Precision and Bias of Spectrocolorimeters

INTRODUCTION

The development and application of color management requires the exchange of device profiles that contain the information on how to map device dependent color coordinates to the device independent color coordinates of the PCS and back out to a second set device dependent coordinates. Generally, the embedded profile was created by the image originator and the output profile was created by the image utilizer. For the PCS to correctly map the input color data to the output color data the measurement systems of the two profiles must be identical, or at least compatible and equivalent. This white paper is taken from a series of two papers in the literature that describe how to assess inter-instrument agreement of spectrocolorimeters. It also gives as an example, the results of a recent inter-comparison study of four bidirectional, handheld spectrocolorimeters.

Much of the methodology reported here and in the literature is taken from either ASTM Standard Practice E 2214 or DIN Standard 55600. In each standard, the terminology of instrument performance must be explained. There are four basic terms in the metrology of spectrocolorimetry. They are:

- *Repeatability*. Repeatability is how well an instrument can repeat identical measurements. Repeatability can be quantified in terms of short, medium, and long times between measurements. These types of repeatability are taken over seconds or minutes (short); hours (medium); and days, weeks, or longer (long). Intuitively, repeatability can be thought of as the degree to which an instrument makes self-consistent measurements.
- *Reproducibility*. Reproducibility is similar to repeatability except that some aspect of the measurement conditions have changed. This might be the operator, the instrument, or some other condition. The intuitive understanding of reproducibility is the degree to which an instrument makes consistent measurements even when conditions are slightly changed.
- *Inter-instrument and Inter-model Reproducibility*. These are special cases of repeatability where instruments of identical design (inter-instrument) or different design (inter-model) are compared.
- *Accuracy*. Accuracy is how closely an instrument can conform to the accepted value for a given sample. "Accepted" values are usually provided by a high-accuracy laboratory, such as a national standards body. From these samples, accuracy is independently determined for the wavelength scale and the radiometric scale.

Goals of this White Paper

The primary goal of this paper is to document the methods of traditional color-difference based instrument evaluation. For this comparison we will report on the level of agreement between and among three typical hand-held, portable, bidirectional (45:0) spectrocolorimeters. These instruments are typical of those instruments that meet the requirements of ISO 13655 for the assessment of the color of printed images.

While most profiles are produced from instruments with some form of automated readings system, it is unlikely that these results are significantly biased by significant operator errors. Analysis of measurement data in historical data sets has shown that it is the instrument and the specimen and not the operator which are identified as the main sources of variance in reproducibility.

Reflecting media

The International Organization for Standardization (ISO) standard ISO 13655 specifies how color measurements and calculations for use in Graphic Technology are to be conducted. ISO 13655 requires that instrument geometry be either 0°:45° or 45°:0° and that all calculations of tristimulus values be achieved using the CIE 1931 standard colorimetric observer, which assumes a 2 degree field of view. ISO also calls for using the CIE D50 illuminant and defines spectral weighting functions derived from this observer and illuminant which should be used.

This specification presents some difficulties when making measurements intended for use in developing the characterization data required for the construction of ICC profiles, however.

Experimental Procedure

Instruments. All measurements and testing were completed on the entire set of three spectrophotometers. The set three hand-held bidirectional instruments. The instruments were standardized immediately prior to each measurement set. With the exception of some hand-held units that automatically power off when idle, all instruments were left powered on for the duration of the study.

Test Procedures. In the original study there were three types of reproducibility were quantified. Glossy ceramic tiles (Ceram Research CCS-II™ also known as BCRA tiles), matte sintered polytetrafluorethylene (Labsphere Spectralon™), and printing ink on synthethic paper (nitrocellulose flexible packaging inks on gloss coated, archival synthetic papers). In this white paper only the twenty ink prints made on a precision gravure proofing press will be included. Results for the other materials can be seen in the full paper in the literature.

