Received: 19 April 2023 | Revised: 1 November 2023 | Accepted: 16 November 2023

DOI: 10.1002/col.22913

RESEARCH ARTICLE

COLOR RESEARCH AND APPLICATION WILEY

Building a metric of color reproduction difference by combining multiple observers in a modular online experiment

Gregory High | **Peter Nussbaum** | **Phil Green**

Norwegian Colour and Visual Computing Laboratory, Norwegian University of Science and Technology, Gjøvik, Norway

Correspondence
Gregory High, Norwegian Colour and Visual Computing Laboratory, Norwegian University of Science and Technology, Postboks 191, Gjøvik NO-2802, Norway.
Email: gregory.high@ntnu.no

Funding information
Norges Teknisk-Naturvitenskapelige Universitet

Abstract

A web-hosted online experiment was previously developed to find the visual difference between four reproduction gamuts using direct magnitude estimation (Proc. IS&T 29th Color and Imaging Conf, 2021:317–322). In order to increase the size of the data set, but without overburdening observers, a modular approach was adopted. The original methodology was therefore extended across 10 linked sub-experiments to make comparisons between some 36 gamuts, which were designed to exhibit a variety of different gamut shapes, contrast ratios, and substrate colors within the constraints of a desktop display. In addition to each set of test images, a common normalization set was included in all sub-experiments in order to adjust each observer's choice of modulus to a global average observer, and thus combine the results into a larger data set. Finally, an interval scale was inferred from the normalized magnitude data using a categorical judgment approach to calculate scale values. The fitted data revealed a power function close to a square-root between the interval and magnitude scales.

KEYWORDS

cross media color reproduction, gamut mapping, online experiment, soft proofing, visual difference

1 | INTRODUCTION

There is current interest and activity on the subject of "consistent color appearance." This has resulted in a CIE Technical Committee on the subject, with coordinated work between several research groups.[1,2] Visual consistency across a set of reproductions is desirable, but an exact appearance or colorimetric match may not be possible due to differences in substrates, colorants, and viewing conditions. In particular, output media gamuts and the gamut mapping strategies employed govern the color differences between reproductions.

The concept of consistent color appearance has been demonstrated to be valid,[3] with some commercial gamut mapping products shown to give more consistent sets of reproductions when producing outputs across a wide range of device gamuts.

Assessing visual similarity is problematic, since it is a multi-dimensional problem. However, the overall magnitude of visual difference between pairs of reproductions

is more practical to assess, and this is the motivation for the present work.

1.1 | Motivation

The visual difference between color reproductions was previously investigated by High et al.[4] Whereas much work exists in the field of image difference and image quality, our previous study was motivated by differences in color gamut volumes and the visual differences between the resulting color reproductions.

The work started out as a lab-based experiment under controlled viewing conditions. Reproduction images were created by rendering reference images to a small number of output print gamuts, including their various media white points. The rendered images were prepared as simulated prints on a color managed display. Observers were then asked to rate the visual difference between pairs of reproductions using a method of direct magnitude estimation.

In response to the global pandemic an online version of the experiment was also developed. Following a similar format, the color reproductions were prepared as soft-proof images ready for use on an sRGB display. Online observers then gave a numerical response to the visual difference between reproductions using a slider in a web-based user interface (UI) (see Figure 1) for a visualization.

Results from the lab-based and online experiments, which may be thought of as "controlled" and "uncontrolled," were in agreement following a "group means scale normalization," as outlined by Engeldrum.[5, p.148] For online participants, the normalization also incorporated differences in their viewing conditions and display specifications.

FIGURE 1 Estimating visual difference between reproductions – a visualization of the online experiment's user interface on different devices.

In both modes, observers were able to judge the gamut mapped images in terms of "overall difference" (rather than looking for small structural differences). Analysis of the pilot confirmed a correlation between the ratios of the gamut volumes and the resulting visual differences between reproductions (where the larger differences in gamut volumes resulted in greater visual differences).

It may therefore be possible to predict likely visual difference based upon knowledge of two output gamuts. However, for this to be modeled adequately it will first be necessary to build a data set consisting of comparisons between many gamuts which are different in terms of size, shape and media white points (or paper colors in the case of prints).

1.2 | Aims and objectives

In our pilot, just four gamuts led to six combinations of gamut pairings, which were then multiplied by the number of test images. The challenge of building a larger dataset lies in the potential number of comparisons to be made, and the time required by observers to make them. However, the online experiment, combined with the group means scale normalization, offers the potential to combine the efforts of many observers across different blocks, in other words to create a modular online experiment.

Our aim is therefore to generate a comprehensive set of visual differences, which will result from the various gamut volumes used to make the reproduction images. Using the method described, these will fall along a common continuum, and are expected to form a ratio-like magnitude scale. In practice, the resulting magnitude scale is unlikely to be linear to an equal interval scale, with the relationship between them expected to be close to a power function.[5, p.151] For use in further work, it will be beneficial to convert the magnitude data to an interval scale, prior to modeling visual difference from gamut volume metrics or comparing it to existing image difference metrics.

1.3 | Preparation of test gamuts, ICC profiles and soft-proofing images

The content for this web-based experiment naturally has to work within the color gamut constraints of a standard RGB display. This means that the output gamuts used here and test images rendered to them must fall within

the boundary of the sRGB gamut.[6] A brief description of test gamuts and the image preparation methods is given below in Section 4.2. However, for a more detailed account of the preparation and rationale, we encourage readers to view the online Supplementary material. In particular, the generation of test gamuts, including choice of gamut shapes and simulated paper colors, steps taken to create ICC color profiles, gamut mapping strategy and the preparation of test images, is described more fully.

2 | BACKGROUND

2.1 | Color differences in images

Color differences between reproduction images, or between a reference and reproduction, can result from the gamut mapping operations that are needed when two imaging devices produce a different range of colors. The primary constraint on color reproduction is the available color gamut of each reproduction medium (where media white, maximum available contrast, and maximum chroma at each hue angle are some of the main limitations). Given those limitations, the subjective accuracy or pleasantness of the reproductions can be maximized using an appropriate gamut mapping strategy, though differences in their appearance will still remain apparent.

Previous work on color difference in images includes a CIE technical report,[7] which noted that whilst individual color differences can be measured and calculated, the distribution of those differences between images is rarely normal. Usually it is high chroma pixels which suffer the greatest compression when being mapped to a smaller gamut, with Uroz et al.[8] finding that observers often identified these worst-case color differences first, particularly when they occurred in recognizable features within the image.

Systematic differences (in lightness, chroma or hue) were also found to be more perceptible than random color changes. This was consistent with Hong and Luo,[9] who illustrated the effect using reproductions where a modification had been applied either globally or at selective hue angles. Though the two examples had similar mean color differences from the reference image, the selectively edited reproduction was perceived as having far greater visual difference. These apparent chroma differences at selective hue angles were also analogous to the visual differences seen between reproductions when their output gamuts differ in shape as well as volume.

Color image difference is also an important determiner of image quality. Typically, objective image quality metrics are developed to correlate with subjective data which compare multiple different reproductions against a reference image (a full-reference approach). Early image quality metrics based on structural similarity, such as SSIM,[10] did not do particularly well at predicting difference between gamut mapped images (since the mapping is typically a monotonic compression in a uniform color space, and does not result in strong structural differences within the image). The SSIM approach has subsequently been extended to include color difference terms, with the aim of better predicting differences between pixelwise gamut mapped images[11] or to optimize image-specific spatial gamut mapping operations.[12]

However, the reproduction-to-reference approach does not consider the likely scenario of two reproductions being viewed side-by-side, each of which could be judged as having an equal magnitude of difference from the source image, but in different aspects. The resulting difference between the two reproductions is therefore unrelated to the reference image, and is not well predicted by the similarity metric.

Furthermore, the reproductions can appear on different media with various white points (or paper colors in the case of prints). In side-by-side viewing, the visual difference between reproductions is also a function of the combination of media as well as differences between output devices. It is therefore helpful to think of the resulting visual difference in terms of "color reproduction difference" rather than image difference.

