Can Colorimetry Predict a Match Between Proof and Press Sheet?

John Seymour

Keywords

Appearance, CIE, Colorimetry, correlation, pressproof

Abstract

In the 2006 IPA (The Association of Graphic Solutions Providers) proofing roundup, a group of judges scored the degree of match between press sheets and the proofs printed to simulate these sheets. In addition, colorimetric measurements were made of test targets printed on those same press sheets and similarly printed proofs. This data, objective and subjective, provides an excellent opportunity to investigate the correlation between objective and subjective evaluations of images.

The first step in evaluating the data was to check how well the judges agreed amongst themselves. Obviously if their agreement is poor, it is fruitless to even consider predicting the visual assessment.

The ΔE color differences were computed for each of the 1,617 patches in the test targets for each of the 22 proofing systems. Various techniques were used to distill these color differences down to a single objective quality number for each of the proofing systems, including standard descriptive statistics, averages of various collections of patches. Each of these various distillations of ΔE_{ab} values was then correlated against the judges' scorings.

Description of the IPA Roundup

In the past, vendors participating in the IPA Roundup had been given printed images to match. In the 2007 Roundup, the vendors were only given test targets to adjust their proofing. Most of the vendors received either "very good" or "excellent" ratings from the judges, demonstrating that it is possible to successfully proof by the numbers.

In the first step of the roundup, there was a press run that printed a collection of test images along with a test target with 1,617 patches. The test targets and test images were cut into separate sheets.

Each of the vendors was provided with one of the printed test targets. A scanning spectrophotometer was used by the vendor in order to profile their proofing system. The vendor then printed a proof of the test target, as well as a proof of the sheet with a collection of test images.

The proof and the press sheet of test images were then shown to a group of 17 judges. These judges provided scorings based on visual assessment of the degree of match between the proof and press sheet.

Figure 1 – The test image shown to the judges

Meanwhile, the proof of the test target of 1,617 patches was measured with a spectrophotometer.

Thus, there were subjective scorings of the match between press sheet and proof and there were objective measurements indicative of that same match.
For more details about the roundup, refer to Sharma, 2006a, 2006b.

The Subjective Data

Each of seventeen judges scored each of 32 proofing systems on the quality of the match. Scores were given between 1 and 10 for each of the following questions:

1. Color hue accuracy. Correct rendering of individual hues ("selective" colors).
2. Gray Balance: Accurate rendering of neutral and grayscale images.
3. Flesh tone Reproduction. Correct rendering of flesh tone color and smoothness.
4. Correct rendering of shape, detail, and tonal transitions.

Thus, each judge gave 128 separate ratings. How well do the judge's ratings agree with each other?

Agreement among the Judges

It would be hard to interpret a comparison of each judge to every other judge. This would yield 136 correlation coefficients. There would simply be too many numbers to assess. Instead, I chose to compare the ratings of each judge against the average of the other sixteen judges.

