EuroRV3: EuroVis Workshop on Reproducibility, Verification, and Validation in Visualization (2017)
K. Lawonn, N. Smit, and D. Cunningham (Editors)

A Crowdsourced Approach to Colormap Assessment

Terece L. Turton[1], Colin Ware[2], Francesca Samsel[1], David H. Rogers[3]

[1] Center for Agile Technology, University of Texas at Austin, TX, USA
[2] Center for Coastal and Ocean Mapping, University of New Hampshire
[3]Data Science at Scale Team, Los Alamos National Laboratory, Los Alamos, NM, USA

Abstract

Despite continual research and discussion on the perceptual effects of color in scientific visualization, psychophysical testing is often limited. In-person lab studies can be expensive and time-consuming while results can be difficult to extrapolate from meticulously controlled laboratory conditions to the real world of the visualization user. We draw on lessons learned from the use of crowdsourced participant pools in the behavioral sciences and information visualization to apply a crowdsourced approach to a classic psychophysical experiment assessing the ability of a colormap to impart metric information. We use an online presentation analogous to the color key task from Ware's 1988 paper, Color Sequences for Univariate Maps*, testing colormaps similar to those in the original paper along with contemporary colormap standards and new alternatives in the scientific visualization domain. We explore the issue of potential contamination from color deficient participants and establish that perceptual color research can appropriately leverage a crowdsourced participant pool without significant CVD concerns. The updated version of the Ware color key task also provides a method to assess and compare colormaps.*

Categories and Subject Descriptors (according to ACM CCS): H.1.2 [Models and Principles]: User/Machine Systems—Human Factors H.5.2 [Information Systems]: User Interfaces—Evaluation/methodology H.m [User/Machine Systems]: Miscellaneous—Colormapping

1. Introduction

User evaluation is a critical step in the design and development of tools and applications for visualization. While some experimental approaches can only be done within an in-person laboratory setting, online studies on platforms such as Mechanical Turk (MTurk) [AMT] provide easy participant access at reasonable cost. Over the past decade, research within psychology, linguistics and other behavioral sciences have studied the use of crowdsourced participant pools [BKG11, MS12, LRR15, PC14]. Across fields, researchers have sought to validate MTurk as a research platform by deliberately replicating classic experiments, exploring both reproducibility issues and providing insight into crowdsourced demographics [CMG13, SOJN08]. In this short time, crowdsourcing has become an accepted research paradigm within the behavioral sciences.

As early as 2008, Kittur, Chi and Suh [KCS08] explored the use of micro-task markets such as Mturk for user studies and researchers in information visualization have compared classic visualization in-person experiments to online results [ARPDC14, HYFC14, HB10]. MTurk has been successfully leveraged for wide-ranging research questions [BVB*13, KBB*15] including color-based studies [KLT*15, LFK*13, LH13, SPG*15]. Use of Mturk has become sufficiently normalized in these varied fields that tools specifically designed to facilitate crowdsourced studies have been developed [EK16, LRA16, OJ15, TBR17].

The ubiquitous use of color in scientific visualization presents specific issues when doing online studies. An online presentation has a significant increase in ecological validity at the expense of control over monitor and viewing conditions. The issue of color vision deficiency (CVD) is particularly worrisome when doing online studies directly involving color. In order to understand the impact of color vision deficiencies, we used the Farnsworth D-15 color cap arrangement test to populate a CVD group of participants. We then reproduced Ware's classic color key identification experiment on univariate colormaps [War88]. Using updated stimuli and an experimental design setup specifically for online presentation, we compare three groups of participants, described more fully in Section 2.3 to determine the potential impact of CVD participants:

UM Usual Mturker participant pool.
WO Women only: Very low probability of CVD contamination.
CVD CVD group: drawn from a series of studies requesting participants with CVD.

This paper has three contributions. First, a statistical comparison of the above participant groups across multiple colormaps leads us

DOI: 10.2312/eurorv3.20171106

to conclude that MTurk can provide an acceptable research platform for user studies in scientific visualization with minimal impact from potential CVD contamination. Second, a qualitative comparison to the Ware 1988 study is used to validate a crowdsourced approach to psychophysical studies of this type. Third, this updated version of the Ware color key task provides a methodology for assessing the ability of a colormap to impart metric information thus providing a meaningful way to compare and chose an appropriate colormap for qualitative tasks.

