

Evaluation of Reference Materials for Standardization of Spectrophotometers

John Seymour*

Keywords

spectrophotometer, calibration, inter-instrument agreement, standardization

Abstract

This paper starts by surveying papers that have quantified the level of agreement between instruments. The results are consistent in that there is not as much agreement as we would like. “Standardization” has been suggested as a means for improving intra-model agreement. In standardization, measurements of a set of samples are taken with two instruments. The measurements are compared, and this comparison determines the parameters in an equation to compensate for differences. In this way, measurements from one spectrophotometer can be adapted to improve the intra-model agreement.

The set of BCRA tiles is invaluable for spectrophotometer evaluation in that it has a wide range of rich colors, the tiles are readily cleaned, and (except for some well characterized anomalies with temperature) very stable. A set of tiles can be measured by a standards lab, and it can be assumed that the tiles, if properly handled, will not change characteristics over a long period of time, perhaps a few years. As a result, they are widely used as a means to validate spectrophotometers (to establish the accuracy of a spectrophotometer) and to quantify the degree to which it agrees with other spectrophotometers.

This set of tiles has also been used and recommended for use as a way to standardize one instrument to another. However, the agreement between instruments may not always be seen. Determination of correction coefficients can be an unstable mathematical problem and hence can lead to standardization that does not improve agreement. Worse yet, seemingly proper standardization can significantly degrade agreement.

In the main section of this paper, an experiment is performed where standardizations are performed among three spectrophotometers. Four different mathematical models are investigated and four different sample sets (including the BCRA tiles) are used to determine parameters for the standardizations.

The BCRA tiles are shown in these experiments to be less than robust with some of the mathematical models. One particularly simple set of samples – a set of paint samples – proved to be less prone to inducing large errors. An explanation of this is put forth, with the eventual aim of the development of a set of samples which can be used as a robust standardization set.

Another finding of this particular experiment is that nonlinearity is perhaps underappreciated as a source of disagreement. Unfortunately, the distinction between wavelength alignment and nonlinearity is the area where the BCRA tiles are most deficient.

Nomenclature

Before delving into the topic, a few definitions are in order. The term “inter-instrument agreement” has been commonly used to refer to the measurement agreement between two spectrophotometers of different make and/or model. This common use goes against the recommendations from the relevant standards committees, so the preferred term “intra-model agreement” will be used.

Similarly, the word “calibration” is commonly used to describe the correction of measurements from one instrument to match those of another. The relevant standards committees draw a distinction between operations performed in the factory by a spectrophotometer manufacturer (“calibration”) and operations performed by the end user (“standardization”). Again, this paper will use the term preferred by the standards community rather than the colloquial term.

Gap between expectations and performance

It is important for color measurement devices to agree. If the color of a blouse in a catalog does not match the color of the actual blouse, the garment is likely to get returned. Specific colors are associated with brand colors and product recognition, so these colors must be accurately reproduced. It is naturally expected that ads for foods and the

packaging around the foods should have the appealing color that was originally intended. These color matches are only possible if color measurement instruments all along the supply chain agree with each other. As such, print buyers often specify that printed colors be within a few ΔE of the intended color.

The common rule of thumb for statistical process control is that measurement error must not take up more than 30% of the tolerance window, and ideally should not take up more than 10%. Making a tolerance window too tight is counterproductive. Bad product may pass inspection, while good product may not.

As an example, the main ISO standard for printing (the 12647 series) calls for tolerance windows of generally 4 or 5 ΔE_{ab} . Given this, if two spectrophotometers agree to within 1.2 ΔE_{ab} , they are acceptable, although agreement to within 0.4 ΔE_{ab} is preferred. If a tolerance window of 2.0 ΔE_{ab} is demanded, then the spectrophotometers must agree to within 0.6 ΔE_{ab} .

A special report from IFRA [Williams, 2007] laid out somewhat tighter expectations for accuracy of color measurement devices.

Inter-instrument agreement is usually indicated by a colour difference value between two instruments or between a master instrument and the average of a group of production instruments. Although various ways are used to describe this colour difference, a common value is the average or mean value for a series of twelve British Ceramic Research Association (BCRA) Ceramic Colour Standards Series II (CCS II) ceramic tiles. A value of 0.3 ΔE_{ab} is acceptable.

Summary of intra-model agreement studies

The IFRA recommendations are consistent with print buyers who wish to maintain tolerances of 1.0 ΔE_{ab} , but are these expectations realistic?

In a recent report to CGATS [Seymour, 2011], a summary was given of seven recent studies on agreement between instruments. The results are consistently larger than one might expect. All the studies reviewed show that an agreement to within 0.3 ΔE_{ab} is an unreasonable expectation. Here is a synopsis from that report.

Nussbaum [Nussbaum, Sole, and Hardeberg, 2009] compared various color measurement devices of different make using a set of BCRA tiles. Measurements were compared against reference values supplied by a national standards organization. Quoting from their TAGA 2009 paper:

Overall, it can be observed that almost all instruments produce values greater than 0.5 ΔE_{ab} for all 14 tiles measured.

Eight of the nine instruments in Nussbaum's study had errors over 2.0 ΔE_{ab} on at least one of the BCRA tiles.

Greg Radencic [Radencic, Neumann, and Bohan, 2008] of PIA led a similar study that was also presented at TAGA 2009. They compared eight spectrophotometers from four manufacturers on a Lab-Ref card. Instruments were compared against the median measurement of all instruments. Here is what they had to say:

When measuring the standard reference material with the eight instruments the results show that all eight instruments did not measure all twelve colors the same. In many cases the difference between any two instruments was documented to be greater than 1 ΔE_{ab} . Problematic colors on the Lab-Ref for the instruments were instrument specific. It was shown that all instruments had problematic colors, meaning that each instrument was unable to measure all colors within 1 ΔE_{ab} of the median.

This paper reported color measurement differences of up to 10 ΔE_{ab} !

A paper by Wyble and Rich [2007] appeared in the peer-reviewed journal Color Research and Application in 2007. This paper compared both benchtop and handheld spectrophotometers using a variety of statistical techniques. Among the three handheld devices they compared, they found an average agreement on BCRA tiles of from 0.73 ΔE_{ab} to 1.68 ΔE_{ab} . The agreement among measurements on ink was slightly worse.

The ICC published a white paper [ICC, 2006] that summarized a comparison of a small set (three) of instruments. Here is a quote from this paper:

When comparing individual, single readings the differences between two identical handheld instruments exhibited an average of 0.47 and a maximum of 1.01 ΔE_{ab} units, when the individual readings were averaged.

Fred Dolezalek of FOGRA [Dolezalek, 2005] presented results of his own study to the ISO TC130 committee in 2005. He took measurements from three instruments on 46 printed patches on five different stocks. Half of his measurements showed differences greater than 1.0 ΔE_{ab} , and one in five of the measurements disagreed by more than 2.0 ΔE_{ab} .

All the previous studies were done under laboratory conditions, with instruments that were carefully calibrated and maintained in pristine condition. The VIGC (the Flemish Innovation Center for Graphic Communication) took this test out into the field in 2008, testing twenty instruments that were in daily use by printing companies [Hagen, 2008]. Hagen reports that:

The VIGC study revealed deviations up to 3.77 ΔE_{ab} for specific colors. On average the deviation per instrument of all 13 patches is 1.56 ΔE_{ab} .

