MBBE BWC 2: Preregistration

Kaitlyn Fallow (kmfallow@uvic.ca)

2018-09-12

Contents

Background

Background information is mostly copied from the pre-registration for a preceding pilot study, for which I hope to post summary data shortly (this will be located at https://osf.io/h7sjr/).

We have observed a remarkably robust conservative response bias in more than 20 recognition memory experiments using full-colour scans of masterwork paintings as stimuli (e.g.,https://osf.io/7bz9g/, https://osf.io/dksab/). A smaller number of more recent experiments we have conducted using photos of faces and of assorted indoor and outdoor scenes suggests this phenomenon may generalize to at least some subset of other kinds of visually rich, complex images. However, the underlying source of these effects remains mysterious. None of these experiments have included manipulations with a well-known tendency to produce conservative responding, such as a skewed new:old ratio on the recognition test (it has always been 1:1 in these experiments) or a payoff structure that preferentially rewards correct "new" responses. Conservative responding cannot be said to be the default in the old/new, study/test recognition paradigm, or even in our experimental setup specifically; for example, we have conducted experiments with words as stimuli that were methodologically identical to some of those mentioned above in all other respects, and observed no such bias. Additionally, we have thus far not found support for any of our hypotheses regarding the underlying

mechanism, such as the possibility that participants perceive paintings as highly memorable relative to words and are therefore more apt to be cautious in endorsing them as "old".

Although we have now observed this materials-based bias effect with 3 image types and under a range of experimental conditions, the boundary conditions at both the experiment and stimulus levels remain unclear. The existing literature does not support claims as broad as response bias always being conservative when images are used as stimuli, or even always being more conservative for pictures than words (e.g. Fawcett, Quinlan, & Taylor, 2012, who report roughly equally conservative responding to pictures and words in most conditions; Osth, Dennis, & Kinnell, 2014, who found a conservative bias with scene photos but not faces or fractal images). Thus, an important step toward understanding what may be driving the effects described above is establishing which image features are necessary and/or sufficient to produce conservative responding under our usual experimental conditions. Any number of stimulus- or set-level characteristics could conceivably be important to understanding the pattern of responding we have observed with images in these experiments, from low-level visual features to higher-level properties like inter-item similarity, image distinctiveness, emotionality, or perceived familiarity. The current project begins at the low level by investigating the role of image colour. There has long been interest in the role of colour in memory, but the focus has generally been on broad questions of whether it improves accuracy or otherwise leads to enhanced memory, with relatively less attention paid to recognition parameters like false alarms and response bias. A non-comprehensive survey of some of the literature in this area suggests accuracy operationalized in terms of d', Br (HR-FAR), or hit rates tends to be better for colour compared to grayscale photos of various categories (e.g., Bredart, Cornet, & Rakic, 2014; Gegenfurtner & Rieger, 2000; Spence, Wong, Rusan, & Rastegar, 2006; Wichmann, Sharpe, & Gegenfurtner, 2002), although these effects are neither large (5-10% in these studies) nor without exception (e.g., Nijboer, Kanai, de Haan, & van der Smagt, 2008).

Most of the above studies did not compare response bias between conditions, but both that did reported significantly more conservative responding to grayscale/monochrome images than to colour images (Bredart, Cornet, & Rakic, 2014; Nijboer et al., 2008). While Wichmann, Sharpe, and Gegenfurtner (2002) did not statistically test response bias nor report the associated means, rough estimates of c on the basis of mean hit and false alarm rates suggest it may also have been more conservative for black and white relative to colour images in their studies (although it cannot be assumed c calculated in this way will map perfectly onto the mean of subject-level c estimates). A pilot study conducted in Spring 2018 set out to test the following hypotheses with paintings as stimuli (quoted from that pre-registration): * We expect mean response bias to be significantly conservative for both colour paintings (consistent with previous results) and grayscale paintings * Based on the apparent trend suggested by previous literature exploring the role of colour in memory, we expect to observe higher average sensitivity (d') and hit rates for images presented in colour than those presented in grayscale. * Based on prior research comparing colour vs. grayscale photographs, we hypothesize thatresponse bias will be significantly more conservative for grayscale paintings. However, this literature being fairly scant (and the author's initial hunch being that the opposite would occur!), this hypothesis is a more tentative one.