2. Color Key Task

The original Ware color key task, denoted Experiment 3 in the 1988 paper, presented a subject with a data set in various colormaps. A set of crosshairs indicated a specific location on the image and participants were asked to identify which of 16 equally spaced color keys was closest in color to the data at the center of the crosshairs. Only the 12 central keys were populated. The colormap range is [0.0, 1.0], hence each of the 16 keys spans a range 0.0625 wide.

2.1. Experimental Task

The online version followed a similar experimental design. To avoid issues of running code on remote participant computers, a purely image-based approach was used. A synthetic scalar field was generated to which colormaps were applied. The field was constructed by summing multiple Gabor functions into a two dimensional array with randomly varying amplitudes, wavelengths, orientation and centers [SW04]. For this example, the main spatial frequencies were between 32 and 116 pixels. There were 60 stimuli images generated, each with a set of crosshairs. The location of the crosshairs in each stimuli image was distributed such that there were five stimuli images corresponding to each of the 12 central color keys. The same set of 60 stimuli was repeated in each of the study colormaps. A sample stimuli image can be seen in Figure 1. The subject task was again to chose the color key most similar to the color at the center of the crosshairs. The study was coded within the JavaScript API of Qualtrics survey software, utilizing the *Key Task* module of the Evaluation Toolkit [TBR17].

2.2. Colormaps

In order to compare a crowdsourced approach with the original in-person study, we chose four colormaps similar to those in the Ware 1988 paper: RA, GP, SAT, RG; we included two colormaps considered contemporary standards: CW, VI; and included two more recent colormaps from the Data Science at Scale (DSS) team at Los Alamos National Laboratory: BOD, YGB. Colormap images and example stimuli are available in the supplemental material.

RA Rainbow, from ParaView [AGL05].
GP A perceptually uniform grey scale using CIElab $L*$.
SAT Monotonically increasing in saturation, grey to red.
RG A divergent colormap going from red to green.
CW A divergent cool/warm (blue to red) colormap [Mor09].
VI Viridis. A colormap with good uniformity and designed to be more CVD-safe [vS15].
BOD An extended cool/warm from deeper blues into oranges.

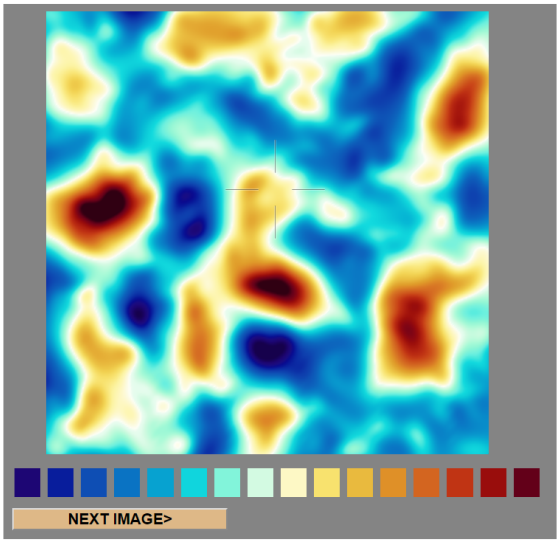

Figure 1: *The stimuli pattern, rendered in the BOD. The participant task was to click on the key closest in color to the color at the center of the crosshairs. This image was also used as the validation question.*

YGB Similar in spirit to Viridis, but designed to go through a wider range of hue and value.

2.3. Participant Groups

We used three orthogonal groups of participants. The *Women-Only* (WO) group was designed to be effectively free of color vision deficiencies. This research utilized TurkPrime, "a versatile crowdsourcing data acquisition platform for the behavioral sciences" [LRA16]. A useful feature of TurkPrime is a gender consistency score that tracks Mturker responses to a gender identity question. We required a response that was 100% consistently female. With a CVD rate of 0.5% and a sample size of 180 women, we estimate one possible CVD participant. Participants were also asked to self-select out of the study if they knew they had any type of CVD.

The *CVD* group was chosen based on a series of studies requesting CVD participants. The study itself was an online presentation of the Farnsworth D-15 color cap arrangement test for CVD [CJJ93] (color caps and survey available in supplemental material). CVD is itself a spectrum and a non-trivial percentage of people with milder variations of red/green color vision deficiencies will pass standard tests such as Ishihara plates or the FD-15 test. Given the non-specificity of standard CVD testing, the main purpose of the FD-15 test was not so much to establish a color vision issue but rather to present the subject with an appropriate test for CVD while allowing the participant to self-identify as CVD. Participants were asked to provide the formal type of their CVD, if known, and/or to describe their CVD issues. A training and validation task asked them to arrange six color caps ranging from black to light grey and then presented them with the 15 color caps of the Farnsworth arrangement test. Participants self-identifying as CVD were included in the CVD group if their answers showed that they understood the

task and spent a minimally reasonable amount of time on the validation and task. Thus, this "CVD" group is not guaranteed to be completely populated by CVD participants but is certainly much more highly populated with CVD subjects than either of the other two groups. The CVD group contained a maximum of 298 Mturkers during these studies.

