https://doi.org/10.2352/issn.2169-2629.2021.29.317

Estimating visual difference between reproduction gamuts: moving our pilot study from the lab to online delivery

Gregory High, Peter Nussbaum, Phil Green; The Norwegian Colour and Visual Computing Laboratory; Norwegian University of Science and Technology, Gjøvik, Norway

Abstract

Images reproduced for different output devices are known to be limited in the range of colours that can be reproduced. It is accepted that reproductions made with different print processes, and on different substrates, will not match, although the overall reproduction appearance can be optimized using an output rendering. However, the question remains: how different are they visually? This paper reports on a pilot study that tests whether visual difference can be reduced to a single dimensional scale using magnitude estimation. Subject to recent Covid restrictions, the experiment was moved from the lab to an online delivery. We compare the two methods of delivery: in-person under controlled viewing conditions, and online via a web-based interface where viewing conditions are unknown.

Introduction

There is current interest and activity on the subject of 'consistent colour appearance', relating to visual consistency between differing colour reproductions [1].

Consistent colour appearance

A computer science definition of consistency usual means that all copies of data are identical [2], whereas a broader definition may be that which does not contain contradiction. However, the notion of visual consistency is more subjective in nature. Whilst visual consistency across a set of reproductions is desirable, an exact appearance or colorimetric match may not be possible due to differences in substrates, colorants, and viewing conditions.

A standard definition of visual consistency does not yet exist, and there is no standard way to assess whether a set of colour reproductions has a consistent colour appearance. It is therefore difficult to assess similarity, since this is a multi-dimensional problem. However, it may be easier to assess the overall magnitude of visual difference between pairs of reproductions, or even a set of reproductions.

There is already a body of work on gamut mapping, colour difference and image difference, and its application in print reproduction. One recent addition has been a set of 'Characterized Reference Printing Conditions' (CRPCs) that covers the working gamuts of seven commercial print processes, from newsprint through to a wide colour gamut printer (CRPC1-CRPC7) [3]. However, the chosen reference gamuts may be thought of as somewhat benign, since they use the same colorants, avoid non-neutral substrates, and are visualized as a well-behaved 'Russian doll' of concentric gamut volumes, with primaries and secondaries at very similar hue angles. The visual difference between outputs to these CRPCs might be expected to be roughly correlated to the ratio of their gamut volumes. However, this is likely to be highly image dependent, since images featuring neutral colours might be expected to be least affected in this regard, whilst high chroma images would be expected to suffer from the limitation of smaller output gamuts.

A pilot study under controlled viewing conditions

With this in mind we developed a straightforward experiment based on magnitude estimation of visual difference. The initial pilot was lab-based or in a similarly controlled environment. Using a colour managed display and controlled viewing conditions we ran a pilot experiment with 18 observers in total (10 were regular lab observers, whilst 8 were known 'expert' observers attending the CIC25 conference in Lillehammer).

Moving the experiment to online delivery

The recent global pandemic has made the continuation of lab-based psychophysical experiments almost impossible, with such challenges as working in close proximity to observers, cleaning between lab sessions, and difficulties when handling physical samples. This has necessitated the moving of some research work online. However, a controlled viewing environment is not easy replicated outside the lab.

Delivering an experiment online requires several key components that may be addressed individually or using a combined commercial solution: programming and building the experiment; hosting the experiment online; a recruitment platform for managing observers; and data capture and reporting [4].

With this in mind, we proceeded to develop our lab-based experiment into a web-hosted experiment.

Method

Four candidate gamuts were selected based on widely adopted ICC profiles, with a good progression of gamut volumes and lightness contrast (see Fig. 1 and Table 1).

