Notes on the application of the standardized residual sum of squares index for the assessment of intra- and inter-observer variability in color-difference experiments

Manuel Melgosa,[1,*] Pedro A. García,[2] Luis Gómez-Robledo,[1] Renzo Shamey,[3] David Hinks,[3] Guihua Cui,[4] and M. Ronnier Luo[4]

[1]*Optics Department, Faculty of Sciences, University of Granada, 18071 Granada, Spain*
[2]*Statistics and Operations Research Department, Faculty of Sciences, University of Granada, 18071 Granada, Spain*
[3]*Department of Textile Engineering, Chemistry, and Science, North Carolina State University, 2401 Research Drive, Raleigh, North Carolina, 27695-8301, USA*
[4]*Department of Colour Science, University of Leeds, Leeds LS2 9JT, UK*
**Corresponding author: mmelgosa@ugr.es*

Received October 28, 2010; revised March 19, 2011; accepted March 20, 2011; posted March 22, 2011 (Doc. ID 137322); published April 29, 2011

The standardized residual sum of squares index was proposed to examine the significant merit of a given color-difference formula over another with respect to a given set of visual color-difference data [J. Opt. Soc. Am. A **24**, 1823–1829, 2007]. This index can also be employed to determine intra- and inter-observer variability, although the full complexity of this variability cannot be described by just one number. Appropriate utilization of the standardized residual sum of squares index for the assessment of observer variability is described with a view to encourage its use in future color-difference research. The main goal of this paper is to demonstrate that setting the F parameters of the standardized residual sum of squares index to 1 results in a loss of essential properties of the index (for example, symmetry), and is therefore strongly discouraged.

OCIS codes: 330.1690, 330.1730.

1. INTRODUCTION

In a previous paper [1], the standardized residual sum of squares (*STRESS*) index was proposed as an appropriate measure of the performance of a color-difference formula with respect to a given set of visual data. This index can also be employed to determine the statistical significance of the difference between performances of two color-difference formulas with respect to the same visual data. Color-difference formulas are designed to provide a numerical description of color pairs, under fixed viewing and illumination conditions, such that the numerical value closely correlates with the magnitude of the visually-perceived difference.

The *STRESS* index can also be employed to determine observer variability, which is a well-known influential factor in most color-difference experiments [2]. In experimental studies, two types of observer variability are usually described: observer accuracy or inter-observer variability (deviation between mean results from each observer against the mean results of a panel of observers), and intra-observer variability or observer repeatability (deviation among results of a given observer in replicated trials in an experiment). Recent reports have included well-controlled studies involving small suprathreshold color differences, indicating surprisingly large inter- and intra-observer variability [3–5] while the discrepancy between average visual results and predictions made by current advanced color-difference formulas was also relatively large (~30%) [6,7]. Fortunately, the *STRESS* index can be used to compute and compare both observer variability and performance of a color-difference formula. In practice, this comparison is very useful since any effort garnered to improve a color-difference formula and reduce its *STRESS* value beyond the corresponding observer variability *STRESS* values may be futile.

The *STRESS* index has been employed by different researchers [4,8–12], but, in some cases, in the computation of intra- and inter-observer variability, the F parameters included in *STRESS* definitions have been assumed to be equal to 1. Since many visual experiments employ the same visual scale for all observers (for example, a given anchor pair, or gray scale), it may be plausible to think that F parameters are only arbitrary scaling factors in the calculation of intra- and inter-observer variability. However, as shown here, this assumption is not correct and F parameters are factors that are designed to minimize the *STRESS* index. The main goal of this report is to demonstrate that setting the F parameters of the *STRESS* index to 1 results in a loss of essential properties of the index, and it is therefore strongly discouraged. Moreover, the results of a visual experiment employing 31 color-normal naïve observers who assessed a set of 10 color pairs on two separate occasions [13] confirmed that the average values of the F parameters for the calculation of intra- and inter-observer variability are sometimes significantly different from 1.