Of these, only the first hypothesis was supported. False alarm but not hit rates were significantly higher in the grayscale condition, and both this result and corresponding analyses of the signal detection measure c were consistent with the possibility of participants adopting a more liberal approach to endorsing grayscale than colour paintings. However, it is also possible the higher visual interstimulus similarity of the grayscale set simply made these items more confusable with each other and thus more likely to feel familiar/elicit a "studied" response without any need to assume a decisional mechanism. Perhaps most importantly, this pilot study was constrained by a very small sample and the need to exclude more participants than usual for near- or below-chance performance (d' < 0.2), such that these basic analyses should be interpreted with caution, and finer-grained analyses (e.g., looking at how hit and/or false alarm rates change over time, generating ROCs and determining whether it is appropriate to calculate corrected measures of bias and sensitivity) were not advisable. The current study aims to replicate these patterns with a larger sample and explore them in more detail.

Methods

Participants

Participants will be UVic undergraduates who take part in research for bonus course credit. Intended sample sizes are ~90 per group in the betweens-ubjects condition and ~80 in the within-subjects group. These estimates were based on G*Power calculations determining the sample sizes needed to achieve 90% power to detect $d = 0.5$ (the smallest effect size of interest) in two-tailed independent- ($N = 86$/group) and paired-sample t-tests ($N = 44$) respectively (with alpha = 0.05), such as those comparing mean c between colour conditions. We are overshooting the latter estimate substantially to ensure better per-item *n*s for the purposes of item-level analyses (more detail below). Testing will occur in sessions ranging from one participant at a time up to as many as 20 (past experience suggests most sessions will involve groups of 5-15 people).

Materials

Stimuli are 198 high-resolution colour scans of masterwork paintings (originally obtained from Jeffrey Toth), and luminance-matched grayscale versions of each generated using IrfanView (Skiljan, 2016; https://www.irfanview.com/). These can all be viewed in the Stimuli folder accompanyting this pre-registration. The experiment will be administered on desktop PCs with widescreen monitors, using E-Prime software.

Study and test lists are randomized anew for each subject such that individual paintings will vary across subjects with respect to study status, study/test position, and, in the within-subjects group, colour condition (colour or grayscale). In this group, exactly 50% of both studied and new paintings will be in each colour, and the same painting will never appear in both colours for the same participant. Similarly, paintings studied in colour will never be tested in grayscale or vice versa. The two colour conditions will be randomly intermixed for these participants with no constraints on the number of consecutive trials in a given colour condition.

Procedure

Participants will begin by studying 96 paintings bookended by 6 additional primacy/recency buffers (3 of each) for 1 s each with a 900-ms ISI (including a 500-ms central fixation cross). Paintings are presented in the centre of the screen at a maximum size of 50% of the vertical and horizontal dimensions of the display and are resized proportionally based on the vertical maximum to accommodate widely varying image dimensions (e.g., portraits vs. landscapes). Participants are informed at the beginning of the study phase that they will be seeing a series of paintings of various types, including landscapes, portraits, and still lifes [in colour and/or black and white, as applicable] and that the task is to try their best to remember each. Following the study phase there is a short filler task wherein subjects are asked to judge whether a series of numbers and letters, some of which have been rotated clockwise or counterclockwise to varying degrees, have been horizontally flipped or not. This task is designed to last approximately 5 min including time to read the instructions, with the actual task set to display as many trials as a given participant can complete in 3.5 min. There is a pause screen between this task and the recognition test to ensure the verbal test instructions are given to all participants at once.

The recognition test comprises all studied images and an equal number of new images. Participants will be told they are going to see another series of paintings, some of which will be from the study list and some of which will be new, and that they will be asked to make studied/not studied judgments on a 6-pt confidence weighted scale (from 1 = definitely not studied to 6 = definitely studied). This scale will remain onscreen throughout the test, which participants will complete at their own pace. At the end of the test participants are asked a few additional questions (see instructions_wording), only one of which is critical for our purposes (as detailed in the exclusion criteria section below; the rest are for curiosity's sake and we have not previously gotten around to doing much with the responses, but may do so in the future). In group sessions participants are asked to wait until everyone is finished at the end of the experiment to avoid distracting those who are still working on the test, and are then told about the purpose of the experiment as a group.

