How well can you expect two spectros to agree?

John Seymour, Nov 4, 2011

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 should 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.

In this study, I have compiled the results from seven different studies on how well one color measurement instrument agrees with another. The results are consistent, and perhaps surprising.

Summary of results

In a special report from IFRA, Andy Williams laid out the 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.*

All the research I have found unfortunately shows that this is an unreasonable expectation.

Peter Nussbaum (of Gjøvik University College, Norway) has recently published two papers where he compared measurement of various color measurement devices of different make on a set of BCRA tiles against reference values supplied by a national standards organization. Quoting from his 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 had errors over 2.0 ΔE_{ab} on at least one of the tiles.

Greg Radencic of PIA led a similar study that was presented at TAGA 2009. He borrowed eight spectrophotometers directly from four manufacturers, and compared measurements on a Lab-Ref card. Instruments were compared against the median measurement of all instruments. Here is what he 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 reports color measurement differences of up to 10 ΔE_{ab}.

The ICC published a white paper in 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 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. Eddy Hagen of the VIGC recently took this test out into the field, testing twenty instruments that are currently in use. He 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}.*

In 2006, X-Rite and GretagMacBeth merged, 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. The first part of this paper described the degree to which their new stable 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 ranged from 0.27 ΔE_{ab} to 1.08 ΔE_{ab}. Under the IFRA recommendation, only this one particular pairing of instruments can be considered acceptable, and then only just barely, and then only under the laboratory conditions for 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 ranges 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.

What to do?

First and foremost, expectations must be set properly. One should not expect two spectrophotometers 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 agreement of 1.0 ΔE_{ab} is not uncommon.

This realignment to reality has an implication when it comes to setting reasonable tolerances for the color of a printed work. 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%.

As an example, let us say that the measurement devices of a print buyer and of a printer differ by 0.5 ΔE_{ab} on a given critical color. Given the results of these studies, this is an optimistic expectation. With this amount of disagreement, a tolerance window smaller than 1.6 ΔE_{ab} is counterproductive. If the particular critical color is one where a discrepancy of 1.0 ΔE_{ab} is more typical, a tolerance window greater than 3.3 ΔE_{ab} is necessary.

One effort that will help manage expectations is for the industry to use a common terminology for specifications. Currently, the meaning of the phrase "inter-instrument agreement" could mean either the degree to which two instruments of the same make and model agree, or it could mean the degree to which two instruments from different manufacturers agree. There is not a common agreement of whether measurements should be averaged before comparing two instruments. There is a confusing plethora of terms that are used to describe specifications: repeatability, accuracy, precision, inter-instrument agreement, and intra-instrument agreement. When specifications and tolerances are given, it is often not clear whether average performance is being described or worst case. Neither is it clarified what samples are to be used.

To combat this, I have introduced a new work item for CGATS to develop a document that defines a common language.

The evolution of standards and standards labs will also help with this effort. As recently as 1997 (Verrill et al.), the national laboratories that serve as the official standard for color measurement showed surprisingly large differences in measurements. Only half of the measurements between these labs agreed to within 0.5 ΔE_{ab}. Some discrepancies were as large as 2.0 ΔE_{ab}. This has improved, and as a result, improvement has trickled down.

One recent change to the ISO 13655 standard has to do with the measurement of fluorescent materials, such as paper with UV brighteners. Danny Rich presented results of a study on the effect of fluorescence to CIE Division 8 in 2005. In this study he found that the difference between including the UV component and excluding the UV component can be on the order of 3.0 ΔE_{ab} for measurements of solids and can be on the order of 6.0 ΔE_{ab} for halftones. This has led to clarifying ISO 13655 so as to more completely specify the UV component of the lighting in a spectrophotometer.

Various researchers (Chung et al. 2002, Rich 2004, and Nussbaum 2011) have described ways to improve the agreement of two instruments. The second part of the X-Rite white paper shows results from their efforts (described as XRGA) to bring all their instruments into better agreement. While their effort has not improved the best case agreement, their worst case pairings of instruments is now 0.60 ΔE_{ab} on average, and the worst 95th percentile is now 1.36 ΔE_{ab}.

I remain cautiously optimistic about the use of post-facto calibration techniques such as these. I hope to see the XRGA results repeated outside of X-Rite, and I hope to see tests include specialty colors. My concern was raised in a paper from 1997, where I showed that over-zealous calibration will make the agreement between instruments worse when the calibration set differs from the samples measured in real life.

Conclusions

Currently, there is a disconnect between the level that color measurement devices agree, and the expectations of how well they are expected to agree. Education and consistent terminology can help close this gap. In addition, efforts are underway to continuously improve the level of agreement between instruments.

Bibliography

Chung, Sidney Y., K.M. Sin, and John H. Xin, *Comprehensive comparison between different mathematical models for inter-instrument agreement of reflectance spectrophotometers*, Proc. SPIE 4421, 789 (2002)

Dolezalek, Fred, *Interinstrument agreement improvement, Spectrocolorimeters*, TC130 2005

Hagen, Eddy, *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 white paper 22, *Precision and Bias of Spectrocolorimeters*, 2006

Nussbaum, Peter, Jon Y. Hardeber, and Fritz Albregtsen, *Regression based characterization of color measurement instruments in printing applications*, SPIE Color Imaging XVI, 2011

Nussbaum, Peter, Aditya Sole, Jon Y. Hardeberg, *Consequences of Using a Number of Different Color Measurement Instruments in a Color Managed Printing Workflow*, TAGA 2009

Radencic, Greg, Eric Neumann, and Dr. Mark Bohan, *Spectrophotometer inter-instrument agreement on the color measured from reference and printed samples*, TAGA 2008

Rich, Danny, *The effects of fluorescence in the characterization of imaging media*, CIE Division 8 meetings, July 2005

Rich, Danny, *Graphic technology — Improving the inter-instrument agreement of spectrocolorimeters*, CGATS white paper, January 2004

Seymour, John, *Why do color transforms work?*, Proc. SPIE Vol. 3018, 1997

Verrill, J F, P C Knee, and A R Hanson, *Study of improved methods for absolute colorimetry*, NPL report QM 130, Feb 1997

Williams, Andy, *Inter-instrument agreement in colour and density measurement*, IFRA special report, May 2007

X-RITE, *The new X-Rite standard for graphic arts (XRGA)*, CGATS N1163, 2011