2. STRESS DEFINITIONS AND PROPERTIES

Let us suppose that we are given a set of n objects, and a method of determining the dissimilarity in any given pair. Metric

multidimensional scaling (MDS) is a procedure that can be employed to obtain a configuration of n points in a p-dimensional space, usually Euclidean, such that each point uniquely represents each object, and, for all pairs of objects, the Euclidean distance between two points approximates the corresponding dissimilarity. For a given configuration, the approximation error in representing the dissimilarity between two objects can be defined. A *loss function* is a weighted (and possibly normalized) sum of squares of the approximation errors for all pairs of objects. *Raw STRESS* is the most elementary MDS *loss function*, as it simply accumulates the total squared representation error. One of the most frequently employed *loss functions* is the *normalized STRESS*, or *Kruskal's STRESS* [14,15].

The *normalized* or *Kruskal's STRESS* can be applied to the conventional color-difference domain, where the main goal is to achieve a color-difference formula whose results correspond to visually perceived color differences, for a given set of $i = 1, \ldots, N$ color pairs, under fixed illumination and viewing conditions. Specifically, for the color pair i, we can designate the visually perceived color difference (i.e., the response of the human visual system) by ΔV_i, and the computed color difference (i.e., the result provided by a color-difference formula) by ΔE_i, defining a normalized percentage *STRESS* index [1] as follows:

$$STRESS = 100\left(\frac{\sum(\Delta E_i - F_1\Delta V_i)^2}{\sum F_1^2\Delta V_i^2}\right)^{1/2} \quad \text{with}$$
$$F_1 = \frac{\sum \Delta E_i^2}{\sum \Delta E_i \Delta V_i}. \tag{1}$$

As reported in [1], the next Eqs. (2) and (3) lead to the same results based on Eq. (1), and therefore, they can be considered as alternative definitions of *STRESS*

$$STRESS = 100\left(\frac{\sum(F_2\Delta E_i - \Delta V_i)^2}{\sum \Delta V_i^2}\right)^{1/2} \quad \text{with}$$
$$F_2 = \frac{\sum \Delta E_i \Delta V_i}{\sum \Delta E_i^2} = \frac{1}{F_1}, \tag{2}$$

$$STRESS = 100\left(\frac{\sum(\Delta E_i - F_3\Delta V_i)^2}{\sum \Delta E_i^2}\right)^{1/2} \quad \text{with}$$
$$F_3 = \frac{\sum \Delta E_i \Delta V_i}{\sum \Delta V_i^2}, \tag{3}$$

where $F_i (i = 1, 3)$ are three different scaling factors, which will be designated here as F parameters.

As proved in Appendix A, the F parameters shown in Eqs. (1)–(3) are not only arbitrary scaling factors to correlate perceived and computed color differences, but specific parameters that minimize their corresponding *STRESS* functions. Therefore, it is clear that by assuming that $F_1 = F_2 = F_3 = 1$, the *STRESS* definition would be strongly modified with adverse practical consequences, as indicated below.

The *STRESS* index defined in Eqs. (1)–(3) can be also used to determine intra- and inter-observer variability in a given experiment. Specifically, for the assessment of intra-observer variability, ΔV_i and ΔE_i must be replaced by the visual responses of a given observer in two different assessment sessions. If several trials are conducted, the mean observer response from all trials may be compared against responses from individual trials. For the assessment of inter-observer variability, ΔV_i and ΔE_i must be replaced by the mean assessment responses of one observer and the mean responses obtained from all observers. The final values for intra- and inter-observer variability would thus be the mean values from all observers participating in the experiment. Additionally, the standard deviation of these values can be used to express variability among observers. Incidentally, if raw visual responses are grade points based on standard gray scales, these values must be converted into visual differences ΔV_i using an appropriate fit. It must be noted, however, that the arrangement of gray pairs in the current scales used in the ISO [16] and AATCC standards [17] is not perceptually linear, and neighboring contrast pairs contain color differences that do not progress in an arithmetical fashion; rather the differences are perceptually geometric [18]. For this reason, often a polynomial fit has been employed to correlate gray scale grade points against a color-difference formula, usually CIELAB [19]. The development and application of a perceptually linear scale, for use in a manner similar to that of the ISO or AATCC Gray Scale, was reported recently that showed a statistically significant reduction in assessment variability among observers [20].

