Eurographics Conference on Visualization (EuroVis) 2017
B. Kozlíková, T. Schreck, and T. Wischgoll (Editors)

Short Paper

ETK: An Evaluation Toolkit for Visualization User Studies

Terece L. Turton[1], Anne S. Berres[2], David H. Rogers[2], and James Ahrens[2]

[1] Center for Agile Technology, University of Texas at Austin, TX, USA
[2]Data Science at Scale Team, Los Alamos National Laboratory, Los Alamos, NM, USA

Abstract

This paper describes the design and features of the Evaluation Toolkit (ETK), a set of JavaScript/HTML/CSS modules leveraging the Qualtrics JavaScript API that can be used to automate image-based perceptual user evaluation studies. Automating the presentation of the images can greatly decrease the time to build and implement an evaluation study while minimizing the length and complexity of a study built within Qualtrics, along with decreasing the possibility of error in image presentation. The ETK modules each focus on automating a specific psychophysical or experimental approach. Because each module is an extension or plug-in to a Qualtrics question, the resultant study can be easily used in a laboratory setting or in a crowdsourced approach. We present the open source repository of ETK with the six modules that currently make up the toolkit and invite the community to explore, utilize, and contribute to the toolkit.

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

1. Introduction

User evaluation, be it qualitative or quantitative, is a critical step in the design and development of a visualization system or technique. Each approach has its challenges and the difficulty of carrying out rigorous and effective evaluation is well-documented [Car08, For10, Pla04]. Qualitative evaluation, by its nature, often relies on expert interviews, walk-throughs, and user feedback. Quantitative evaluation focuses on measurable quantities: how quickly can a user/subject identify a feature; how accurately can information be transferred and often utilizes psychophysical approaches as in [RKPC99] or [War88].

Consider the particular challenges within scientific visualization – how to balance the ever-growing size of the data with the need of the scientist to not lose vital scientific information. In-situ approaches, compression, sampling of data, rendering choices can all impact a visualization and the information available from it. The evaluation of such factors is amenable to standard psychophysical techniques [EE99, Fec89]. This might include finding a discrimination threshold on compressed or sampled data via a 2-Alternative Forced Choice (2AFC) or a Method of Adjustment approach, or evaluating rendering options via A/B choice experiments or through task-based approaches.

Perceptual user evaluation in visualization can often be reduced to a set of visualization artifacts that are simply images with varying levels of a stimulus applied (e.g., varying levels of data compression) and/or different experimental conditions applied (e.g., colormap or rendering method used).

1.1. Contribution

The contribution of this paper is the Evaluation Toolkit (ETK), an extensible toolkit that embodies standard psychophysical techniques to run perceptual experiments based on images as the visualization artifacts. Using this toolkit, a wide range of scientists can leverage the psychophysical approaches embodied in the modules. Embedded into Qualtrics survey software [Qua], ETK modules streamline the user study implementation process. The resultant study can be used either in a laboratory setting or launched in a crowdsourced approach. Through the online repository of ETK modules, we invite the community to test the toolkit, suggest upgrades and additional modules, or contribute modules themselves.

2. Backgroud

Rigorous evaluation is a laudable goal, yet researchers in visualization may not always have the necessary experimental background to easily design and implement studies [For10]. Additionally, the uniqueness of many visualization systems is a challenge to developing a general evaluation approach. For example, the Hierarchical Visualization Testing Environment (HVTE) of Andrews and Kasanicka [AK07] was built specifically for their Hierarchical Visualisation System to compare multiple hierarchical browsers.

DOI: 10.2312/eurovisshort.20171131

Mackay et al.'s Touchstone [MABL*07] took a much more general approach as both a repository for HCI-related experiments and as an experiment development tool. A more recent addition from Aigner, Hoffmann, and Rind [AHR13] is EvalBench, a flexible library developed to evaluate interactive visualization artifacts.

The behavioral sciences also provide tools to facilitate the implementation of experimental design. PsiTurk [psi] provides an experiment exchange with experiments relevant to behavioral sciences. The tool jsPsych [dL15] is a JavaScript library to facilitate building online behavioral experiments one trial at a time. OpenSesame [MST12] provides a comprehensive approach to building an experiment for the social sciences. Some of these experimental design tools from the behavioral sciences could be adapted to visualization research by researchers willing to learn them.

