Does sadness impair color perception? Thorstenson et al.'s plan to find out is flawed

Alex O. Holcombe, University of Sydney

Nicholas J. L. Brown, University of Groningen

Patrick T. Goodbourn, University of Sydney

Alexander Etz, University of Texas at Austin

Sebastian Geukes, University of Münster

Abstract

In their article reporting the results of two experiments, Thorstenson, Pazda, & Elliot (2015a) found evidence that perception of colors on the blue-yellow axis was impaired if the participants had watched a sad movie clip, relative to participants who watched clips designed to induce a happy or neutral mood. Subsequently, these authors retracted their article (Thorstenson, Pazda, & Elliot, 2015b), citing a mistake in their statistical analyses and a problem with the data in one of their experiments. Here, we discuss a number of other methodological problems with Thorstenson et al.'s experimental design, and also demonstrate that the problems with the data go beyond what these authors reported. We conclude that repeating, with minor revision, one of the two experiments, as Thorstenson et al. (2015b) proposed, will not be sufficient to address the problems with this work.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1536v1 | CC-BY 4.0 Open Access | rec: 25 Nov 2015, publ: 25 Nov 2015

Introduction

The reviewers for *Psychological Science*, when evaluating a manuscript for publication, are asked to consider whether the claims made in the article are justified by the methods used (Eich, 2014).

Based on two experiments, Thorstenson, Pazda, and Elliot (2015a) claimed that a state of sadness—induced by watching a short film clip—impairs performance on a specific perceptual task: discrimination of colors along the blue–yellow axis, but not the red–green axis. This conclusion is interesting because it is specific to a single dimension of color space; poor performance on tasks generally, or low willingness to cooperate with an experimenter, would be an unsurprising effect of sadness.

In their retraction notice (Thorstenson et al., 2015b), the authors acknowledged that their data did not support the conclusion that impairment was specific to one part of color space. They also described an anomaly in the histogram of the data of Experiment 2. In our sections below entitled "A confounded comparison" and "Perceptual impairment or change in bias?", we detail other problems with the way the experiments were conducted, the choice of stimuli, and the measures chosen. It is these problems that lead us to believe that even a re-done Experiment 2 will not justify Thorstenson et al.'s conclusion. In an Appendix, we describe a number of statistical issues, going beyond the single problem mentioned in the retraction notice, and report some other anomalies with the dataset, which further undermine confidence in the way the experiments were carried out and in the resulting findings. We offer this analysis in the hope that it will lead to better work in this area in the future.

A confounded comparison

When designing an experiment comparing two conditions, one strives to make the factor of interest the *only* difference between the conditions. Thorstenson et al. contrasted two film clips, one of which was intended to cause the participants to feel sad. The clips should have been chosen to avoid any other differences (on average) in their effect on the participants. The "sadness" clip of Experiment 1 is an excerpt from the animated Disney movie *The Lion King*,

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1536v1 | CC-BY 4.0 Open Access | rec: 25 Nov 2015, publ: 25 Nov 2015

with an unusual lighting that gives the impression of daylight filtered through dust, while the "happiness" clip is a warmly-lit, indoor recording of the comedian Bill Cosby. The "sadness" clip used in Experiment 2 is the same Lion King excerpt, converted from color to grayscale, and the "neutral" clip is a grayscale film of sticks appearing on top of one another at different orientations, converted from color to grayscale.

Unfortunately, differences in mean color and the color variability in these clips may have differently affected subsequent perception of blue and yellow versus red and green. For example, the contrast along the blue–yellow axis might have been greater in the sadness clips. Such a difference would result in reduced sensitivity to blue–yellow (Krauskopf, Williams, & Heeley, 1982). The use of grayscale clips does not eliminate this issue. An analysis of the HSB values in the movie files posted online indicates that the mean color of the grayscale clips is bluish-reddish, with some saturations approaching 5%. These grayscale clips, therefore, may have had differences in average color as well as color contrast that were uncontrolled. To resolve the issue, the colors displayed on the laboratory screen must be measured with a colorimeter. The authors should have made these measurements and reported them in their paper, in order to provide an indication of whether simple contrast adaptation specific to each color axis would occur when viewing the clips. In the absence of a report of these measurements or any mention of the issue, it appears that Thorstenson et al. (2015a) did not take the appropriate steps to eliminate the possibility that a classic process in color perception could explain the results.

Blue, yellow, green and red are all defined relative to a white point that is quite flexible in the human visual system. Just as one can adjust the white balance of a camera, for humans the point considered to be the center of color space changes depending on the palette of colors we are confronted with (Webster & Silverman, 2008). Unfortunately, Thorstenson et al. displayed their test stimuli in a manner unsuited to controlling the participant's white point. Color perception experiments typically use a neutral grey or white background to provide a white-point reference for participants, alongside the test stimulus. Thorstensen et al.'s use of full-field color, without any simultaneous reference stimulus, makes categorization of desaturated patches problematic. In such circumstances, the participants' white points may be more dependent on the color content of the movie clip they viewed previously, which as mentioned previously appears to have been uncontrolled. In addition, the lack of a grey or white reference stimulus may cause

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1536v1 | CC-BY 4.0 Open Access | rec: 25 Nov 2015, publ: 25 Nov 2015

participants to more often be completely unable to judge the stimulus color. In such circumstances, responses may be particularly prone to influence by cognitive factors or by priming (García-Pérez & Alcalá-Quintana, 2012).