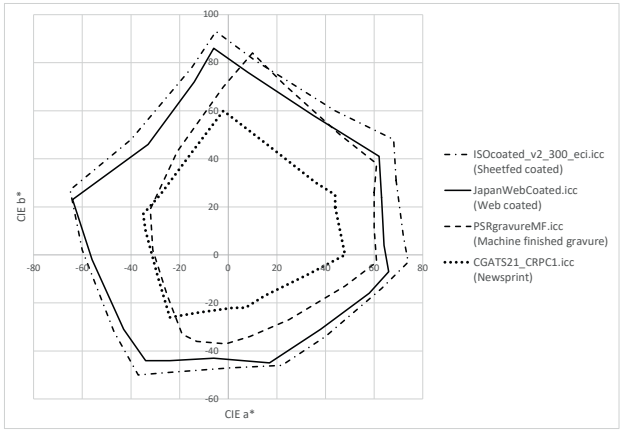

Figure 1. *Projection of printer gamut volumes derived from the four test ICC profiles in the CIELAB UCS.*

Eight test images were selected from the sRGB ISO SCID image set [5]. Two additional synthetic images were included,

Table 1. Four test ICC printer profiles, showing their gamut volumes, white points and black points in the CIELAB UCS.

Profile Name	Gamut Volume	Whitepoint L*	Whitepoint a*	Whitepoint b*	Blackpoint L*	Blackpoint a*	Blackpoint b*
ISO Coated v2 300%	402279	95	0	-2	9	0	2
JapanWebCoated	281370	90	0	-1	10	0	-1
PSRgravureMF	173298	89	-1	4	18	0	-1
CGATS21_CRPC1	84280	85	1	5	32	0	1

also prepared for sRGB: a gamut boundary descriptor based on Green's GBD [6], and a 288-patch chart based on the X-Rite Eye-One Scan Target 1.4 (see Fig. 2). The sRGB source images were then transformed to each of the output print profiles using the ICC perceptual rendering intent, before being converted to the calibrated display RGB using the ICC absolute colorimetric rending intent (consistent with a soft-proofing strategy).

Figure 2. *Ten test images: eight sRGB SCID images [5], plus two synthetic test targets.*

Thus, for each image we obtained six pair comparisons between each of the four renderings. An example of the visual difference between output renderings may be seen in Fig. 3. Only the reproductions were viewed, and the reference sRGB originals were not shown.

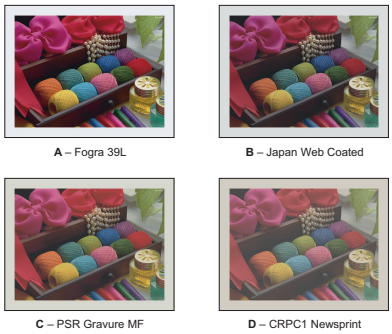

Figure 3. *Visualization of images reproduced with the four test ICC profiles (where A has the largest gamut volume and D has the most limited gamut).*

Judging visual difference on a calibrated display under controlled viewing conditions

The experiment was delivered on a BenQ SW320 display (32"), calibrated to a D50 white point at 160cd/m^2. This device's native colour gamut was very close to AdobeRGB, therefore well suited to print simulation, and it was capable of producing a display black point of just 0.26cd/m^2. The room was dimly lit, in agreement with the P2 viewing condition for graphic arts soft proofing in ISO 3664:2009 [7].

In full screen mode, a multipage PDF (colour managed specifically to our calibrated display) was presented, each page containing two postcard-sized reproductions on a mid-grey background. The large display afforded an extended neutral background to the stimuli, and was comparable to viewing prints in a large viewing booth. The document featured randomized right/left presentation of the images, and the PDF pages were shuffled between each observer session.

Images were prepared following the guidelines in [8]. Images were presented at 146mm x 114mm, including a 9mm border representing the unprinted substrate colour. An 18mm gap was placed between each pair of reproductions. Observers were seated at a desk edge approx 80cm from the display face, but they were free to move slightly, as they might in front of a hardcopy viewing booth.

The purpose of the experiment was to derive data along a single dimension of visual difference (ΔV). Following a training and familiarization session, each observer was asked to estimate the magnitude of visual difference between each pair of reproductions, rated on a scale of 0 to 9. The score was given verbally by the observer, and recorded manually by the researcher.