Figure 2 – Agreement among the judges

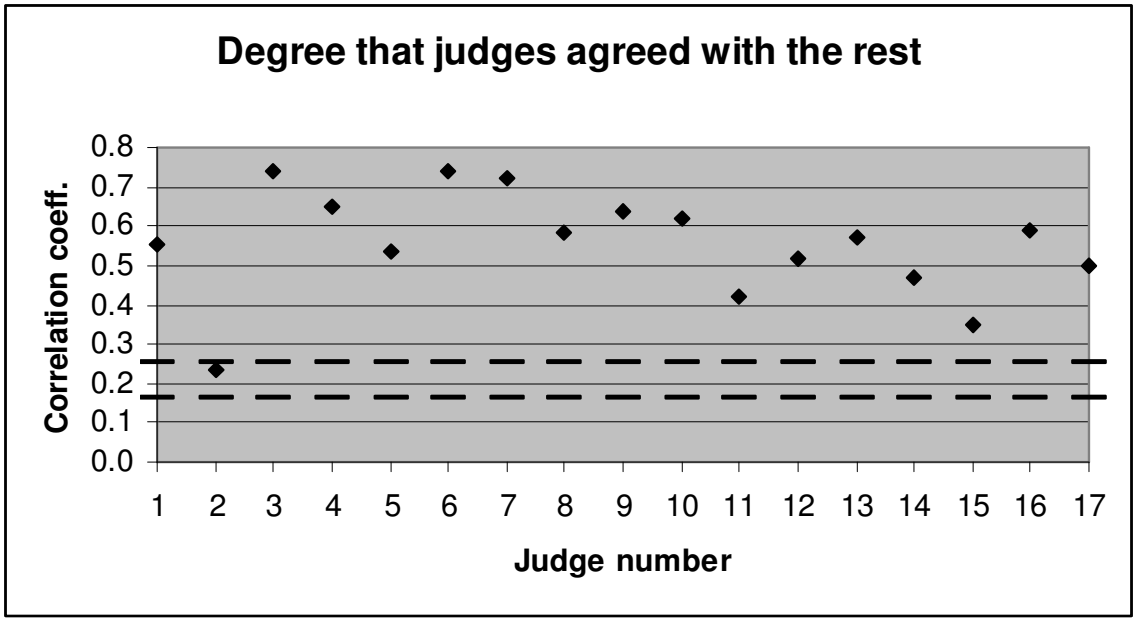

Figure 2 shows the correlation coefficient of each of the judges when compared against the average of the other judges. The lower dotted line is the 10% confidence level. There is a 10% chance that a blindfolded judge rolling a set of dice could show this high of a correlation to the rest of the judges. The dotted line above that is the 1% confidence level. There is only a 1% chance that random data could match this well.

With the exception of Judge 2, all the judges agreed with one another with much better than a 1% confidence level. The correlation coefficient for this judge is over three standard deviation units outside the correlation coefficients of the other 16 judges. Judge 2 was clearly an outlier. I decided to exclude Judge 2.

Note: There were three individual cases where, for whatever reason, data was not entered into the spreadsheet for a given judge and vendor combination. For example, Judge 5 had no scores for Vendor 24. In these cases, I used the average over the other judges for that particular vendor.

I decided to use the average of the scores from the sixteen remaining judges as my best estimate of the subjective scores for each vendor.

Estimate of the Standard Error

How reliable is this estimate of the subjective scores?

With Judge 2 excluded, I computed the standard deviation of scores for each of the 128 questions. The average of these 128 standard deviations was calculated to be 1.32. This number is an estimate of the reliability of any single judges' scores. Since sixteen judges were averaged, the reliability of the average is a factor of four better, four being the square root of sixteen. The standard error is 0.33, which is to say (assuming Gaussian distribution of errors) that there is a 68% chance that the "true" subjective score for any question is within 0.33 of the estimate I have arrived at through averaging.

The overall standard deviation of the averages, which is to say, the standard deviation of the 128 averaged scores, is a measure of how much variability there is from one vendor's scores to the next. This standard deviation is 0.90.

Thus, it is expected that about 37% of the differences between vendors can be attributed to variability among the judges.

Conclusion of Subjective Results

My conclusion is that, with the exception of one judge, the correlation between the scores from the different judges is quite good. The RoundUP committee who conducted the roundup are to be commended for a well run experiment.

I have every reason to believe that the average of the sixteen remaining judges is a very reliable measure of the degree of visual match between each of the sets of sheets. The average of the sixteen judges will be used as the subjective measure of agreement between proof and press sheet for the rest of this analysis. For each vendor, I thus had an average score for each of the four questions, and also an average of the four questions.

The Objective Data

Ten of the proofing systems were soft proofing systems. Since the colorimetric data from the monitors is somewhat less reliable, I have considered in this analysis only the measurements from the twenty-two hard copy proofing systems.

The colorimetric data consisted of L*a*b* measurements of 1,617 patches in an IT8.7/4 test target, measured on a press sheet and also on a proof sheet. Two measurements of each sheet were made. I averaged the replicate measurements for the twenty-two press sheets and for the twenty-two proof sheets.

I performed only a rudimentary check for outliers. The ΔE was computed between each pair of replicate measurements.
NOTE: Unless otherwise noted, I used $\Delta E_{76.}$, entirely for ease of computation.

The mean ΔE over all pairs of measurements was tiny (0.066 ΔE), but the maximum (1.01 ΔE) is suspect. Unfortunately, with only two measurements made, it is not possible to decide which of the two measurements should be discarded. I opted to take the average of the two as being the correct answer, without discarding any outliers.

General Descriptive Statistics

I next computed the color difference between the test targets printed on the press and the test targets printed by the proofing systems. This resulted in 1,617 ΔE values. We need some way to distill this collection of color errors down to a manageable size.