The *Usual Mturkers* (UM) group is required to pass a set of restrictions that are commonly used in online research studies conducted by the authors. These include:

- Located in English-speaking countries. This helps to ensure participants are sufficiently fluent in English to understand the task.
- Typical Mturk performance requirements: $>$ 100 micro-tasks completed with $>$ 95% work accepted.
- Not a member of an author-maintained exclusion group. This group of Mturkers has demonstrated that they either do not understand a typical visualization task or have not been a faithful participant in some previous study. This group has usually failed a validation at some point. At the time of this writing, there were 292 Mturkers in this general exclusion group.
- Not a member of an author-maintained CVD exclusion group. This group consists of any Mturker taking an author study who has ever self-identified as CVD, 441 Mturkers currently.

2.4. Procedure

Each participant was given an explanation of the task, asked to do a validation question, Figure 1, and then saw a randomly chosen subsample of the 60 stimuli for a single colormap. The WO and UM groups saw 20 stimuli images and were limited to completing the study for a single colormap. Given the limited number of CVD participants, the CVD group was asked to do 25 stimuli images and allowed to complete the study for up to four different colormaps. Total number of trials for each colormap varied from 480 to 540 for the WO group; from 500 to 560 for the UM group; and from 275 to 550 for the CVD group. Participants who were unable to correctly answer within ± 1 key of the correct key on the validation question were removed from the study. This can be compared to the Ware 1988 Experiment 3 with 12 participants and two trials per key (12 keys) for 288 trials per colormap.

3. Experimental Results

3.1. Data Analysis

As discussed in the Introduction, this experimental design allows us to address multiple questions. Our analysis uses the mean absolute error. Each stimuli has a *ground truth* answer: the actual value at the center of the crosshairs. In an approach analogous to the Ware 1988 paper, we calculate the absolute error for each stimuli response as: $absError$ = |center of estimated key - ground truth|. The mean of the absolute error can be used to compare colormaps, averaging across all keys. We will summarize the results of the statistical analyses. Note: We acknowledge that various fields may prefer a confidence interval (CI) approach versus a null hypothesis significance testing (NHST) approach. In the interest of brevity and considering the potential familiarity of the average reader with each approach, we chose to present the NHST results in the paper. We invite the reader to see the supplemental materials for greater detail and a summary of the confidence interval approach. We do note that both CI and NHST approaches are in agreement.

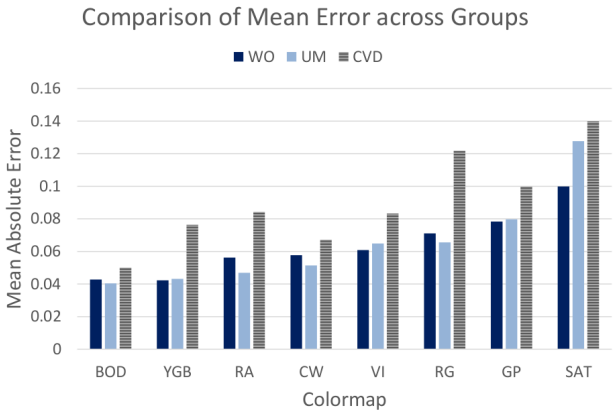

Figure 2: *Mean absolute errors for the eight tested colormaps (as noted) and the three participant groups, Women Only (dark blue), Usual Mturkers (light blue) and CVD (patterned grey).*

Table 1: *TukeyHSD p-values for each Group and Colormap*

Colormap	WO-UM	WO-CVD	UM-CVD
BOD	0.62	0.27	0.031
YGB	0.91	p<0.001	p<0.001
RA	0.068	p<0.001	p<0.001
CW	0.62	0.24	0.030
VI	0.87	p<0.001	p<0.001
RG	0.62	p<0.001	p<0.001
GP	0.36	0.066	0.56
SAT	p<0.001	p<0.001	0.87

3.2. CVD Impact

We assess the CVD impact by doing a comparison across all three participant groups for each of the eight colormaps tested. Mean absolute errors are shown in Figure 2 for each colormap and subgroup. This plot highlights the difference in the response of the CVD group particularly for colormaps with potential CVD issues.