For a typical gamut mapping strategy (where the same strategy may be used for multiple images), visual difference between reproductions is primarily a function of the gamut volumes' constraints, which in turn may be more apparent in some images than others. As a metric, color reproduction difference (the magnitude of overall visual difference between gamut mapped images) offers an alternative approach to quantifying the color image differences that regularly occur in color imaging and graphic arts applications.

2.2 | Magnitude estimation and observer normalization

Direct magnitude estimation has long been used to establish scales based on physical stimulus intensities and their perceived magnitude, with the relationship generally following a power function.[13] When asking observers to give a numerical response, the researcher typically allows each participant to use their own range of magnitudes. This might be relative to a reference stimulus, to which the participant will give an "anchor" value, resulting in magnitude data along a ratio scale.

A practical alternative is to ask participants to use a generic scale, such a 0 to 100 (see Rowe in Pointer

et al.).[14] Depending on the nature of the stimulus, this bounded scale has the disadvantage that observers might commit to using the maximum response in a way that leads to clipping or compression of response values. Other researchers have therefore opted for a reference stimulus with an observer-assigned value, and then using an open-ended scale.[15,16] It is also possible to use magnitude estimation without a reference ("absolute scaling"),[5, p.140] with each observer forming their own internal reference, or modulus.

Each individual observer's scaling response will therefore differ depending on their modulus and working range. The resulting data may be normalized by adjusting each observer's choice of modulus using a "group means scale normalization" (GMSN) procedure, as outlined by Engeldrum,[5, p.148] (A similar method was previously used by Luo et al.,[16, p.172] to normalize magnitude data relating to the perception of colorfulness). The normalization is performed in the log domain (which assumes that the observers' sense of magnitude follows a power function). The method consists of calculating the mean averages for the log scores of each stimulus. The log responses of each observer are then normalized to the mean log scores by applying an offset and slope derived using a least squares fit. Post-normalization, the resulting mean averages of the stimuli's log scores are then exponentiated to give the normalized scores (essentially a geometric mean approach).

In our pilot experiment,[4] observers estimated the visual difference between pairs on a nominal scale of 0 to 100 using an on-screen slider, with the resulting scale value updated in real time. It was felt that this was an intuitive approach, avoided lengthy explanation and training, and helped keep visual attention on the display.

A reference difference was not provided during the experiment, as this was deemed too prescriptive. However, prior to the experiment, training pages (containing thumbnail scatter proofs arranged on a gray background) were shown to give an overview of likely reproduction differences. A live training session then followed, which deliberately included pairs of reproductions with small, medium and large differences. Within the experiment, each observer formed their own modulus. Participants' scaling responses also differed depending on their use of an internal working range.

2.3 | Specific considerations for online and uncontrolled experiments

Moving a visual study outside the controlled conditions of the lab poses some additional problems. Whilst participation may be increased by attracting a wider range of observers, a primary concern is the contribution of different and unknown viewing conditions and display equipment, as well as the indirect relationship with the observers, which might increase the overall variability in the results.

2.3.1 | Controlled versus uncontrolled visual experiments

In a print-based experiment Zuffi et al.[17] ran a paired comparison study in the lab and elsewhere using a portable ring-bound book of reproductions. They broadly concluded that data collected in uncontrolled lighting conditions could substitute lab-based data. However, it was noted that there was a degree of image dependence, with some print reproductions proving problematic under artificial light sources.

By the late 2000s improvements in display technology combined with near-universal web-access had facilitated some high quality online visual studies. In 2009 Sprow et al.[18] produced an HTML-based image quality experiment together with a hosted database of test images. Pairs of reproduction images alongside an sRGB reference were presented on-screen at a fixed pixel resolution. Observers then selected the best reproduction with a mouse click. The lab-based experiment was performed on calibrated sRGB reference displays in standard viewing conditions according to ISO 3664.[19] By contrast, observers in uncontrolled viewing conditions used their own computer equipment, and researchers had little control over the display settings used. Much like the print-based study, the display-based experiments revealed good agreement between observer judgments on the web and in the lab.

As noted elsewhere by Zuffi et al.,[20] Katoh's 1998 conclusion that observers tend to be more adapted to their CRT displays and less to the ambient light in the room[21] suggests that viewing environment plays less of a role in online tasks compared to similar work using hard copy reproductions. In the intervening years displays have become far brighter, with many now automatically adjusting their luminance to compensate for ambient conditions. It follows that observers' adaptation to their displays can remain strong across a range of viewing conditions, and not just in a dimly lit environment. A similar conclusion is drawn by Sprow et al.,[18] noting that for side-by-side image comparison the visual differences remain relative across a wide range of display types, settings and viewing conditions.

2.3.2 | Increased observer variability in uncontrolled studies

The increase in observer variability for online studies means that a greater number of participants may be needed to return data with similar uncertainty to a lab-based experiment. Zuffi et al.[20] found in their print-based study that the ratio of uncontrolled to controlled sample sizes needed to provide equivalent variance ranged from 1.2 to 3.9, but that this ratio was very image dependent. In our earlier pilot[4] we found that in a study using both lab-based displays and online displays an average ratio of approximately 1.4 uncontrolled to controlled observers resulted in the same uncertainty. That would equate to a target of 21 uncontrolled observers when compared to the classic minimum of 15 controlled observers in Rec. ITU-R BT.500-13[22] (though more recently this ITU requirement has been increased to 24).

2.3.3 | Commercial observer recruitment solutions and GDPR requirements

The move to online experiments has prompted the development of commercial and academic solutions offering a range of services which fall into three broad categories: experiment builders (design and programming software); observer recruitment and management platforms; and online hosting platforms.[23] Any combination of these may be used, depending on the requirements and existing in-house resources of the researchers. Clearly, one of the advantages of an online study is that it can reach a global audience, with many observers completing tasks in parallel, and at times convenient to the participants. A corollary to this is that the management of observers becomes very complex. When personal data is captured, the General Data Protection Regulations (GDPR)[24] add an extra burden to researchers in terms of data security, informed consent and the ability of individual participants to opt out at a later date. This requires extra data management in order to associate experimental results with specific observers, and for that data to be held securely. Some countries and institutions[25] have a further requirement to register studies with a central agency when they capture any personal data.

It is challenging to comply fully with these requirements in an online study. In a lab-based experiment the researcher will typically present each observer with both written and verbal information about the study, before obtaining a signature on a consent form. However, for casual online observers this administrative overhead can become a barrier to participation, or take a disproportionate amount of time compared to the experiment itself. Anonymizing the data at the point of capture may be a practical workaround.

Online visual experiments, such as paired comparisons, tend to be of a "short stimuli design,"[26, p.20] with observers cycling through a large number of stimuli. This is usually self-paced, but a rule-of-thumb is that the duration should be about 20–30 min in total. In larger studies this limitation either requires observers to return for multiple sessions, or else calls for a far greater total number of participants.

Therefore, to fulfill an extended online study it is important to remove the barriers to observer recruitment, participation, completion, and retention. It may be that commercial recruitment platforms such as MTurk[27] and Prolific[28] will be required to manage a large number of participants whilst remaining GDPR compliant.[23]

The unsupervised nature of online experiments changes the relationship with the observer, but the use of paid participants can also change the dynamic. There is often a suspicion that paid workers will produce lower quality results compared to volunteer observers (who are typically students with an interest in the subject). However, for comparable unsupervised online experiments (where paid and volunteer participants received the same instructions, and also faced similar out-of-lab distractions) studies have shown that experienced MTurkers demonstrate greater attentiveness than their unpaid counterparts.[29] What has changed over more recent years is a broadening of the pool of available online workers, whilst for text-based surveys the built-in attention checks have become a less effective indicator of poor quality data.[30] For online surveys repeated from 2013 to 2020, it was found that low levels of proficiency in the language used within the study was the major source of increasingly unusable responses.

Filtering observers by language fluency is a challenge, particularly when recruiting and training a large number of observers across many different countries. Even for visual studies, the instructions and training are typically text-based and are open to misinterpretation. There is also a risk that a participant will simply proceed without reading the instructions at all. Paid or unpaid, online studies run the risk of observers misunderstanding the task, a problem that would be recognized easily in a supervised lab-based study.