I distilled the 1,617 ΔE values down to the following seven statistics: mean ΔE, standard deviation of ΔEs, median ΔE, third quartile ΔE, 90^{th} percentile ΔE, 99^{th} percentile ΔE, and the maximum ΔE. The 90^{th} percentile is that ΔE value that is larger than the error of all but 162 of the patches. The 99^{th} percentile is that ΔE that exceeds all but the worst 16 of the patches.

Comparison among the Seven Statistics

Are all seven of these statistics really needed, or do some of them tell pretty much the same thing about the underlying collections of errors? One might expect, for example, that the mean and the median for the errors would track

each other fairly well. If one vendor has a higher mean than the other, one would expect the median also to be higher. This is indeed the case.

In Table 1, I show the correlation coefficient between the seven statistics. It can be seen that there is an excellent correlation between all pairings of mean ΔE, median ΔE, third quartile and 90th percentile. In other words, for this data set, it makes no difference whether we describe the collection of errors in terms of mean, median, third quartile, or 90th percentile. Since the mean has the highest correlations to the others, I have chosen to use this as a measure of the overall color error for a given vendor.

Table 1 also shows that the 99th percentile, the maximum and the standard deviation from another group of statistics that are well correlated. That is to say, they provide essentially the same information amongst themselves.

This group has a comparably smaller correlation to the first group of statistics. This suggests that there are at least two independent measures of the color errors for a given vendor: one for the overall error, and the other expressing something about the extreme errors. Whatever information is in this group of statistics, it is apparently distinct from the information provided by the mean. Knowing the mean does not tell a great deal about the 99th percentile.

I chose the 99th percentile as the measure for the extreme errors. I did not choose the maximum since the expected value of the maximum depends on the sample size (a target with less patches would be expected to have a smaller maximum error even if the overall statistics were the same.

Table 5 – Coated stocks in the web offset experiment

	Mean	Stdev	Median	Third Q	90th %	99th %
Mean						
Stdev	0.673					
Median	**0.978**	0.509				
Third Q	**0.992**	0.591	**0.989**			
90th %	**0.974**	0.800	**0.909**	**0.952**		
99th %	0.551	***0.975***	0.375	0.455	0.685	
Max	0.388	***0.903***	0.212	0.285	0.525	***0.934***

The mean and the 99th percentile of the ΔEs will be used as the objective measures of agreement between proof and press sheet.

Note: This sort of problem is common in the industry. One starts with a large collection of ΔE values, and we wish to reduce this large collection of numbers to a tractable set. Various approaches have been tried. The mean and standard deviation have been suggested, but also maligned, since this may lead one to believe that 68% of a data set is within one standard deviation of the mean. It has been pointed out that this assumption is only true for normal data, and ΔE values are known to be poorly described by the normal curve.

Another approach is to use the cumulative distribution function. I first saw this applied to ΔE data by Michael Rodriguez of R.R. Donnelly. David McDowell has been a strong proponent of this tool.

This analysis suggests that, *at least for the purposes of this data set*, the use of mean and 99th percentile may adequately describe ΔE data.

Agreement between Subjective and Objective Data

Comparison against Overall Statistics

I computed the correlation coefficient between judges' scores and the descriptive statistics of the colorimetric differences (see Table 2). The overall average vendor rating ("Ave") was correlated against, as well as the average for each of the four questions.

Table 2 – Agreement of ΔE values with judges' scores

	Mean ΔE	99th %
Ave	-0.327	-0.191
Q1	-0.262	-0.195
Q2	-0.260	-0.207
Q3	-0.190	-0.156
Q4	-0.422	-0.099

The fact that all these correlations are all negative is to be expected. A pair of images with a higher ΔE (that is, larger color error) should result in a lower score from the judges. A negative correlation means that one goes up while the other goes down.

However, the magnitudes of the correlations almost all fall just short of being statistically significant. A value of -0.200 is not very significant. A blind judge

with a set of dice would be expected to get a correlation coefficient of -0.200 or better about once out of every five tries.

Looking at the individual correlations, it can be seen that the most statistically significant correlation is between the score for question 4 and the mean ΔE. This correlation coefficient is close enough to the cut off to be called significant at the 5% level.

Actually, this is the only statistically significant correlation in the table. The table below of significance levels for $n = 22$, is taken from Snedecor (1980).