2.1. User Study Implementation as an Online Survey

Another approach to streamline the implementation and data-collection aspects of an evaluation is through a survey builder. Geared towards the business world, Qualtrics [Qua] is a well-known survey platform. Its powerful survey development functionality can be equally exploited by academic researchers to develop research experiments. With academic licensing, Qualtrics is usually free, it's easy to learn, and it fulfills the personal data security requirements of human-based academic research. With a wide range of question types, it has become a common study implementation platform in the behavioral sciences.

While Qualtrics provides the functionality to include graphics, it does so on an image by image and question by question basis. Images must be uploaded into the graphics library and then individually loaded into the appropriate place within a question format. Upon upload, Qualtrics generates a random URL/name for each image. The time to upload the many images and create the large number of questions needed by psychophysical image-based approaches can become laborious. The randomized names create an additional difficulty. For images with a high degree of similarity, as one might find with a typical set of 2-alternative forced choice stimuli images, error-checking a study requires checking each random name against the original uploaded version. With potentially dozens of images in an image-based experiment, this becomes onerous.

The Qualtrics JavaScript API is a natural solution to automate the process of image presentation. The Evaluation Toolkit (ETK) is a series of JavaScript/HTML/CSS modules that leverage the Qualtrics JavaScript API to automate the implementation and presentation of image-based perceptual experiments.

2.2. Mturk: Crowdsourcing User Evaluation

Following in the footsteps of the behavioral sciences [BKG11, CMG13, MS12], crowdsourcing user evaluation is rapidly gaining traction in visualization (e.g., [BVB*13, HYFC14, HB10, LH13, OJ15, WTS*17]). Crowdsourcing has many advantages over traditional laboratory studies. Study participants are easily recruited through crowdsourcing sites and generally encompass a much broader demographic than a typical university participant pool [PCI10, PC14]. Crowdsourcing can be both cost and time effective. The fast turn-around time available with this method makes it possible to incorporate user evaluation early in the process of designing a visualization. Researchers must of course balance ecological validity with losing some amount of control over participants and monitor viewing conditions. Thus, not all user evaluations are appropriate for a crowdsourced approach. However, by casting standard psychophysical techniques into purely image-based experiments, the power of crowdsourcing can potentially be harnessed for a wide range of perceptual experiments in visualization. MTurk itself provides an easy linkage to outside survey software sites such as Qualtrics [Qua].

3. The Evaluation Toolkit: ETK

The Qualtrics JavaScript API allows a user to expand the question functionality available within Qualtrics through a standard set of methods called on the question object. A wide range of predefined functions and properties can be used, for example, to write out question responses, hide/show the "next" button to move on to the next question, or interact with the question container or question choices. The question-level JavaScript works with question-level HTML and a survey-level CSS file to create the question text, question actions, and modify the look and feel as needed.

There are currently six modules within ETK that encompass a range of psychophysical or general experimental approaches:

2AFC 2-Alternative Forced Choice module: a module that presents the stimuli images for a 2AFC experiment, each compared to a baseline image.

MoA Method of Adjustment module: a module that allows a subject to cycle back and forth through a set of stimuli images and make a choice of a specific image, e.g., at a threshold.

RRC Round Robin Comparison module: a module that presents each possible pair of images in an A/B choice experiment.

C2A Compare 2 Arrays: a module that compares each image in an array to its counterpart in a second array. This can be used to present specific A vs. B comparisons or as a 2AFC approach where the baseline image varies as a function of stimuli level.

CC Click Counting: a module that counts the number of times the displayed image is clicked.

KT Key Task: a flexible module that displays a series of individual stimuli images along with a set of answer keys; keys can be coded by color or by name/text.

Each module creates an image container (or a canvas in the case of the CC module) within the Qualtrics question container. The image container is used to display either a single stimuli image or a pair of comparison images as appropriate to the module. The ETK modules handle the randomizations required by the experimental approach. For a module displaying a list of single stimuli images such as KeyTask, image order is randomized. For modules displaying a pair of images, such as 2AFC, RRC or C2A, both the order in which the comparison pairs are displayed is randomized along with the left/right order in which the images appear within their container. The MoA module requires an ordered set of stimuli images. As a check for potentially bad participants, a flag is set when a participant only chooses one side (left or right) of all comparison

pairs. The functionality of the 2AFC module is extended to allow multiple baselines and multiple images for each stimuli level. If those multiple options are populated, then the choice for each baseline or stimuli is appropriately randomized. Where relevant, a user flag determines if a randomly chosen subset of images is shown or if the full set is shown.