Since each participant only saw a subset of the possible stimuli, we use a mixed model approach to assess whether the differences between the three groups are statistically significant. For each colormap, we performed a mixed model ANOVA on the log transform of the absolute error, assessing the interaction effects of the group and the participant ID. For the perceptual greyscale (GP), the main effect for (Group) showed no significance at the 0.05 level ($F(2,1367) = 2.553, p = 0.078$). For the other colormaps, the ANOVA revealed significant differences at the $p < 0.05$ level for cool/warm and blue/orange and at the $p < 0.001$ level for all others. For a post-hoc test, we performed a Tukey HSD [Tuk49] for each colormap across the three groups: WO, UM and CVD. The resulting p-values can be seen in Table 1. Note that the WO

group, a group designed to have minimal possibility of CVD contamination, is statistically indistinguishable from the CVD group for the greyscale, as should be expected. It is also statistically indistinguishable for the cool/warm divergent and the blue/orange divergent. There is an impact due to CVD for the other colormaps. However, there is no significant statistical difference between the WO group and the UM group at the $p < 0.05$ level except for the saturation colormap.

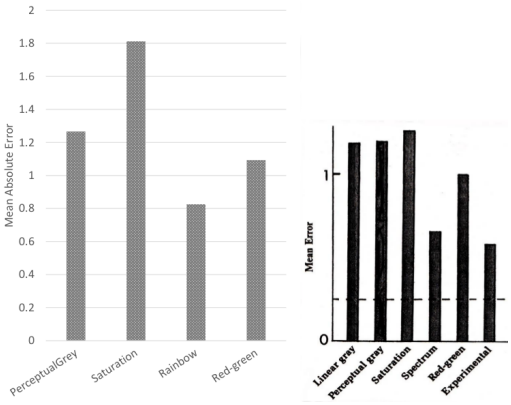

Figure 3: *The mean absolute error for four colormaps in the current experiment are compared to the results of the analogous colormaps from Ware 1988. Left: current experiment; right: Figure 10 from the Ware 1988 Experiment 3. Note mean errors have been scaled by the number of keys (16).*

3.3. Qualitative Comparison to Ware 1988

We next consider the four colormaps that were analogous to colormaps used in the Ware 1988 paper: RA, SAT, GP and RG. Given the results of Section 3.2, we combine both the women-only data and the usual Mturker data. While we can only do a qualitative comparison as data sets and colormap RGB values are not identical, the trends are very similar, Figure 3. The smallest mean error is seen in the Rainbow (Spectrum) colormap, followed by the red/green, then the grey, with the saturation (SAT) colormap faring the worst.

3.4. Colormap Comparison

Lastly, we assess the ability of all eight colormaps to carry metric information. Combining the WO and UM groups, there are over 8000 trials across 419 unique Mturkers. Given the large number of unique participants, we expect minimal effect from participant influence. A one-way ANOVA was thus conducted to compare the effect of colormap on the log transform of the absolute error. The ANOVA revealed significant differences between the eight colormaps, $(F(7, 8344) = 182.2, p < 0.001)$. A post-hoc Tukey HSD showed significant differences between many of the tested colormaps ($p < 0.05$). Figure 4 shows the mean absolute errors for the combined datasets (WO and UM) for all eight colormaps. Black bars above the means indicate colormap groupings with no statistically significant differences.

4. Conclusions

Carrying out user evaluations in a crowdsourced environment is becoming a norm in visualization. The inability to control for color vision deficiencies is a valid concern of researchers and reviewers alike. By comparing multiple subject groups, one with a very low probability of contamination, one highly populated by CVD participants together with a typical group of Mturkers, we explored this issue, finding significant differences between CVD and non-CVD participants across a variety of colormaps with the exceptions of a purely luminance-based greyscale, the standard cool/warm and the DSS blue/orange divergent.

When comparing the non-CVD group with a more typical group of participants, the only significant difference we find between the usual group and the non-CVD group is for the saturation colormap – a colormap demonstrably poor at conveying metric information. From these results, we conclude that, with reasonable precautions to minimize potential colorblind issues, Amazon Mechanical Turk can provide a valid research platform for color-based studies. These reasonable precautions could include avoiding colormaps susceptible to CVD issues or actively excluding CVD participants by choosing only women or by developing and maintaining an exclusion list of self-identified CVD Mturkers.