Table 3 – Levels of statistical significance for n = 22

Significance level	*r*
10%	0.360
5%	0.423
2%	0.492
1%	0.537

While this may sound like wonderful news, it needs to be tempered with the following observation. A correlation coefficient of -0.422 means that ΔE values explain about 9% of the of the judge's observations, since

$$1-\sqrt{1-0.422^2}=0.093\,.$$

Question 4 asked the judges to assess "Correct rendering of shape, detail, and tonal transitions." The high correlation to this question is unexpected, since the wording on question 4 is the least specific to "color". One might have expected question 1, which asks about color hue accuracy, to be more closely correlated.

Taking Popularity into Account

It is not surprising that the 99th percentile does not show a strong correlation to visual matches. The 99th percentile is decided upon by only pixel of the 1,617 color patches. There is no way of knowing whether any of these sixteen worst case CMYK combinations were even printed on the page of images that the judges assessed. The more popular a CMYK value is in the image that was assessed, the more we would expect a color error to influence judges' scores.

Software was written to determine the popularity of each of the 1,617 CMYK values in the visually assessed images. NOTE: Ken Elsman of Global Graphics was kind enough to give me a beta version of the TATOO program to perform this analysis; however, another person in my work group, Adam Nelson, had already written his software to perform this. The CMYK image files from the image sheet were first decimated down to 100 DPI. This gave an image with roughly three million pixels.

Then each of the CMYK values in this image was compared against each of the 1,617 CMYK combinations in the test target to determine which patch was closest in CMYK value. Essentially, this is a paint-by-number process, using the 1,617 patches as the palette. In this way, a histogram was produced showing the number of occurrences of each of the patches in the assessed image.

The utility of this analysis is predicated on the assumption that the color error in the image for a given pixel can be estimated by looking at the color error of the closest patch from the test target.

To give a few highlights from the histogram, the most popular patch was white paper, accounting for almost 18% of the pixels. The second most popular CMYK combination was a dark blue (CMYK = 100, 100, 70, 60) with 3.2% of the pixels. A dark gray (CMYK = 70, 70, 70, 40) came in second with 3%. The most popular twenty CMYK values accounted for over half the pixels in the image.

I correlated the judge's assessments against a number of statistics derived using these popularities. The assessments were correlated against:

1. The ΔE for paper alone,
2. The ΔE for the two most popular CMYK values, not including paper,
3. The ΔE for the four most popular CMYK values, not including paper,
4. The ΔE for the eight most popular CMYK values, not including paper,
5. The ΔE for the sixteen most popular CMYK values, not including paper,
6. The ΔE for the 32 most popular CMYK values, not including paper,
7. The ΔE for the 64 most popular CMYK values, not including paper,
8. A weighted average of the ΔEs of all the patches, not including paper, where the weights were the popularities of each CMYK value, and
9. A weighted average of the ΔEs of all the patches, including paper, where the weights were the popularities of each CMYK value.

Another statistic I would have liked to correlate against is not the average of the most popular patches, but the worst case ΔE over some set of popular patches.

The table below shows the correlation coefficients: The three levels of shading highlight those correlations which are significant at the 10%, 5%, and 1% levels (0.360, 0.422, and 0.537, respectively).

Table 4 - Agreement of popular ΔE values with judges' scores

	% of pixels	Overall score	Q1	Q2	Q3	Q4
Paper	17.7%	-0.399	-0.300	-0.249	-0.430	-0.346
2 most pop	3.2%	0.131	0.133	0.266	0.031	0.029
4 most	11.3%	-0.013	0.014	0.132	-0.039	-0.153
8 most	18.4%	-0.151	-0.123	0.021	-0.123	-0.305
16 most	29.7%	-0.254	-0.236	-0.065	-0.184	-0.407
32 most	41.2%	-0.369	-0.348	-0.208	-0.213	-0.534
64 most	52.0%	-0.400	-0.379	-0.223	-0.247	-0.560
W, no p	83.3%	-0.329	-0.301	-0.178	-0.199	-0.478
Weight	100.0%	-0.402	-0.342	-0.231	-0.319	-0.492

The first perhaps surprising finding is that the color error for paper correlated well with the judges responses, especially against question 3. The ΔE values for paper were in some cases somewhat sizeable, with the average being 0.83 ΔE, and four of the 22 systems had greater than 1.5 ΔE. I verified that these measurements were real by looking at replicate measurements of the proof.