X-Rite and GretagMacBeth merged in 2006, and the new company ran into a unique problem. Both companies had gone to great pains to make sure that all of their own instruments agreed with one another. But when the two companies joined, they were now responsible for making sure that instruments which had previously been calibrated through a different process would agree.

X-Rite submitted a white paper to CGATS in 2010 [X-Rite, 2010]. The first part of this paper described the degree to which their new collection of six instruments agreed. They measured 46 patches printed with CMYK inks on nine different substrates, and compared measurements between all pairings of the six instruments.

Again, the results might well be surprising. The average agreement between instruments (after X-Rite calibration) ranged from 0.27 ΔE_{ab} to 1.08 ΔE_{ab} . Under the IFRA recommendation, only one particular pairing of instruments can be considered acceptable, and then only just barely, and then only under the laboratory conditions of the X-Rite test. All measurements used the same black backing, the temperature was held within 1° of 23° C, and the relative humidity was controlled to 65%.

The X-Rite study also reported the 95% percentile agreement for the pairings of instruments, which ranged from 0.48 ΔE_{ab} to 2.94 ΔE_{ab} . One can expect that one out of twenty measurements will disagree by at least that much.

Study	Number of Instruments	Samples Measured	Reference	Errors
Nussbaum	9	BCRA Tiles	NIST Standard	8 of 9 >2.0 ΔE_{ab}
Radencic	8 Two each from four different manufacturers	Lab-Ref card	Median of all instruments	All >1.0 ΔE_{ab} , Max. 10 ΔE_{ab}
Wyble and Rich	3	BCRA tiles and ink	Paired comparison	Avg. 0.73 ΔE_{ab} to 1.68 ΔE_{ab}
ICC	3 Three units of the same model	Gravure printing	Identical model	Avg. 0.47 ΔE_{ab} , Max. 1.01 ΔE_{ab}
Dolezalek	3	46 patches 5 stocks	Paired comparison	50% >1 ΔE_{ab} , 20% >2 ΔE_{ab}
Hagen	20 Field study of in-use instruments	13 patches	GretagMacBeth NetProfiler card	Avg. 1.56 ΔE_{ab} , Max. 3.77 ΔE_{ab}

X-Rite	6 One of each of their models	46 patches 9 substrates	Paired comparisons	0.27 ΔE_{ab} to 1.08 ΔE_{ab}
--------	-------------------------------------	----------------------------	--------------------	---

Table 1 - Summary of the results from seven intra-model agreement studies

It's difficult to distill these studies down to a single number, but it is fair to say that 1.0 ΔE_{ab} is a reasonable estimate for average disagreement, and that seeing disagreement greater than 2.0 ΔE_{ab} is not uncommon. Based on the 30% rule of tolerance windows for process control (the weaker limit), setting a tolerance window for color that is less than 3.3 ΔE_{ab} is aggressive. For some pairs of instruments, even a tolerance of 5 or 10 ΔE_{ab} might be too tight.

The lack of solid agreement between spectrophotometers poses a problem for printers. Regardless of how well they are printing, the agreement between instruments may be the limiting factor.

What to do?

Wyble and Rich [2007] provided pointed advice for instrument manufacturers and for the standards community:

It is thus incumbent upon the instrument and standards community to try to identify and re-engineer or model and correct these remaining systematic differences out of their instruments before releasing them onto an unsuspecting user population.

The second part of the X-Rite white paper shows results from such correction efforts within X-Rite. This effort, primarily a manufacturing effort, has been termed "XRG". Surprisingly, the most significant change to their manufacturing process was that they standardized on white measurements from one national standards body. While XRG has not greatly improved the best case agreement, the worst case pairings of instruments improved from 1.08 to 0.60 ΔE_{ab} on average, and the worst of the various 95th percentile comparisons was reduced from 2.94 down to 1.36 ΔE_{ab} .

While these are encouraging results, further work is required. Until this happens, expectations must be set properly. One should not expect two spectrophotometers of different make and model to *always* agree to within 0.3 ΔE_{ab} , and should not even expect them to agree within 0.3 ΔE_{ab} *on average*. Under laboratory conditions, an average disagreement of more than 1.0 ΔE_{ab} is common.

One way to deal with the issue of disagreement between color measurement devices is to mandate that only one model of instrument be used. This can reduce variability since instruments within the same family tend to agree better than instruments from different families.

Coca Cola is one company that has taken this route. They require that all the providers in the supply chain use a specific model of color measurement device. This places a burden on various suppliers, since they may need different devices to serve different customers. This could be looked at as just the cost of doing business with Coke. On the other hand, it may not be possible. Most printing presses today use an onpress color measurement or control device. These devices are necessarily different from the handheld devices used elsewhere in the supply chain. Printers thus have a conundrum. Should they utilize onpress devices to improve makeready time and color consistency, knowing that the onpress device may (in effect) drive them to the wrong target color?

Standardization as a means to bridge the gap

Various researchers have defined ways to standardize one spectrophotometer so as to come into better agreement with another instrument.

One of the typical assumptions is that instruments vary because of differences in wavelength calibration and in spectral bandwidth. Robertson [1986] explained that the errors induced by wavelength misalignment and bandwidth differences are roughly proportional to the first and second derivatives of the spectrum (with respect to wavelength), respectively. Thus, to account for these sources of disagreement, a standardization equation should include terms for an offset, gain, and the first and second derivative.

Berns and Reniff [1997] described an ab initio method (that is, one based on the underlying physics, as opposed to an empirical curve fitting method) for standardization. Their method made the assumption that the differences between two instruments are caused largely by differences in white and black calibration and by differences in wavelength alignment. Their experimental results are based on the BCRA Series II tile set.

Rich and Martin [1999] and Rich [2004] proposed a standardization method that added spectral bandwidth to the list of physical source of disagreement. A mathematically equivalent method was developed by Chung et al. [2002].

Van Aken [2001] described a method that added nonlinearity to the set of sources that Berns considered.

Nussbaum et al. [2011] reported experimental results from a method to standardize measurements of one spectrophotometer to another. Unlike the ab initio methods of the other researchers, their method was empirical. In their paper, a 3 X 5 polynomial is used to standardize $L^*a^*b^*$ values, rather than spectra.

Theoretical background

Correction formulas

The aforementioned ab initio standardization methods assume that black level, white level, nonlinearity, wavelength misalignment, and bandwidth differences are the major sources of disagreement between two spectrophotometers. As an example of how such a correction could be performed, I revisit the work of Rich [2004].

The model from Rich and Martin [1999] and Rich [2004] expressed the correction between measurements from one instrument to another as the sum of four terms. In the equation, $R_m(\lambda)$ represents the measured reflectance values of the instrument to be corrected, and $R_c(\lambda)$ represents the corrected reflectance of this instrument.