Among important properties of the *STRESS* index [1] are its limitation in the range of 0–100 (the value 0 indicating perfect agreement between visual and computed results, as desired in practical applications) and its symmetry: swapping ΔV_i and ΔE_i does not influence the index. However, by setting $F_1 = F_2 = F_3 = 1$ to compute intra- or inter-observer variability, the following three undesirable consequences are noted: (1) The upper theoretical limit for *STRESS* values is no longer set to 100; (2) the *STRESS* values from Eqs. (1) and (3) are not identical; (3) *STRESS* changes when swapping ΔV_i with ΔE_i (i.e., the *STRESS* index becomes asymmetric). This last consequence is particularly unacceptable when determining intra- or inter-observer variability since the same intra-observer variability must be obtained regardless of which two trials from each observer are considered as ΔV_i or ΔE_i; also, the same inter-observer variability must be obtained whether the mean results from a group of observers are considered as ΔV_i or ΔE_i. In summary, by assuming that $F_1 = F_2 = F_3 = 1$ and using the *STRESS* formulas given in Eqs. (1) and (3), two different "*pseudo-STRESS*" indices (one for $F_1 = 1$ and another for $F_3 = 1$) are obtained, while the true *STRESS* index proposed by multidimensional scaling is lost.

Sources of intra- and inter-observer variability can be separated into several categories, but in general they are due to systematic and random errors. In some studies, *pseudo-STRESS* ($F_1 = 1$) has been used to express the combination of the two sources of error. In visual assessment studies, systematic variation may originate from consistency in the use or interpretation of the scale employed. Hence, *STRESS* and F values of each observer may be useful when comparing observers' variability. To make the results from different experiments comparable, they must be evaluated in the same manner, however.

3. VISUAL EXPERIMENT

In this section, a practical example is provided to demonstrate the results and issues associated with setting the F parameters in *STRESS* definitions to 1 in the computation of intra- and inter-observer variability.

The results of an experiment involving visual assessment of 10 color pairs by a group of 31 nonexperienced observers with normal color vision in two separate sittings were reported previously [13]. These 10 color pairs were provided to examine the performance of the most recent CIE-recommended color-difference formula, CIEDE2000 [21]. In this experiment [13], visual assessments were performed in a GretagMacbeth SpectraLight III viewing cabinet using a five-step gray scale (ISO 105-A02). This viewing cabinet operated with a good D65 simulator constituted by two 750 W tungsten halogen lamps, with measured correlated color temperature of 6263 K and general color rendering index of 95 [22]. In different days, each observer performed two assessments for each color pair.

Using the average results of all observers, the *STRESS* values for the CIEDE2000 color-difference formula (28.4) and the intra- and inter-observer variability (48.3 and 35.4, respectively) were determined. These results indicate a relatively good performance for the CIEDE2000 formula as well as large observer variability, in agreement with previous literature [3–7]. More specifically, because of the lower *STRESS* value for the performance of CIEDE2000 formula than those corresponding to intra- and inter-observer variability in this experiment, it is concluded that modification of the CIEDE2000 formula, to improve its correlation with visual responses, may not be warranted. Table 1 shows the average and standard deviation of F_1 and F_3 parameters for each observer in this experiment, as well as the results based on testing the $F_1 = 1$ and $F_3 = 1$ hypotheses using a two-tailed one-sample t test. Based on results shown in Table 1, at a 95% confidence level, it can be concluded that in this experiment the hypothesis $F_1 = 1$ is rejected, although the hypothesis $F_3 = 1$ is not.