A suggestion to the proofing system vendors is that they could improve their score by simply selecting paper that better matches the press sheet. The importance of paper is often underappreciated. To quote (Wales, 2005),

> Paper contributes to color reproduction in saturated tones to the degree it reflects light, in the mid tones to the degree it scatters light, in the highlight tones to the degree it deviates from color-neutral, and in all tones with its unique dot gain.

In a follow-up conversation with Trish, I learned that it is not possible, in general, to use a typical web offset stock in a proofing device, since the goals for the chemistry are completely different. Web offset inks are oil based and ink jet inks are water based. She informed me of a company called "The Whole Proof" (TWP, Intl) which offers a process by which actual press paper can be conditioned for ink jet printing.

Before considering this a big issue, however, reducing the color error in paper by one ΔE could improve the overall score by only 0.2 points. The dependence on paper only represents 8% of the differences in the scoring of the judges. (This

number (8%) was computed from the correlation coefficient by using Equation 1 in Appendix A.)

It may be fruitful to convert all the CIELAB values into paper relative values and repeat the correlation analysis. This will *perhaps* allow us to separate the effect of the paper from the effect of the ink on the paper.

Question 4 once again showed the best correlations. The best correlation was between this question and the average of the 64 most popular. With this correlation (r = -0.560), 17% of the differences in judges' scores are explained by this particular measurement of color error.

Evaluation of Gray Balance

Question 2 asked about gray balance. One would expect there to be a good correlation between some measure of the fidelity of gray patches on the proof and the scores for this question. To investigate this, I chose three different methods to segregate patches that were to be considered "gray". The ΔE for the segregated patches were then averaged for each vendor, and this average was correlated against the judges' scores.

In the first method, the "gray average" statistic is the average of the ΔE values for the 100 patches with the smallest chroma. This corresponds roughly to all patches where the C* value is less than 5.0. These 100 patches represented about 15% of the image.

Patches with low chroma may not necessarily be deemed "gray", since pure white and pure black have zero chroma. In the second method for segregating gray patches, patches were selected as gray if the L* value was between 40 and 60, and the chroma was less than 6. This yielded 19 patches and about 5.4% of the image.

A third method attempts to incorporate the printer's definition of gray balance, meaning areas where the three process colors are balanced. I selected the 64 patches in the test target where cyan, magenta and yellow tone values were all equal. This represented almost 36% of the image.

The "saturated average" statistic averages at the other end of the "spectrum". This is the average ΔE for the 100 patches with the highest chroma value. This corresponds roughly to patches with C* above 60.0.

Table 5 - Agreement of ΔE values of gray patches with judges' scores

	Gray ave	Near (50,0,0)	CMY bal	Sat. ave
Ave	-0.190	-0.438	-0.332	-0.228
Q1	-0.186	-0.447	-0.308	-0.199
Q2	-0.084	-0.156	-0.180	-0.218
Q3	-0.162	-0.260	-0.256	-0.077
Q4	-0.228	-0.702	-0.410	-0.313

It is surprising that the judges' answers to question 2 (the gray balance question) correlated poorest with all three measures of color accuracy of gray patches. *Of all the questions, the question about gray balance correlated poorest with all three measures of gray patches!*

Question 4, which so far has correlated best with other objective color error measurements, shows once again the highest correlation: -0.702. Roughly 30% of the differences in judges' scores can be explained by the error in the gray patches. For an experiment of this type, this is a quite good result.

Why do gray value measurements seem to correlate better with a question dealing with image quality in general? Dr. Hunt (1991) gives a clue that might explain the conundrum. "The sharpness of images depends much more on luminance than on chrominace content of the image…" It may be that getting the gray values correct will improve the sharpness. Then again, it is hard to imagine that a paltry 5.4% of the image can have that much effect on perceived sharpness.

Caveat: My selection of gray patches (for "near [50, 0, 0]) is admittedly somewhat arbitrary. I chose the cutoff values so as to optimize the correlation coefficient, so there is some fear of cherry picking. As a more extreme example of cherry picking, I could have chosen the single patch CMYK = (40, 40, 3, 3) and reported the fabulous correlation coefficient of -0.787. This is not so extraordinary, however. With 1617 patches to choose from, the expected best correlation with purely random data is ±0.662.

