A Crowdsourcing System for Integrated and Reproducible Evaluation in Scientific Visualization

Rickard Englund^{1*}

Sathish Kottavel^{1†}

Timo Ropinski^{2‡}

¹ Interactive Visualization Group, Linköping University, Sweden

² Visual Computing Group, Ulm University, Germany

Figure 1: The proposed system's workflow is divided into three phases. In the first study-centered phase, a questionnaire design is created or selected, and combined with visual stimuli to generate trials. The second phase involves conducting the study by submitting it to the crowd from where the response data is collected and stored in a centralized database. In the third result-centered phase, the study results are analyzed and a report is generated.

ABSTRACT

User evaluations have gained increasing importance in visualization research over the past years, as in many cases these evaluations are the only way to support the claims made by visualization researchers. Unfortunately, recent literature reviews show that in comparison to algorithmic performance evaluations, the number of user evaluations is still very low. Reasons for this are the required amount of time to conduct such studies together with the difficulties involved in participant recruitment and result reporting. While it could be shown that the quality of evaluation results and the simplified participant recruitment of crowdsourcing platforms makes this technology a viable alternative to lab experiments when evaluating visualizations, the time for conducting and reporting such evaluations is still very high. In this paper, we propose a software system, which integrates the conduction, the analysis and the reporting of crowdsourced user evaluations directly into the scientific visualization development process. With the proposed system, researchers can conduct and analyze quantitative evaluations on a large scale through an evaluation-centric user interface with only a few mouse clicks. Thus, it becomes possible to perform iterative evaluations during algorithm design, which potentially leads to better results, as compared to the time consuming user evaluations traditionally conducted at the end of the design process. Furthermore, the system is built around a centralized database, which supports an easy reuse of old evaluation designs and the reproduction of old evaluations with new or additional stimuli, which are both driving challenges in scientific visualization research. We will describe the system's design and the considerations made during the design process, and demonstrate the system by conducting three user evaluations, all of which have been published before in the visualization literature.

Index Terms: H.3.4.1 [Information Systems]: Systems and Software—Performance evaluation (efficiency and effectiveness); I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism—Color, shading, shadowing, and texture

*e-mail: rickard.englund@liu.se

†e-mail: sathish.kottavel@liu.se

‡e-mail: timo.ropinski@uni-ulm.de

1 INTRODUCTION

Research on computer-supported data visualization now dates back roughly thirty years, in which the community is working on efficient and effective visualization algorithms for different application areas. While this research led to many high impact results in form of visualization algorithms used in today's data analysis software, the prediction of the efficiency and the effectiveness of visualization algorithms from a user's perspective is still not possible [7]. This missing predictability is acknowledged by the community, and led to the demand of including user evaluations in submitted visualization research papers, in order to demonstrate the benefits of the developed algorithms [40]. As a consequence, recent literature reviews show a clear trend towards evaluations that investigate the performance and the experience of the user [29]. However, very few of these reported evaluations are of quantitative nature, despite the fact that in general science depends on quantitative studies [5]. Isenberg and colleagues have investigated the use of evaluations in scientific visualization, and found that evaluations in this area are time- and resource-intensive to design, conduct, and analyze [23]. They state, that these are the main reasons for the shortage of evaluations in scientific visualization. Additionally, many of the few quantitative user evaluations reported in the scientific visualization community involve problems. First, as many visualization researchers have a technical education, often the skill set for correctly performing, analyzing and reporting the results of quantitative evaluations leaves space for improvement [12]. Second, often the data acquired during an evaluation is not available, or only available in proprietary formats, which makes the transfer and the reproducibility of the results difficult [23]. As reproducibility is a basic foundation for science [11, 39], it is important to improve this aspect for research within the scientific visualization community. Third, as conducting a controlled user evaluation can be a tedious task, the development and evaluation of visualization algorithms often follows a linear pattern, whereby a summative user study is conducted at the end. This does usually hamper to integrate the evaluation results into the development process, and thus may hinder the improvement of the developed algorithms based on the latest findings.