Color is not the only possible confounding difference between the two types of clips. The clips likely also differed in interest, action, and other features. It is difficult to know whether such features might have affected participants' color perception. It certainly is possible for such differences to bias the participants' responses when they are uncertain of the stimulus' color. Of course, any two individual clips will have featural differences. Because of this, good experimental design requires using a large set of clips, assessing the various featural differences between the stimuli, and either matching the two groups of clips carefully on their features, or modeling them as random effects in a mixed-effects model (Wells & Windschitl, 1999).

Perceptual impairment or change in bias?

Thorstenson et al. (2015a) concluded that sadness "impair[s] color perception on the blue–yellow color axis" (p.1). But a signal detection theory analysis would be needed to show whether the decrease in accuracy found was due to color perception indeed being impaired (i.e., there had been a decline in sensitivity along the blue–yellow axis), or whether the judgments of the sadness group were instead biased away from blue and yellow. Studies of perception have long used this type of analysis to distinguish between a change in perceptual ability and a change in, say, cognitive bias to press the blue or yellow button rather than the red or green one (Green & Swets, 1966). Unfortunately, Thorstenson et al's plan to simply re-do Experiment 2 would not allow for the appropriate analysis. In Experiment 2, participants were tested in only two trials for each stimulus. Much more data would probably be needed to determine whether the participants' ability to discriminate the colors declined, rather than the accuracy change being attributable to a decline in the participants' bias toward pressing the blue or yellow button (instead of the red or green). For the four-alternative categorization task used by Thorstenson et al. (2015a), a multivariate extension of signal detection theory should be used, such as general recognition theory (Ashby & Townsend, 1986).

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1536v1 | CC-BY 4.0 Open Access | rec: 25 Nov 2015, publ: 25 Nov 2015

The analyses of Thorstenson et al., and also those that we have suggested above, assume statistical independence of participants' accuracy on the red–green stimuli and the blue–yellow stimuli. Unfortunately, however, this assumption may be wrong. Participants' accuracy on one axis might affect their guessing strategy on another. In Experiment 1 for example, accuracy was very high on the blue–yellow axis, suggesting that many participants may have had a clear color percept of the blue or yellow stimuli, but been less certain about the red and green stimuli. If so, when an unclear patch came up and they guessed, they may have been unlikely to guess blue or yellow, in an effort to balance their responses across the available options (many participants may have correctly guessed that the stimuli were roughly equally distributed among the four categories). This would artifactually improve performance on the red and green stimuli. Modeling this phenomenon, however, would be difficult. Even if we had access to the raw responses (rather than the summary data provided by Thorstenson et al.), it would be difficult to estimate the participants' guessing strategy. To avoid this problem in a future version of this experiment, we suggest that Thorstenson and colleagues should consider adopting the two-alternative forced choice design commonly used in psychophysics.

Thorstenson et al. are not the only researchers to have used bias-prone measures of perception to support claims that some non-perceptual state can influence perception. In a recent paper, Firestone & Scholl (2015) cataloged many other examples, and provided useful discussion.

Conclusion

While we strongly support the decision of Thorstenson et al. (2015b) to retract their article (Thorstenson et al., 2015a) on the basis of the problems they noted with Experiment 2, it appears that the basic methodology of both of their experiments is flawed. As Thorstenson and colleagues move forward, together with others who seek to assess whether mood and other factors can influence perception, they have an opportunity to bring their work up to modern standards of statistical and psychophysical rigor. Doing so for experiments like those of Thorstenson et al. would involve: 1) Careful control of the visual differences between the movie clips, or, better, mood induction via non-visual stimuli such as an audio recording of a story; 2) The use of many movie clips or recordings, and mixed-effects analysis to address differences

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1536v1 | CC-BY 4.0 Open Access | rec: 25 Nov 2015, publ: 25 Nov 2015

that cannot be eliminated between any two clips or recordings; 3) A baseline measurement of color perception; 4) An analysis based on signal detection theory.

There are further issues specific to Thorstenson et al.'s (2015a) dataset that were not described in their retraction. Some of these issues affect Experiment 1, which Thorstenson et al. indicated that they plan to re-publish. We describe and discuss these issues in Appendix A.

Contributions

AOH coordinated the team and wrote most of the main section of the article. NJLB started the discussion (on Twitter), wrote the R code to perform the detailed analysis of the dataset, and wrote most of the Appendix. PTG contributed to the analysis, the points about color vision, and Figure 4. AE contributed to the discussion of stimuli problems and (lack of) necessary stimulus sampling. SG contributed to the analysis, contributed early versions of some of the R code, and helped revise the manuscript.