While I acknowledge the possibility that I may be picking cherries, the pixels averaged were not a widely disparate collection, but rather a clump of similar pixels. It is also expected that these colors should be important for matching images.

This is analogous to an eyewitness seeing someone in a green Jaguar perpetrating a crime. That in itself is not enough evidence to convict any particular owner of a green Jaguar. On the other hand, if it can be established

that a particular owner of a green Jaguar was seen in the neighborhood of the accident around the time of the crime, then the evidence is quite a bit more compelling.

Evaluation of Flesh Tones

The judges were given one question, question 3, which specifically asked about flesh tones. I chose two methods to segregate flesh tones. As before, the ΔE values for the segregated patches were averaged and correlated against the judges' scores.

The first method to select flesh tones was to identify patches that would qualify as flesh tone if seen in isolation. How to define these? The Macbeth Color Checker has two relevant patches, one called "light skin", and the other called "dark skin". The CIELAB values of these two patches were measured. All patches that were within 15 ΔE of a line between these two patches were identified as flesh tone for this method. This yielded 75 patches which cover 7.6% of the image.

The second method is perhaps a bit closer to what humans actually do. Ken Elsman used Photoshop to select all the flesh tones from the original CMYK image, by cutting out faces and hands (see Figure 3). An image with just these pixels was passed through the TATTO program and the resulting CMYK values were reported.

Figure 3 – Elsman's flesh tones

The resulting set of 195 patches included a wide range of values, ranging from very nearly white to very nearly black. These highlight and shadow pixels are interpreted as flesh tone by the human eye from the context of the image. The 195 patches cover 35% of the image.

Table 6 shows the correlations against judges' scores.

Table 6 - Agreement of ΔE values of flesh tone patches with judges' scores

	Elsman	Macbeth
Ave	-0.406	-0.365
Q1	-0.369	-0.336
Q2	-0.192	-0.095
Q3	-0.305	-0.255
Q4	-0.549	-0.600

Comparing the two sets of correlation values, the Elsman flesh tone set has marginally stronger correlations in four of the five cases, but not so much that this method of selecting flesh tones could be construed as being a clear winner. This is not to say that I have proven the two to be equivalent. This experiment is just not sensitive enough to distinguish between the ways of defining "flesh tone".

Once again, it is odd that this set of data did not correlate better with question 3. Question 4 once again matches best against colorimetric measurements.

Combining Sets of Patches

There were a number of sets of patches that were somewhat successful in predicting the judges' scores. Combining these sets of patches as a multi-variate regression may prove more successful in correlating subjective and objective measures.

When doing multivariate regression, it is important that the predictors, in this case the various averages of ΔE values, be uncorrelated. If there is a high correlation between them, then multivariate regression becomes mathematically unstable.

Many of the averages of ΔEs that were successful in this test are somewhat correlated, so the sets that could be used were limited. I chose the "near (50, 0,

0)" set and "paper". The following combination was a marginal improvement over the near (50, 0, 0) set by itself:

$$\Delta E_{ave} = 0.1\Delta E_{paper} + 0.9\Delta E_{gray}$$

This special averaging of the colorimetric differences had a correlation coefficient of -0.730, so that 32% of the variation is explained.

I think this is a long walk on a rickety pier for a short drink.

Limitations

While some of the results of this analysis have been quite good, there has been a recurring theme that ΔE statistics don't correlate well with the questions that we would expect them to correlate with. There is also the nagging issue that the answers to many of the questions did not correlate well with anything.

From discussions with a number of people, I have an even larger number of explanations for what could be done to improve the experiment.

Possibly the largest issue is that the judges were shown sheets with several images on them. Restricting the view to one image would help reduce confusion. Asking the judge about specific parts of the image (how does the sweater match?) may also help.

The types of color errors were not well controlled. Better results would be had if we were to decide ahead of time to darken flesh tones in one area, for example, or add some tint to the gray tones.

The questions perhaps were not interpreted as we would expect. Perhaps more careful wording or calibration of the judges beforehand could have tilted the correlations more toward the expected questions.

The color errors on the whole were small enough as to be in some cases near the limits of discernibility by human and machine. Larger color errors might improve the correlations.

No use was made in the analysis of the position of pixels. It is expected that contiguous pixels with color errors in the same general direction would be more discernible than those same errors scattered across the image.