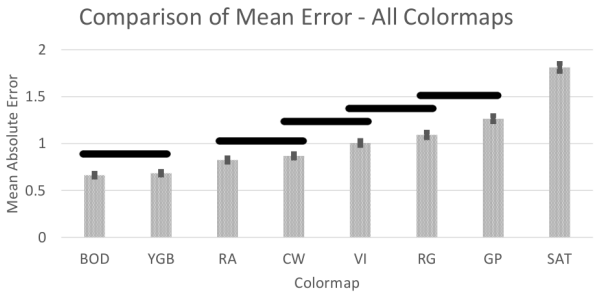

Figure 4: *The mean absolute error (scaled) for the eight tested colormaps (combined WO and UM groups). Error bars indicate the standard error of the mean. The horizontal bars above the means indicate colormap groups which are not statistically separated based on the Tukey HSD analysis.*

We also assessed the ability of a colormap to carry metric information. A Tukey HSD comparison across colormaps found that the DSS blue/orange divergent and the DSS yellow/green/blue colormaps both provide an improved ability to impart metric information compared to some common standards. While the rainbow colormap performs well for qualitative tasks, its well-known flaws [BI07, Mor09] argue against its use. The results presented here provide tested alternatives to the rainbow when choosing a colormap for a metric task.

Acknowledgments

This material is based upon work supported by Dr. Lucy Nowell of the U.S. Department of Energy Office of Science, Advanced Scientific Computing Research under Award Numbers DE-AS52-06NA25396, DE-SC-0012438, and DE-SC-0012516. The authors would like to thank Dr. Roxana Bujack and Dr. James Ahrens.

References

[AGL05] AHRENS J., GEVECI B., LAW C.: ParaView: An end-user tool for large-data visualization. In *Visualization Handbook*, Hansen C. D., Johnson C. R., (Eds.). Butterworth-Heinemann, Burlington, 2005, pp. 717–731. 2

[AMT] Amazon Mechanical Turk Website. www.mturk.com/mturk/welcome. 1

[ARPDC14] ABDUL-RAHMAN A., PROCTOR K. J., DUFFY B., CHEN M.: Repeated measures design in crowdsourcing-based experiments for visualization. In *Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization* (New York, NY, USA, 2014), BELIV '14, ACM, pp. 95–102. URL: http://doi.acm.org/10.1145/2669557.2669561, doi:10.1145/2669557.2669561. 1

[BI07] BORLAND D., II R. M. T.: Rainbow color map (still) considered harmful. *IEEE Computer Graphics and Applications 27*, 2 (2007), 14–17. 4

[BKG11] BUHRMESTER M., KWANG T., GOSLING S.: Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science 6* (2011), 3–5. 1

[BVB*13] BORKIN M., VO A., BYLINSKII Z., ISOLA P., SUNKAVALLI S., OLIVA A., PFISTER H.: What makes a visualization memorable? *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (2013), 2306–2315. 1

[CJJ93] CJ B., JC G., J H.: Comparison of the Farnsworth-Munsell 100-hue, the Farnsworth D-15, and the Anthony D-15 desaturated color tests. *Archives of Ophthalmology 111*, 5 (1993), 639–641. doi:10.1001/archopht.1993.01090050073032. 2

[CMG13] CRUMP M., MCDONNELL J., GURECKIS T.: Evaluating amazon's mechanical turk as a tool for experimental behavioral research. *PLoS ONE 8*, 3 (2013). 1

[EK16] ERLEWINE M. Y., KOTEK H.: A streamlined approach to online linguistic surveys. *Natural Language & Linguistic Theory 34*, 2 (2016), 481–495. URL: http://dx.doi.org/10.1007/s11049-015-9305-9, doi:10.1007/s11049-015-9305-9. 1

[HB10] HEER J., BOSTOCK M.: Crowdsourcing graphical perception: Using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2010), CHI '10, ACM, pp. 203–212. URL: http://doi.acm.org/10.1145/1753326.1753357, doi:10.1145/1753326.1753357. 1

[HYFC14] HARRISON L., YANG F., FRANCONERI S., CHANG R.: Ranking visualizations of correlation using weber's law. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (Dec 2014), 1943–1952. doi:10.1109/TVCG.2014.2346979. 1

[KBB*15] KIM N. W., BYLINSKII Z., BORKIN M. A., OLIVA A., GAJOS K. Z., PFISTER H.: A crowdsourced alternative to eye-tracking for visualization understanding. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems* (New York, NY, USA, 2015), CHI EA '15, ACM, pp. 1349–1354. URL: http://doi.acm.org/10.1145/2702613.2732934, doi:10.1145/2702613.2732934. 1