Crowdsourcing is a technique to gather information from a large group of people online, whereby participants are usually paid a small amount of money for each completed task. It is often used to perform tasks that are simple for humans but hard for computers such as object recognition and image assessment. Kittur et al. investigated the possibility to use crowdsourcing to perform user evaluations [25]. They could show that, with correct questionnaire design, user evaluations performed online, using Amazon's

Mechanical Turk to recruit participants, have comparable results to evaluations done by experts. Similarly, Heer and Bostock investigated how crowdsourcing can be used to evaluate graphical perception by replicating previously conducted lab experiments [21]. Their results indicate that the time needed to conduct a user study using crowdsourcing is greatly reduced compared to lab experiments. They further report that the costs required for crowd sourcing evaluations is lower by a factor of 6-9 as compared to lab experiments. Since the number of participants is not constrained by the available lab equipment, crowd sourcing studies also allow for a much higher number of participants. However, regarding the statistical confidence Heer and Bostock [21] find, that it is comparable to the one achieved in lab experiments, and Kittur et al. [25] report that the positive correlation is statistically significant. An additional benefit is that participants of crowdsourced user studies can participate at any given time, and is not limited to the opening hours of the lab. While there has not yet been any work published where crowdsourcing has been utilized for evaluating scientific visualization research, crowdsourcing has been successfully utilized for user studies in information visualization [9, 10, 16, 17, 19, 24, 38, 47].

In this paper, we address the downsides observed in the context of quantitative user evaluations in scientific visualization research. Therefore, we propose a system for evaluating scientific visualization algorithms, which minimizes the overhead for conducting such evaluations, while at the same time assists the user in setting up such studies and in analyzing and reporting their results. Due to the inherent benefits of crowdsourcing services, we adopt this approach to simplify the conduction of quantitative user evaluations in scientific visualization. By allowing spawning of evaluations directly from the algorithm development framework, we hope to increase the amount of formative evaluations performed by researchers during their development process and use the results to improve their algorithms, instead of only performing a summative evaluation at the end of the whole process. Thus, the presented system has been designed to lower the barrier for conducting crowdsourced user evaluations by supporting questionnaire design, trial generation, the submission and analysis of studies. To further support reproducibility, all relevant study data is stored in a centralized database, which supports to make all study data publicly available and thus enables other researchers to analyze the existing results, rerun already conducted studies, or reuse the results within the context of a different study. We believe that once the presented system is available to the community, it fills in an important gap in visualization research, as it lowers the barrier for performing quantitative user evaluations in a reproducible manner.

2 RELATED WORK

In this section, we discuss prior work related to the proposed system, whereby we outline quantitative evaluation, crowdsourcing and evaluation systems, all with a special focus in scientific visualization.

Quantitative evaluation. Naturally, the lack of prediction in visualization design, leads to an increased value of user evaluations, which is amongst others indicated by the spawn of focused workshops as well as tutorials [6], and the emphasis of evaluation as a key research challenge [30]. Consequently, many authors address this problem by formalizing the evaluation process in visualization [5, 14, 37, 40, 43], whereby several evaluation scenarios exist that involve the user [29]. However, in this paper we focus on User Performance (UP) and User Experience (UE) evaluations conducted in the form of quantitative studies. In visualization research, these types of studies can be used to address perceptual and comprehensibility questions [20], whereby low-level tasks such as compare, contrast, associate, distinguish, rank, cluster, correlate or categorize can be considered [40]. Unfortunately, the number of quantitative user evaluations is still rather low [23], despite the fact that these evaluations can be used to address a broad variety of research questions, reaching from aesthetics [41], over spatial comprehension [18], and visual averaging [17] to memorability [3].