2.3.4 | Pre- and post-screening of observers

The ITU documents Rec. ITU-R BT.500[22] and Rec. ITU-T P.913[26] (subjective assessment of video quality) provide a useful framework for preparing display-based experiments with a large number of stimuli. There are many

aspects analogous to running an online experiment, including issues surrounding pre- and post-screening of observers.

It is certainly acceptable to pre-screen observers[26, p.18] for acuity, color vision deficiency, and so forth (though this is difficult to implement in a remote setting). Post-screening and the elimination of a subject's data is more problematic.[26, p.19] It may be appropriate to perform a "soft rejection"[26, p.27] by applying a weighting based on an observer's bias and consistency. Alternatively, a rejection criteria may be based on a threshold value of linear correlation to the mean observer[26, Annex A]. However, both these methods are designed to sharpen discriminal dispersion, and assume that an observer consensus is inherent to the data.

For online studies (which are typically viewed remotely and are unsupervised), there is also the chance that observers perform differently to their peers in a supervised experiment, or in a way that is not anticipated by the designer of the experiment. In an earlier online experiment Zuffi et al.[31] noted the occurrence of errors that did not happen in an equivalent supervised experiment. These tended to be behavioral differences (such as completing a task too quickly or too slowly) or else returning spurious responses (null responses, or multiple extreme responses). These indicated a lack of engagement by the unsupervised observer, and were treated as criteria for discarding those records post hoc.

3 | REFLECTIONS ON THE PILOT ONLINE EXPERIMENT

The online version of the experiment was initially a response to the global pandemic, with a web-hosted application developed to deliver it for remote observers. The experiment was created using the builder application "PsychoPy,"[32] and the UI was designed to scale to almost any display size (see Figure 1). This was achieved by specifying each element's size relative to the display's pixel height (that is, the test images were knowingly re-sampled). PsychoPy also generated a JavaScript version of the experiment using the associated PsychoJS library,[33] and this was then pushed to the pay-per-observer web hosting solution "pavlovia.org."[34]

3.1 | Extending the online experiment

In our pilot experiment[4] only four industry-standard output profiles were used to render the simulated output images, which resulted in six permutations of gamut comparisons using pairs of reproduction images. Increasing the number of gamuts would mean that the number of possible comparisons would increase exponentially. It would therefore be necessary to make only selective gamut comparisons in order to limit the number made. There was the added challenge of recruiting and retaining participants for a multi-part online experiment, as well as a practical limit on the number of comparisons an individual observer might be expected to make in a single session without jeopardizing completion. Therefore, in order to obtain a larger data set we developed an extended modular version of the previous experiment with a far greater number of observers.

4 | METHOD

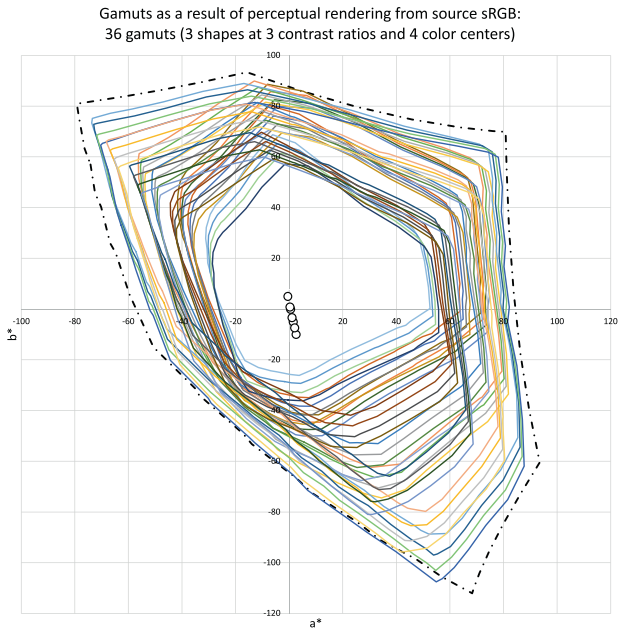
4.1 | Recruitment of observers

Observers were recruited via email invitation from groups of colleagues and students with a known interest in color imaging. Ten different invitations with unique URLs were issued ad hoc in order to achieve 20+ observers for each of the 10 sub-experiments. To comply with local GDPR requirements each participant self-generated a nickname, thereby pseudonymizing their data. Further non-attributable information and feedback was captured using a secure online form at the end of the experiment.

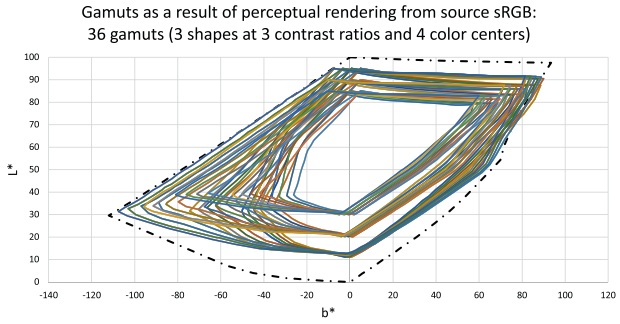
4.2 | Choice of reproduction gamuts

A total of 36 test gamuts were prepared featuring 3 different shapes, 3 contrast ratios, and 4 simulated media white color centers. A matching set of ICC color profiles was generated using ArgyllCMS[35] with a compression-type gamut mapping algorithm, optimized to reproduce colors from an sRGB source. The perceptual rendering intent transform within each profile was then used to map source colors into an addressable reproduction gamut which fell within the sRGB display gamut, suitable for our online experiment. Please see Figure 2 for a visualization of the 36 addressable reproduction gamuts. (Note: a very small number of colors around the display blue primary could potentially be clipped when simulating the largest gamut with a bluish media white point (see bottom of Figure 2A). All other addressable gamuts fall entirely within sRGB.

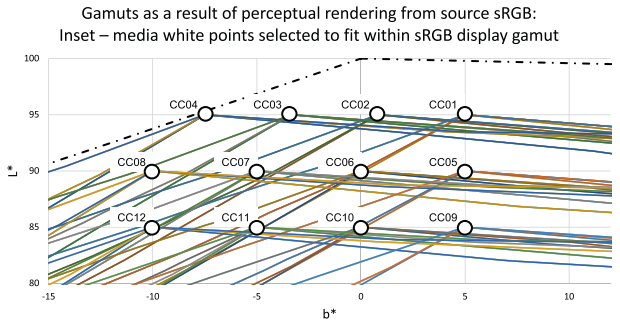
The image set was limited to eight sRGB source images (a mix of color pictorials, a grayscale portrait, and a high chroma GBD image), each prepared with a white border. These were then rendered to the 36 output gamuts using the ICC profiles with the perceptual



(A) a*b* projection of 36 test gamuts designed to fit within sRGB display gamut



(B) b*L* projection of 36 test gamuts



(C) Inset of b*L* projection of 36 test gamuts. Their 12 media white points fall within the sRGB display gamut

FIGURE 2 36 test gamuts. These are the addressable gamut volumes achieved using the perceptual rendering intents of 36 ICC profiles, designed to fit within the sRGB display gamut. ○ Media white points, — · — · sRGB display gamut.

rendering intent. Lastly, the reproduction images were converted to sRGB using a media-relative transform, with the border appearing as a simulated substrate color relative to the display white point. This resulted in 288 soft-proof images ready for use in the web-based experiment.

Given the number of possible permutations, it was important to identify a sub-set of gamut comparisons that could then be divided up across our modular experiment, with the aim that observers could complete each session in approximately 25 min.

4.2.1 | Main image set

The 36 gamuts were divided into nine groups of four (with each four-group providing six possible comparisons). The gamuts in each group were selected at random, but had to include at least one of each gamut shape, contrast ratio and color center. A tenth group of four was formed, providing six additional comparisons.

Using these gamut combinations with our eight test images gave 480 image-pair reproduction differences, which would divide neatly into 48 for each of the 10 sub-experiments.