Appendix A

Statistical shortcomings

Thorstenson et al.'s (2015a) crucial claim was that there was a difference in performance between their two measures, namely color perception along the blue–yellow axis and color perception along the red–green axis. However, these authors provided no statistical test of a difference in the effect of the film clip on blue–yellow compared to red–green. This problem was widely discussed on blogs and on PubPeer (PubPeer, 2015), and was acknowledged by Thorstenson et al. (2015b) in their retraction notice. Thorstenson et al. (2015a, p. 4) pointed out that the difference between red–green and blue–yellow color perception, such that "sadness influenced chromatic judgments about colors on the blue–yellow axis, but not those on the red–green axis," is critical to ruling out "the possibility that sadness simply led to less effort, arousal, attention, or task engagement". The demonstration of such a difference implies a statistical interaction between the "emotion condition" and "color axis" factors. However, the authors did not report the results of such an interaction in either of their experiments. When we (and the

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1536v1 | CC-BY 4.0 Open Access | rec: 25 Nov 2015, publ: 25 Nov 2015

authors of various blogs, such as Areshenkoff, 2015) tested this interaction with the published data, we found that it was not statistically significant: Experiment 1, $F(1, 125) = 3.51$, $p = .06$; Experiment 2: $F(1, 128) = 0.40$, $p = .52$. Thorstenson et al. (2015b) in their retraction notice reported a z test (for unknown reasons, they did not use a conventional statistical interaction) to test the same issue.

A further potential source of error is that Thorstenson et al. (2015a) did not record the color perception performance of their participants before the film clips were shown in either experiment. It was apparently considered sufficient to randomize the participants to watch one of two film clips; presumably the reasoning was that this randomization would guarantee that there was no significant difference in baseline performance between the two groups. However, even if this assumption were to be confirmed, the two groups would necessarily differ at baseline, even if by only a small amount, and such a difference could have an effect on the outcome given the relatively small sample sizes involved (Saint-Mont, 2015). We believe that it would have been useful for these differences to be measured and included in the subsequent analyses, given that Thorstenson et al.'s hypothesis was that sadness would "impair" (i.e., reduce, compared to a previous state) participants' color perception. In addition, using a change score for each participant can increase statistical power by reducing the contribution of variation among participants to the error term.

Finally, we note that Thorstenson et al.'s (2015a) experimental design assumes the complete independence of participants' accuracy on the two sets of stimuli (red–green and blue–yellow). We discuss a possible violation of this assumption in our "Perceptual impairment or change in bias?" section above.

Anomalies and strange patterns in the data

Large numbers of participants with identical scores. We observed a strange pattern in the data for the blue–yellow axis in Thorstenson et al.'s (2015a) Experiment 2. Specifically, a very large number of participants (53 out of 130) had a score of exactly 50%, corresponding to 12 out of 24 correct responses, with every other number of correct responses (10, 11, 13, 14, etc) being achieved by a much smaller number of participants. This is illustrated in Figure 1, where the

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1536v1 | CC-BY 4.0 Open Access | rec: 25 Nov 2015, publ: 25 Nov 2015

spike at the 50% level is clearly visible. This problem was one of the reasons given by Thorstenson et al. (2015b) for retracting their article.