I could have used ΔE_{2000} rather than ΔE_{94} to better approximate human perception of color differences. I confess that this is entirely due to laziness on my part. On the other hand, Figure 4 of Sharma (2006c) seems to hint at some sort of equivalence between the different color difference equations.

The use of 0/diffuse geometry may be a better approximation to the illumination in the light booth. I unfortunately do not have gloss measurements from all the proofs to suggest whether this might be a factor.

There is a potential metamerism issue in that the spectrophotometer was using idealized D50 filters for the calculation of XYZ, whereas the viewing booth used F8 lighting, which is an approximation to D50.

I could repeat all the calculations using paper relative L*a*b* values.

There is a tacit assumption throughout this analysis that color errors between the pairs of test targets will approximate the color errors of corresponding CMYK values in the images. The variability across and down the form in printing press sheets is often underappreciated. In Siljander (2001), a downsheet variation of 0.07D for a solid magenta patch caused by ink starvation was reported for two press tests. (See Graph 7, page 69.)

Ideally, I would have liked to directly compare colorimetric measurements of the images themselves. Unfortunately, the technology for colorimetric scanning without pesky metamerism issues is still on the horizon.

I understand from Trish Wales that the press sheets did have fluorescent whitening agents. This raises problems with the whole test, since 1) the proof sheets may or may not have FWAs, 2) the UV content of viewing booths is not well standardized, and 3) the vendor had the option of making measurements with or without the UV filter in the scanning spectrophotometer.

Conclusions

With the exception of one judge, the judges in the 2006 IPA roundup were in very good agreement among themselves. From this I conclude that the variables of the human comparison were well controlled.

Various means for distilling down the colorimetric data were tried. The following correlation coefficients were significant at the 1% level:

- The average ΔE for the 64 patches most predominant in the image when correlated against question 4
- The average ΔE for the 19 patches near gray when correlated against question 4.
- The average ΔE for patches representing flesh tone when correlated against question 4.

A few additional correlation coefficients are statistically significant at the 5% level, most notably:

- The overall ΔE of all the patches when correlated against question 4
- The ΔE of the paper when compared against question 3.

Surprisingly, the scores for question 4, having to do with "Correct rendering of shape, detail, and tonal transitions", showed the strongest correlation to colorimetric errors. Even more surprisingly, the scores on the gray balance question did not show strong correlations against averages of gray patches, and the scores on the flesh tone question did not show strong correlations against averages of flesh tone patches.

Larry Warter offered the following succinct explanation of this enigma:

> The viewing experts know when things match but they can't explain why. This is even more true when they see images that don't match.

That said, the average ΔEs, computed through various distillations, could only account for about one-third of the differences between vendors, and then really only on one of the four questions. Roughly one-third of the variation may be due to noise in our measurement of the subjective scores, which is to say, disagreements among the judges. The final third is from unknown sources.

Given the number of limitations, it is expected that a series of experiments aimed at specifically addressing this question would prove instructive. CGATS subcommittee 3, task force 1 will be working in this direction.

Acknowledgements

This work was inspired by CGATS subcommittee 3, task force 1, whose charter is the "*Development of a method based on colorimetric measurements which will estimate the probability that hardcopy images reproduced by single or multiple systems, using identical input, will appear similar to the typical human observer.*"

I would like to thank the members of CGATS for their critical review. I am grateful for suggestions from Dave McDowell as to how to proceed with the analysis, to Richard Goodman for encouragement and perspective, and to Ken Elsman, Larry Warter, Bill Birkett, Danny Rich, John Daugherty, and Ray Cheydleur for specific comments. Comments from Trish Wales were also appreciated.

I appreciate the direct contributions of Ken Elsman to this work in providing data from the TATOO program as well as flesh tone data. I also appreciate the help of Adam Nelson in crunching image data.

I would like to commend the IPA RoundUP committee for running a very well controlled experiment. I would like to thank Steve Bonoff of the IPA for granting permission to use the data, and to Ray Cheydleur for making the data available to me.