4.2.2 | Additional image set with specific gamut differences

An additional set of gamut comparisons was made which deliberately featured differences in one criteria only (either contrast difference or media white difference).

Three of the 36 gamuts were again selected, but this time differing only in contrast ratio, having the same gamut shape, and media white points which shared a similar chromaticity (CC02, CC06, and CC10 in Figure 2C). This gave three gamut comparisons which represented differences in contrast ratio only.

Additionally, four gamuts featuring the same shape and contrast ratio, but differing in media whites (CC05, CC06, CC07, and CC08 in Figure 2C), gave six possible comparisons to represent differences in media whites only.

Using these extra gamut combinations with our eight test images gave 72 image reproduction differences, or approximately 8 per sub-experiment.

4.2.3 | Normalization image set

The present modular experiment extends the use of the GMSN to normalize the magnitude data from different image sets, in order to create a super-set of observer visual differences. An added benefit is that each observer session gets treated as a different participant, and the effect of any differences in equipment and

viewing conditions between blocks is also normalized. This also removes the need for each observer to complete all image sets, which is very useful in a large scale online experiment where observer retention cannot be guaranteed.

Based on experience gained from the pilot, a small subset of stimuli was found to accurately predict the offset and slope of the normalization for a larger set. The best predictors were pairs with either large or small differences, toward both the extremes of the working range.

For the present modular study 15 reproduction image pairs were selected, with 8 pairs exhibiting large visual differences, and 7 pairs having quite small visual differences. This normalization image set was then included in each of the sub-experiments for the purpose of the GMSN, with the resulting intercept and offset being calculated against a global average observer before being applied to normalize the observers' other data in each of the sub-experiments.

4.2.4 | Repeatability image set

In order to assess intra-observer variability, 6 image pairs (approximately 10% of the main image set) were repeated at the end of each sub-experiment.

4.3 | Hosted online experiment

All 10 standalone sub-experiments followed the same format, including introduction and training, and a main body of stimuli that contained test images, the common normalization set, and a repeatability check.

4.3.1 | Screening and introduction

The first few web pages were designed to introduce observers to their task, and to provide the contact details of the researchers. Observers were asked to self-declare that they had color-normal vision, with a link provided to an online CVD test for those in doubt.

It was expected that users' web browsers would color manage the UI and image content to their displays, so that it would appear consistent with an sRGB display. Participants were therefore asked to use Apple Safari, Microsoft Edge or Google Chrome web browsers. (There are some occasional issues with older versions of Mozilla Firefox, whereby sRGB content is handled as device RGB, and this lack of color management could cause images to appear over-saturated on a WCG display. Therefore, observers were asked not to use that platform).

Observers then confirmed their chosen nickname, and were invited to take part in an online questionnaire at the end of the session. Participants had to actively give consent at this point in order to move forward. A flowchart of the introduction and the following training session is included in Figure 3.

4.3.2 | Observer training

After obtaining the observer's consent to proceed an image familiarization task took place. Four consecutive web pages showed thumbnail scatter proofs arranged on a gray background, giving a preview of likely reproduction differences. A live training session then followed, which introduced the form of the main experiment. Using the experiment's UI (see Figure 1) participants were asked to rate the difference between pairs of reproduction images from 0 to 100, where "0" represented no visual difference and "100" represented the maximum visual difference. Six pairs of reproductions followed, with the observer free to gain experience with the slider control and progression tools. This phase of the training deliberately included pairs with large, small, and medium visual differences, to further inform the observer's internal scale. Finally, before the main experiment, the written instructions were repeated.

> Using your skill and judgment, please give each pair a rating from 0 to 100 (where "0" represents no visual difference, and "100" represents the maximum visual difference).

4.3.3 | Main body of experiment

The main part of each sub-experiment contained 48 image pairs from the main set, 8 image pairs from the specific-difference set, and the 15 images of the normalization set. This gave a total of 71 reproduction pairs. Images from each group of four gamuts were kept together in the sub-experiments and compared "in-the-round." This meant each image was downloaded once and used in at least three comparisons per sub-experiment, and this greatly reduced the memory footprint and download time of the web application. For each sub-experiment the left–right presentation and running order of pairs was randomized.

The experiment was concluded with a repeatability check consisting of 6 image pairs from the preceding main image set. A flowchart of the experiment is included in Figure 4.

5 | RESULTS

A total of 219 observer sessions were completed across the 10 sub-experiments, each gaining at least

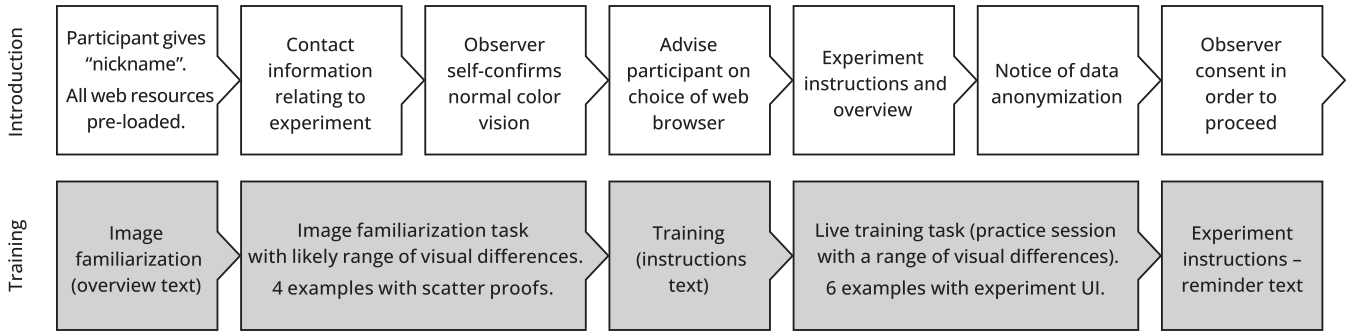


FIGURE 3 Flowchart of website pages which contain the introduction and training sessions that precede the main body of the experiment.

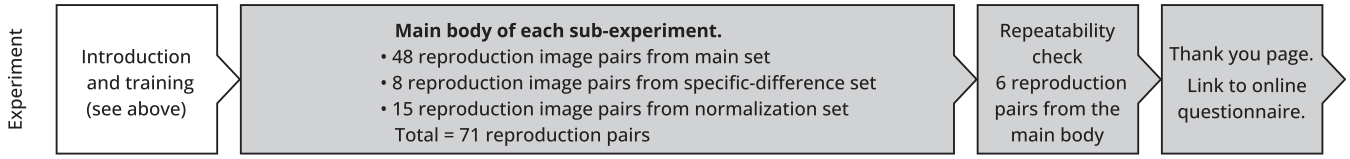


FIGURE 4 Flowchart of website pages containing the main experiment.

20 participants. The pseudonymized nicknames used by each participant allowed the cross-referencing of 134 separate observers who took part, with the vast majority of participants completing either one or two sessions. The average time taken to complete an experiment was 23 min 48 s, with the average response time for each image comparison 9.41 s.

5.1 | Online questionnaire about observers' equipment and viewing conditions

A link to a separate secure online questionnaire was provided upon completion of each experiment. A total of 209 questionnaires were completed. The information given reflected each session rather than each observer, since some observers may have used different equipment on each occasion. A breakdown of the participants' demographics and equipment used is given in Table 1.

Participants were also asked about their viewing conditions. Although categories were suggested the responses were given as free text. The mix of lighting types and lighting levels is visualized in Figure 5. We see that dim lighting levels were provided mainly by warm LED sources, mixed light sources often provided a perceived medium level, whilst daylight and office fluorescent sources tended to feature in brighter viewing conditions.

5.2 | Unexpected observer behavior in the online experiment

Although the online experiment contained written instructions, training and on-screen reminders, it was still possible for an observer to misinterpret the task. This was in contrast to an in-person experiment, where any such misunderstanding would likely be challenged by the supervisor during the briefing and training sessions.