The experiment was initially conducted using ten research colleagues. The experiment was then relocated to attract additional observers at the CIC25 conference in Lillehammer, and was completed by eight 'expert' attendees following the same format as before.

Online experiment

The recent pandemic forced us to consider how the work could be moved online, and the decision was made to develop an application that could be deployed to a website.

'PsychoPy' [9] was chosen as our experiment builder application, as it offers a mix of pre-determined graphical elements together with bespoke coding elements. It is flexible enough to scale the resulting user interface to different sized displays (albeit by down-sampling images), by specifying each element's size as a fraction of the display's height (see Fig. 4).

Figure 4. *Visualization of user interface on different devices.*

PsychoPy also generates a JavaScript version of the experiment, using the PsychoJS library [10], and this is automatically pushed to the web hosting solution 'pavlovia.org'.

The primary challenge for online deployment is one of colour management. Remote observers' setups will feature many display types, each in a different calibration state. Its resolution and physical dimensions are unknown, nor are the viewing distance, viewing angle, and viewing environment. Of particular concern is the increasing use of wide gamut display types. Our strategy was therefore to prepare all images for an sRGB display, and with an ICC colour profile embedded in each image. The print-simulation images from the previous phase were reused, converted to sRGB using the media relative rendering in-

tent. Since the images originated in sRGB only minimal gamut clipping was expected from this conversion.

Images were saved as JPEGs (high quality / low compression), and at a resolution suitable for the largest expected desktop display. These images might be expected to be colour managed by each individual web browser, to keep the appearance of the sRGB encoding rather than using native display RGB values. However, some legacy systems may contain older browser applications that do not colour manage or do so in a way that is inconsistent, and therefore the state of colour management is unknown. On the other hand, older displays tend to be manufactured close to the sRGB standard, and so may still give good appearance matches.

Observer details and feedback were captured by a secure online form, in line with local GDPR requirements.

In the user interface, the estimate of visual difference for each pair was captured using an adjustable slider (with numerical feedback showing a score from 0 to 100 — see Fig. 4). The GUI elements were light grey on a mid-grey background so as to avoid any unwanted stimulus effects. Placing all the controls on the display adjacent to the test images allowed the observers to keep their eyes on the experiment, rather than constantly glancing at their keyboards. Right/left position and running order were both randomized by the web application. 27 observers responded to an email invitation and participated.

Results

Results were obtained from the three observer groups (two in-person under controlled viewing conditions scaling differences from 0 to 9, and a third group via the online hosted experiment scaling differences from 0 to 100).

Estimation of visual difference under controlled viewing conditions

An estimate of visual difference (for each pair of gamuts) was obtained under the controlled viewing conditions (estimated on a scale of 0 to 9). The 18 observers were divided into two groups, with 10 'Lab Researcher' observers and 8 'CIC Expert' observers (see Fig. 5).

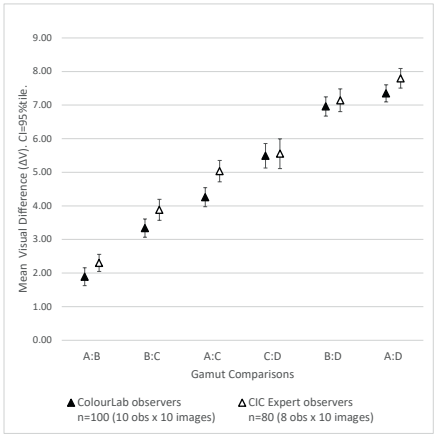

Figure 5. *Comparison of two observer groups under controlled viewing conditions.*

The two groups gave remarkably similar results, with the 'CIC Experts' consistently scoring the visual differences with a slightly higher value. Experienced observers may therefore be more aware of appearance difference. In this way we see the potential for both individual observers as well as groups to use different internal scales, which Engeldrum refers to as the 'observer modulus' [11] .