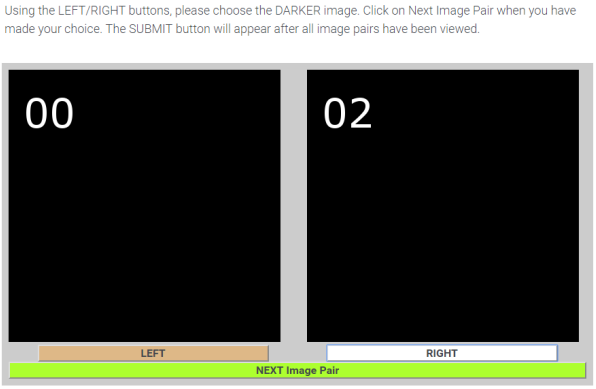

Figure 1: *The 2AFC Demo uses the trivial case of determining the threshold for discriminating between a grey box and the black baseline. The numbers on each image are for code-checking/development and indicate which of the multiple baselines and stimuli images has been randomly chosen. The user selection buttons are highlighted when a choice has been made (white). The* Next Image Pair *button is only active after an initial choice has been made (green).*

An image URL and image arrays provide the image input handles for the user and allows the researcher to host images on one's own server. This avoids the Qualtrics-imposed name change for uploaded images. Output is via Qualtrics *Embedded Data* variables. Image names and subject choices are saved to Qualtrics embedded variables. In the case of the 2AFC module, the randomization for A/B vs. B/A is flipped as needed so that the output already has that randomization unfolded, obviating the need to know the actual presentation order for the analysis. In an actual study, one module might be implemented in several different questions, allowing multiple conditions to be presented to the subjects (for example, using different colormaps or visualization rendering techniques in each set of stimuli images). The Qualtrics Survey Flow can be used to order and randomize multiple questions to control the set of conditions presented to each participant to accommodate either within-subject or between-subject experiments.

Each ETK module includes a tailored README and a set of trivial example images that can be used to explore the module functionality before user-specific implementation. Figure 1 shows one comparison for the 2AFC demo using the example images. Each module also includes a screenshot of the Qualtrics survey flow to illustrate the necessary embedded variables to output subject responses. Each module consists of the necessary files that must be embedded within the Qualtrics survey or question. The user need only make minor modifications, such as changing images names, sizes and URL, to customize the JavaScript, HTML and CSS files for their specific study. Modules can be downloaded from `github.com/ascr-ecx/etk`. A tutorial is available in the supplemental material and the ETK website, `www.etklab.org` provides demos and documentation.

4. Use Examples

ETK was a natural outgrowth from the evaluation of scientific visualizations conducted over the course of our project. The ease of general study implementation within Qualtrics makes it an accessible research platform. The straightforward coupling between Qualtrics and Mturk allows quick and easy study implementation in Qualtrics and participant recruitment from Mturk. We illustrate the value of ETK within that process with examples from our research.

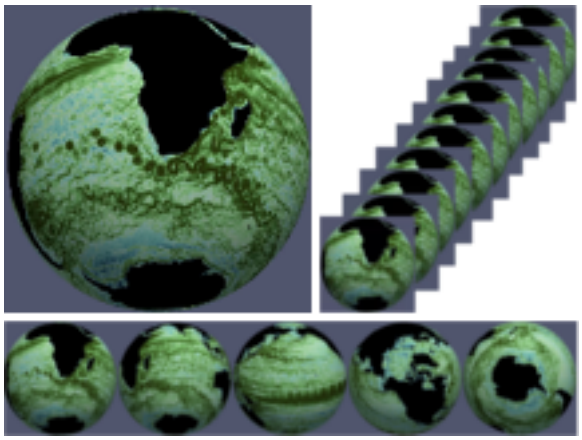

Figure 2: *A 2AFC discrimination threshold experiment for video compression needs to compare the baseline (uncompressed) image (top left) with many stimuli levels (top right). In the ETK 2AFC module, the images are condensed to an array of image names. Five regions across the globe (bottom) average over multiple conditions.*