The initial responses of all observers were examined for unusual behavior, and the general correlation between each observer and the group mean was checked. Out of the 219 observer sessions, four were found to have a negative correlation with their peers, and this was also expressed as a negative slope in the group means scale normalization. A likely explanation was that these participants inverted their working scale during some or all of their responses.

A further behavioral trait was noticed in a small number of observers who appeared to rely heavily on giving either minimum or maximum scores. The nominal available working range in the UI was 0 to 100, and so it was expected that each observer would find their own modulus and working range within that. Some participants may have scored certain visual differences to be very high or very low, and were therefore expected to return at least some values of 0 or 100. This was reflected in most observers' behavior (see Figure 6).

We see that the number of observers giving an increasingly high proportion of 0 and 100 s approximates

TABLE 1 Responses to online questionnaire following each observer session.

Age group (category)		Gender (free text)		Imaging professional (Y/N)		Device used (category)		Browser (free text)		Display size (free text)	
18–24	59	M	124	Yes	123	Laptop	107	Edge	24	<12″	2
25–34	73	F	82	No	86	Laptop with External Display	39	Chrome	156	12″ to 17″	89
35–44	33	X	3			Desktop PC	60	Safari	26	>17″ to 24″	62
45–54	21					Mobile	3	Other	3	>24″	46
55–64	17									Do not know	10
65 and over	6										
Total	209	Total	209	Total	209	Total	209	Total	209	Total	209

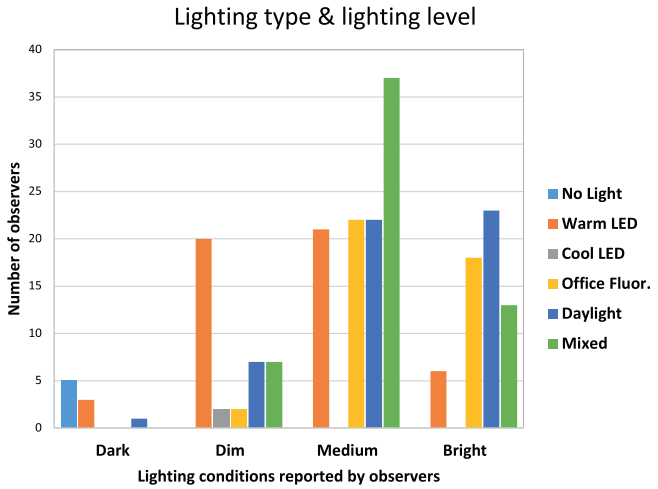


FIGURE 5 Participants' descriptions of lighting types and lighting levels.

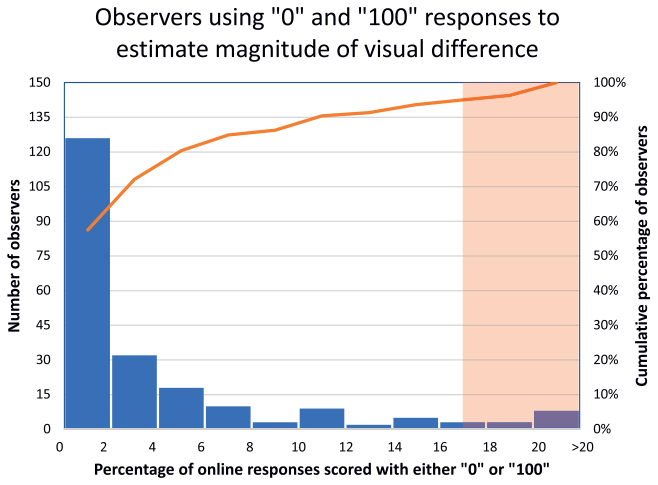


FIGURE 6 The proportion of responses scoring either 0 or 100. Observers within the 95th percentile used the 0 or 100 scores in less than 17% of their responses. A small number of observers were found to rely on the minimum and maximum scale values.

a rapid decay function, with the majority of observers returning very few minimum or maximum ratings. By way of explanation, it is possible that a few observers misinterpreted the task as a binary test of acceptability rather than a magnitude estimation of difference. Observers within the 95th percentile used the 0 or 100 scores in less than 17% of their responses. Excluding the four observers identified previously for inverting the scale, 11 out of the remaining 215 observers were found to exceed this limit, and they may be thought of as having misunderstood the task.

5.3 | Effect of internal inconsistency on the normalization process

We utilized a common set of reproduction pairs to normalize the modulus of individual observers across all 10 sub-experiments. The efficacy of the normalization relies on congruity between the common normalization set and the main image set in each sub-experiment, both of which should fall along the same continuum of visual difference. However, when an individual observer is internally inconsistent their responses to the normalization set may not be characteristic of their responses to the main image set. This results in a slope and offset obtained from the normalization process which, when applied to the main body of images, produces adverse results. Since an inconsistent observer's slope and offset is fitted and applied in the log domain, the normalization could, in the antilog domain, create negative values for low scores or high scores which are orders of magnitude greater than the original 0 to 100 scale. Conversely, for observers who are consistent throughout their experiment it may be assumed that the normalization would have a beneficial effect, and that either the mean difference from their peers would be reduced (an improvement in inter-observer agreement) or else

for those observers already close to the average, it would have practically a null effect.

5.3.1 | Analyzing the effect of the group means scale normalization

We therefore examine the effect of internal inconsistency on the normalization of results in a modular experiment. Using only the stimuli from the main image sets, we calculate the mean difference between each individual observer's responses and their sub-experiment's mean observer. This is performed on the pre-normalized magnitude data. A similar calculation is made on the post-normalized magnitude data, where each observer's log-normalized scores are exponentiated for the purpose of comparison. It is expected that the GMSN should improve inter-observer agreement, and therefore reduce this difference. The reduction or increase in mean difference is visualized in Figure 7.

We see that, in the antilog domain, the plot of changes in mean difference between individual observers and the mean observer in their sub-experiment as a result of the GMSN appears normally distributed. For the majority of observers the GMSN does indeed produce a reduction or close to a null effect. For a small number of observers we see that the GMSN has increased the mean difference between them and their sub-experiment's group average.

5.3.2 | Comparison of normalization performance with inter-observer variability

Excluding those participants who exhibited unusual behavior (see Subsection 5.2), we hypothesize that those observers outside the 95th percentile in Figure 7 (13 out of the remaining 204 observers for whom the GMSN did not work well) are internally inconsistent. We therefore use the six reproduction pairs repeated at the end of each experiment in a test of observer repeatability, and their intra-observer STRESS is calculated according to Melgosa et al.[36, p.73] Observers are divided into two groups based on the 95th percentile of normalization performance in Figure 7, and a one-tail Welch's *t*-test (a two sample *t*-test assuming unequal variance) is used to compare their STRESS scores. We find a significant difference in intra-observer STRESS between the consistent grouping within the 95th percentile (mean STRESS 31.6) and the inconsistent grouping above the 95th percentile (mean STRESS 49.1); $t(17)=7.55$, $p = 3.999\text{E-}07$.

The STRESS results of the two groups are visualized as a boxplot in Figure 8. We conclude that the performance of the GMSN for each participant is determined by their intra-observer variability. Since our repeatability image set contains just six pairs it may be thought of as a limited snapshot of intra-observer variability. However, the GMSN performance, which is based on a larger number of stimuli, acts as an alternative indicator of observer internal consistency.

5.4 | Filtered versus unfiltered magnitude data

We compare results for two versions of the observer data. We first consider the combined data from all 219 observers across the 10 sub-experiments. We then exclude observers who exhibited unusual behavior (inverted the scale, or exceeded the 95th percentile in usage of 0 and 100 scores) or who were deemed to be internally inconsistent

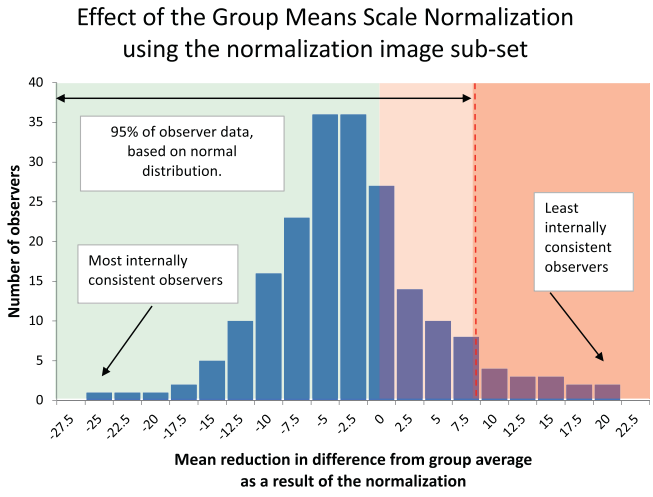


FIGURE 7 Effect of internal inconsistency on the normalization process.