Estimation of visual difference in the web-based online experiment

An estimate of visual difference (for each pair of gamuts) was obtained from 27 observers using the web-hosted online experiment (this time estimated on a scale of 0 to 100).

Comparing modes of delivery – Raw scores

We plot the mean results from the lab-based and online phases for each gamut comparison (see Fig. 6) using the combined data from all ten images. Since the two phases used different scales of magnitude we perform a linear scaling to adjust previous scores (0 to 9) to align with the new scale (0 to 100).

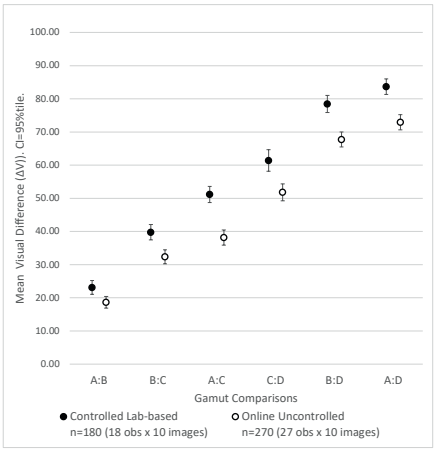

Figure 6. *Raw scores – Comparison of two viewing modes – lab-based controlled vs. online uncontrolled.*

Results obtained online follow a similar pattern to the previous phase, though a difference in observer modulus is apparent. Online observers judged the reproduction differences to be smaller, and used a smaller range of scores overall. Since each group of observers may be using a different internal scale we will apply a normalization technique to the data later in this paper.

Uncertainty in the raw scores

For each response in the raw data we calculate the confidence internal at the 95% percentile. Looking at the 60 image comparisons individually we see a pattern in the uncertainty of the data. For comparison, a second order trendline is applied to the data for each experimental mode (see Fig. 7). Along the abscissa we see the difference in the range of values used by each group that we saw in Fig. 6.

The distribution of uncertainties may be thought to reflect the ease or difficulty with which observers make their judgements. In both experimental modes the observers find judging small differences to be easiest, and the largest differences to be moderately easy. However, the greatest uncertainty, and therefore the greatest difficulty, is found when judging visual differences in the middle of the range.

Comparing modes of delivery – Group means scale normalized scores

Given the apparent difference in range between the observer groups, the combined results may therefore be better served by adjusting each observer's choice of modulus using a 'group means scale normalization', as outlined by Engeldrum

[11, p.148], and used by Luo et al. [12] to scale observer judgements of colourfulness. The method consists of calculating the mean average of the log scores (essentially a geometric mean approach). The log responses of each observer are then normalized to the mean log scores with an offset and a first order gradient derived from a least squares fit. The mean averages of the normalized log scores are then exponentiated to give the normalized scores.

In Fig. 8 we see the normalized data from the two experimental modes. Each individual observer's data has been fitted to the group average, and so when comparing the two sub-groups we see that both now use almost exactly the same range of values, and that the data are very similar throughout that range with only a small significant difference in the A:C gamut comparison.

Inter-observer STRESS

A measure of observer variability will better differentiate the two modes of delivery. From the results described above, the inter-observer standardized residual sum of squares index (STRESS) was calculated according to Melgosa et al. [13, p.73] (please see Table 2).

For the raw scores we can see that the inter-observer STRESS for the two in-person groups (non-expert and expert) is comparable, with mean STRESS values 22.19 and 18.98 respectively. However, we can see that inter-observer variability is far higher for the online experiment, with a mean STRESS value of 29.16. By way of comparison, we note that most advanced colour difference formulas produce STRESS values somewhere in the range of 20 to 30 [13, p.74].

Table 2. Inter-observer STRESS for different observer groups.