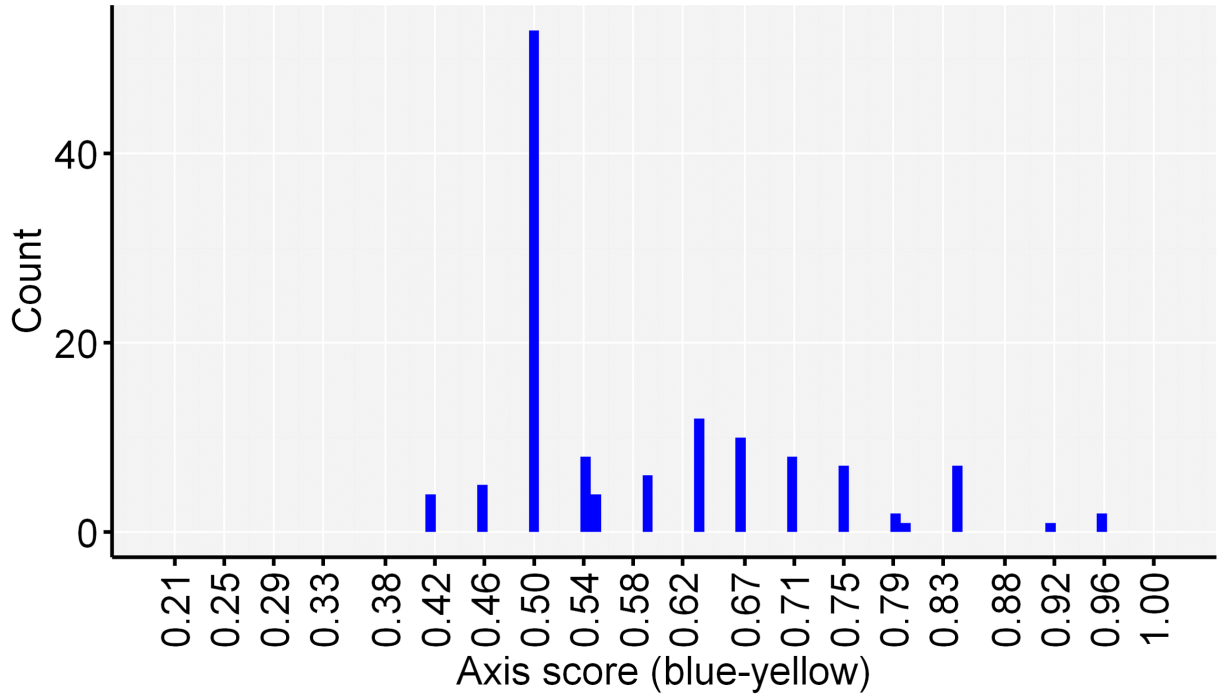


Figure 1. Number of occurrences of each recorded blue–yellow axis score (expressed as a fraction of 24 attempts) recorded by Thorstenson et al. (2015a) in Experiment 2. The pairs of bars close to each other at 0.54 and 0.79 correspond to cases where the fractions appear to have been incorrectly rounded by hand.

Closer examination of the per-color patch data, supplied by Christopher Thorstenson at our request (see "The datasets" section), shows that of the 53 participants scoring exactly 50%, 49 (i.e., 37.7% of all participants in Experiment 2) had identical scores for both colors, namely 6.0 (100%) for blue and 0.0 (0%) for yellow. (In the patch data each correct observation counts for a half-point, so that scores for each color range from 0.0 to 6.0 in increments of 0.5; thus, a score of 6.0 corresponds to 12 correct responses out of 12.) We are at a loss to explain this phenomenon, which affects the two experimental conditions equally (26 of the 49 participants with this 12–0 split were in the neutral condition, versus 23 in the sadness condition). There seems no reason to suppose that the undergraduate participants in this experiment would have

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1536v1 | CC-BY 4.0 Open Access | rec: 25 Nov 2015, publ: 25 Nov 2015

been markedly less sensitive to yellow than those in Experiment 1. However, even if their ability to distinguish the color yellow was affected by some environmental factor, or if they had been accidentally (perhaps due to a software problem) shown, say, a gray patch instead of a yellow one, their expected score for yellow would be 1.5 (i.e., 3 correct identifications out of 12 attempts) by chance alone.

Manual calculation of percentages. A further concern with Thorstenson et al.'s (2015a) published dataset is that the color perception values for both axes in Experiment 2 appear in some cases to have been converted from counts to percentages by hand, rather than having been calculated by a computer. These percentage values, reported to two decimal places, ought to be the result of dividing the number of successful attempts on each axis (i.e., the total number of correct identifications of red or green patches for the red–green axis, and the total number of correct identifications of blue or yellow patches for the blue–yellow axis) by 24. For example, an examination of the patch scores shows that participants #4 and #5 both scored a total of 6.5 for blue and yellow patches combined, corresponding to 13 correct identifications out of 24 on the blue–yellow axis. However, in the published dataset file, participant #4 has a value of 0.54 for the corresponding percentage variable BY_ACC, whereas participant #5 has a value of 0.55 for the same variable (the true value of 13/24 being *0.541666*). It is difficult to imagine how this might have occurred if these percentages had been generated by a computer. These observations suggest that there has been some form of manual intervention in the dataset, with the attendant risks of accidentally introducing other inaccuracies due to typing errors, incorrect cell selection, etc. The data for at least 26 of the 130 participants in Experiment 2 were affected by this issue (the exact number depends on whether the smaller or larger calculated number for the percentage is considered to be correct); no errors of this kind were observed in the dataset for Experiment 1.

The dataset discussed above was replaced by new files on September 14 and 15, 2015, about two months after the original files were posted, according to the change log for the project page at `https://osf.io/sb58f/` . The new files were accompanied by the comment "Rounding errors may result in slightly different values. Updated data to reflect 3 decimals, accordingly." Examination of these files shows that some, but not all, of the percentage values for the color axis scores are indeed now reported to three, rather than two, decimal places. However, this does not appear to have been done correctly or consistently. We give two examples, both concerning

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1536v1 | CC-BY 4.0 Open Access | rec: 25 Nov 2015, publ: 25 Nov 2015

the data file corresponding to Experiment 1 (although similar problems affect the file for Experiment 2 as well):

1. In the original file, the values on the blue–yellow axis for participants #5 and #11 (cells D4 and D9) both contained the value 0.67, which corresponds—within the limits of precision—to these participants' total scores of 16 out of 24 correct responses for blue and yellow patches combined. In the new file, cell D4 has been changed to 0.665, but cell D9 has not.
2. In the original file, the values on the red–green axis for participants #31 and #50 (cells C28 and C47) contained 0.80 and 0.79, respectively. These participants both scored a total of 19 out of 24 attempts for the red and green patches combined. Since 19/24 = *0.791666*, the value of 0.80 for participant #31 was incorrect. In the new file, this cell (C28) has been "corrected," but only to 0.795, which is still (a) less close to the true value than 0.79, and (b) different to the value for participant #50, even though both participants had the same number of correct responses.

