Color naming and color matching: A reply to Kuehni and Hardin

Pendaran Roberts & Kelly Schmidtke

Forthcoming in Review of Philosophy and Psychology. The final publication is available at Springer via http://link.springer.com/article/10.1007/s13164-014-0225-0

Abstract. We recently conducted an experiment to show that a lot of the empirically measured disagreement cited to support the premise that there is *mass* perceptual disagreement about the colors, a premise often cited by philosophers, is due to conceptual factors. Kuehni and Hardin object to how we measured disagreement and to various aspects of our experimental design. In this reply, we defend our study.

1 Introduction

In our (2012) paper, we conducted an experiment designed to show that a lot of the measured disagreement used by philosophers to support the premise that there is *mass* perceptual disagreement about the colors, roughly what we called 'P-Disagreement,' is due to conceptual factors. There are two types of tasks used to measure disagreement about color: naming tasks and matching tasks. A difference between naming tasks and matching tasks is that the former but not the latter requires participants to have color concepts associated with their color terms. We discussed two conceptual factors that may be relevant to naming tasks: (1.) the variation in the color concepts (e.g. the concept of unique red) associated with color terms (e.g. the term 'unique red'), and (2.) the broadness of the color concepts associated with color terms. A relevant conceptual factor not explicitly discussed in our article is the ability of people to apply their color concepts. If naming tasks result in more disagreement than matching tasks, then it would seem that the additional disagreement must be due to

conceptual factors that arise because of the unique requirement of naming tasks. Our (2012) experiment shows that naming tasks result in more disagreement.

Kuehni and Hardin (henceforth "K&H") take issue with our experiment (2014). They object to the way we measured disagreement, specifically that we did not use the coefficient of variation, and to various aspects of our experimental design: that our samples were not perceptually uniformly spaced, that our samples were collapsed into units different from the standard, and that what we defined as our standards for unique green and unique red are not really unique green and unique red. In this reply, we defend our experiment from these criticisms. We respond to their criticism about the way we measured disagreement (Sect. 2) and then to their criticisms of our experimental design (Sect 3.).

2 Measuring disagreement

K&H say that we should have used relative standard deviation (i.e. coefficient of variation, COV). We doubt this. We provide two arguments by analogy and next a statistical one. Our first argument by analogy is from Livers (1942). He looks at a temperature analysis for two samples, A and B. First when the samples are measured in Fahrenheit he finds the COV for A to be greater than B. When the same information is then analyzed using Celsius the results reverse. In contrast, whether the results are analyzed in Fahrenheit or Celsius he finds the standard deviation for A to be the same as for B. Livers stresses that "the standard deviation is not changed if a shift of origin [the scale in this case] takes place, whereas the arithmetic mean is changed by precisely the amount of the shift. This means that we can obtain any value we wish for the coefficient of variation by simply changing the origin from which the variates are measured" (p. 893). This demonstrates an unfortunate consequence of the COV that by altering the scale you alter its outputs. In our experiment, we altered the scale by using units different from the standard, and in so doing large numbers became small

numbers. So, given this, it is entirely reasonable that we did not use the COV. (Whether using units different from the standard is problematic will be addressed in Sect. 3.)

Let us now look at our second argument by analogy. Say a person has two sets of five objects that vary in mm length (A = 9, 17, 5, 6, 13; B = 125, 30, 135, 130, 80). Are the objects' lengths more similar in A than in B? The objects in A have a mean length of 10 mm (range = 5-17), and the objects in B have a mean length of 100 mm (range = 30-135). The standard deviation for A is 5.00 (95% descriptive confidence interval 5.62 to 14.38), and the standard deviation for B is 44.86 (95% descriptive confidence interval 60.68 to 139.32). The COV for A is 50.00%, and the COV for B is 44.86% (a confidence interval cannot be constructed from COV's). If the COV is the proper index, we must conclude that the objects' lengths are less similar in A than in B, but this conclusion is dubious. The mean for A being *smaller than* for B is irrelevant to how similar the lengths are in A compared to B under any quotidian sense of the relevant question. Under any quotidian sense, what is relevant is that there is a lot less unqualified variation around the mean for A than for B. Hence, under such a sense, we should conclude, "yes, the objects' lengths are more similar in A than in B." Thus, one ought to accept that relative variation is not always the right indicator of variation, even when the means widely differ. Similar to this example with object lengths, the question for our experiment can be put as follows: are people's choice responses more similar in matching than naming tasks? As such, it is entirely reasonable for us not to have used the COV. It is wholly reasonable to think that our mean participant responses should not have altered how we understand the variation around the means in the tasks.

