun polyester fabric were used. Figure 1 shows the location of each sample in a CIE a* b* plane.

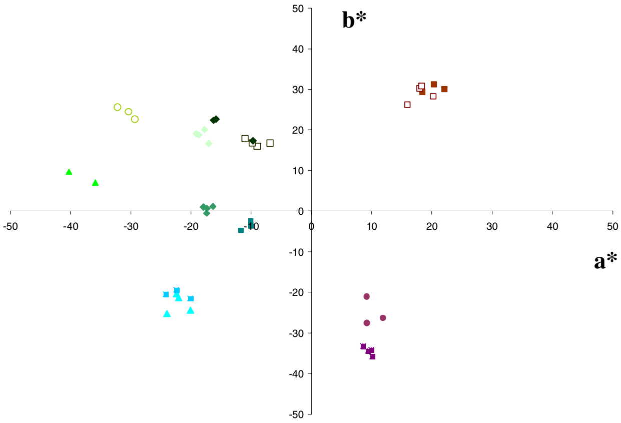

Figure 1. *Graph showing the location of dyed samples in the CIE a*b* plane.*

Each sample was cut to precise 2”×2” dimensions and mounted onto custom manufactured plastic holders. The sample mountings used precision cut PVC as backing and all the components were uniformly spray painted to a L* of 74 which is approximately equivalent to Munsell N 7.25. Each sample mounting could slide in a bar on a custom designed display easel, as shown in Figure 2. With this setup sharp dividing lines were produced with no shadows.

Test samples were measured spectrophotometrically 3 times using a Datacolor International SF600 spectrophotometer with the following setup: specular included, UV included, illuminant D_{65} and 10 degree standard observer. Each measurement was based on an average of 4 readings. The average of 3 measurements was then taken. The sample pairs had an average $DECMC_{(2:1)}$ of 1.74, with a range of 0.48-4.52 and varied in lightness, chroma and hue.

Sample viewing

The easel was viewed at a 45° angle and was located in GretagMacbeth Spectralite III standard lightbox, illuminated with a filtered tungsten daylight simulating lamp with a correlated color temperature of 6500±100K and constant illuminance of approximately 1400 lx in the middle of the display board. All extraneous light was eliminated. The light source was carefully controlled during the experiment in order to minimize variability at constant room temperature.

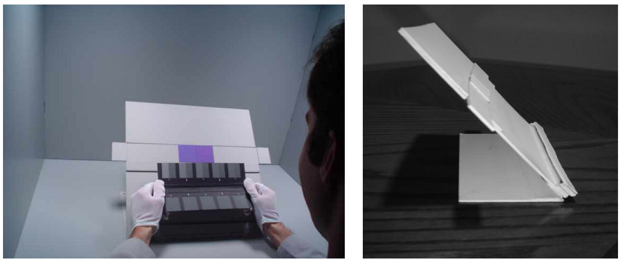

Figure 2. Custom set up for visual assessment of color difference using an AATCC gray scale.

The observers wore a mid-grey lab coat and a pair of mid-grey gloves. The samples were placed by the experimenter who also wore a mid-grey laboratory coat. At the beginning of the experiment the observer's eyes were adapted to the light source for 3 minutes by looking at the empty viewing booth during which the experiment was explained to the observer.

Psychophysical method

In this experiment an AATCC Gray Scale for visual assessment of change of shade [9] was used as a guide for assessing the perceptual differences in color. For each sample pair the question asked was: "Which grey scale difference is in closest agreement with the difference between the displayed sample pair? The result can be between two steps, such as 3-4." In the current experiment 50 observers participated, 25 naïve (mostly students of North Carolina State University, tested for normal color perception using the Neitz test [10], of which 11 were females and 14 were males) and 25 expert observers (industrial-colorists from the U.S. textile industry, including 10 females and 15 males). Each observer sat in front of the box so that he/she could move the reference grey scale freely. Each naïve observer assessed the differences 3 times on separate days. Each expert observer assessed the sample set once. The same light box, sample presentation, and sample sets were used in all cases.

Results

A total of 3100 assessments were made using 31 sample pairs; this represents a subset of the actual samples to be used in the international visual replication experiment. For the first analysis, the raw data was analyzed as grade units. The AATCC gray scale for color change consists of 9 steps of color difference defined by the CIE 1976 L*a*b* (CIELAB) formula. A grade of 5 is assigned by the observer when the sample pair presents no perceivable color difference. A grade of 1 corresponds to the largest lightness difference on the scale. Figure 3 shows the average results in grey scale grade units for the three repetitions carried out by the naïve observers and the average grade for expert observers.

Figure 3 shows that the average gray scale rating for the expert observers is below that for the three repetitions carried out by the naïve observers. This suggests that expert observers on average assess chromatic differences to be the perceptual equivalent of 8.95% higher gray scale lightness difference steps than the naïve observers and are stricter in their assessment of color differences and tend to judge with tighter tolerances.