4.1. 2AFC Discrimination Threshold

We first consider the case of Video compression of image databases [BTRA16, BTP*17]. The discrimination threshold at which subjects can distinguish the effect of compression is a task that naturally lends itself to an image-based crowdsourced approach. The uncompressed baseline is compared to a series of compressed stimuli images, Figure 2. In an initial study, the baseline plus 11 stimuli levels were used. Five regions across the globe represent five possible conditions, leading to 60 total images. Previous experience in implementing a 2AFC experiment in Qualtrics, without using an ETK module, has shown us that, for a study with roughly 50 images across three conditions, it can take up to a day to implement the necessary set of Qualtrics questions based on the stimuli set and baseline images, error check the image names/links and set up the randomizations within Qualtrics necessary to do a full 2AFC ramp. With ETK, the image URL goes into a variable, names go into an array. Image error checking consists of simply scanning the array code for typos. Randomization of image pairs and ordering automatically takes place within the module. Implementing an ETK question is a matter of 30 minutes of effort. Copying one Qualtrics question quickly generates questions for each of

the other regions (conditions) and 10 minutes are needed to edit the JavaScript for each new question to point to the correct image URLs. While the full study including questions such as consent, training, and demographics will take additional time, that is time that must be spent regardless of how the image questions are created. This study was crowdsourced on Mturk.

The pilot study with 11 stimuli levels of varying compression indicated the general location of the discrimination threshold. We then added a few additional images around the threshold to better map out the discrimination ramp, Figure 3. Because we were utilizing an ETK module, updating the experiment to include additional stimuli levels consisted of editing the five JavaScript files to add a few additional lines of code in each of the image arrays – a matter of 10 minutes of effort before the study was ready to relaunch.

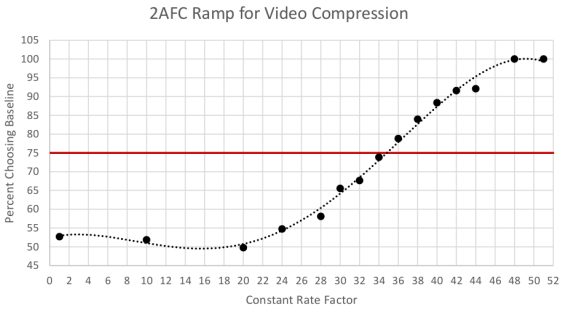

Figure 3: *The 2AFC discrimination ramp for the constant rate factor (CRF) for video compression. The polynomial fit crosses the threshold (75% line) just above CRF=34.*

4.2. Key Task

A second example is from a recent paper [TWSR17] that utilized the key task module from ETK. This study deliberately replicated Ware's 1988 *Univariate Color Sequences* paper [War88] in order to study issues of color vision deficiency(CVD) and Mturk as a platform for perceptual studies in visualization. The study included 60 stimuli images for each of eight colormaps – 480 images in total. The key task module inserts a stimuli image into the image container and generates the matching set of keys, Figure 4, outputting which key was chosen along with the identifying information of which stimuli image was shown. An array of 60 stimuli images is again easily error-checked by simply scanning the JavaScript array. Randomization and presentation of a subset of the full stimuli images were handled automatically by the ETK module. While we do not have a comparison of how much time it would take to upload, insert into questions, and error check 480 images within Qualtrics without an ETK module, the prospect is daunting. The relative agreement between [War88] and [TWSR17] provided validation of Mturk as a research platform while quantifying the impact of CVD and assessing a more recent set of colormaps.