Assuming ΔE_i and ΔV_i represent the first and second trials, when $F_1 = 1$ and $F_3 = 1$, *pseudo-STRESS* values of 62.5 and 55.2, respectively, are obtained for the intra-observer variability. Analogously, considering ΔE_i as the results from a given observer and ΔV_i as the average results of the group, when $F_1 = 1$ and $F_3 = 1$, two *pseudo-STRESS* values of 46.5 and 43.5, respectively, are obtained for inter-observer variability. Bearing in mind that the *STRESS* values for intra- and inter-observer variability were 48.3 and 35.4, respectively, it can be concluded that in this experiment the *pseudo-STRESS* (which we consider not useful) and *STRESS* values (which we consider useful) are considerably different. Moreover, *pseudo-STRESS* values greater than 100 were obtained for three observers in the computation of the intra-observer variability, and for two observers in the computation of the inter-observer variability. In summary, the results based on the assumption of $F_1 = 1$ or $F_3 = 1$ lead to *pseudo-STRESS* indices, which are considerably different from those using the true *STRESS* index shown in Eqs. (1)–(3) and therefore are not recommended for use.

Table 2 lists the inter-observer variability results of 31 observers who took part in this experiment [13]. It can be seen that the mean *pseudo STRESS* ($F_1 = 1$) value of 46.5 is larger than the mean *STRESS* value of 35.4. In the calculation of the *STRESS* index, the F_1 value for observers varies from 0.69 to 2.04, with a mean of 1.17. Observers 3 and 12 exhibit low *STRESS* values (19.8 and 22.2, respectively) and Observer 6, a high *STRESS* value of 59.8. In addition, although Observers 16 and 4 performed well in one of two assessments, they show high *STRESS* values and can be considered inconsistent. Figure 1 plots visual assessment results of Observers 3 and 14 against the mean results calculated from all 31 observers. Both observers gave similar inter-observer variability performance (approximately 20 *STRESS* units) but they significantly vary in terms of distribution of their response around the mean. This suggests that while the *STRESS* index is very useful in providing an objective measure of determining variability among groups of observers as well as individual observer's repeatability of assessments, it does not exhibit the full complexity of observer behavior during visual assessments.

Table 1. F_1 and F_3 Values for Intra- and Inter-Observer Variability in Experiment Described in [13]; p Values Indicate Result of Testing Hypotheses $F_1 = 1$ and $F_3 = 1$

	F_1			F_3		
	Average	Standard Deviation	p value	Average	Standard Deviation	p value
Intra-observer	1.33	0.61	0.006	0.94	0.29	0.246
Inter-observer	1.17	0.34	0.007	1.00	0.31	1.000

Table 2. Inter-Observer Variability for Each Observer in Experiment Described in [13]

Observer	*STRESS*	Pseudo-*STRESS* ($F_1 = 1$)	F_1
1	26.5	39.9	0.69
2	22.2	25.0	1.12
3	19.8	20.0	1.03
4	50.5	51.0	1.08
5	24.3	24.3	1.00
6	59.8	102.3	2.04
7	44.1	56.5	1.39
8	45.9	46.6	0.91
9	26.9	27.4	0.95
10	39.8	48.9	1.31
11	30.2	31.8	0.89
12	22.2	22.8	1.05
13	49.5	49.8	1.07
14	19.7	41.8	1.38
15	22.1	22.6	1.05
16	50.2	50.3	1.03
17	22.6	32.4	0.76
18	38.7	59.3	1.49
19	32.7	33.3	1.06
20	29.5	95.5	1.95
21	21.1	21.5	0.96
22	33.0	43.4	1.30
23	52.6	52.8	1.05
24	20.2	20.9	0.95
25	68.4	69.4	0.84
26	20.3	90.3	1.90
27	60.8	62.0	1.15
28	48.3	54.3	1.28
29	38.2	80.4	1.77
30	31.1	32.6	1.10
31	27.3	31.2	0.84
Mean	**35.4**	**46.5**	**1.17**

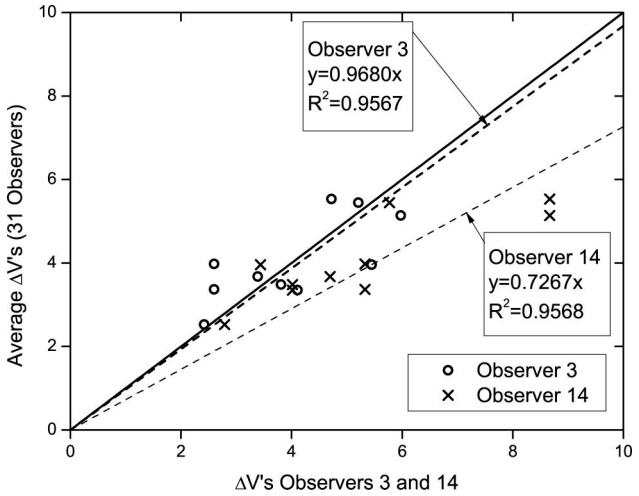

Fig. 1. Plot of visual differences of 10 color pairs for Observers 3 and 14, against average visual results of all observers participating in the experiment described in [13].