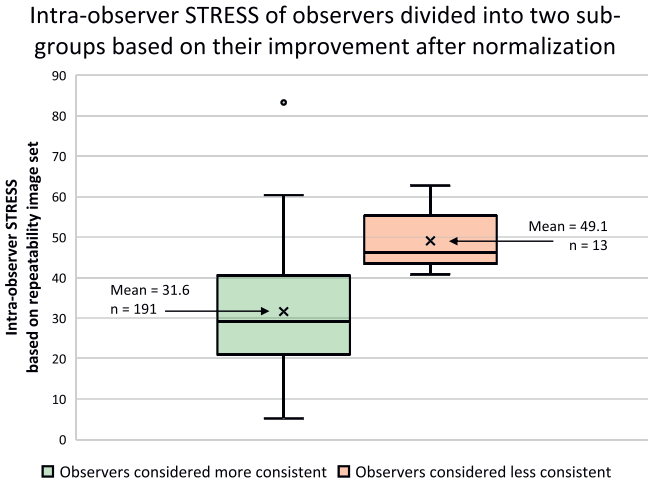


FIGURE 8 Intra-observer STRESS of participants considered consistent (below the 95%tile of normalization performance) or inconsistent (above the 95%tile of normalization performance).

(exceeded the 95th percentile of improvement as a result of the GMSN). One hundred ninety-one observers remain. As an additional step following normalization, any spurious individual observer scores are, in the antilog domain, clipped to the working range of 0 to 100. The exclusion of some observers means that the GMSN used by each version will be fitted to a slightly different mean observer, resulting in two different normalized magnitude scales. Across the 552 reproduction image pairs there is a mean difference between the two scales of 2.14, with a maximum difference of 15.36. More interesting is the reduction in standard deviation. The unfiltered version has a mean SD of 19.87 (max SD = 50.56) whereas the filtered version has a lower mean SD of 16.86 (max SD = 41.27).

The filtered results, based on 191 observers, and with a minimum of 16 observers in any one sub-experiment, are therefore taken forward for the creation of an interval scale.

5.5 | Test of bimodality

The normalized observer magnitude data, in the antilog domain, is expected to be approximately normally distributed with a discriminal dispersion around a mean for each image comparison. We perform a test of bimodality on the filtered dataset to confirm this behavior. The log-normalized observer data is therefore exponentiated, and for each reproduction pair we calculate a bimodality coefficient (BC) using the bimodalitycoeff.m[37] Matlab function. Based on the skewness and kurtosis of the data a BC is calculated in the range of 0 to 1, where a threshold value of 0.555 represents the value for a uniform distribution, and a BC >0.555 indicates bimodality. Of the 552 image comparisons, only 25 exceed this threshold. Upon visual inspection of their histograms, those distributions with a BC >0.555 tend to be data clustered around a low mean magnitude, with one or two outliers at higher magnitudes. The mean BC for the dataset is 0.3981 (min = 0.1594, max = 0.7501). We conclude that the data is not bimodal in nature.

6 | ESTIMATING AN INTERVAL SCALE FROM THE NORMALIZED MAGNITUDE DATA

The normalized magnitude data cannot be assumed to behave in a way consistent with an interval scale (since a set increment of the magnitude scale may not represent an equal visual difference throughout the range).

Whereas a true ratio scale would exhibit an increase in observer uncertainty proportional to the increase in stimulus magnitude, an interval scale would exhibit approximately constant uncertainty throughout its range. Magnitude estimation is expected to produce a ratio scale, but a real-world experiment may produce a magnitude scale that behaves somewhere between the two definitions described above.

Engeldrum[5, p.151] suggests using a category judgment approach to confirm the relationship between the magnitude and interval scales. This assumes Condition D of Torgerson's Law of Categorical Judgment (that is, for an interval scale the discriminal dispersion, or observer

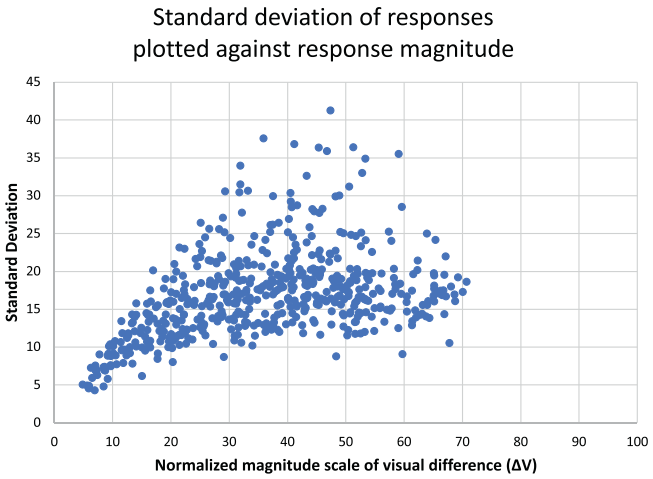


FIGURE 9 Standard deviation increases with magnitude of response, consistent with a ratio scale.

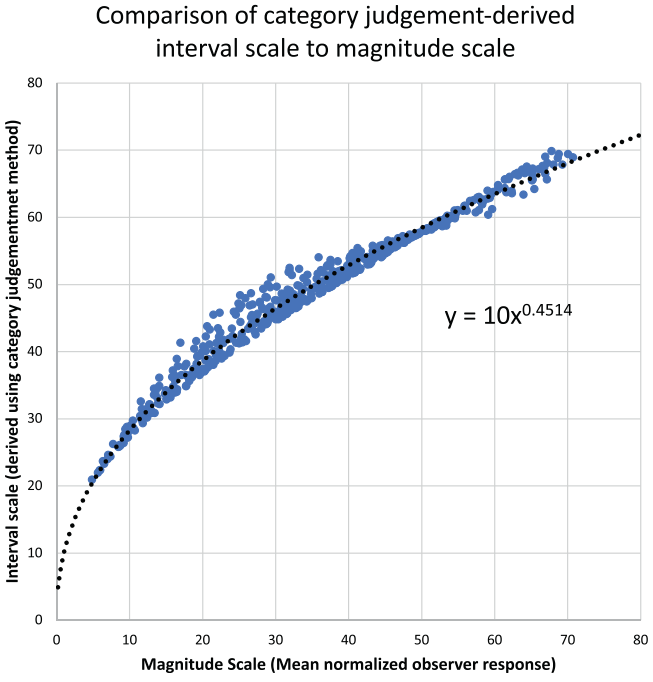


FIGURE 10 Scale values are plotted against magnitude data, and fitted with a power function.

uncertainties, would be expected to be equal across all categories). Typically, category judgment data would be used to create a frequency matrix (the number of times a stimulus is placed in each category), with z-score scale values calculated from the cumulative proportions using a least squares fit matrix operation.[5, p.133] For the present experiment, since the GMSN returns a geometric mean together with a standard deviation, a cumulative proportion matrix can be calculated from the magnitude data, with the number of categories determined by the level of precision required. The estimated scale values are not a perfect monotonic transformation of the magnitude data, but the underlying relationship may be used to derive a fitting function. Engeldrum[5, p.151] suggests using a power function to fit the interval scale to the magnitude scale, with the exponent set to one or less. Bartleson[38] suggests an approximation using a square root function for magnitude data that are prothetic (that is, the standard deviations increase in proportion to the magnitude of the data). A visualization of our normalized magnitude data reveals that the uncertainty does generally increase with the magnitude (see Figure 9).