	Lab-based/calibrated display				Online	
	Non-expert observers Inter-observer STRESS		CIC Expert observers Inter-observer STRESS		Online observers Inter-observer STRESS	
Raw Scores	No. of obs.	10	No. of obs.	8	No. of obs.	27
	Min	13.68	Min	14.22	Min	20.73
	Max	30.71	Max	26.75	Max	53.66
	Mean	22.19	Mean	18.98	Mean	29.16
Group Means Normalized Scores	No. of obs.	10	No. of obs.	8	No. of obs.	27
	Min	16.18	Min	14.00	Min	19.45
	Max	29.76	Max	28.93	Max	50.98
	Mean	22.76	Mean	22.82	Mean	27.62

Applying the group means scale normalization to the data gives only a modest reduction in the STRESS, demonstrating that although the technique normalizes the range of the observer modulus it does not greatly reduce uncertainty in the data.

Number of online observers required to give results comparable to their lab-based counterparts

Given the increased uncertainty, it may be necessary to expand the number of observers in an online study relative to its lab-based counterpart. Using the group means scale normalized data (0 to 100) across all images, we find that the average confidence interval at the 95th percentile for our 18 lab-based observers is 6.71, whereas the confidence interval for our 27 online observers is 6.43. We can iterate the number of observers needed to give a comparable confidence interval, and find that for this particular experiment approximately 25 online observers will generate a comparable confidence interval to our 18 lab-based observers.

Display size and its effect on the responses of online observers

Given the potential range of viewing setups, it might be expected that display size would have a marked impact on the responses of online observers. Via an online questionnaire the users' display sizes and types were recorded, and the results divided into two size groupings: $<= 16$" which consisted mainly of laptops and a tablet, and > 16" which included larger desktop displays from 24" up to 32" models.

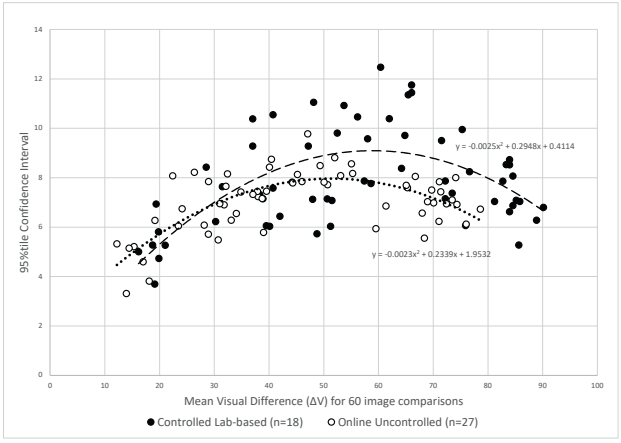

Figure 7. *Uncertainty through the range of visual differences based on raw scores – lab-based controlled vs. online uncontrolled.*

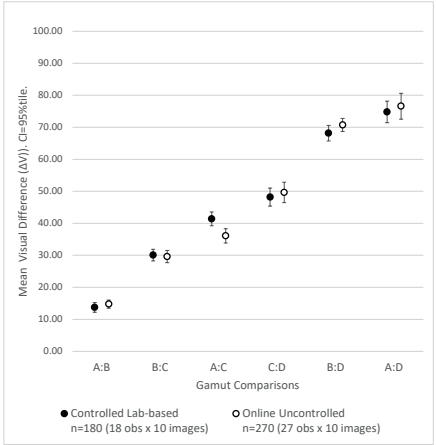

Figure 8. *Group means scale normalized scores – Comparison of two viewing modes – lab-based controlled vs. online uncontrolled.*