Hypotheses

Subject-level signal detection analyses (primary hypotheses in bold)

- **Mean response bias (as quantified by the signal detection measure c) will be significantly conservative for both grayscale and full colour paintings across the board (i.e., in the within-subjects condition and for both single-colour groups)**
  - **Response bias will be significantly *more* conservative for full colour paintings than grayscale paintings in the within-subjects condition, consistent with the results of an earlier pilot study (note that we did not have a strong prediction on this front going into that pilot study)**
    - A secondary and more tentative hypothesis based on the pilot study results is that there will be a significant interaction between item status (old or new) and image colour on the proportion of "studied" responses in the within-subjects condition, whereby the mean false alarm rate will be higher for grayscale than colour images in the within-subjects condition, but mean hit rates will be fairly similar. We did not anticipate this interaction originally and it was only directional in the pilot study, so we hope to gain further insight with a larger sample.
  - In the between-subjects case, response bias will be significantly *more* conservative in the full colour group, similar to the within-subjects results (this hypothesis is also tentative; we expect this difference to be directionally smaller than that in the within-subjects condition as this seems to be an important variable in recognition memory, but are not planning our sample sizes around detecting such an interaction).
- Mean sensitivity (quantified as the signal detection measure d') will be significantly higher for colour paintings in the within-subjects condition, and in the full colour group relative to the grayscale group.

ROC analyses

- We expect zROC slopes to be lower than 1 (specifically, between 0.5 and 0.8) in all conditions, consistent with both our previous results and the norm in recognition memory data.
- We have no specific predictions as to whether or not there will be differences in these slopes as a function of image colour, but do plan to look at this in an exploratory fashion.
- We expect response bias measures corrected for unequal variance (ca and ce) will produce the same pattern of results as predicted above for uncorrected C, but
  - The pattern of results for corrected sensitivity measures (da and de) may well differ; d' is highly sensitive to fluctuations in the old:new variance ratio. Here we have no specific predictions, but are interested to see whether these measures paint a substantially different picture given the still predominant reliance on d' in studies exploring the role of colour in recognition memory.

Test quartile analyses

- We expect response bias estimates (c, ca, and ce) to increase (become more conservative) from the first 48-item test quartile to the last quartile in all conditions, and sensitivity estimates (d', da, and de) to decrease.
  - Consistent with our previous results, we expect corresponding quartile-level analyses with raw response proportions will show a decline across quartiles for hit rates, while false alarm rates remain fairly stable or increase slightly at most.
- We will investigate the possibility of colour-based differences in quartile-level patterns but have no *a priori* hypotheses here.

Planned analyses

Our primary dependent measures of interest will be hit and false alarm rates (determined by collapsing our 1-6 response scale into a binary one, with 1-3 = "new" and 4-6 = "old"), and corresponding signal detection theory (SDT)-based measures of response bias (c) and sensitivity (d'). Ceiling and floor rates will be replaced according to Macmillan and Kaplan (1985) to enable calculation of c and d'. Given long-known issues with the application of c and d' to recognition memory data, namely the tendency for data to be inconsistent with the assumption of equal variance of the old and new item distributions (conditions under which c and d' become confounded), we will also conduct receiver operating characteristic (ROC) analyses with the goal of estimating the extent of this violation in our data and calculating corrected measures of sensitivity (da and de) and bias (ca and ce). The plan is to apply these corrections at the participant level based on participant-level ROCs, but if results suggest these curves are ill-fitting (e.g., if a lot of people have primarily used the extreme ends of the confidence scale) we will instead apply corrections based on aggregate by-condition ROCs. These alternate measures will be subject to the same analyses as c and d', and all will be reported; for brevity, I will refer only to c and d'.