Our magnitude data is then fitted to an interval scale using the category-derived estimated scale values. The exponent and the intercept of the power function are calculated by finding an optimal linear fit in the log–log space. The category-derived estimated scale values are plotted against the magnitude data in Figure 10, together with the fitted power function. An exponent of 0.45 is found to give the best fit, which is consistent with Bartleson's generalized square root and other examples cited by Engeldrum. In addition, a constant is applied to scale the highest used value on the interval scale value to 70, which is approximately the same as the highest used value on the magnitude scale. Each unit of the resulting interval scale does not represent an attribute JND[39] (representing a threshold of perceptibility in a single visual attribute) though the scale itself may prove to be a useful tool in determining a quality JND in terms of multi-dimensional differences between reproductions.

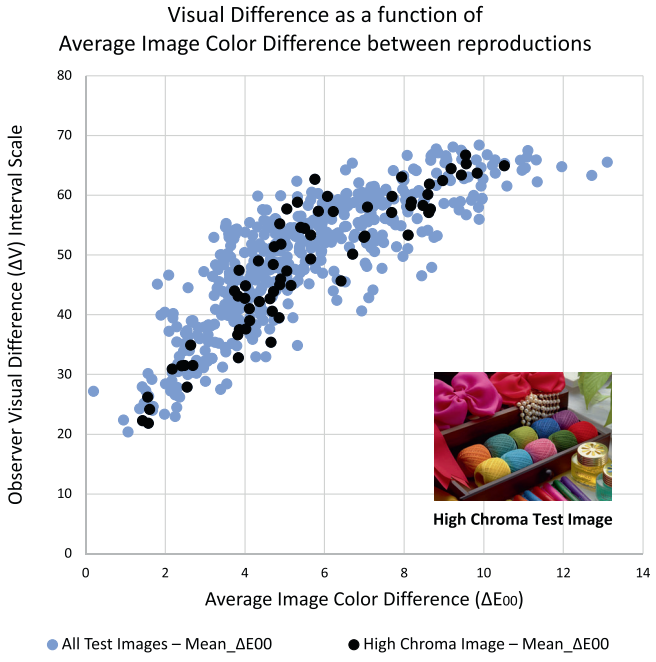


FIGURE 11 Observer visual difference using the interval scale is plotted against the average image color difference between pairs of reproductions.

7 | DISCUSSION

The combined data set includes visual difference scores for 552 side-by-side reproduction comparisons. On the resulting interval scale these data range from approximately 20 to 70 (see Figure 10). The zero point on the interval scale is arbitrary, but we keep all scale values positive in order to avoid negative values of difference.

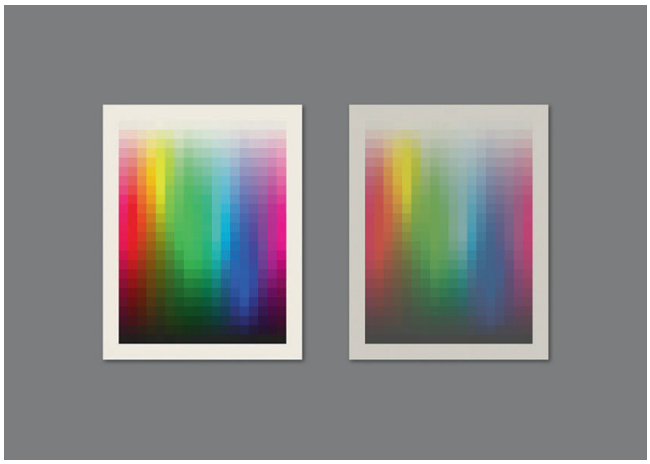
Our experiment did not include any null-difference pairs (two instances of the same reproduction). Every pair had at least some image-difference, and this was done to keep the experiment as simple as possible and avoid any "trick" pairings. That could have complicated the training and briefing phases, and potentially undermined online observer confidence. There was also concern that very small differences could become meaningless when viewed in uncontrolled conditions.

It is therefore possible that a similar experiment with small and null differences conducted under controlled viewing conditions could establish an indicative zero

TABLE 2 Coefficient of determination between interval scale of observer visual difference and calculated color differences.

Image comparison	Excl. paper colored border		Inc. paper colored border	
Color difference calculation	Mean ΔE_{00}	95%tile ΔE_{00}	Mean ΔE_{00}	95%tile ΔE_{00}
	R^2	R^2	R^2	R^2
All images	0.70	0.54	0.65	0.51
High chroma image (example)	0.81	0.69	0.67	0.53

(A) ΔV of 30 on interval scale



(B) ΔV of 45 on interval scale

(C) ΔV of 60 on interval scale

FIGURE 12 Examples of small, medium, and large visual differences between color reproductions.

point, but great care would have to be taken to normalize this with the present data set.

It is interesting to compare the resulting interval scale of reproduction difference with the mean color differences calculated between the pairs of reproductions. In Figure 11 we see that increasing visual difference approximately correlates with larger mean color differences. Across all images (with the exception of grayscale images) the mean color difference provides a better correlation than higher percentiles (where the higher percentiles represent larger color differences between any two reproductions). We also see that including the paper-colored border in the calculation reduces the correlation, though its presence in the experiment clearly has an impact on visual difference (see Table 2) for summary statistics.

In this study, we attracted participants by email for over 200 observer sessions. However, for any greater number it would have been more practical to use an online recruiting platform to manage observers. In this study, we targeted observers with a known interest in color imaging, and who were expected to understand the concepts presented in the training phase. However, there was still the challenge of unexpected observer behavior that differed from what would be accepted in the lab.

For illustrative purposes, Figure 12 visualizes three example image pairs at interval scale values of 30, 45, and 60. These examples include pairs with known differences in contrast ratio and substrate color, though by its nature, the reproduction here lacks the scale and the extended gray background of the experiment's UI.

8 | CONCLUSIONS

Within the constraints associated with web-based soft-proofing it was possible to gather magnitude estimation data based on overall visual difference between simulated color reproductions. We presented a method whereby a group means scale normalization was used to normalize observer modulus across several linked experiments by way of a common normalization image set. However, the success of the normalization process was found to be reliant on the observers' internal consistency (that is, low intra-observer variability). It was also found that a categorical judgment approach could be used to compare the distribution of the magnitude data with an interval scale, and provide a fitting function. In this experiment the relationship between the scales was close to a square root.

9 | FUTURE WORK

The reproduction gamuts, created with characterization data and ICC profiles, will allow a range of gamut comparison metrics to be calculated, including differences in gamut volume, shape, contrast and substrate color difference.

When used in conjunction with the present interval scale we expect these metrics to provide a means of modeling the systematic causes of visual difference in the side-by-side comparison of color reproductions.

AUTHOR CONTRIBUTIONS

Gregory High: Conceptualization, Methodology, Software, Investigation, Visualization, Writing – Original Draft. **Phil Green:** Writing – Review & Editing, Supervision. **Peter Nussbaum:** Writing – Review & Editing, Supervision.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the time, care and efforts of all the online observers, including colleagues and students from NTNU Colourlab, Yamagata University, Rochester Institute of Technology, Linköping University and beyond.

DATA AVAILABILITY STATEMENT

Research data are not shared.

ORCID

Gregory High https://orcid.org/0000-0001-5315-0772

REFERENCES

[1] CIE TC8-16. Consistent Colour Appearance. Accessed 2021. https://color.org/resources/consistentappearance.xalter

[2] CIE. TC 8-16: Consistency of Colour Appearance within a Single Reproduction Medium. International Commission on Illumination (CIE). Accessed 21 September 2023. https://cie.co.at/technicalcommittees/consistency-colour-appearance-within-single-reproduction-medium

[3] Mattuschka M, Kraushaar A, Tröster P, Wittmann J. Consistent colour appearance – a novel measurement approach. 27th Color and Imaging Conference, 27:314–319. Society for Imaging Science and Technology. 2019. doi:10.2352/issn.2169-2629.2019.27.57

[4] High G, Nussbaum P, Green P. Estimating visual difference between reproduction gamuts: moving our pilot study from the lab to online delivery. Proceedings of IS&T 29th Color and Imaging Conference, 2021:317–322. Society for Imaging Science and Technology, 2021. 10.2352/issn.2169-2629.2021.29.317

[5] Engeldrum PG. *Psychometric Scaling: A Toolkit for Imaging Systems Development*. Imcotek press; 2000.