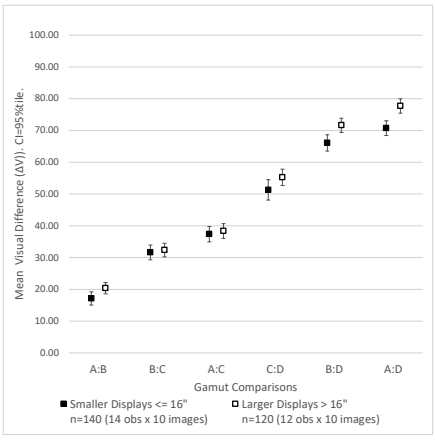

Figure 9. *Effect of display size on online responses – raw scores.*

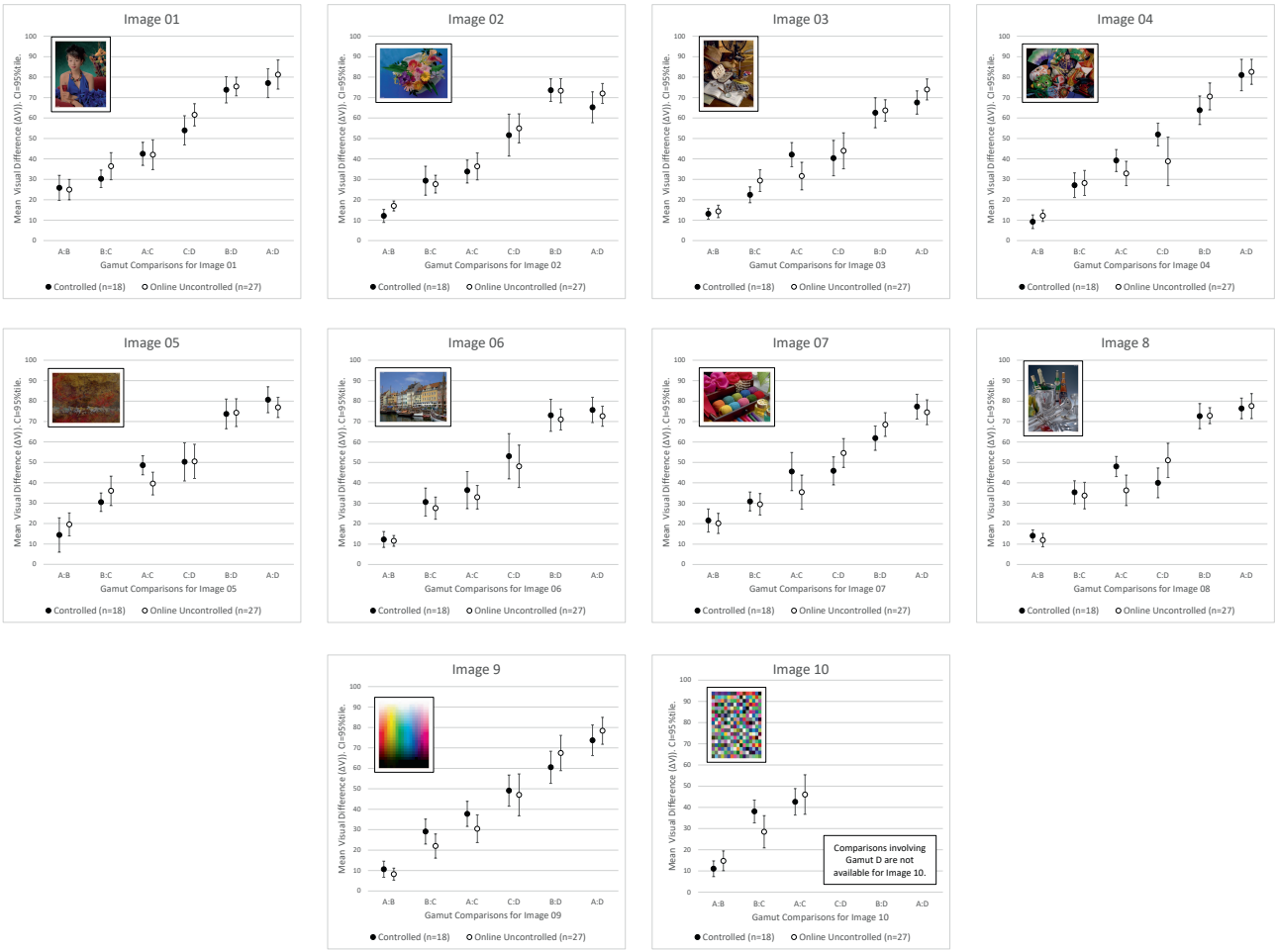

Figure 10. *Image dependency and experimental mode – group means scale normalized data.*