Selected Bibliography

Hunt, R
"Why is Black-and-White so Important in Color?," from Recent Progress in Color Science, TAGA Proceedings, 1991

Sharma. Abhay, Tom Collins, Ray Cheydleur, Steve Smiley
2006a "IPA Proofing RoundUP Results," IPA
2006b "IPA Proofing RoundUP Webinar," IPA

Sharma. Abhay, Tom Collins, Ray Cheydleur, Steve Smiley, Florian Suessel
2006c Visual and Colorimetric press to proof matching using the new GRACoL reference printing condition, TAGA Journal, Vol 3

Siljander, Roger, and Richard S. Fisch
"Accuracy and Precision in Color Characterization," 2001, TAGA Proceedings, p 57 - 78

Snedecor, George, and William Conchran,
"Statistical Methods," Seventh Edition, Iowa State Press, 1980

Wales, Trish
" Paper, the Fifth Color: How Brightness, Gloss, and Fluorescence Affect Color Reproduction," GATF World, December 2005

Appendix A – A Bit about the Correlation Coefficient

I have made much use of correlation in this paper, by which I mean "Pearson's correlation coefficient". This is a number that assesses how well one array of data tracks with another array of data. The correlation coefficient is thus a good way to assess whether twenty-two ratings from judges show agreement to colorimetric measurements from those twenty-two sheets.

Note: Pearson's correlation coefficient was first used by Francis Galton, which is yet another example of Stigler's law of eponymy: "No scientific discovery is named after its original discoverer."

The formula for the correlation coefficient, *r* between two arrays of data, x_i, and y_i, is as follows:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n\sigma_x\sigma_y}$$

The average and standard deviation of the *x* values are $\bar{x}$ and σ_x, respectively, and similarly for *y*.

A high correlation coefficient does not mean that the values of one data set are the same as the values in the other data set. A high correlation coefficient means that when a value in one data set is higher than the average, the corresponding value from the other data set is likely to be high as well. Thus, measurements in inches of a set of parts on a conveyor belt will correlate excellently with measurements in centimeters of those same parts. Similarly, if you add seventeen to all the measurements in inches, the correlation coefficient will still be excellent.

The correlation coefficient is a number that is always between -1 and 1. If the value is exactly 1.0, that means that there is perfect tracking between the two data sets. A correlation coefficient of 1.0 means that it is possible to multiply one of the data sets by some positive constant and add another constant and arrive at the second set of data. That is to say, there is a perfect linear relationship between the two data sets.

If the correlation coefficient is -1.0, then there is also perfect tracking between the two data sets, only this time, one data set goes down whenever the other goes

up, and vice versa. One can arrive at the second data set by multiplying one data set by a negative constant and adding another constant.

A correlation coefficient of 0.0 means that there is no correlation between the zigs of one data set and the zags of the other. No matter what you multiply one data set by and no matter what you add, you cannot get the two data sets in any closer agreement.
Note: Unless, of course, you multiply both data sets by zero. This, of course, is silly. Mathematicians may be bad historians, but they don't deal well with silly.

If the correlation coefficient is between 0 and 1, then a little work is required to determine just how significant the result is. It is important to realize that the level at which a correlation coefficient is considered statistically significant depends upon the number of data points. The higher the number of data points, the smaller the threshold.

As an example, let's say that we have only two data points: $\{x_1, y_1\}$, and $\{x_2, y_2\}$. It is always possible to fit a line perfectly through two data points, so the correlation coefficient between data sets with two points will always be 1.0 or -1.0. Obviously, finding such a good correlation with two data points proves nothing about how well two data sets track.

So, for any number of data points, there is a significance threshold. I stated above that:

For n = 22 (this is the number of vendors, and hence the length of data that gets correlated), the $10^{\%}$ significance level for the correlation coefficient is 0.360. This means that randomly generated data has a 10% chance of correlating this well. The 5% significance level is 0.423, and the 1% significance level is 0.537. A complete table of significance levels can be found, for example, in Snedecor, (1980).

As n grows larger, the 10% significance level decreases. This means that using more vendors would allow us to more conclusively state that the relationship between the objective and the subjective matches is statistically significant.

That, of course, is only part of the story. A correlation coefficient that is only slightly above the threshold demonstrates that there is a relationship between the two, but it also suggests that the relationship may not be that strong. That is to say, there are other factors involved.

To get a gauge on how strong the relationship is, we look at a side calculation based on the correlation coefficient. The formula $1-\sqrt{1-r^2}$ tells us how well the differences in one data set explain the differences in the other data set. If, for example, $1-\sqrt{1-r^2}=0.4$, that means that 40% of the standard deviation in one data set is related to the standard deviation in the other data set. The other 60% must be explained by something else.