The short-term repeatability data are provided for two samples: a disk of pressed PTFE and the each instrument's own calibration standard. Calibration standards are typically glossy tiles. For those instruments whose standard tile is unavailable (eg: the GretagMacbeth Spectroeye utilizes an internal standard) we use a different glossy standard. The goal of using multiple samples was to determine if repeatability is related to sample gloss.

As stated above, E2214 specifies that difference be calculated from the first measurement, not from the sample average. The result is that some of the calculated statistics will not be centered about the mean.

Data Collection. All instruments used in this study have the capability to report spectral reflectance factor. For our calculations, we always used spectral reflectance factor from 400 to 700 nm, samples every 10nm. This wavelength range was used even for the instruments that report a larger range. To avoid possible variations in the methods for calculating colorimetric coordinates, all color values were calculated from these 31 reflectance factor points in an identical spreadsheet.

Discussion of Results

Instruments are identified only by a letter. For comparison, Table I gives the usual, average, maximum and RMS differences cited in the manufacturer's literature for each instrument. Table II shows the multivariate statistical results for the Ink proofs.

Reproducibility
Ink Proofs

ΔE_{ab}	H_h	I_h	J_h
Ink1	0.76	1.13	1.85
Ink2	0.74	1.28	2.09
Ink3	1.03	1.55	2.67
Ink4	0.59	1.36	1.89
Ink5	1.22	1.09	2.41
Ink6	0.59	0.82	1.45
Ink7	1.01	0.83	1.88
Ink8	0.67	0.78	1.39
Ink9	0.95	1.14	2.13
Ink10	0.70	0.70	1.39
Ink11	0.50	0.53	0.97
Ink12	1.31	1.20	2.89
Ink13	0.52	0.53	0.98
Ink14	0.75	0.59	1.29
Ink15	0.84	0.72	1.43
Ink16	0.65	0.96	1.62
Ink17	0.84	0.84	1.51
Ink18	0.55	0.88	1.49
Ink19	0.78	1.38	2.38
Ink20	1.81	2.11	3.92
Average	0.84	1.02	1.88
Maximum	1.81	2.11	3.92
RMS	0.90	1.09	2.01

Repeatability
PTFE

	H_h	I_h	J_h	K_h
ΔE_{ab}	0.11	0.22	0.23	0.07
ΔE_{94}	0.11	0.22	0.23	0.07
ΔE_{00}	0.07	0.20	0.16	0.06
Hotel. ΔE	0.24	0.44	0.20	0.22
Hotel. ΔR	1.33	1.42	1.30	1.41
$\Delta X(2\sigma)$	0.27	0.16	0.62	0.17
$\Delta Y(2\sigma)$	0.29	0.18	0.65	0.18
$\Delta Z(2\sigma)$	0.31	0.35	0.77	0.17
$\sum\{\Delta W(2\sigma)$	0.88	0.69	2.05	0.51
$\sigma^2_x+\sigma^2_y+\sigma^2_z$	0.06	0.04	0.35	0.02
$\Delta R\ 2\sigma\,(440)$	0.31	0.36	0.74	0.16
$\Delta R\ 2\sigma\ (560)$	0.29	0.16	0.66	0.17
$\Delta R\ 2\sigma\ (650)$	0.28	0.12	0.65	0.26
average σ_R	0.15	0.11	0.34	0.10
weighted ΔR	0.23	0.20	0.34	0.18
Hotel. ΔR λ=400-460	1.98	2.36	2.48	2.73
Hotel. ΔR λ=470-630	1.12	1.05	1.17	1.18
Hotel. ΔR λ=640-700	2.31	1.92	2.82	3.31

Table 1 – Traditional Repeatability and Reproducibility Results – Average, Max, RMS CIELAB ΔE^*, ΔE_{94}, ΔE_{00} (Hotel = Hotelling's T^2 statistic). Instrument K is the same model as instrument J. Repeatability is based on 50 readings on different days.