The first term β_0 is the difference in photometric zero (PMZ). The second term, $\beta_1 R_m(\lambda)$, accounts for a difference in white level. The third term, $\beta_2 \frac{dR_m(\lambda)}{d\lambda}$, accounts for a wavelength shift between the two instruments. The fourth and final term, $\beta_3 \frac{d^2 R_m(\lambda)}{d\lambda^2}$, will account for any difference in spectral bandwidth. Together, the formula is

$$R_c(\lambda) = \beta_0 + \beta_1 \cdot R_m(\lambda) + \beta_2 \cdot \frac{dR_m(\lambda)}{d\lambda} + \beta_3 \cdot \frac{d^2 R_m(\lambda)}{d\lambda^2} \quad (1)$$

This formula is to be applied on a wavelength-by-wavelength basis. There are to be one set of the four β parameters at each wavelength. Just to clarify, the formula would more correctly be written as

$$R_c(\lambda) = \beta_0(\lambda) + \beta_1(\lambda) \cdot R_m(\lambda) + \beta_2(\lambda) \cdot \frac{dR_m(\lambda)}{d\lambda} + \beta_3(\lambda) \cdot \frac{d^2 R_m(\lambda)}{d\lambda^2} \quad (2)$$

Equation 2 represents one set of assumptions about the source of disagreement between two spectrophotometers. Naturally, other assumptions could be made. The differences between two instruments that are important depend, of course, on the specific instruments chosen. Equation 2 does not take into account a number of possible sources of disagreement, such as nonlinearity, fluorescence, aperture size (causing a sensitivity to lateral diffusion into the substrate), and goniophotometric differences.

In choosing equation 2, the assumption was made that these other factors are minor contributors. An equation that accounts for mild nonlinearity can be created, for example, by adding a term that is quadratic in $R_m(\lambda)$. For example

$$R_c(\lambda) = \beta_0(\lambda) + \beta_1(\lambda) \cdot R_m(\lambda) + \beta_4(\lambda) \cdot (1 - R_m(\lambda)) \cdot R_m(\lambda) \quad (3)$$

This equation could be restated with the final term being simply $(R_m(\lambda))^2$. The formulation in Equation 3 is preferred, however, since it clearly separates $\beta_4(\lambda)$ as the nonlinearity parameter.

Implementing this theory

Two comments must be made to translate these equations into a software solution.

The spectral derivatives are not directly available. A “spectral derivatometer” is not a standard piece of laboratory equipment. When these correction methods are employed, the first and second derivatives may be estimated through formulas such as

$$\frac{dR_m(\lambda)}{d\lambda} \approx \frac{R_m(\lambda+\Delta) - R_m(\lambda-\Delta)}{2\Delta} \quad (4)$$

$$\frac{d^2 R_m(\lambda)}{d\lambda^2} \approx \frac{R_m(\lambda+\Delta) - R_m(\lambda) + R_m(\lambda-\Delta)}{\Delta^2} \quad (5)$$

Here, it is assumed that Δ is the sampling interval available. For typical hand-held devices, the data available to the user is at $\Delta=10$ nm, although finer resolution is often available inside the instrument. (Note that this is one specific difference between standardization and calibration. The manufacturer is potentially able to do a better job.)

The model is trained with measurements of a set of reference standards. A set of samples (for example, the BCRA tiles) is measured with both instruments to train the standardization, that is, to determine the values of the standardization parameters. Here are the variables, with the index i added to indicate the sample number.

$R_m(i, \lambda)$ - the measured values on the test instrument to be corrected,

$R_r(i, \lambda)$ - the reference values measured with the instrument to be standardized to, and

$\beta_0(\lambda)$, $\beta_1(\lambda)$, $\beta_2(\lambda)$, and $\beta_3(\lambda)$ - the standardization coefficients.

The formula from Rich [2004] may be rewritten as

$$R_r(i, \lambda) = \beta_0(i, \lambda) + \beta_1(i, \lambda) \cdot R_m(i, \lambda) + \beta_2(i, \lambda) \cdot \frac{dR_m(i, \lambda)}{d\lambda} + \beta_3(i, \lambda) \cdot \frac{d^2R_m(i, \lambda)}{d\lambda^2} \quad (6)$$

This collection of equations can be written in matrix form:

$$\begin{bmatrix} R_r(1, \lambda) \\ R_r(2, \lambda) \\ \vdots \\ R_r(n, \lambda) \end{bmatrix} = \begin{bmatrix} 1 & R_m(1, \lambda) & \frac{dR_m(1, \lambda)}{d\lambda} & \frac{d^2R_m(1, \lambda)}{d\lambda^2} \\ 1 & R_m(2, \lambda) & \frac{dR_m(2, \lambda)}{d\lambda} & \frac{d^2R_m(2, \lambda)}{d\lambda^2} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & R_m(n, \lambda) & \frac{dR_m(n, \lambda)}{d\lambda} & \frac{d^2R_m(n, \lambda)}{d\lambda^2} \end{bmatrix} \begin{bmatrix} \beta_0(i, \lambda) \\ \beta_1(i, \lambda) \\ \beta_2(i, \lambda) \\ \beta_3(i, \lambda) \end{bmatrix} \quad (7)$$

Or in a more compact form as

$$\vec{R}_r(\lambda) = M(\lambda) \cdot \vec{\beta}(\lambda) \quad (8)$$

The least squares solution for the coefficient vector $\vec{\beta}(\lambda)$ is given by

$$\vec{\beta}(\lambda) = (M^T(\lambda) \cdot M(\lambda))^{-1} M^T(\lambda) \cdot \vec{R}_r(\lambda) \quad (9)$$

Note: Equations 1 – 3 are all wavelength dependent. The values of the four β parameters are to be determined at each wavelength. If both instruments utilize spectral gratings, the wavelength shift may be a smoothly changing parameter with respect to wavelength, for example, the wavelength shift with respect to wavelength may follow the relation

$$\beta_2(\lambda) \approx k_1 + k_2\lambda + k_3\lambda^2 \quad (10)$$

Thus, it may prove more stable to perform the regression globally, at all wavelengths at once, using k_1 , k_2 , and k_3 as global regression parameters. On the other hand, many common instruments are filter-based, with one distinct bandpass filter at each wavelength. If this is the case, then the parameters $\beta_2(\lambda)$ and $\beta_2(\lambda + \Delta)$, for example, must be taken as independent variables.

Similarly, it may be tempting to apply a nonlinearity correction globally, that is, one parameter for all wavelengths. This is supported by the fact that the largest sources of nonlinearity might be in the analog circuitry directly following the detector and before the A/D converter. In many designs, there is a single copy of this analog circuitry through which all measurements flow.

Tempting as this may be, this falls into the “don’t try this at home” category. The nonlinearity depends on the voltage coming from the detector, and not directly on the reflectance values provided to the user. Reflectance values are computed by dividing measurements of the sample by measurements of the white reference. The relationship between voltage at the detector and reflectance depends upon, among other things, the spectral output of the illumination and spectral sensitivity of the sensor. This is another case where factory calibration can potentially be better than user standardization.

Stability, intuitive discussion

What is needed in a sample set for standardization? At one extreme [Berns and Reniff, 1997] have suggested that a single cyan tile from the BCRA tile set is adequate for standardization. Rich and Martin [1999] have recommended a more cautious approach of using a large number of samples.

It is intuitively obvious that in order to determine a wavelength shift between two instruments, the spectra in the reference set cannot all be spectrally flat. At least some samples in the reference set need to have a significant derivative with respect to wavelength. Similarly, some of the spectra must have a significant second derivative if one is to reliably determine a difference in bandpass between two instruments.

The graph below shows the first derivative of reflectance with respect to wavelength for a set of thirteen BCRA tiles. Larger derivatives (either negative or positive) are beneficial for discerning a wavelength shift between two instruments.