Large differences in skewness between experiments. An examination of the distribution of the scores for the two color axes reveals considerable differences between Experiment 1 and Experiment 2. In Experiment 1, the distribution for both axes was substantially negatively skewed, with the majority of participants correctly identifying almost all of the patches for all four colors (Figure 2, left-hand side). In Experiment 2, the score distribution was different for each axis. For the red–green axis (Figure 2, right-hand side, top panel) the scores were approximately normally distributed: roughly similar numbers of participants achieved each possible score, with a small number having very low or very high scores. In contrast, the blue–yellow axis was positively skewed, displaying the "spike" discussed previously (Figure 2, right-hand side, bottom panel).

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1536v1 | CC-BY 4.0 Open Access | rec: 25 Nov 2015, publ: 25 Nov 2015

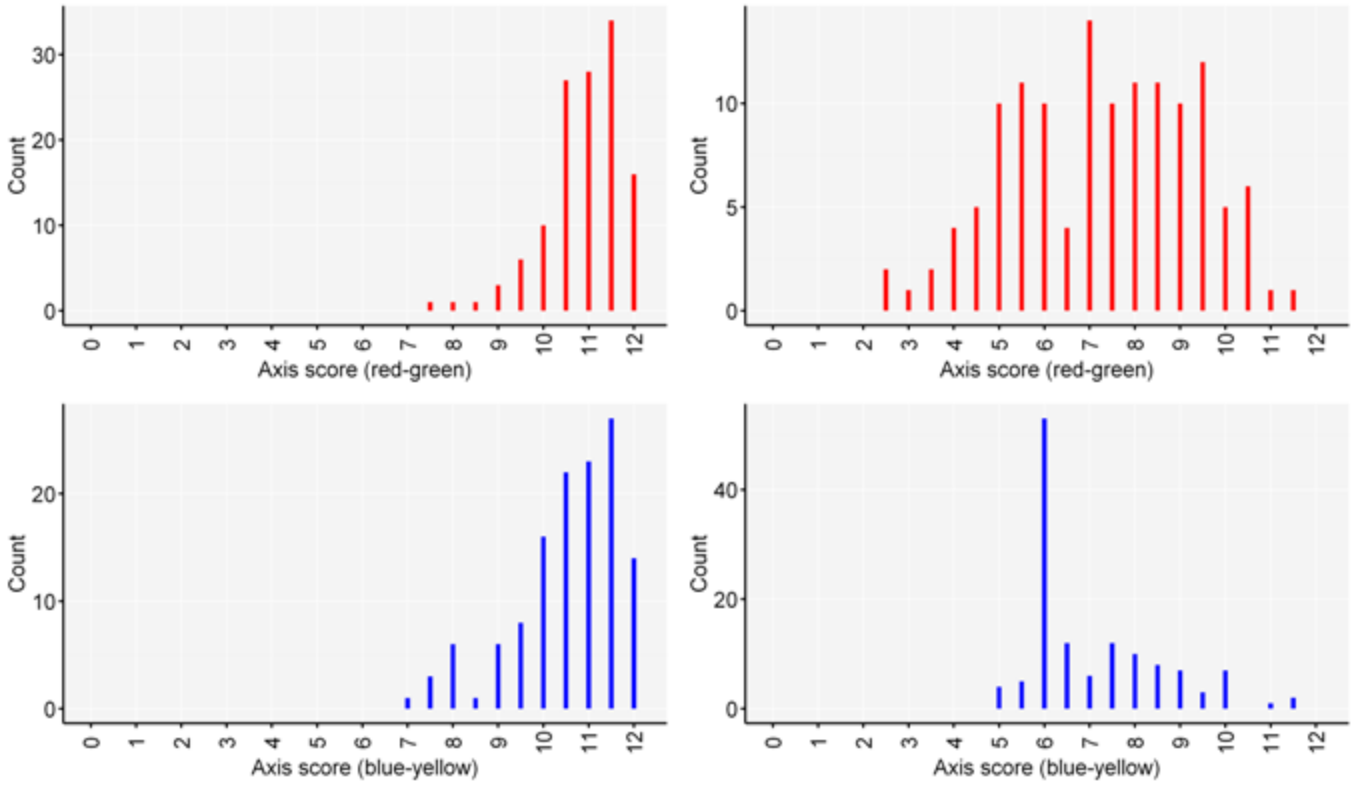


Figure 2. Distribution of per-axis scores for Thorstenson et al.'s (2015a) Experiment 1 (left) and Experiment 2 (right). The range of scores on the X-axis is 0–12, reflecting Thorstenson et al.'s scoring scheme of 0.5 points per correct answer, with 12 trials per color and two colors per axis.