Other types of experiments that can be easily implemented using ETK include standard A/B choice experiments such as those in [SKP*16]. This can be implemented either using the Round Robin Comparison module or by hard-coding the specific comparisons within the Compare 2 Arrays module. The color counting

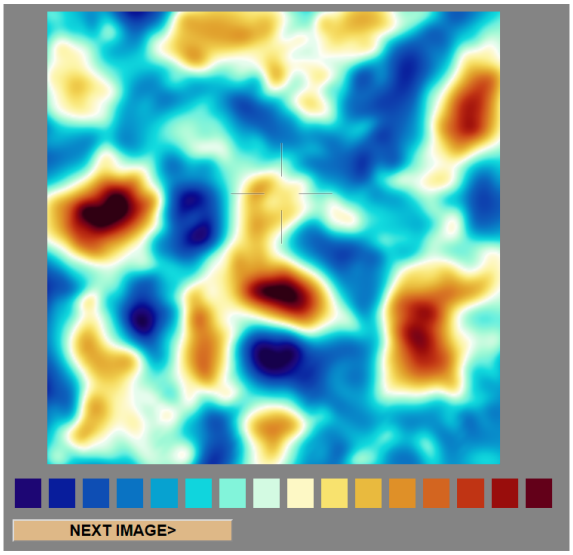

Figure 4: *Example stimuli image for key task experiment.*

approach of [SPG*15] is covered in the functionality of the Click Counting module.

5. Conclusions

We have presented ETK: the Evaluation Toolkit, a new approach to developing user evaluation studies for visualization. As ETK is based solely on images as the visualization artifacts, it can't provide the flexibility for interactive evaluation that one might find in a tool such as EvalBench. However, it simplifies the study implementation process and avoids the step by step implementation needed by tools such as OpenSesame or jsPysch. By taking advantage of the ease of study design within Qualtrics, ETK allows researchers with a wide range of experimental backgrounds to easily implement perceptual image-based experiments. Placing ETK within the visualization evaluation patterns context of Elmqvist and Yi [EY12], ETK can be used in a pilot study; it can provide a study for the complementary participants pattern enabling comparison between experts and a more general demographic; and it excels at presenting individual trials for an experiment.

We are just beginning to explore the range of visualization experiments where ETK can provide a solution and invite other researchers to explore, suggest, and develop new ideas and solutions to share with the community. We invite you to explore the ETK website, `www.etklab.org` for live demos and documentation, and download the toolkit from `github.com/ascr-ecx/etk`.

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 and DE-SC-0012516. The authors would like to thank Dr. Colin Ware, Dr. Phillip Wolfram, and Divya Banesh.

References

[AHR13] AIGNER W., HOFFMANN S., RIND A.: Evalbench: A software library for visualization evaluation. In *Proceedings of the 15th Eurographics Conference on Visualization* (Chichester, UK, 2013), EuroVis '13, The Eurographs Association & John Wiley & Sons, Ltd., pp. 41–50. URL: http://dx.doi.org/10.1111/cgf.12091, doi:10.1111/cgf.12091. 2

[AK07] ANDREWS K., KASANICKA J.: A comparative study of four hierarchy browsers using the hierarchical visualisation testing environment (hvte). In *Information Visualization, 2007. IV '07. 11th International Conference* (July 2007), pp. 81–86. doi:10.1109/IV.2007.8. 1

[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. 2

[BTP*17] BERRES A. S., TURTON T. L., PETERSEN M., ROGERS D. H., AHRENS J. P.: Video compression for ocean simulation image databases, June 2017. Accepted to EnviroVis 2017: EuroGraphics Workshop on Visualization in Environmental Sciences. 3

[BTRA16] BERRES A., TURTON T. L., ROGERS D., AHRENS J.: VideoDB: Reducing large image databases through video encoding and video compression. LA-UR-16-24358, 2016. 3

[BVB*13] BORKIN M. A., VO A. 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 (Dec. 2013), 2306–2315. URL: http://dx.doi.org/10.1109/TVCG.2013.234, doi:10.1109/TVCG.2013.234. 2

[Car08] CARPENDALE S.: Information visualization. Springer-Verlag, Berlin, Heidelberg, 2008, ch. Evaluating Information Visualizations, pp. 19–45. URL: http://dx.doi.org/10.1007/978-3-540-70956-5_2, doi:10.1007/978-3-540-70956-5_2. 1

[CMG13] CRUMP M., MCDONNELL J., GURECKIS T.: Evaluating amazon's mechanical turk as a tool for experimental behavioral research. *PLoS ONE 8*, 3 (2013). 2

[dL15] DE LEEUW J. R.: jspsych: A javascript library for creating behavioral experiments in a web browser. *Behavior Research Methods 47*, 1 (2015), 1–12. doi:10.3758/s13428-014-0458-y. 2