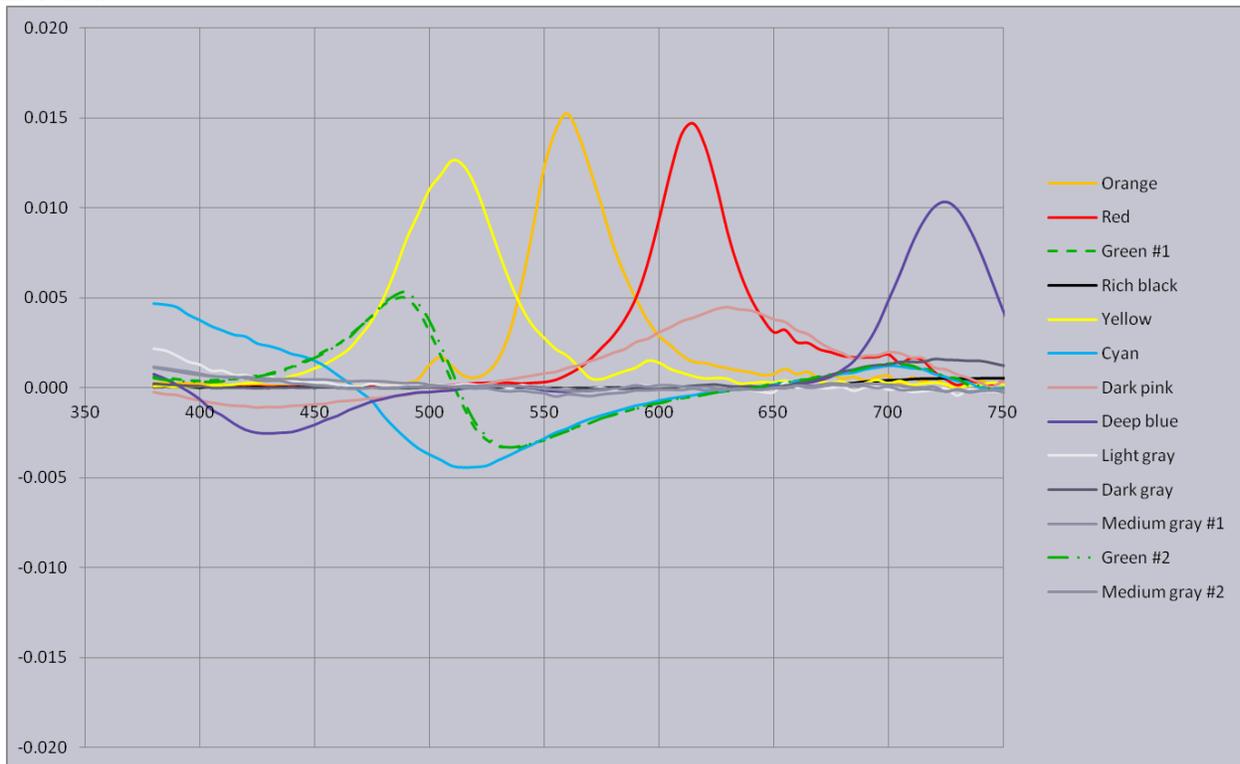


Figure 1 – First derivatives of the spectral of the BCRA tiles

As can be seen, the wavelength discrimination below 470 nm and between 650 nm and 700 nm is not as good as in other areas of the spectrum. This gives an intuitive understanding of the issue. It suggests that the ability to discern wavelength shift at 610 nm (where the red derivative peaks) is about seven times that at 680 nm, where none of the tiles have appreciable derivatives.

It can also be seen that beyond 600 nm, none of the tiles have an appreciable negative derivative. It may be difficult in this range to discern between the scaling parameters and the wavelength related parameters. To illustrate why this is important, Figure 2 demonstrates that a wavelength shift may mimic the effects of nonlinearity. The blue line in the graph is the spectrum of the orange BCRA tile. The red line is that same spectrum, shifted to the left by 10 nm. Note that, this would be a huge discrepancy, but it provides a clearer exposition of the idea. The green line is the result of a simple parabolic nonlinear transform applied to original spectrum, without any shift. While there are still noticeable differences, overall the two modified spectra are quite similar.

Let's say the spectrum being measured had a negative slope. The important point is that this same shift to the left would appear to have the opposite nonlinearity. Thus, having both a negative and a positive derivative allows one to distinguish between a wavelength shift and a nonlinearity.

This same sort of problem exists when trying to assess differences in spectral bandwidth.

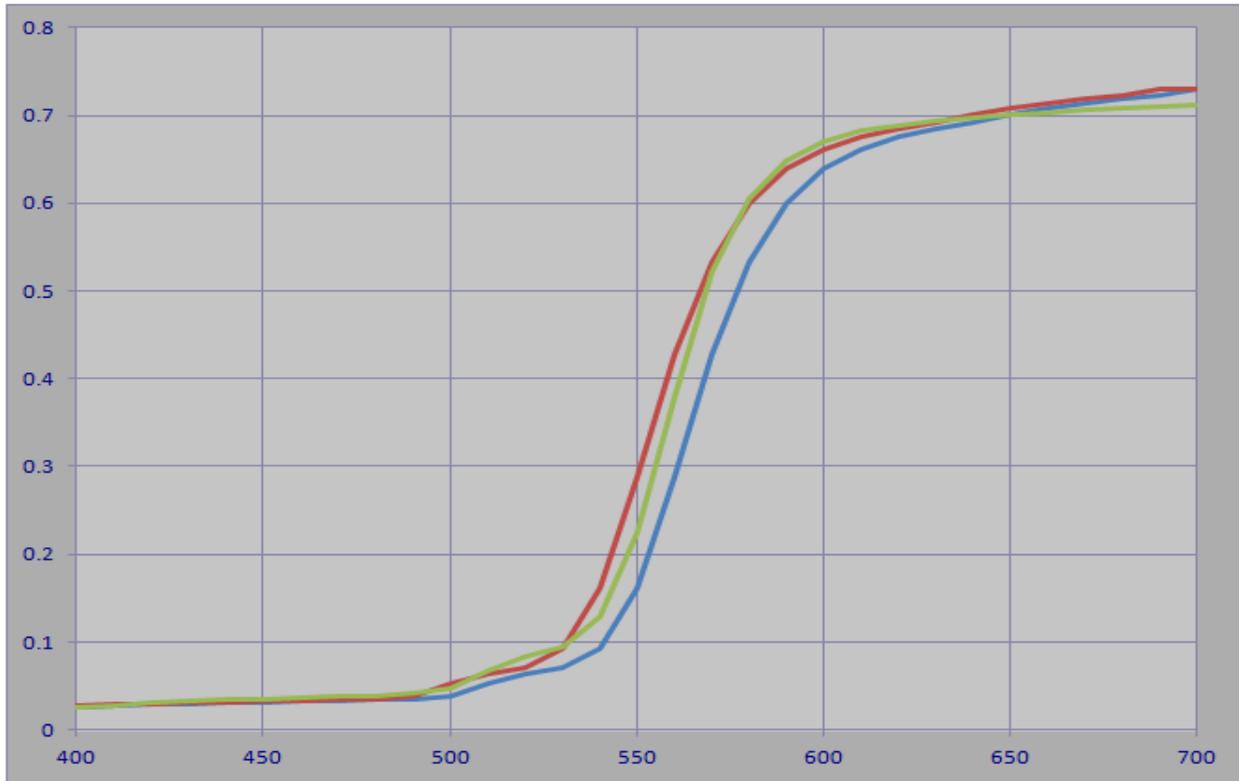


Figure 2 – Ambiguity of wavelength shift and nonlinearity

Rules for a good set of standardization samples

To select an ideal sample set for standardization of one spectrophotometer to another, there are a few simple rules that must be kept in mind.