Confirmatory analyses

- c and d' will be calculated at the participant level, and separately by colour condition in the within-subjects group. These estimates will be subject to appropriate t-tests, namely:
  - Paired tests in the within-subjects condition and two-sample tests in the between-subjects condition (not assuming equal variance between conditions) to test for differences as a function of image colour; although we do have directional hypotheses, this line of inquiry is still in its early stages, so we will stick with the conventional two-tailed approach (setting alpha = 0.05);
  - Four one-sample tests to examine the hypothesis that c will be significantly conservative (i.e., greater than 0) in all conditions. Because this hypothesis is explicitly directional and firmly grounded in our prior research (which suggests moderate-to-large effects in terms of the difference from neutral), these will be one-tailed tests (using a stricter alpha level of 0.01 for each test);
- Hit and false alarm rates will be analyzed via 2(true item status: old or new) x 2(colour: grayscale or full colour) ANOVAs (the colour factor being a repeated measure in the within-subjects case) using the proportion of "studied" responses as the dependent variable to examine the hypothesis that the effect of colour is primarily on false alarms, not hits. The effect of primary interest here is therefore the interaction - we expect this to be significant at the .05 level, representing a greater colour-based difference on responding to new items than old items (which we will further characterize via post hoc comparisons).

(Mostly) exploratory analyses

- c, d', and hit and false alarm rates will all also be calculated at the test quartile level for each participant and condition and analyzed via 2 (colour) x 4 (quartile) ANOVAs. This is partially confirmatory (w.r.t. the main effects of quartile) but the colour component is entirely exploratory.
- ROCs will be calculated at the subject level (and possibly at the condition/group level, collapsing across subjects) for reasons mentioned above, and possibly at the item level for more exploratory purposes. These will be fit using the PCA-based method described by John Vokey (2016).

Item-level analyses using generalized linear mixed modeling (GLMM)

Item-level estimates of c and d' will be obtained via GLMM, with a formulation of the model originally developed and tested by Reinhold Kliegl and Max Rabe (e.g., Rabe, 2018; general approach described in Chapter 3 of Knoblauch & Maloney, 2012). These may also be calculated along with item-level hit and false alarm rates using the coarser method of collapsing across subjects for the sake of comparison.

Very broadly, the current iteration of the model used for this purpose treats c and d' as predictors of the response variable and allows these parameters to vary at both the subject and item level. The response variable is, in this case, dichotomized as old/new, which is reflected in the use of a binomial link function in the fitting procedure. The full within-subjects version of the model includes a total of 20 parameters: four fixed effects (corresponding to c and d' for the two stimulus types, in this case two colour conditions), eight random effects (allowing for both subject- and item-level variability in both c and d' for each stimulus type), and eight parameters representing correlations among the random effects. The model for datasets from experiments that used only one stimulus type is much less complex with only eight parameters in total (2 fixed, 4 random, and 2 correlation parameters). I am new to these analyses myself so have so far stuck with the original formulation, but in this case plan to test iterations of varying complexity lest our data not support the use of the full model(s). Also, the baseline iteration assumes equal variance of the underlying old and new item distributions, so alternatives that do not make this assumption may also be tested with these data.

I plan to calculate within-study rank correlations for these item-level estimates (e.g., looking at whether paintings that are better discriminated when presented in colour are also better discriminated when presented in grayscale) as well as correlations with estimates for the same images obtained from prior studies. Depending on these results I may explore some finer-grained questions, e.g., whether these patterns differ between landscapes and portraits, with the goal of further refining future hypotheses.

Null effects?

- I am in the process of learning how to better characterize results when evidence is insufficient to reject the null hypothesis, and expect to use Bayes factors and/or equivalence testing methods toward this purpose.