Using the patch-level data, we broke the two-color axis scores down into individual colors, as shown in Figure 3. For Experiment 1, the per-color data more or less followed the pattern of the two-color axis of which each color was a part (Figure 3, left-hand side); this was also true for the red–green axis in Experiment 2 (Figure 3, right-hand side, top two panels). However, an even stranger pattern emerged for the blue–yellow axis in Experiment 2 (Figure 3, right-hand side, bottom two panels). Of the 130 participants, 106 (81.5%) scored a maximum 6.0 (corresponding to 12 correct responses) for blue, while 56 (43.1%) scored zero for yellow. The observed "spike" at 50% (i.e., 12 out of a possible 24 correct responses) for the blue–yellow axis is thus mostly explained by people who had a perfect score for of 12 blue, while completely failing to recognize yellow patches at any saturation and thus obtaining a score of 0.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1536v1 | CC-BY 4.0 Open Access | rec: 25 Nov 2015, publ: 25 Nov 2015

Figure 4 plots Thorstenson et al.'s (2015a) participants' performance for each color, broken down further into the proportion of correct responses for each saturation level. (Recall that participants were asked to identify colors at each of six different levels of saturation.) In Experiment 1, this resulted in an apparent ceiling effect, with mean color-accuracy performance already reaching 90% or more at the third-lowest color saturation level (.10) and leveling off as saturation increased thereafter (Figure 4, left-hand side). In Experiment 2, the ceiling effect disappeared for the red–green axis, for which scores on both colors improved approximately linearly with increasing color saturation (Figure 4, right-hand side, top two panels); however, on the blue–yellow axis, the effect of the split between the two colors is once again clear, with the ceiling effect became even more pronounced for blue, while the proportion of scores for yellow is low even at the highest color saturation level (Figure 4, right-hand side, bottom two panels).

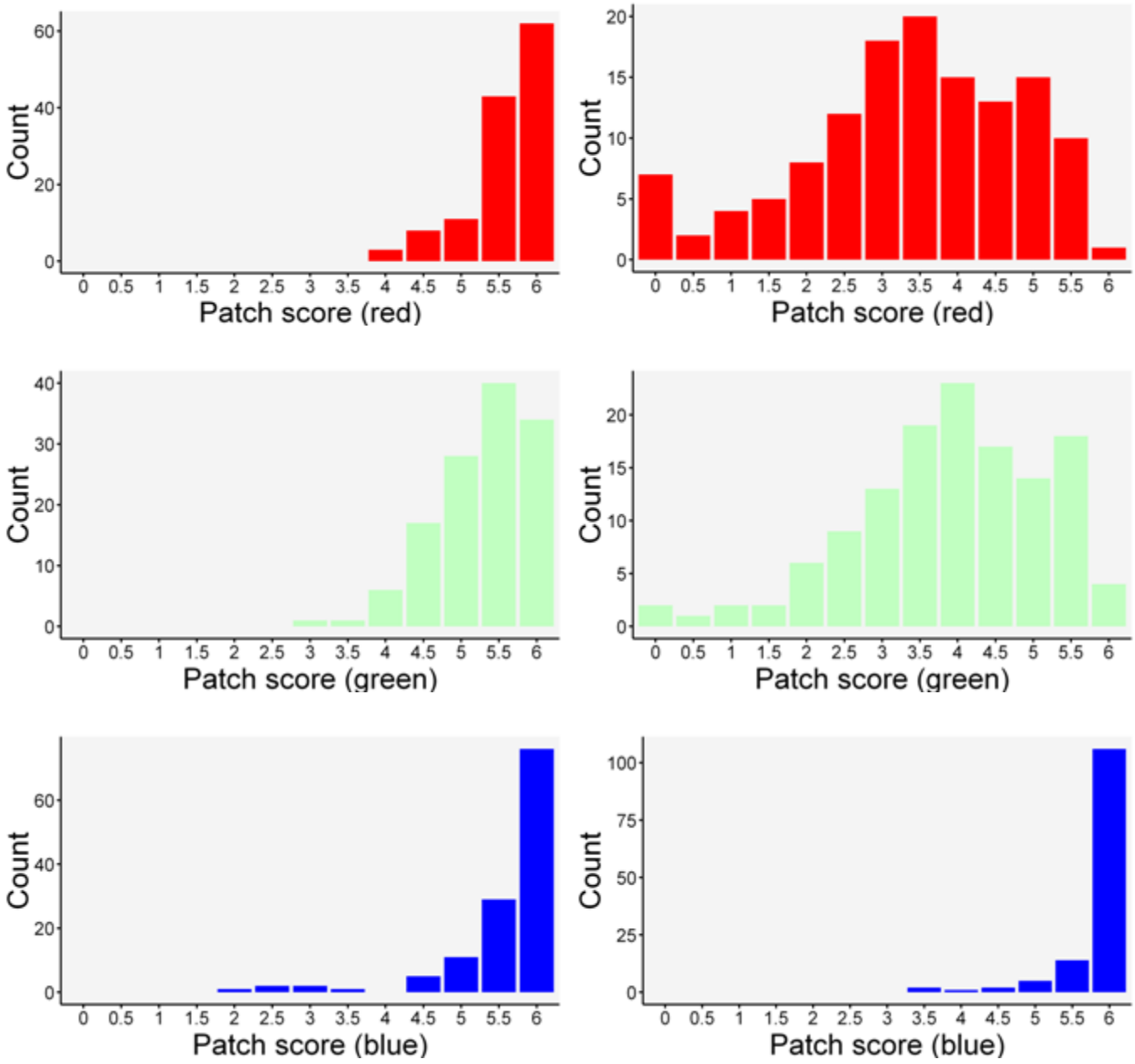
It is difficult to imagine what might have caused these remarkable results in Experiment 2. Thorstenson et al.'s (2015a) Method section for this experiment suggests that the only change that was made from Experiment 1 was the nature of the film clips that were shown to participants. The differences for both axes (and, indeed, for all four colors) between Experiments 1 and 2—regardless of the film clip watched by participants—are puzzling, given that both samples were apparently drawn from the same population of undergraduates and hence ought not to differ widely in their physiological characteristics. Because the color characteristics of the two sets of film clips were apparently not well-controlled, one possible explanation for this discrepancy is differential adaptation of the color mechanisms in the visual system, which adds to our concern of a confound (see our section "A confounded comparison?"), but we have difficulty believing that this could account for such a substantial difference between the two experiments.