1. In the absence of any effects other than gain and offset, a few samples with widely differing reflectance values are sufficient for standardization of one instrument to another.
2. If one instrument has a wavelength shift with respect to the other, samples with a large wavelength derivative in all parts of the spectrum are necessary.
3. If there is a difference in linearity between two instruments, a sample set with multiple reflectance values is necessary.
4. In the presence of a potential wavelength shift and a difference in linearity, a sample set with both positive and negative wavelength derivatives is necessary.

Experiment

Description of the test

This is a test of standardization of one spectrophotometer to another. One goal is to compare the effectiveness of one set of standardization samples versus another. Another goal is to investigate the effectiveness of various parameters that could be used in the standardization. This experiment used three different spectrophotometers to measure four different sample sets. Regression was used to perform standardizations of one instrument to another, and the results were analyzed.

Four sample sets were used in this study. They included:

BCRA series II tiles (13 samples)

Coated Pantone “primaries” from the first few pages of a Pantone guide (19 samples)

Six cards pulled from a coated Pantone guide, with seven gradations on each, including the cards with rhodamine, orange 021, green, process blue, and violet. To be clear, the rhodamine card included Pantone 230 to 235; the orange card included Pantone 1485 to 1545, and so on. (42 samples)

Six sample cards of matte paints from Behr: 150B (four shades of red), 230B (four shades of orange), 430B (four shades of green), 580B (four shades of blue), 640B (four shades of violet), and 750F (four shades of gray) (total of 24 samples).

One characteristic of the sample sets is that they differ greatly in gloss. The tiles are very glossy, the Pantone samples are moderately glossy, and the paint samples are quite matte. This choice was intentional, since goniophotometric differences (differences in illumination or collection angle) can appear to be differences in offset and/or gain. Since the apparent offset and gain differences depend on the gloss of the sample, this sample set provides a means for diagnosing this potential issue.

Measurements of all samples were made with three spectrophotometers:

Gretag SPM50 which is slightly over 20 years old and can no longer be factory calibrated.

XRite 939 which was recently calibrated to the XRGA certification.

XRite SpectroEye which was nearing its due date for calibration, and had not been XRGA certified.

The SpectroEye was selected as the instrument to standardize to. One set of regressions was performed to standardize the Gretag SPM50 to the SpectroEye. These two instruments are of similar design, but presumably, the SPM50 has “aged”. The second set of regressions standardized the XRite 939 to the SpectroEye. These two instruments are of dissimilar design, the first being filter-based and the second grating-based. In addition, the SpectroEye has not been XRGA certified. It is my understanding that this instrument is one that will change the most as a result of the recertification.

While all three instruments have been carefully taken care of, it is expected that there should be moderate (but normal) disagreement between them.

The comparisons of measurements prior to standardization are summarized below. All numbers in the table are in ΔE_{ab} . The numbers above are the median color difference of all the samples in the set. The numbers below are the 90th percentile. The results are not unexpected, with medians in the range from 0.5 to 1.0 ΔE_{ab} , and 90th percentiles in the range from 1.0 ΔE_{ab} to 2.0 ΔE_{ab} .

	BCRA	Pantone primaries	Pantone ramps	Behr ramps
SPM50 to SpectroEye	0.35 (1.84)	0.66 (1.69)	0.49 (1.60)	0.63 (1.72)
XRite 939 to SpectroEye	0.71 (1.26)	0.86 (2.00)	0.90 (1.73)	0.46 (0.93)

Table 2 - Median and 90th percentile color differences of sample sets before standardization

Regression was performed with four different combinations of parameters. All four of the regression techniques included the basic parameters of offset, gain, and wavelength shift. The four techniques covered all combinations of a) including or not including a correction of bandpass of the instrument, and b) including or not including a correction for nonlinearity.

Regression technique 1: Intra-model differences were modeled as offset, gain, and wavelength shift.

Regression technique 2: Intra-model differences were modeled as offset, gain, wavelength shift, and bandpass correction.

Regression technique 3: Intra-model differences were modeled as offset, gain, wavelength shift, and nonlinearity.

Regression technique 4: Intra-model differences were modeled as offset, gain, wavelength shift, bandpass correction, and nonlinearity.

For any given instrument pairing and regression technique, each of the four sample sets were used for regression. The parameters from the regression were then used to standardize the readings from that instrument on all sample sets to match the corresponding SpectroEye measurements. These standardized readings were compared against the corresponding SpectroEye readings and color differences (ΔE_{ab}) were determined. The sample set readings were summarized as median ΔE_{ab} and 90th percentile ΔE_{ab} .

(Thus, regression was performed a total of 128 times, with all combinations of four regression techniques, two instrument pairings, four standardization sets, and four test sets.)

Summaries of all of the results are shown in the Appendix.

Summary of results

In the Appendix, each combination of sample set and regression technique has been given a “rating” indicating percentage improvement from the non-standardized measurements to the standardized measurements. These ratings have been collected in the table below. The columns 1.1 through 4.2 are the combinations of regression technique (the first number) and instrument pairing (the second number). Thus, the column labeled “3.2” is the results using the third regression technique to standardize the XRite 939 to the SpectroEye. As before, the upper number is the median of the color differences and the lower number (in parentheses) is the 90th percentile.

One important finding is that a substantial number of standardizations resulted in little or no improvement, and many actually made intra-model agreement worse. The worst case was using the Pantone primaries and regression technique 2 to standardize, and applying this standardization to measurements on the BCRA tiles. In this case, a modest original instrument disagreement (90th percentile of 1.84 and 1.26 ΔE_{ab}) more than doubled to 3.70 and 4.01 ΔE_{ab} .

This should be taken as a cautionary note. *Well intentioned standardization using a seemingly reasonable training set and a recommended mathematical model will often be a worthless and misleading exercise. At worst, it can fail miserably.*

The numbers in the chart have been highlighted by color to rate the success of the standardization. A number has been highlighted with green to indicate a “good” standardization where the standardization lowered the color difference by at least 20%. An orange number indicates a “feckless” standardization, where the change in color error was between $\pm 20\%$. A red number highlights instances where the standardization process worsened the agreement by at least 20%.

Regression set	1.1	1.2	2.1	2.2	3.1	3.2	4.1	4.2
BCRA	1% (-8%)	26% (0%)	0% (10%)	13% (-1%)	4% (0%)	19% (-5%)	2% (-7%)	14% (19%)
Pantone primaries	-42% (-8%)	-25% (-28%)	-77% (-28%)	-33% (-49%)	-69% (-38%)	-20% (-42%)	-82% (-38%)	-26% (-38%)
Pantone ramps	-8% (24%)	9% (2%)	-34% (18%)	2% (-1%)	-10% (0%)	12% (10%)	-31% (-9%)	-2% (3%)
Behr ramps	13% (46%)	39% (53%)	-22% (34%)	24% (41%)	20% (50%)	30% (47%)	-15% (36%)	24% (46%)

Table 3 – Summary of results of standardization tests

The results are further summarized by counting the number of instances in each category, first for each of the sample sets, and then by each of the regression techniques.