The average scale rating for each pair was compared for each pilot experiment's repetition and each pilot's repetition was compared to the experts' ratings. A t-test, results of which are summarized in Table 1, was used to evaluate any statistical difference between each repetition.

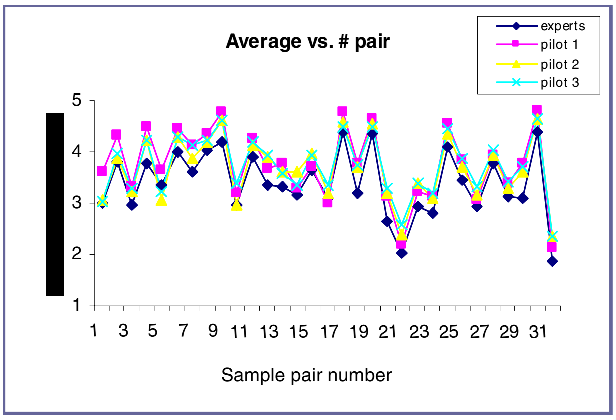

Figure 3. Average results in grade units for the visual assessments

Table 1. Summary statistics for pilot assessments carried out by naïve and expert assessors.

Group	t	p	Significance
Pilot 1 vs. Pilot 2	-2.086	0.046	Border line statistical difference at 95% confidence interval
Pilot 2 vs. Pilot 3	2.877	0.007	Statistically different at 95% confidence interval
Pilot 1 vs. Pilot 3	-0.558	0.581	No significant difference at 95% confidence interval

As the results indicate there is a statistical difference between assessments in pilot study 1 and 2, and also between pilots 2 and 3 since the value of p in the t test for these assessments is below 0.05. However, the value of p in the comparison of pilots 1 and 2 is very close to 0.05 and therefore may indicate borderline statistical significance. In addition, no significant difference was observed for the assessments between the first and third pilots. This suggests that training (i.e. repetition of the experiment) had no significant effect on the assessment of small color differences for the naïve observers under the conditions employed in this study. If there was a significant training effect pilot 3 would have produced statistically more consistent response data compared with pilot 1. Interestingly, this finding is in contrast to recent data reported by Mangine using a paired comparison test in which training naïve observers via repeat assessments produced more consistent results [11]. The Mangine experiment used the same dyed substrate as the present work, but a different experimental procedure and a different observer group and hence the disparity in training is an example of the importance of the effect of the visual assessment method employed and viewer panel on the results produced. Due to constraints in availability and geographic location, the expert observers performed the experiment only once. Results of a t-test comparison between average data for each of the three identical naïve assessments and the average experts' assessments showed statistical difference at the 95% confidence interval in all cases, as shown in Table 2.

Table 2. Summary statistics for naïve pilots vs. expert observers

Group	t	p	Significance
Pilot 1 vs. Experts	-10.883	<.0001	Statistically significant
Pilot 2 vs. Experts	-8.485	<.0001	Statistically significant
Pilot 3 vs. Experts	-10.741	<.0001	Statistically significant

To test the potential effect of non-normal distribution of the data, a Wilcoxon signed ranking test [12] was performed using the software package, STATCRUNCH. The results of this analysis were in agreement with the statistical significance t-test shown in Table 2

The standard deviations from the average gray scale rating in the three pilot studies involving naïve observers as well as those from expert assessors were analyzed within the group of naïve observers and between naïve and expert observers. In order to compare results the ΔSD for each pair sample assessed by naive observers was calculated between pilot 1 and pilot 2, pilot 2 and pilot 3, and, pilot 1 and pilot 3. A positive ΔSD represents an increase in the variability, whereas a negative ΔSD represents a decrease in the variability. Table 3 summarizes the mean ΔSD for each comparison.

Table 3. Mean delta standard deviations for the gray scale rating in pilot studies employing naïve observers

ΔSD between pilots 1 & 2	0.0785
ΔSD between pilots 2 & 3	-0.0618
ΔSD between pilots 1 & 3	0.0168

A t-test was performed for each case to evaluate if such variability within the three repetitions is statistically significant. A summary of the t-tests is shown in Table 4. Results shown in Table 4 indicate that there is a statistical difference between the mean SD differences between pilots 1 and 2 and pilots 2 and 3. However, if there is any trace of training within the observers, the ΔSD between pilots 1 and 3 should be negative and such a difference should be statistically significant. As can be seen in Tables 4 and 5, the t-test results show no training effect and, in addition, the difference in the standard deviation values is not statistically significant.

Table 4. Summary statistics for the differences in mean delta standard deviations for naive observer assessments in three repetitions.

Group	t	p	Significance
Pilot 1 vs. 2	3.396	0.001	Statistically significant
Pilot 2 vs. 3	-2.966	0.005	Statistically significant
Pilot 1 vs. 3	0.674	0.505	Statistically not significant

The inter-observer variability between naïve and expert observers was also analyzed. The mean ΔSD in each case was calculated using the same approach described above, and are shown in Table 5. As already explained a positive mean standard deviation difference value signifies that there is a higher degree of variability within the expert observers.