Given that the extreme blue–yellow scores in Experiment 2 were obtained from participants in both the neutral and sadness conditions, a further possibility is that simply watching grayscale film clips for a few minutes was sufficient to substantially distort participants' color vision (on the blue–yellow axis only). However, if Thorstenson and colleagues had noticed such a finding, they would presumably have mentioned it in their article, and perhaps alerted colleagues in the University of Rochester's Department of Pharmacology and Physiology to this remarkable finding. Otherwise, we are left with two possible conclusions: either around 40% the participants

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1536v1 | CC-BY 4.0 Open Access | rec: 25 Nov 2015, publ: 25 Nov 2015

in Thorstenson et al.'s (2015a) Experiment 2 all had the same problem with their vision (which was not shared by any of the participants in Experiment 1), or some form of equipment failure or other technical problem caused this unusual pattern of values to be recorded. In any case, it seems likely that Thorstenson et al. failed to notice this anomaly when examining their data prior to performing their statistical analyses.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1536v1 | CC-BY 4.0 Open Access | rec: 25 Nov 2015, publ: 25 Nov 2015



PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1536v1 | CC-BY 4.0 Open Access | rec: 25 Nov 2015, publ: 25 Nov 2015

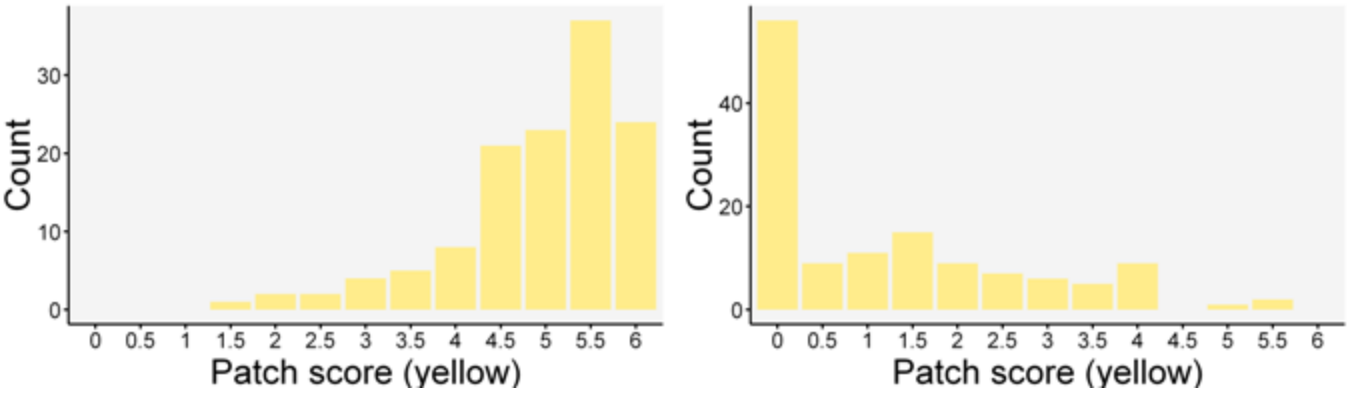


Figure 3. Per-color scores for each color patch for Thorstenson et al.'s (2015a) Experiment 1 (left) and Experiment 2 (right).