4. CONCLUSION

The *STRESS* index [1] can be employed to evaluate the performance of any color-difference formula with respect to a given set of visual data, and to measure intra- and inter-observer variability in color-difference experiments. Efforts to achieve color-difference formulas with lower *STRESS* values than those corresponding to intra- and inter-observer variability should be considered meaningless. The F parameters in *STRESS* definitions are the result of a minimization process, and not arbitrary scaling factors between perceived and computed color differences. Therefore, fixing $F_1 = F_2 = F_3 = 1$ in *STRESS* definitions in Eqs. (1)–(3) implies a radical change in the definition and properties of the *STRESS* index, which is unacceptable when a proper measurement of intra- or inter-observer variability is desired. It may also be useful to compare each individual observer's *STRESS* and F_1 values to obtain a better understanding of variability among observers. Although some authors have employed $F_1 = 1$ in a few recent papers, we do not feel that our current results imply that all research previously done must be revised. In particular, we can state that the *STRESS* values provided in [7] for the experimental data sets employed at CIEDE2000 development are correct, and derived from the corresponding values (not 1) of the F parameters involved in *STRESS* definition. Current research on color differences must probably deal with more interesting problems such as the reliability of the experimental data employed to develop and test new color-difference formulas [23], including color-difference formulas for complex images (CIE TC8-02). Anyway, in using *STRESS*, the assumption of $F_1 = F_2 = F_3 = 1$ must be avoided in the future.

APPENDIX A

Let

$$A = \sum \Delta E_i^2; \qquad B = \sum \Delta E_i \Delta V_i; \qquad C = \sum \Delta V_i^2. \quad \text{(A1)}$$

From the *STRESS* definition in Eq. (1), it can be seen that

$$STRESS^2 = 10^4 \left(\frac{\sum (\Delta E_i - F_1 \Delta V_i)^2}{\sum F_1^2 \Delta V_i^2} \right) = 10^4 \left(1 + \frac{A - 2F_1 B}{F_1^2 C} \right). \quad \text{(A2)}$$

Minimizing $STRESS^2$ with respect to F_1 results in

$$\frac{\partial STRESS^2}{\partial F_1} = 0 \Rightarrow 10^4 \frac{-2BF_1^2 C - 2F_1 C(A - 2F_1 B)}{F_1^4 C^2} = 0 \Rightarrow$$

$$F_1 = \frac{A}{B}. \quad \text{(A3)}$$

The F_1 obtained in Eq. (A3) is simply the one shown in Eq. (1). Finally, it can be proved that using F_1, the $STRESS^2$ function achieves a minimum:

$$\frac{\partial^2 STRESS^2}{\partial F_1^2} \left(\frac{A}{B} \right) = \frac{2B^2}{CA} > 0. \quad \text{(A4)}$$

Analogously, it can be proved that the F_2 and F_3 parameters minimize their corresponding *STRESS* functions defined in Eqs. (2) and (3).

ACKNOWLEDGMENTS

Research Project FIS2010-19839 from the Ministerio de Educación y Ciencia (Spain), with the European Regional Development Fund (ERDF), provided support for this work. Samples were kindly provided by Dr. D. H. Alman (DuPont Automotive Products, Troy, Michigan, USA).