[KCS08] KITTUR A., CHI E. H., SUH B.: Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2008), CHI '08, ACM, pp. 453–456. URL: http://doi.acm.org/10.1145/1357054.1357127, doi:10.1145/1357054.1357127. 1

[KLT*15] KIM J., LEKSIKOV S., THAMJAMRASSRI P., LEE U., SUK H.-J.: Crowdcolor: Crowdsourcing color perceptions using mobile devices. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services* (New York, NY, USA, 2015), MobileHCI '15, ACM, pp. 478–483. URL: http://doi.acm.org/10.1145/2785830.2785887, doi:10.1145/2785830.2785887. 1

[LFK*13] LIN S., FORTUNA J., KULKARNI C., STONE M., HEER J.: Selecting semantically-resonant colors for data visualization. In *Proceedings of the 15th Eurographics Conference on Visualization* (Chichester, UK, 2013), EuroVis '13, The Eurographs Association; John Wiley & Sons, Ltd., pp. 401–410. URL: http://dx.doi.org/10.1111/cgf.12127, doi:10.1111/cgf.12127. 1

[LH13] LIN S., HANRAHAN P.: Modeling how people extract color themes from images. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2013), CHI '13, ACM, pp. 3101–3110. URL: http://doi.acm.org/10.1145/2470654.2466424, doi:10.1145/2470654.2466424. 1

[LRA16] LITMAN L., ROBINSON J., ABBERBOCK T.: Turkprime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods* (2016), 1–10. doi:10.3758/s13428-016-0727-z. 1, 2

[LRR15] LITMAN L., ROBINSON J., ROSENZWEIG C.: The relationship between motivation, monetary compensation, and data quality among us- and india-based workers on mechanical turk. *Behavior Research Methods 47*, 2 (2015), 519–528. doi:10.3758/s13428-014-0483-x. 1

[Mor09] MORELAND K.: Diverging color maps for scientific visualization. In *Proceedings of the 5th International Symposium on Advances in Visual Computing, Part II* (2009), ISVC '09, pp. 92–103. 2, 4

[MS12] MASON W., SURI S.: Conducting behavioral research on amazon's mechanical turk. *Behavior Research Methods 44*, 1 (2012), 1–23. 1

[OJ15] OKOE M., JIANU R.: Graphunit: Evaluating interactive graph visualizations using crowdsourcing. *Comput. Graph. Forum 34*, 3 (June 2015), 451–460. URL: http://dx.doi.org/10.1111/cgf.12657, doi:10.1111/cgf.12657. 1

[PC14] PAOLACCI G., CHANDLER J.: Inside the turk: Understanding mechanical turk as a participant pool. *Current Directions in Psychological Science 23*, 3 (2014), 184–188. 1

[SOJN08] SNOW R., O'CONNOR B., JURAFSKY D., NG A. Y.: Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (Stroudsburg, PA, USA, 2008), EMNLP '08, Association for Computational Linguistics, pp. 254–263. URL: http://dl.acm.org/citation.cfm?id=1613715.1613751. 1

[SPG*15] SAMSEL F., PETERSEN M., GELD T., ABRAM G., WENDELBERGER J., AHRENS J.: Colormaps that improve perception of high-resolution ocean data. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems* (2015), CHI EA '15, pp. 703–710. doi:10.1145/2702613.2702975. 1

[SW04] SWEET G., WARE C.: View direction, surface orientation and texture orientation for perception of surface shape. In *Proceedings of Graphics Interface 2004* (School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada, 2004), GI '04, Canadian Human-Computer Communications Society, pp. 97–106. URL: http://dl.acm.org/citation.cfm?id=1006058.1006071. 2

[TBR17] TURTON T. L., BERRES A. S., ROGERS D. H.: ETK: An evaluation toolkit for visualization user studies, June 2017. Accepted into EuroVis 2017: 19th EG/VGTC Conference on Visualization. 1, 2

[Tuk49] TUKEY J.: Comparing individual means in the analysis of variance. *Biometrics 5*, 2 (1949), 99–114. 3

[vS15] VAN DER WALT S., SMITH N.: Matplotlib documentation update. bids.github.io/colormap/, 2015. 2

[War88] WARE C.: Color sequences for univariate maps: Theory, experiments and principles. *IEEE Computer Graphics and Applications 8*, 5 (1988), 41–49. 1