From the raw scores given by the online observers, we can see little difference between the two size groupings (see Fig. 9), the only significant difference being amongst the largest visual differences of gamut comparisons B:D and A:D. But overall, they are remarkably similar. Therefore, we conclude that display size has little effect on observers when they are judging visual difference between images on a single display. Any differences in observer modulus would be removed using the normalization method described previously.

Image dependency

Using the group means scale normalized data, we see that the similarity between experimental modes seen in Fig. 8 is repeated when we look at individual images (see Fig. 10). No one image exhibits a significant difference between modes.

However, we do see a slight difference in the responses given to the gamut comparisons for individual images. Images 04 and 09 provide the widest range of visual differences (approx. 10 to 80), whereas image 01 and image 07 accentuate the smallest visual difference for gamut comparison A:B (approx. 20). Some graphs also show a flatter response for larger visual differences, with images 02, 06 and 08 giving similar differences for gamut comparisons B:D and A:D.

Overall, the trend based on the comparison of gamut renderings is consistent across all the images. However, if we wish to distill visual difference between reproductions down to a function of gamut volume, it is important that a wide variety of images is used.

Ratio of gamut volume as a predictor of visual difference between reproductions

We calculate the ratio of gamut volumes described in Table 1. Using the combined normalized scores, the average visual difference for each gamut comparison (combining results from all ten images) is then plotted against the gamut volume ratio. The resulting chart shows a clear non-linear trend — the greater the volume ratio the greater the visual difference (see Fig. 11). As a visual guide a third order trendline is applied to the data.

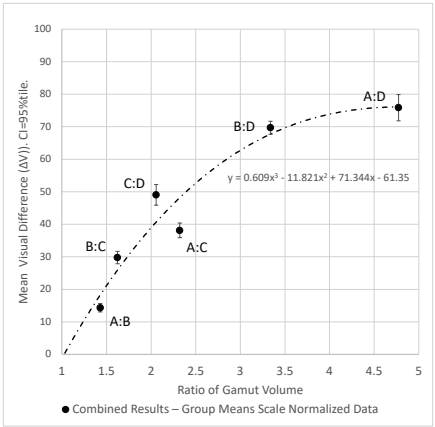

Figure 11. *For six gamut comparisons, the visual difference is plotted against the gamut volume ratios (see also Fig. 3 for examples).*

However, our results are not perfectly monotonic, since by rank order the comparisons A:C and C:D are in reversed positions. This may be due to gamut C (PSRgravureMF) being a different shape to the other gamuts, with proportionally more coverage of the red/orange colour centres (see Fig. 1). Volume ratio alone is therefore insufficient to accurately predict visual difference.

Discussion

Moving work online requires a considerable investment in time, with builder apps and hosting services having a significant learning curve. However, builder apps like PsychoPy offer near-infinite possibilities to customize the user interface, and once mastered can be used to deliver lab-based experiments as well as online hosted tasks.

In addition, each experiment requires thorough testing, and will need to be piloted on various hardware/operating system/web-browser combinations should the exact end user setup be unknown.

Observer recruitment can be difficult, and it is important to find observers who show the same level of care as would be expected for an in-person experiment.

The biggest benefit to the organizer/researcher is not having to be present for each observer session! This facilitates the upscaling of the experiment, once it has been found to perform satisfactorily in a pilot.

In this present pilot we have shown that a magnitude estimation task can be performed online. One drawback, however, is that the presentation of pairs of stimuli is a very inefficient process, with comparison combinations increasing exponentially as the number of gamuts increases. The limiting factor is the time an online observer might be expected to sustain their level of concentration.