Regression set	Bad	Feckless	Good
BCRA	0	15	1
Pantone primaries	14	2	0

Pantone ramps	2	13	1
Behr ramps	1	3	12

Table 4 – Frequency of success and failure of standardization sets

The industry standard sample set, the BCRA tiles, resulted almost always in feckless results. Based on the results of this experiment, there is no point to using BCRA tiles to train the standardization of one instrument to another, except of course, if the instrument will be used strictly to measure BCRA tiles.

From Table 4 it can be seen that the Behr paint samples are the most stable standardization set. Nearly all standardizations led to a substantial improvement. The Pantone primaries were by far the worst. Nearly all the standardization with the Pantone primaries resulted in substantially worse intra-model agreement.

Regression technique	Bad	Feckless	Good
Offset, gain, shift	3	8	5
Offset, gain, shift, band	6	7	3
Offset, gain, shift, nonlin	3	10	3
Offset, gain, shift, band, nonlin	5	8	3

Table 5 – Frequency of success and failure of standardization models

Table 5 breaks the experiment down by regression technique. From this table it can be seen that, if an arbitrary standardization set is chosen, the simplest regression technique is most likely to give good results. This same technique is also one of the least likely to give bad results.

Another interesting result is that both standardization techniques that included a bandpass correction were less stable. This could be because any bandpass related errors were insignificant, or it could be that the sample sets were not rich enough in wavelength transitions to adequately gauge bandpass differences. At any rate, at least for these training sets, bandpass correction via the second derivative is ill-advised.

While the simplest regression technique was preferable in general, when the Behr paint ramps were used as the standardization set, the regression technique including offset, gain, wavelength shift, and nonlinearity correction performed best. Three of the results were good, and the fourth was at the edge between good and feckless. This is an important result. While many of the previous researchers have ignored the effect of nonlinearity, it would appear from analysis of these three instruments that nonlinearity could be a larger contributor to disagreement than bandpass.

Then again, in a personal conversation, Harold Van Aken has pointed out that lateral diffusion is another physical difference that can masquerade as a nonlinearity issue for some samples (notably BCRA tiles and printing on translucent substrates). Light spreads a large distance in some samples and will exit perhaps a millimeter or so away from where it hit the sample. Because of this, the reflectance that is measured on such samples depends on the aperture size and the relationship between the area illuminated and that measured.

This effect has been described in an early TAGA paper [Spooner, 1991]. This paper also describes a laundry list of phenomena that may cause disagreement between spectrophotometers.

Discussion

Selection of the Behr set of samples was not based on an exhaustive analysis of a spectral database of paints. Very little analysis went into the selection. The selection was based on finding a set of color ramps that roughly had the property of mixing negative and positive derivatives. The initial thought was to include green and purple, since (spectrally) these two are opposite. One has positive derivatives where the other has negative. The gray ramp was added to bolster the nonlinearity correction. Red, orange, and blue were added for no other good reason than a desire to fill out the set.



Figure 3 – The Behr paint samples

It is certain that this set of samples is not optimal. It may be that the red, orange or blue cards add little to the effectiveness of the set. If they are only marginally helpful, it might be beneficial to leave them out. It may not be necessary to have four ramps; maybe just the two extreme ones (the lightest color and the darkest color in the ramp) are enough. Then again, maybe additional ramps may be advantageous. More to the point, a different set of pigments altogether might be preferred.

Determination of the optimal set of samples is beyond the scope of this project, but some brief comments are in order. As shown in Figure 4, the gray samples in the Behr sample set do not provide a very good sampling of reflectance values. If four gray samples were to be used, it would be preferable to have white and deep black samples, along with samples of roughly 30% and 60% reflectance. Additional gray samples and/or better spacing of gray samples would provide better substantiation of the effect of nonlinearity.

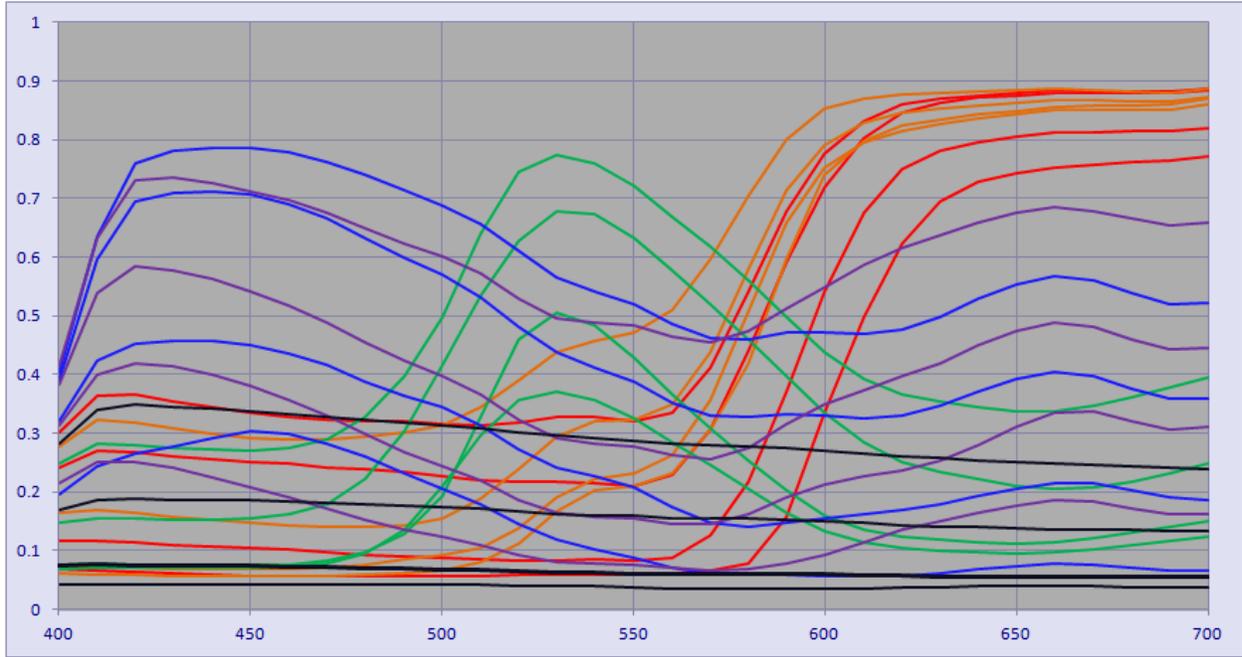


Figure 4 – Spectra of the Behr samples

Figure 5 shows the first derivatives of the 24 samples in the Behr paint sample set. It can be seen that the purple, blue and green samples provide negative derivatives to balance against the positive derivatives.