[6] Anderson M, Motta R, Chandrasekar S, Stokes M. Proposal for a standard default color space for the internet – sRGB. In IS&T 4th Color and Imaging Conference, 4:238–245. Scottsdale, AZ, USA: Society of Imaging Science and Technology. 1996. doi:10.2352/issn

[7] CIE 199:2011. Methods for Evaluating Colour Differences in Images. Vienna: Commission Internationale de l'Eclairage. 2011.

[8] Uroz J, Luo R, Morovic J. Perception of colour differences in large printed images. In: MacDonald L, Luo MR, eds. *Colour Image Science: Exploiting Digital Media*. John Wiley & Sons, Ltd; 2002:49-73.

[9] Hong G, Luo MR. Perceptually-based color difference for complex images. 9th Congress of the International Colour Association, 4421:618–621. SPIE. 2002. doi:10.1117/12.464761

[10] Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. *IEEE Trans Image Process*. 2004;13(4):600-612. doi:10.1109/TIP.2003.819861

[11] Lissner I, Preiss J, Urban P, Lichtenauer MS, Zolliker P. Image-difference prediction: from grayscale to color. *IEEE Trans Image Process*. 2013;22(2):435-446. doi:10.1109/TIP.2012.2216279

[12] Preiss J, Fernandes F, Urban P. Color-image quality assessment: from prediction to optimization. *IEEE Trans Image Process*. 2014;23(3):1366-1378. doi:10.1109/TIP.2014.2302684

[13] Stevens SS. Matching functions between loudness and ten other continua. *Percept Psychophys*. 1966;1(1):5-8. doi:10.3758/BF03207813

[14] Pointer MR, Ensell JS, Bullock LM. Grids for assessing colour appearance. *Color Res Appl*. 1977;2(3):131-136. doi:10.1002/col.5080020308

[15] Pointer MR. The concept of colourfulness and its use for deriving grids for assessing colour appearance. *Color Res Appl*. 1980;5(2):99-107. doi:10.1002/col.5080050212

[16] Luo MR, Clarke AA, Rhodes PA, Schappo A, Scrivener SAR, Tait CJ. Quantifying colour appearance. Part I. Lutchi colour appearance data. *Color Res Appl*. 1991;16(3):166-180. doi:10.1002/col.5080160307

[17] Zuffi S, Brambilla C, Eschbach R, Rizzi A. Controlled and uncontrolled viewing conditions in the evaluation of prints. Proceedings of SPIE 6807, Color Imaging XIII: Processing, Hardcopy, and Applications, 6807:680714. San Jose, CA, United States: SPIE. 2008. doi:10.1117/12.766398

[18] Sprow I, Baranczuk Z, Stamm T, Zolliker P. Web-based psychometric evaluation of image quality. Image Quality and System Performance VI, 7242:95–106. San Jose, CA, United States: SPIE. 2009. doi:10.1117/12.805313

[19] ISO 3664:2009. Graphic Technology and Photography – Viewing Conditions. International Standard. Geneva: International Organization for Standardization. 2009.

[20] Zuffi S, Brambilla C, Eschbach R, Rizzi A. A study on the equivalence of controlled and uncontrolled visual experiments. Proceedings of SPIE 7241, Color Imaging XIV: Displaying, Processing, Hardcopy, and Applications, 7241:724102. San Jose, CA, United States: SPIE. 2009. doi:10.1117/12.805854

[21] Katoh N, Nakabayashi K, Ito M, Ohno S. Effect of ambient light on the color appearance of softcopy images: mixed chromatic adaptation for self-luminous displays. *J Electron Imaging*. 1998;7(4):794-806. doi:10.1117/1.482665

[22] Recommendation ITU-R BT.500-13. *Methodology for the Subjective Assessment of the Quality of Television Pictures*. International Telecommunication Union (ITU); 2012. https://www.itu.int/rec/R-REC-BT.500-13-201201-S/en

[23] Sauter M, Draschkow D, Mack W. Building, hosting and recruiting: a brief introduction to running behavioral experiments online. *Brain Sci*. 2020;10(4):251. doi:10.3390/brainsci10040251

[24] gdpr.eu. General Data Protection Regulation (GDPR) Compliance Guidelines. Accessed 5 April 2023. https://gdpr.eu/

[25] Sikt. Norwegian Agency for Shared Services in Education and Research. Accessed 5 April 2023. https://sikt.no/en/home

[26] Recommendation ITU-T P.913. *Methods for the Subjective Assessment of Video Quality, Audio Quality and Audiovisual Quality of Internet Video and Distribution Quality Television in Any Environment*. 3rd ed. International Telecommunication Union (ITU); 2021. https://www.itu.int/rec/T-REC-P.913-202106-I
[27] Amazon Mechanical Turk. Amazon Mechanical Turk. Accessed 6 October 2023. https://www.mturk.com/
[28] Prolific. Prolific Quickly Find Research Participants You Can Trust. Accessed 6 October 2023. https://www.prolific.com/
[29] Hauser DJ, Schwarz N. Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behav Res Methods*. 2016;48(1):400-407. doi:10.3758/s13428-015-0578-z
[30] Marshall CC, Goguladinne PSR, Maheshwari M, Sathe A, Shipman FM. Who broke Amazon mechanical Turk? An analysis of crowdsourcing data quality over time. Proceedings of the 15th ACM Web Science Conference 2023, 335–45. WebSci'23. New York, NY, USA: Association for Computing Machinery. 2023. doi:10.1145/3578503.3583622
[31] Zuffi S, Scala P, Brambilla C, Beretta G. Web-based versus controlled environment psychophysics experiments. Image Quality and System Performance IV, 6494:56–63. San Jose, CA, United States: SPIE. 2007. doi:10.1117/12.703926
[32] Peirce J, Gray JR, Simpson S, et al. PsychoPy2: experiments in behavior made easy. *Behav Res Methods*. 2019;51:195-203. doi:10.3758/s13428-018-01193-y
[33] Bridges D, Pitiot A, MacAskill MR, Peirce JW. The timing megastudy: comparing a range of experiment generators, both lab-based and online. *PeerJ*. 2020;8:e9414. doi:10.7717/peerj.9414
[34] Open Science Tools Ltd. Pavlovia. Accessed 8 February 2023. https://pavlovia.org/
[35] Gill G. Argyll Color Management System. Mac OS. ArgyllCMS. 2021 http://www.argyllcms.com/
[36] Melgosa M, Trémeau A, Cui G. Colour difference evaluation. *Advanced Color Image Processing and Analysis*. Vol 3. Springer; 2013:59-79.
[37] Zhivomirov H. Bimodality Coefficient Calculation with Matlab. Matlab. MATLAB Central File Exchange. 2023. https://uk.mathworks.com/matlabcentral/fileexchange/84933-bimodality-coefficient-calculation-with-matlab
[38] Bartleson CJ. Direct ratio scaling. *Visual Measurements*. Vol 5. Academic Press, Inc; 1984:491-507. Optical Radiation Measurements.
[39] ISO 20462-1:2005. Photography—Psychophysical Experimental Methods for Estimating Image Quality—Part 1: Overview of Psychophysical Elements. International Standard. Geneva, Switzerland: International Organization for Standardization. 2005.

AUTHOR BIOGRAPHIES

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 color appearance."

Peter Nussbaum is an associate professor of color 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 color imaging at the Colour and Visual Computing Laboratory, NTNU, Norway.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: High G, Nussbaum P, Green P. Building a metric of color reproduction difference by combining multiple observers in a modular online experiment. *Color Res Appl*. 2023; 1-16. doi:10.1002/col.22913