Future work

The experimental interface, once it has been produced, can easily be re-used. A greater variety of reproduction gamuts could be used, though the number of gamut permutations is a limiting factor. It may be best to create multiple experiments, each completed within a sensible time frame, with the results being collated and normalized.

Conclusions

A straightforward magnitude estimation of difference for print reproductions offers a single continuum of visual difference (ΔV) along which will lie the difference between any two reproduction.

An assumption of correlation between gamut volume ratio and the magnitude of visual difference is largely borne out by the results, but additional criteria such as substrate colour difference, gamut shape and contrast ratio may improve the correlation further.

The recent pandemic necessitated a move to on-line delivery. A web-based experiment gave comparable results to an in-person study with controlled viewing conditions. However, observer variability was greatly increased for the online work.

Acknowledgements

The authors gratefully acknowledges the time, care and efforts of all the observers. Special thanks goes to Chris Bai and BenQ Corporation for the use of their BenQ SW320 32” Photo Editing Monitor during the lab-based experiment.

References

[1] CIE TC8-16 Consistent Colour Appearance. Consistent Colour Appearance;. Available from: `https://color.org/resources/consistentappearance.xalter`.

[2] Consistency model;. Available from: `https://en.wikipedia.org/wiki/Consistency_model#Slow_consistency`.

[3] ISO 15339-2 Graphic technology — Printing from digital data across multiple technologies — Part 2: Characterized reference printing conditions, CRPC1-CRPC7; 2015. International Organization for Standardization.

[4] Sauter M, Draschkow D, Mack W. Building, hosting and recruiting: A brief introduction to running behavioral experiments online. Brain Sciences. 2020 Apr;10(4).

[5] ISO 12640-2 Graphic technology — Prepress digital data exchange — Part 2: XYZ/sRGB encoded standard colour image data (XYZ/SCID) ; 2004. International Organization for Standardization.

[6] Green PJ. Test target for defining media gamut boundaries. In: Eschbach R, Marcu GG, editors. Photonics West 2001 - Electronic Imaging. SPIE; 2000. p. 105–113.

[7] ISO 3664:2009 Graphic technology and photography – Viewing conditions; 2009. International Organization for Standardization.

[8] CIE TC8-16 Consistent Colour Appearance. Consistent Colour Appearance - test images;. Available from: `https://color.org/resources/r8-13/CCA_test.xalter`.

[9] Peirce J, Gray JR, Simpson S, MacAskill M, Höchenberger R, Sogo H, et al. PsychoPy2: Experiments in behavior made easy. Behavior Research Methods. 2019 Feb;51:195–203.

[10] Bridges D, Pitiot A, MacAskill MR, PeerJ JP, 2020. The timing mega-study: comparing a range of experiment generators, both lab-based and online. PeerJ. 2020 Jul;8.

[11] Engeldrum PG. Psychometric scaling: a toolkit for imaging systems development. Imcotek press; 2000.

[12] Luo MR, Clarke AA, Rhodes PA, Schappo A, Scrivener SAR, Tait CJ. Quantifying colour appearance. Part I. Lutchi colour appearance data. Color Research & Application. 1991 Jun;16(3):166–180.

[13] Melgosa M, Trémeau A, Cui G. Colour Difference Evaluation. In: Fernandez-Maloigne C, editor. Advanced Color Image Processing and Analysis. New York: Springer; 2013. p. 59–79.

Author Biography

Gregory High is a PhD candidate at the Colour and Visual Computing Laboratory, NTNU, Norway. The topic of his PhD research project is ‘A model of consistent colour appearance’.

Peter Nussbaum is an associate professor of colour imaging at the Colour and Visual Computing Laboratory, NTNU, Norway. Dr. Nussbaum received an MSc from the Colour & Imaging Institute, University of Derby, GB, in 2002 and completed his PhD degree in imaging science in 2011 from the University of Oslo, Norway.

Phil Green is Professor of Colour Imaging at the Colour and Visual Computing Laboratory, NTNU, Norway.