One may retort that we used what is essentially a distance metric (units different from the standard). One may go on to say that variability increases as distance increases. For example, if people were estimating the distance between New York and Boston and New York and Paris, it would be expected that there would be more variability in the latter case

than the former. So, the COV would be used to account for this. Our opponent may conclude by saying that, given this analogy, we should have used the COV. In reply, although we do use what is essentially a distance metric, there is a significant difference between our case and the one just described. In the case described, people are being asked to estimate the distance between cities, and it is expected that the larger the distance the more variable the estimates. However, in our experiment, people were not asked to estimate the distance of anything. They were merely asked to name true red or green, or find the matching pair. There is no reason to expect variability to depend on the distance from the standard in our case. The standard was just used to order the samples so we could compare participant responses.

We shall now look at our statistical argument for why the COV would be an inappropriate index of variation for our analysis. For consistency with K&H's analysis we will use the mean COV to make our point, but please note that we are not advocating the use of parametric statistics for our data. Let us first look at the green matching task. If one looks at the units different from the standard, the green matching task COV = (4.5/7.5) *100 = 60%. Compare this to the raw units for Cyan in the green matching task, the raw unit with which there was the most variation in the green tasks, COV = (4.77 / 95) *100 = 5%. Thus, the COV for the units different from the standard metric is larger than the COV for the raw units metric for Cyan (60% vs 5%). Let us now look at the green naming task. Using the units different from the standard for the green naming task the COV = (9.4/17.92) *100 = 52%. Compare this to the raw units for Cyan in the green naming task where the COV = (11.1/83.33) *100 = 13%. Hence, again the COV for the units different from the standard metric is larger than the COV for the raw units for Cyan (52% vs 13%).

The above way the units different from the standard and the raw units for Cyan affect the COV causes a serious problem for using the COV as an index of variation for our analysis. Consider the main question of interest in our experiment: Is variation greater in

matching or naming tasks? The difference between the matching and naming tasks for green if one uses the units different from the standard is 60% vs. 52%. So, using this metric, variation in the matching task is greater. However, if one uses the raw units for Cyan in the green tasks (the raw unit, as we said, with which there was the most variation in the green tasks), the difference between matching and naming is 5% vs. 13%. Thus, using the raw units metric for Cyan in the green tasks, variation in the naming task is greater. Both metrics are derived from the same participant responses, but the units different from the standard and the raw units for Cyan provide opposite findings when using the COV. This is a serious problem for whether we should use the units different from the standard or the raw units.

Most would say that the raw units metric is better, for it is closer to the data. As a general rule, we agree, but there is an issue with using the raw units in our case. Participants' choice responses were to sample green or red colors, not to the C[yan], Y[ellow] or M[agenta] units used to create the samples. So, using raw units artificially decreases variation. Our stimuli differed exclusively in C, Y, or M, so if a participant chose a stimulus that differed in C, he/she could not choose one that differed in Y or M. In the case of green, for example, most chose a stimulus that differed in C (hence, as we said, the variability was greatest for this unit). Thus, there are not large differences in the green tasks for the Y and M units. Nevertheless, even though we do not find large differences for the Y and M units, this does not imply that there was not a great deal of variation in the participants' responses for the green naming task, and indeed there was. So, we have shown 1.) that the COV results reverse dependent on whether one uses units different from the standard or raw units for Cyan, and 2.) that normally using raw units would be better but not in our case. Thus, for our experiment the COV is seriously problematic and should not be used.

The Brown-Forsythe test avoids this problem. As noted in our original paper, using the units different from the standard, the Brown-Forsythe test found a significant difference

between matching and naming tasks. This difference remained for the green tasks but not the red tasks. What happens if we apply the Brown-Forsythe test to the raw units? As a reminder, we expect a decrease in variability when using the raw C, Y, and M units. Indeed, unsurprisingly, we lose the significant difference between matching and naming when using the raw units. However, unlike with the COV analysis, which reverses the findings, the difference goes in the same direction for the green matching and naming tasks for both the units different from the standard ($F(1, 22) = 6.06$, $p < 0.05$) and the raw units for Cyan ($F(1, 22) = 8.08$, $p < 0.01$). Thus, the Brown-Forsythe test does not run into the same worrying issues as the COV: it does not have the worrying consequence that its results go in opposite directions depending on whether one uses units different from the standard or the raw units for the unit, Cyan, for which variability was the greatest in the green tasks.