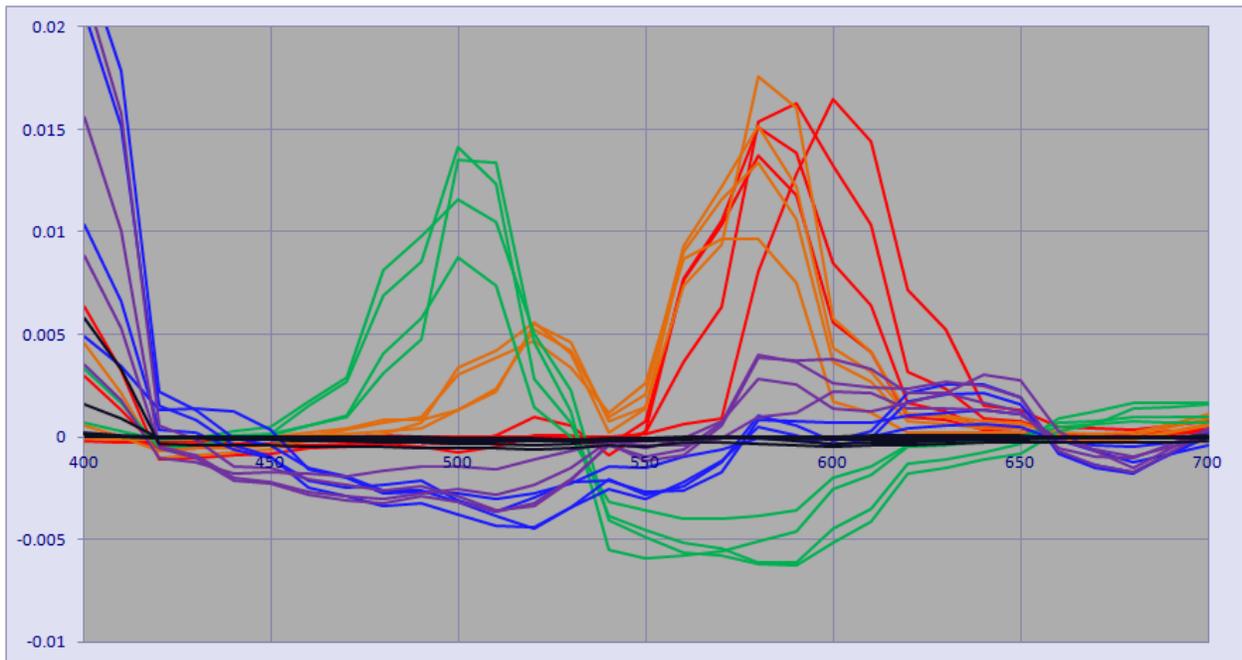


Figure 5 – First derivatives of the spectra of the Behr samples

Comparing this with the comments on Figure 1 (BCRA tiles), where a substantial portion of the spectrum did not have the balancing effect of negative and positive derivatives, this sample set provides good coverage throughout the critical part of the spectrum from 450 nm to 650 nm. This is likely the explanation for the Behr sample set's good performance at standardization.

It would appear that there are some samples that could safely be omitted. Blue and purple have quite similar derivatives. It is likely that elimination of one or the other would not impact the results greatly. Orange may similarly be providing redundant information over the red samples. From this diagram, it would appear that one sample could be selected from each color family without loss of information about the first derivative.

Conclusions

This paper tested various methods for standardizing the measurements from one spectrophotometer to better agree with those of another. Four mathematical models were tested, and four sample sets were tested in standardizing of two instruments to a third.

There were several notable conclusions. The first, and perhaps most important, is that standardization with a seemingly reasonable set of samples and a seemingly reasonable underlying mathematical model can be a worthless endeavor, and can often significantly worsen intra-model agreement.

Second, these tests suggest that nonlinearity may be an appreciable source of disagreement between spectrophotometers.

Third, it was found that the BCRA series II set of tiles is lacking in that it does not provide for reliable differentiation between errors in nonlinearity and wavelength shift. An alternative set of paint samples were found to outperform the BCRA tiles for standardization of one instrument to another. This opens up the possibility that, this sample set or a derivation thereof could be used for standardization. The samples are not as sturdy as the BCRA tiles, but may be acceptable when, for example, the two instruments are in the same room.

Fourth, an explanation has been given for why the paint samples performed better, having to do with the paucity of negative first derivatives in the BCRA set. This suggests that the addition of a few well chosen tiles to the BCRA set may improve their ability to standardize one instrument to another.

Acknowledgements

The author would like to thank several people for their help with this paper. Danny Rich provided encouragement and added some history. Harold Van Aken gave me an appreciation for the effects of lateral diffusion as a source of disagreement. Steve Lachajewski is responsible for the idea that negative and positive derivatives are essential. Jeff Bast did an excellent job of proofreading this paper, and helped assemble information for Table 1.

Bibliography

Berns, R.S., and L. Reniff

1997 "An Abridged Technique to Diagnose Spectrophotometric Errors", Color Research and Application, 22(1), 51(10)

Butts, Ken, Mike Brill and Alan Ingleson

2006 "Reflectance in Perspective - Will Instrument Profiling Give Me Better Measurements?", American Association of Textile Chemists and Colorists, 2006 conference

Chung, Sidney Y., K.M. Sin, and John H. Xin

2002 "Comprehensive comparison between different mathematical models for inter-instrument agreement of reflectance spectrophotometers", Proc. SPIE 4421, 789

Dolezalek, Fred

2005 "Interinstrument agreement improvement", Spectrocolorimeters, TC130

Hagen, Eddy

2008, "VIGC study on spectrophotometers reveals: instrument accuracy can be a nightmare", Oct 10, 2008, <http://www.ifra.com/website/news.nsf/wuis/7D7D549E8B21055CC12574C0004865FC?OpenDocument&0&>

ICC

- 2006 "Precision and Bias of Spectrocolorimeters", ICC white paper 22
- ISO
- ISO 12647-2, 3, and 4
- Nussbaum, Peter, Aditya Sole, Jon Y. Hardeberg
- 2009 "Consequences of Using a Number of Different Color Measurement Instruments in a Color Managed Printing Workflow", TAGA 2009
- Nussbaum, Peter, Jon Y. Hardeber, and Fritz Albrechtsen
- 2011 "Regression based characterization of color measurement instruments in printing application", SPIE Color Imaging XVI
- Radencic, Greg, Eric Neumann, and Dr. Mark Bohan
- 2008 "Spectrophotometer inter-instrument agreement on the color measured from reference and printed samples", TAGA 2008
- Rich, Danny
- 2004 "Graphic technology — Improving the inter-instrument agreement of spectrocolorimeters", CGATS white paper, January 2004
- 2005 "The effects of fluorescence in the characterization of imaging media", CIE Division 8 meetings, July 2005
- Rich, D. and D. Martin
- 1999 "Improved model for improving inter-instrument agreement of spectrocolorimeters", Analytical Chemica Acta, 380, 263-276, (1999)
- Robertson, A. R.
- 1986 "Diagnostic Performance Evaluation of Spectrophotometers", presented at Advances in Standards and Methodology in Spectrophotometry, Oxford, England
- Seymour, John
- 2011, "How well can you expect two spectros to agree?", CGATS N 1273, Nov. 4, 2011
- Spooner, David L.
- 1991, "Translucent blurring errors in small area reflectance spectrophotometer and densitometer measurements", TAGA 1991
- Van Aken, Harold, and Ronald Anderson
- 2000 "Method for maintaining uniformity among color measuring instruments", US patent 6,043,894
- 2003 "Method for maintaining uniformity among color measuring instruments", US patent 6,559,944
- Van Aken, Harold, Shreyance Rai, Mark Lindsay, Richard Knapp
- 2006, "System and method for transforming color measurement data", US Patent 7,116,336
- Verrill, J F, P C Knee, and A R Hanson
- 1997 "Study of improved methods for absolute colorimetry", NPL report QM 130, Feb 1997
- Williams, Andy
- 2007 "Inter-instrument agreement in colour and density measurement", IFRA special report, May 2007
- Wyble, D. and D. C. Rich
- 2007 "Evaluation of methods for verifying the performance of color-measuring instruments. Part 2: Inter-instrument reproducibility", Color Research and Application, 32, (3), 176-194