| | H vs I | | | H vs J | | | I vs J | | |
	DL	Da	Db	DL	Da	Db	DL	Da	Db
A	0.46	0.21	0.54	-1.30	1.90	-1.13	1.76	-1.69	1.67
B	0.38	-0.35	0.29	-1.49	2.37	-0.40	1.87	-2.71	0.69
C	0.45	-0.65	0.05	-0.65	2.40	2.71	1.10	-3.06	-2.66
D	0.57	-0.33	0.63	-1.12	1.53	-1.44	1.69	-1.86	2.07
E	0.23	-0.26	0.30	-0.21	1.39	3.34	0.44	-1.65	-3.03
F	0.12	-0.13	-0.30	-0.91	-0.96	1.54	1.04	0.83	-1.84
G	0.13	-0.51	-0.12	-0.98	-1.95	1.86	1.11	1.44	-1.97
H	-0.03	-0.18	-0.42	-0.01	-1.74	1.07	-0.02	1.56	-1.49
I	0.18	-0.02	-0.39	-0.58	-2.80	1.13	0.76	2.78	-1.52
J	0.05	-0.01	-0.35	-0.44	-2.02	-0.19	0.48	2.01	-0.16
K	0.14	0.09	-0.40	-0.52	-1.22	-0.56	0.66	1.30	0.17
L	-0.08	0.13	-0.30	-3.01	-2.21	-1.90	2.93	2.34	1.60
M	0.07	0.32	-0.30	-1.14	-0.32	-0.87	1.20	0.65	0.56
N	0.06	0.47	-0.23	-1.18	0.35	-1.60	1.24	0.12	1.37
O	0.12	0.63	-0.28	-1.33	0.29	-1.76	1.45	0.35	1.48
P	0.30	0.35	0.26	-1.44	0.62	-1.60	1.74	-0.27	1.86
Q	0.24	0.78	0.08	-1.26	1.02	-1.61	1.50	-0.24	1.70
R	0.34	0.05	-0.02	-1.91	-0.25	-0.66	2.25	0.30	0.64
S	0.62	0.04	-0.08	-3.06	-0.50	-0.61	3.68	0.54	0.53
T	0.30	0.01	-0.06	-5.68	0.37	0.62	5.98	-0.36	-0.69
ave	0.23	0.03	-0.06	-1.41	-0.09	-0.10	1.64	0.12	0.05
a−b− g	0.57	0.08	-0.14	-0.95	-0.06	-0.07	0.98	0.07	0.03
Sigma	0.04	-0.02	0.04	1.62	-0.11	0.51	1.77	-0.25	0.58
	-0.02	0.13	-0.02	-0.11	2.49	-0.10	-0.25	2.67	-0.10
	0.04	-0.02	0.10	0.51	-0.10	2.46	0.58	-0.10	2.64
G	60.31	3.99	-26.16	0.66	0.02	-0.14	0.62	0.05	-0.13
	3.99	7.96	-0.16	0.02	0.40	0.01	0.05	0.38	0.00
	-26.16	-0.16	21.55	-0.14	0.01	0.44	-0.13	0.00	0.41
gE	24.58			0.58			0.67		
t-crit.	0.15			0.98			0.92		

Table II −Table IVb Pairwise contrasts of bidirectional instruments using INK proofs.

As with any large data set, visualization is often difficult. There are multiple approaches to consider, with the focus of each being the evaluation of the metrics themselves, not the performance or comparison of specific instruments. For the BCRA tiles, the results fall into the category what has been traditionally termed very good. That is the average differences will be 0.5 or less. None of the instruments exhibit what has been termed "excellent" where the differences are 0.2 or less. However, manufacturer's literature or publications often cite these two points as the differentiator between average or typical performance in the field and best-case performance in the factory calibration laboratory. It is noteworthy that even the best performing of the instruments – very close to the laboratory level of performance, still show a statistically