3 Experimental design

Our comparisons differed from the standard by different amounts of C, Y or M (exclusively) by 5 unit steps. Our design is such that the samples created by it are not going to be perceptually uniformly spaced. Suppose that one has two colors that are perceptually two units apart, call them X_1 and X_2, and another two colors that are perceptually 10 units apart, call them X_1 and X_{10}. Now suppose that it takes 10 units of M to get from X_1 to X_2 and 5 units of M to get from X_1 to X_{10}. Our experimental design, then, would place X_1 and X_2 10 units apart even though they are only 2 units apart perceptually, and X_1 and X_{10} 5 units apart even though they are 10 units apart perceptually. Thus, our design would tend to overstate the meaningfulness of the disagreement about whether an object is X_1 or X_2 and understate the meaningfulness of the disagreement about whether an object is X_1 or X_{10}.

K&H believe that the fact that our design may overstate the disagreement about some colors and understate it about others is a significant problem for our experiment. We do not

think that this assessment is warranted. We did not design our experiment to establish whether there is more disagreement about whether an object is X_1 or X_2 than X_1 or X_{10}. The sole purpose of our experiment was to test our hypothesis that naming tasks result in more disagreement than matching tasks. Thus, for the considered objection to go through there must be a reasonable expectation that the lack of perceptual uniformity would cause more disagreement in one task type than the other. Is there such an expectation?

Although it seems unlikely, a possibility is that the lack of perceptual uniformity created two or three samples that were equally the best candidates for being true green or red in the naming task thus dividing up participants' responses in this task while having no effect on the matching task where there remained one comparison that best matched the standard. We find little evidence that this farfetched possibility actually obtained. The distribution of our participants' Cyan responses for green naming were platykurtic (flat). For other types of responses, the distributions were more skewed than multi-modal. So, we think that the farfetched possibility in question, without more said in its support, is not a significant problem for our experiment. This said, as our sample sizes were not large, we admit that we cannot be *certain* that the relevant possibility did not obtain and recommend further research.

A related objection that K&H raise is that our design collapses samples that differ in C, Y, or M into units different from the standard. As a result, our participants' responses can only indicate variation in this dimension. Thus, K&H say, "there is no distinction in both Red and Green between yellowish and bluish samples in the data used for calculating means and standard deviations" (2014). Although we admit that this design has limitations, we do not think that it is faulty given our purposes. We were interested in seeing whether there were global differences between task types. Specifically, we wanted to see whether naming tasks result in more disagreement than matching tasks. Thus, for the objection to succeed there must be a reasonable expectation that collapsing samples that differ in C, Y, or M into units

different from the standard would cause more disagreement in one task type than the other. We can think of no obvious reason for why collapsing samples thusly would have this consequence. The comparison items were the same for both task types. So, we do not find the objection under consideration from K&H to be convincing. This said, if there is a reasonable expectation to suspect that collapsing samples as we did would cause more disagreement in one task type than the other, we recommend further research be done.

A final issue that K&H raise is that what we defined as our standards for unique green and unique red are not really unique green and unique red. We have no doubt that our definitions of the standards do not capture unique green and unique red. That this is correct is shown by the fact that our mean for red naming was 14.58 and our mean for green naming was 17.92. These means are more the result of defining the standard the way we did than anything else. Hence, we did not report means in our article. This would be a serious problem if we were trying to figure out what people's unique hue settings are. However, we have no interest in this. The focus of our study was to demonstrate that naming tasks produce more disagreement than matching tasks. So, for the objection to go through, a reasonable expectation that our standards would cause more disagreement in one task type than the other is required. If our standards were not even the correct determinable colors, we could see there being a reason to worry, but this is not the case. This said, if there is a reasonable expectation to be concerned, we look forward to further research on this matter.

References

Kuehni, R. G. & Hardin, C. L. (2014). Color matching and color naming: A response to Roberts and Schmidtke. Doi: 10.1007/s13164-013-0174-z

Livers, J. J. (1942). Some limitations to use of coefficient of variation. *Journal of Farm Economics, 24*(4), 892-895.er, 1942, p. 892-893.

Roberts, P. & Schmidtke, K. (2012). In defense of incompatibility, objectivism, and veridicality about color. *Review of Philosophy and Psychology, 3*(4), 547-558.