REFERENCES

1. P. A. García, R. Huertas, M. Melgosa, and G. Cui, "Measurement of the relationship between perceived and computed color differences," J. Opt. Soc. Am. A **24**, 1823–1829 (2007).
2. CIE, "Parametric effects in colour difference evaluation," CIE Tech. Rep. 101 (CIE Central Bureau, 1993).
3. R. G. Kuehni, "Variability in estimation of suprathreshold small color differences," Color Res. Appl. **34**, 367–374 (2009).
4. R. Shamey, L. M. Cárdenas, D. Hinks, and R. Woodard, "Comparison of naive and expert subjects in the assessment of small color differences," J. Opt. Soc. Am. A **27**, 1482–1489 (2010).
5. S. G. Lee, R. Shamey, D. Hinks, and W. Jasper, "Development of a comprehensive visual dataset based on a CIE blue color center: assessment of color difference formulae using various statistical methods," Color Res. Appl. **36**, 27–41 (2011).
6. R. G. Kuehni, "CIEDE2000, milestone or final answer?" Color Res. Appl. **27**, 126–127 (2002).
7. M. Melgosa, R. Huertas, and R. S. Berns, "Performance of recent advanced color-difference formulas using the standardized residual sum of squares index," J. Opt. Soc. Am. A **25**, 1828–1834 (2008).
8. S. Z. Shen and R. S. Berns, "Evaluating color difference equation performance incorporating visual uncertainty," Color Res. Appl. **34**, 375–390 (2009).
9. J. Ma, H. S. Xu, M. R. Luo, and G. H. Cui, "Color appearance and visual measurements for color samples with gloss effect," Chin. Opt. Lett. **7**, 869–872 (2009).
10. R. S. Berns and B. X. Hou, "RIT-DuPont supra-threshold color-tolerance individual color-difference pair dataset," Color Res. Appl. **35**, 274–283 (2010).
11. S. G. Kandi and M. A. Tehrani, "Investigating the effect of texture on the performance of color difference formulae," Color Res. Appl. **35**, 94–100 (2010).
12. Z. N. Huang, H. S. Xu, M. R. Luo, G. H. Cui, and H. J. Feng, "Assessing total differences for effective samples having variations

in color, coarseness, and glint," Chin. Opt. Lett. **8**, 717–720 (2010).
13. M. Grosman, S. Bračko, E. Muñoz-Ibáñez, L. Gómez-Robledo, R. Huertas, and M. Melgosa, "Una verificación empírica de la mejora de la fórmula de diferencia de color CIEDE2000 respecto a CIELAB," in *Proc. IX Congreso Nacional del Color*, pp. 78–81 (Universidad de Alicante, 2010), ISBN 978-84-9717-144-1.
14. J. B. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," Psychometrika **29**, 1–27 (1964).
15. A. P. M. Coxon, *The User's Guide to Multidimensional Scaling* (Heinemann, 1982).
16. International Organization for Standardization, "Tests for colour fastness–Part A02: gray scale for assessing change in colour," ISO 105-A02 (International Organization for Standardization, 1993), http://www.iso.org.
17. AATCC Committee RA36, AATCC Evaluation Procedure 1, "Gray scale for color change" (AATCC, 2007), http://www.aatcc.org.
18. Fastness Tests Co-ordinating Committee (F.T.C.C.) Publication XI, "The development of the geometric grey scales for fastness assessment," J. Soc. Dyers Colourists **69**, 404–409 (1953).
19. S. S. Guan and M. R. Luo, "Investigation of parametric effects using small colour differences," Color Res. Appl. **24**, 331–343 (1999).
20. L. M. Cárdenas, R. Shamey, and D. Hinks, "Development of a novel linear gray scale for visual assessment of small color differences," AATCC Review **9** (8), 42–47 (2009).
21. CIE, "Improvement to industrial colour-difference evaluation," CIE Tech. Rep. 142-2001 (CIE Central Bureau, 2001).
22. R. Roa, R. Huertas, M. A. López-Álvarez, L. Gómez-Robledo, and M. Melgosa, "A comparison between illuminants and light-source simulators," Opt. Pura Apl. **41**, 291–300 (2008).
23. S. Morillas, L. Gómez-Robledo, R. Huertas, and M. Melgosa, "Fuzzy analysis for detection of inconsistent data in experimental datasets employed at the development of the CIEDE2000 colour-difference formula," J. Mod. Opt. **56**, 1447–1456 (2009).