significant difference at the 5% level. Another way to state these results is "The differences between these instruments is greater than one would expect by chance variation of their readings." The short and medium term precision of these instruments, described in Part 1 of the full paper, is so good that there is almost no hope of ever having two instrument be statistically identical. As indicated above, the arithmetic average of the reflectance curves for each material, as measured by each of the instruments, should produce a most probable estimate of the true reflectance and hence the color of that material. In Figure 1 we have plotted the differences in CIELAB (Δa^*, Δb^* and ΔL^*, ΔC^*) between each instrument and the grand average for each specimen set.
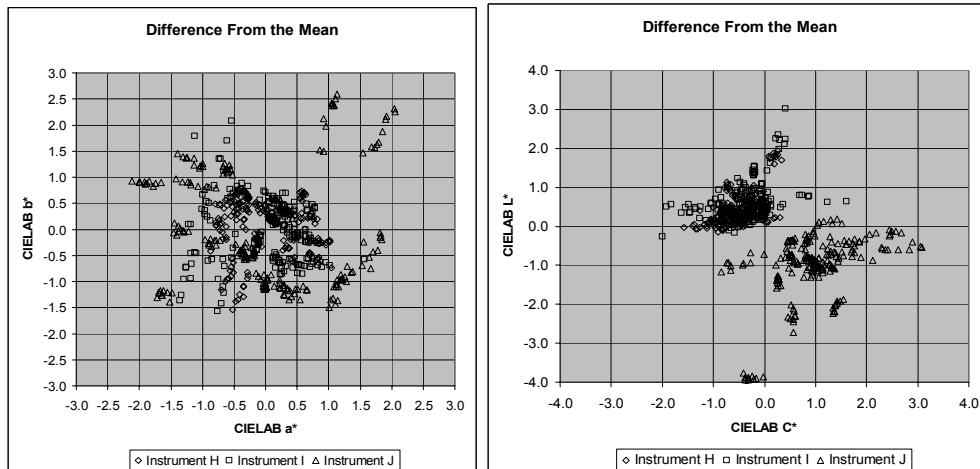

Figure 1 – Difference from the Mean for the 20 Ink Proofs

The MANOVA given in Table III and the two-way contrasts shown in Table II indicate that at least one of the instruments and one of the colors has significantly impacted the results of the study. The two-way contrasts show that all instruments are different with instruments A and B being most similar though still, being statistically different. The critical values are compared to the average of the individual color coordinates.

Effect	Multivariate Tests of Significance Sigma-restricted parameterization Effective hypothesis decomposition					
	Test	Value	F	Effect	Error	P
Intercept	Wilks	0.0	69741.0	3	36.	0.000000
	Hotellng	5811.8	69741.0	3	36.	0.000000
Instr	Wilks	0.39	7.24	6	72.	0.000005
	Hotellng	1.56	9.15	6	70.	0.000000
Color	Wilks	0.0	3809.5	57	108.17	0.000000
	Hotellng	7814.7	4752.8	57	104.	0.000000

Table III – MANOVA Results for INK Proofs Reproducibility

The a*,b* and L*,C* projection plots in Figure 1 further demonstate that even though the instruments have excellent short term repeatability, their medium term reproducibility remains

statistically different and a potentially significant source of difference. The average differences between any two instruments range between 0.8 and 1.9 CIELAB ΔE^* values. This represents up to 80% of a typical production tolerance for high fidelity color reproduction.

The results of this study further emphasize the importance of collecting multiple readings in the development of a color measurement.   When comparing individual, single readings the differences between two idnetical handheld instruments exhibited an average of 0.47 and a maximum of 1.01 CIELAB ΔE_{ab} units. When the individual readings were averaged and then averages compared the results were 0.42 average and 0.72 maximum ΔE_{ab} units. The average across instruments did not change much but the maximum color differences was reduced by nearly 30%.

The use of a bias reduction program designed to correct for various instrument variances can be a significant aid in improving the level of agreement between instruments.