Appendix 1: linear regression including offset, gain, and wavelength shift

Standardization of SpectroEye to 939

	Test set				
Regression set	BCRA	Pantone primaries	Pantone ramps	Behr ramps	Improvement
Before standardization	0.35 (1.84)	0.66 (1.69)	0.49 (1.60)	0.63 (1.72)	
BCRA	0.56 (2.30)	0.47 (2.27)	0.52 (1.42)	0.56 (1.40)	1% (-8%)
Pantone primaries	0.77 (2.02)	0.49 (2.24)	0.79 (1.57)	0.99 (1.57)	-42% (-8%)
Pantone ramps	0.79 (1.56)	0.34 (1.57)	0.34 (0.79)	0.84 (1.27)	-8% (24%)
Behr ramps	0.59 (0.95)	0.54 (1.04)	0.46 (1.18)	0.26 (0.50)	13% (46%)

Standardization of SPM50 to 939

	Test set				
Regression set	BCRA	Pantone primaries	Pantone ramps	Behr ramps	Improvement
Before standardization	0.71 (1.26)	0.86 (2.00)	0.90 (1.73)	0.46 (0.93)	
BCRA	0.29 (0.86)	0.69 (2.61)	0.63 (1.18)	0.58 (1.29)	26% (0%)
Pantone primaries	1.21 (2.05)	0.73 (2.34)	1.04 (1.75)	0.72 (1.44)	-25% (-28%)
Pantone ramps	0.74 (1.36)	0.58 (2.19)	0.39 (0.98)	0.94 (1.28)	9% (2%)
Behr ramps	0.68 (0.88)	0.29 (0.61)	0.63 (0.89)	0.19 (0.38)	39% (53%)

Appendix 2: linear regression including offset, gain, shift, and bandpass correction.

Standardization of SpectroEye to 939

	Test set				
Regression set	BCRA	Pantone primaries	Pantone ramps	Behr ramps	Improvement
Before	0.35	0.66	0.49	0.63	

standardization	(1.84)	(1.69)	(1.60)	(1.72)	
BCRA	0.44 (0.85)	0.62 (1.48)	0.51 (1.56)	0.57 (2.30)	0% (10%)
Pantone primaries	1.36 (3.87)	0.52 (1.44)	0.68 (1.23)	1.21 (2.25)	-77% (-28%)
Pantone ramps	1.06 (2.58)	0.35 (0.77)	0.25 (0.56)	1.22 (1.69)	-34% (18%)
Behr ramps	0.88 (1.85)	0.88 (1.22)	0.70 (1.06)	0.15 (0.41)	-22% (34%)

Standardization of SPM50 to 939

	Test set				
Regression set	BCRA	Pantone primaries	Pantone ramps	Behr ramps	Improvement
Before standardization	0.71 (1.26)	0.86 (2.00)	0.90 (1.73)	0.46 (0.93)	
BCRA	0.36 (1.49)	0.71 (1.62)	0.84 (1.37)	0.63 (1.48)	13% (-1%)
Pantone primaries	1.83 (4.01)	0.63 (1.36)	0.51 (2.03)	0.93 (1.41)	-33% (-49%)
Pantone ramps	1.13 (2.08)	0.49 (1.27)	0.25 (0.79)	1.03 (1.84)	2% (-1%)
Behr ramps	0.69 (1.10)	0.56 (0.80)	0.84 (1.23)	0.13 (0.34)	24% (41%)

Appendix 3: linear regression including offset, gain, nonlinearity, and shift.

Standardization of SpectroEye to 939

	Test set				
Regression set	BCRA	Pantone primaries	Pantone ramps	Behr ramps	Improvement
Before standardization	0.35 (1.84)	0.66 (1.69)	0.49 (1.60)	0.63 (1.72)	
BCRA	0.41 (0.95)	0.53 (1.80)	0.50 (1.31)	0.60 (2.76)	4% (0%)
Pantone primaries	0.87 (1.95)	0.64 (1.58)	0.80 (1.57)	1.30 (4.39)	-69% (-38%)
Pantone ramps	0.61 (1.08)	0.69 (1.89)	0.29 (0.66)	0.76 (3.20)	-10% (0%)
Behr ramps	0.50 (0.81)	0.66 (1.26)	0.44 (1.07)	0.11 (0.26)	20% (50%)

Standardization of SPM50 to 939

	Test set				
Regression set	BCRA	Pantone primaries	Pantone ramps	Behr ramps	Improvement
Before standardization	0.71 (1.26)	0.86 (2.00)	0.90 (1.73)	0.46 (0.93)	
BCRA	0.37 (1.22)	0.91 (2.77)	0.56 (1.08)	0.55 (1.13)	19% (-5%)
Pantone primaries	1.40 (2.41)	0.57 (1.89)	1.02 (2.06)	0.52 (2.05)	-20% (-42%)
Pantone ramps	0.72 (1.34)	0.73 (1.80)	0.39 (0.91)	0.73 (1.24)	12% (10%)
Behr ramps	0.69 (0.92)	0.58 (1.02)	0.60 (0.87)	0.20 (0.32)	30% (47%)

Appendix 4: linear regression including offset, gain, nonlinearity, shift, and bandpass.

Standardization of SpectroEye to 939

	Test set				
Regression set	BCRA	Pantone primaries	Pantone ramps	Behr ramps	Improvement
Before standardization	0.35 (1.84)	0.66 (1.69)	0.49 (1.60)	0.63 (1.72)	
BCRA	0.42 (0.80)	0.66 (1.71)	0.50 (1.55)	0.50 (3.27)	2% (-7%)
Pantone primaries	1.36 (3.90)	0.57 (0.88)	0.62 (1.31)	1.34 (3.41)	-82% (-38%)
Pantone ramps	0.88 (2.53)	0.68 (1.17)	0.23 (0.57)	1.01 (3.20)	-31% (-9%)
Behr ramps	0.85 (1.85)	0.86 (1.26)	0.68 (1.09)	0.06 (0.18)	-15% (36%)

Standardization of SPM50 to 939

	Test set				
Regression set	BCRA	Pantone primaries	Pantone ramps	Behr ramps	Improvement
Before standardization	0.71 (1.26)	0.86 (2.00)	0.90 (1.73)	0.46 (0.93)	
BCRA	0.27 (0.75)	0.83 (1.46)	0.81 (1.35)	0.61 (1.25)	14% (19%)
Pantone primaries	1.60 (3.71)	0.62 (1.03)	0.52 (2.00)	0.94 (1.44)	-26% (-38%)
Pantone ramps	0.98 (1.87)	0.61 (1.09)	0.26 (0.72)	1.16 (2.06)	-2% (3%)

Behr ramps	0.72 (0.92)	0.63 (0.89)	0.77 (1.16)	0.12 (0.24)	24% (46%)
------------	----------------	----------------	----------------	----------------	--------------