Exclusion criteria

- **Art expertise ratings of 5**: At the end of the experiment participants will be asked to rate their own art expertise, with 5 indicating the maximum response of "very above average" familiarity with the paintings presented. Whether art expert participants approach the task differently/would be outliers w.r.t the measures we are interested in is an interesting empirical question but one that would require a separate recruitment effort; in this case, participants who choose this option will be excluded from analysis.
- **d' $<$ 0.2 (in either condition, for within-subjects group)**: An admittedly arbitrary cutoff (and still above chance overall), but performance using this experimental setup is usually quite high, so this is intended to exclude participants who are responding essentially at random (and in the within-subjects case, those who adopt a strategy of allocating attention to one condition).
- **Outliers**: I have in the past excluded outliers on c and d' (defined as 3+ SD from the mean in either direction) from analysis in similar studies, but have come to think there is little justification for this and do not plan to do so here. In the past applying this criterion has never had a substantive effect on the results anyway, but as frequency distributions for all dependent variables will be included with the results, anyone concerned about outliers can evaluate the data structure for themselves.
  - Outliers in the ROC analyses will be dealt with on a case-by-case basis.
  - I have no plans to make response-level exclusions (e.g., very fast or very slow responses) at this stage, but may implement such a criterion if I pursue analyses using response time.

Data peeking plans

The within-subjects pilot study using these stimuli resulted in an undesirably and unusually high number of performance-based (d' $<$ 0.2) exclusions (9 out of 34 participants). However, previous studies using the colour paintings have also had issues with undesirably high numbers of near-ceiling estimates at the participant

and/or item level, especially with respect to near-floor (or even at floor) false alarm rates, so I am hesitant to implement any manipulations intended to boost performance outright. With these problems in mind, I intend to start by running ~20 participants in the within-subjects and grayscale-only conditions and looking at the data. If 4 or more participants in either condition would be excluded from further analysis under the d' < 0.2 criterion, I will consider adjusting the procedure to boost performance. If there are even 2-3 d'-based exclusions per condition at this stage I will probably continue to monitor the data as it comes in for the next 10-20 subjects to make sure exclusion rates do not exceed 15%. If there are 0-1 exclusions I will proceed with data collection and hope this initial group was not disproportionately hypermnesic!

Regardless, the sample sizes discussed above pertain to whatever the *final* procedure ends up being, so if it is changed after this initial data-peeking stage, these early participants will not be included in the final analysis.

References

Brédart, S., Cornet, A., & Rakic, J. M. (2014). Recognition memory for colored and black-and-white scenes in normal and color deficient observers (dichromats). *PLoS ONE, 9*(5), 1–5. http://doi.org/10.1371/journal.pone.0098757

Fawcett, J. M., Quinlan, C. K., & Taylor, T. L. (2012). Interplay of the production and picture superiority effects: A signal detection analysis. *Memory, 20*(7), 655–666. http://doi.org/10.1080/09658211.2012.693510

Gegenfurtner, K. R., & Rieger, J. (2000). Sensory and cognitive contributions of color to the recognition of natural scenes. *Current Biology, 10*(13), 805–808. http://doi.org/10.1016/S0960-9822(00)00563-7

Knoblauch, K., & Maloney, L. T. (2012). Modeling Psychophysical Data in R. http://doi.org/10.1007/978-1-4614-4475-6

Nijboer, T. C. W., Kanai, R., de Haan, E. H. F., & van der Smagt, M. J. (2008). Recognising the forest, but not the trees: An effect of colour on scene perception and recognition. *Consciousness and Cognition, 17*(3), 741–752. http://doi.org/10.1016/j.concog.2007.07.008

Osth, A. F., Dennis, S., & Kinnell, A. (2014). Stimulus type and the list strength paradigm. *The Quarterly Journal of Experimental Psychology, 67*(9), 1826–1841. http://doi.org/10.1080/17470218.2013.872824

Rabe, M. M. (2018). Generalized linear mixed modeling of signal detection theory (Master's thesis). University of Victoria, Victoria, BC, Canada.

Spence, I., Wong, P., Rusan, M., & Rastegar, N. (2006). How color enhances visual memory for natural scenes. *Psychological Science, 17*(1), 1–6. http://doi.org/10.1111/j.1467-9280.2005.01656.x

Wichmann, F. A., Sharpe, L. T., & Gegenfurtner, K. R. (2002). The contributions of color to recognition memory for natural scenes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(3), 509–520. http://doi.org/10.1037//0278-7393.28.3.509