Table 5. Mean delta standard deviation differences in the gray scale ratings for pilots employing naïve observers vs. expert observers

ΔSD between pilot 1 & experts	0.0505
ΔSD between pilot 2 & experts	-0.0281
ΔSD between pilot 3 & experts	0.0337

The results indicate that the variability for observer assessments between naïve pilot 1 and expert assessors was higher for naïve observers. This was also the case for pilot 3 assessments compared to expert observations. However, the variability of observations in naïve pilot 2 assessments was less than that for experts. The magnitude of the differences in all cases, however, was small. To check whether the change in variability is statistically significant, t–tests for the differences were performed. The results are shown in Table 6.

Table 6. Summary statistics for the difference of mean delta standard deviations for naive observers in the three repetition assessments

Group	t	p	Significance
Pilot 1 vs. Experts	1.613	0.117	Not statistically significant
Pilot 2 vs. Experts	-0.905	0.373	Not statistically significant
Pilot 3 vs. Experts	1.212	0.235	Not statistically significant

The results demonstrate that such variability was not statistically significant.

Conclusions

The results obtained in the present study based on the psychophysical method employed demonstrate that the role of training, through repeat assessments, among naïve observers for assessment of small color difference is not statistically significant.

Using this method, it was also found that a statistical difference exists for visual judgments of small color differences between naïve and expert observers.

The results of comparing mean delta standard deviation values for the gray scale ratings in pilot studies employing naïve observers illustrated that the variability of assessments increased from pilot 1 to 2 and decreased from pilots 2 to 3. These changes were statistically significant. However, the differences in variability between pilot 1 and pilot 3 were not statistically significant. While no general conclusions can be drawn from the mean delta standard deviation data, it is clear that observer variability for both naïve and expert assessors should always be taken into account.

The results also indicate that the intra-variability of assessments among naïve observers is not statistically different from that of expert assessors.

These findings highlight the importance of determining the type of observer when designing experiments for the assessment of color differences. Such decisions clearly impact the development and accuracy of color difference equations in use in industry. The assessment of color differences is not only dependent on the uncertainty of the psychophysical interpretations but is also influenced by the observers' prior experience in handling such decisions.

For this experimental method, it is seen that either experts or naïve observers could be used to produce valid visual data, although, on average, a bias toward lower gray scale selections for expert assessors was found.

Acknowledgements

The authors acknowledge the National Textile Center (Grant number: 02-07400; C04NS11) for financial support of this work. We also thank Mr. John Darsey and colleagues at DyStar for generation of the dyed fabric samples, and Dr. Roger Woodard for helpful discussions regarding the statistical analysis.

References

[1] AATCC Test Method 173-1998, CMC: Calculation of small color differences for acceptability. AATCC Technical Manual, pp. 311-315, 2005.

[2] ISO International Standard 105-J03:1995, Textiles -- Tests for colour fastness -- Part J03: Calculation of colour differences

[3] CIE Technical Report: Improvement to industrial colour-difference evaluation. CIE Publication No. 142-2001. Vienna: Central Bureau of the CIE, 2001.

[4] Luo MR, Cui G, Rigg B., The Development of CIE 2000 Colour Difference Formula: CIEDE2000, Color Research and Application, 2001; 26:340-350.

[5] Aspland J.R. and Shanbhag P., AATCC Review, 4 (2004) 26-30.

[6] Gay J., and Hirschler R., Field Trials for CIEDE2000 – Correlation of Visual and Instrumental Pass/Fail Decisions in Industry, 25th Session of the CIE, San Diego, June 25-July 2, 2003.

[7] Noor K., Hinks D., Laidlaw A., Treadaway G., and Harold R., Comparison of the Performance of CIEDE2000 and DE_{CMC}, Book of Papers, AATCC International Conference and Exhibition, Greenville, SC, Sept. 10-12, 2003.

[8] Gibert G.M., Daga J.M., Gilabert E.J., Valldeperas J., "Evaluation of Color Difference Formulae", Coloration Technol., 121 (3) (2005) 147-151.

[9] AATCC Evaluation Procedure 1-1992, Gray Scale for Color Change. pp. 377-378, 2005.

[10] Neitz, J., Manual: Neitz Test of Color Vision. Western Psychological Services. 2001.

[11] Mangine, H., Variability in Experimental Color Matching Conditions: Effects of Observers, Daylight Simulators, and Color Inconstancy. PhD Thesis, Ohio State University, 2005.

[12] http://www.nist.gov/speech/tests/sigtests/wilcoxon.htm.

Author Biography

Lina Cardenas is currently a Ph.D. student at North Carolina State University.

David Hinks is currently an Associate Professor and Program Director of the Polymer and Color Chemistry Program at NCSU.

Renzo Shamey is currently an Assistant Professor in color science & textile chemistry at NCSU.

Rolf Kuehni is an Adjunct Associate Professor at NCSU. He was the AATCC Olney Medal Recipient in 2005.

Warren Jasper is currently an Associate Professor at NCSU.

Melih Gunay received a Ph.D. from NCSU in 2005.