[EE99] EHRENSTEIN W. H., EHRENSTEIN A.: Psychophysical methods. In *Modern Techniques in Neuroscience Research*, Windhorst U., Johansson H., (Eds.). Springer-Verlag, Berlin, 1999, ch. 43, pp. 1211–1241. 1

[EY12] ELMQVIST N., YI J. S.: Patterns for visualization evaluation. In *Proceedings of the 2012 BELIV Workshop: Beyond Time and Errors - Novel Evaluation Methods for Visualization* (New York, NY, USA, 2012), BELIV '12, ACM, pp. 12:1–12:8. URL: http://doi.acm.org/10.1145/2442576.2442588, doi:10.1145/2442576.2442588. 4

[Fec89] FECHNER G. T.: *Elemente der Psychophysik*, 2nd ed. ed., vol. 2. Breitkopf & HÃd'rtel, Leipzig, 1889. 1

[For10] FORSELL C.: A guide to scientific evaluation in information visualization. In *2010 14th International Conference Information Visualisation* (July 2010), pp. 162–169. doi:10.1109/IV.2010.33. 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. 2

[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. 2

[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. doi:10.1145/2470654.2466424. 2

[MABL*07] MACKAY W. E., APPERT C., BEAUDOUIN-LAFON M., CHAPUIS O., DU Y., FEKETE J.-D., GUIARD Y.: Touchstone: Exploratory design of experiments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2007), CHI '07, ACM, pp. 1425–1434. URL: http://doi.acm.org/10.1145/1240624.1240840, doi:10.1145/1240624.1240840. 2

[MS12] MASON W., SURI S.: Conducting behavioral research on amazon's mechanical turk. *Behavior Research Methods 44*, 1 (2012), 1–23. 2

[MST12] MATHÔT S., SCHREIJ D., THEEUWES J.: Opensesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods 44*, 2 (2012), 314–324. URL: http://dx.doi.org/10.3758/s13428-011-0168-7, doi:10.3758/s13428-011-0168-7. 2

[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. 2

[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. 2

[PCI10] PAOLACCI G., CHANDLER J., IPEIROTIS: Running experiments on amazon mechanical turk. *Judgment and Decision Making 5*, 5 (2010), 411–419. 2

[Pla04] PLAISANT C.: The challenge of information visualization evaluation. In *Proceedings of the Working Conference on Advanced Visual Interfaces* (New York, NY, USA, 2004), AVI '04, ACM, pp. 109–116. URL: http://doi.acm.org/10.1145/989863.989880, doi:10.1145/989863.989880. 1

[psi] psiTurk Website. https://psiturk.org/. 2

[Qua] Qualtrics website. www.qualtrics.com. 1, 2

[RKPC99] ROGOWITZ B., KALVIN A., PELAH A., COHEN A.: Which trajectories through which perceptually uniform color spaces produce appropriate color scales for interval data? In *7th Color and Imaging Conference* (1999), pp. 321–326. 1

[SKP*16] SAMSEL F., KLAASSEN S., PETERSEN M., TURTON T. L., ABRAM G., ROGERS D. H., AHRENS J.: Interactive colormapping: Enabling multiple data range and detailed views of ocean salinity. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (New York, NY, USA, 2016), CHI EA '16, ACM, pp. 700–709. URL: http://doi.acm.org/10.1145/2851581.2851587, doi:10.1145/2851581.2851587. 4

[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. 4

[TWSR17] TURTON T. L., WARE C., SAMSEL F., ROGERS D. H.: A crowdsourced approach to colormap assessment, June 2017. Accepted to EuroRVVV 2017: EuroVis Workshop on Reproducibility, Verification, and Validation in Visualization. 4

[War88] WARE C.: Color sequences for univariate maps: Theory, experiments and principles. *IEEE Computer Graphics and Applications 8*, 5 (1988), 41–49. 1, 4

[WTS*17] WARE C., TURTON T. L., SAMSEL F., BUJACK R., ROGERS D. H.: Evaluating the perceptual uniformity of color sequences for feature discrimination, June 2017. Accepted to EuroRVVV 2017: EuroVis Workshop on Reproducibility, Verification, and Validation in Visualization. 2