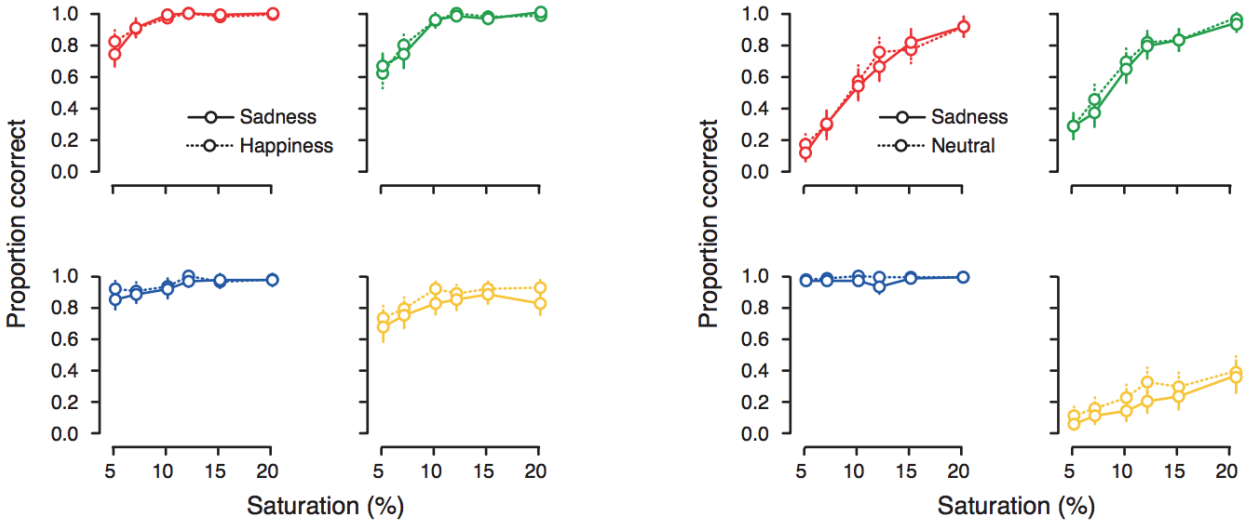


Figure 4. Proportion of correct responses for each of the six color saturation levels and each of the four colors tested, split by experimental condition for Thorstenson et al.'s (2015a) Experiment 1 (left) and Experiment 2 (right).

A note on the datasets

We applaud the decision by Thorstenson et al. (2015a) to publish their data, which led to their published article being awarded *Psychological Science*'s "Open Data" and "Open Materials"

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1536v1 | CC-BY 4.0 Open Access | rec: 25 Nov 2015, publ: 25 Nov 2015

badges. However, the published files were not sufficient for some of our analyses, which required more detailed data that Christopher Thorstenson kindly sent to us. Below, we clarify how we refer to these files and what they contain:

- The **published data**. This term refers to the archive file "Data.zip," which was posted on the Open Science Framework at `https://osf.io/sb58f/` on July 5, 2015. It contains one Excel file for each of the two experiments reported in Thorstenson et al.'s (2015a) article. Each file contains a row for each participant, containing his or her scores on each color axis (red–green and blue–yellow) expressed as a percentage of the 24 attempts made on that axis (two attempts per color for each of the two colors on the axis and for each of six saturation levels). As we have noted in our section entitled "*Manual calculation of percentages*" (which deals with the specific problems caused by this change), an updated version of this archive was uploaded, and the original version deleted, on September 14 and 15, 2015. Unless mentioned otherwise, we refer to the original version of the published data throughout this article.

- The **patch data**. This term refers to two supplementary Excel files (one per experiment), supplied to us by Christopher Thorstenson, with each cell containing a combined score for the participants' two responses for each color and saturation level. The score for each case is either 0.0, 0.5, or 1.0, corresponding to 0, 1, or 2 correct responses.

As our own contribution to open science, we have archived the R code that we used to analyze the data and generate our figures at the Open Science Framework (OSF; `https://osf.io/sbhn9/`). This code works with the original dataset files uploaded to OSF by Thorstenson et al. (which, it appears, have now been deleted), together with the patch data files. We include these data files alongside our R source file.

References

Areshenkoff, Corson N. (2015). "On the importance of plotting; or — Psych. Science will publish anything". [blog post] http://areshenk-research-notes.com/on-the-importance-of-plotting-or-psych-science-will-publish-anything/

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1536v1 | CC-BY 4.0 Open Access | rec: 25 Nov 2015, publ: 25 Nov 2015

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review*, *93*, 154–179. doi:10.1037/0033-295X.93.2.154

Eich, E. (2014). Business not as usual. *Psychological Science*, *25*, 3–6. doi:10.1177/0956797613512465

Firestone, C., & Scholl, B. J. (2015). Cognition does not affect perception: Evaluating the evidence for "top-down" effects. *Behavioral and Brain Sciences*, 1–72.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York, NY: Wiley.

Krauskopf, J., Williams, D. R., & Heeley, D. W. (1982). Cardinal directions of color space. *Vision Research*, *22*, 1123–1131. doi:10.1016/0042-6989(82)90077-3

PubPeer (2015). Sadness Impairs Color Perception. Retrieved November 19, 2015, from https://pubpeer.com/publications/989FBE60680F308F7BF98BB22F3C50

Saint-Mont, U. (2015). Randomization does not help much, comparability does. *PLoS ONE*, *10*, e0132102. doi:10.1371/journal.pone.0132102

Thorstenson, C. A., Pazda, A. D., & Elliot, A. J. (2015a). Sadness impairs color perception. *Psychological Science*. Online publication ahead of print. doi:10.1177/0956797615597672

Thorstenson, C. A., Pazda, A. D., & Elliot, A. J. (2015b). Retraction of "Sadness impairs color perception." *Psychological Science*, *26*, 1822. doi:10.1177/0956797615597672 [*sic*; the retraction notice displays the same DOI as the original article.]

Webster, M.A., & Leonard, (2008). Adaptation and perceptual norms in color vision. *J Opt Soc Am A Opt Image Sci Vis*, *25*(11): 2817–2825.

Wells, G. L., & Windschitl, P. D. (1993). What's in a question? *Contemporary Psychology*, *38*, 383–385. doi:10.1037/033227

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1536v1 | CC-BY 4.0 Open Access | rec: 25 Nov 2015, publ: 25 Nov 2015