Crowdsourcing. As the results of conventional user studies are

affected by the relatively small number of users often having a homogeneous background [5], crowdsourcing has been used in recent years to incorporate perceptual and cognitive aspects in visualization research. Kittur et al. were the first, who have investigated the application of crowdsourcing platforms for user evaluation [25]. A systematic review of crowdsourcing research shows that evaluations are among the top three research foci using crowdsourcing [42]. Heer and Bostock have analyzed the benefits of crowdsourcing when conducting perceptual experiments in the field of visualization [21]. Based on their replication of prior laboratory experiments by using Mechanical Turk, they conclude that crowdsourcing can be used as an alternative, as it gains new insight into visualization design, enables up to an order of magnitude of cost reduction, a faster completion time, and access to a wider population. Due to these benefits several crowdsourcing studies have been conducted to evaluate visualization design in recent years [3, 17, 21]. While most of these use payment to motivate the participants, gameplay has also been investigated as a motivational factor in visualization research [1]. Ahmed and colleagues evaluate the performance of color blending functions by drawing colored circles in three different layers within a purpose-driven game. To obtain insightful feedback, they apply a parameter-sweep analysis, and the game is designed such that the participants are motivated to select the topmost circle, from which they can infer information regarding the blending. As all these studies aim at bringing visualization research forward, Heer and Bostock state that to further excel, future research is needed to develop better tools for crowdsourcing research, which for instance support dynamic task generation and easier access control [21]. A comprehensive survey of existing crowdsourcing systems has been published by Yuen et al. [48]. While software libraries have been proposed to enable easier crowdsourcing for general application cases [32], our proposed system is the first which fills in this gap for conducting user evaluations in scientific visualization research.

Evaluation systems. Due to the prominent role of evaluation, the importance of software that has been developed with the purpose to support this task is widely acknowledged. *EvalBench* has been proposed as a flexible software library that aims at conducting user studies in an easier way [2]. *EvalBench* is written in Java and can be used to evaluate visualization applications in laboratory experiments. To support the evaluation of human-computer-interaction research, *TouchStone* has been proposed as a design platform for exploring alternative designs of controlled laboratory experiments [33]. For the evaluation of information visualizations, Okoe and Jianu have recently proposed an evaluation system exploiting crowdsourcing to evaluate graph visualizations [38]. The system interfaces with crowdsourcing platforms and automatizes some aspects of the evaluation process. Also in areas outside computer science, the importance for software packages supporting the conduction of evaluations is acknowledged. In cognitive science, the *PsychToolbox* is a widely adopted platform for conducting psycho-physics experiments [26]. It supports the generation of visual and auditive stimuli, to be used in perceptual laboratory experiments. To conduct experiments in the social sciences, *OpenSesame* provides an experiment builder that requires a minimum of effort [35]. While all these software packages successfully simplify some or all of the steps involved in conducting an evaluation, to our knowledge no software exists, which enables the conduction of scientific visualization experiments using crowdsourcing.

3 SYSTEM OVERVIEW

The current state-of-the-art of user evaluation in the scientific visualization community has motivated us to develop the proposed system, whereby two key observations were of particular importance. First, the relative low number of quantitative user studies conducted in the scientific visualization community, as reported by Isenberg et al. [23]. Second, the requirement of easy to use and flexible crowdsourcing tools that has been pointed out by Heer and Bostock, as they believe that such tools play an important role beyond simplifying study administration [21]. As Isenberg et al. assume that

Figure 2: Four examples of the various reusable visual perception tasks supported by the proposed evaluation system. The two leftmost images are examples of absolute and ordinal depth tasks. The two rightmost images are examples of localization of critical points tasks and particle tracing tasks. Before launching a study, the researcher can choose one or more of these tasks right from within the evaluation environment.

the low number of quantitative user evaluations results from the involved time and resource efforts as well as the challenges during participant recruitment, we believe that the development of crowdsourcing platforms for evaluating scientific visualization algorithms is the key to this problem. Furthermore, such a system can assist researchers in proper reporting, and making the acquired study data available to the public, both are main ingredients for reproducible evaluations.

With these observations in mind, the proposed system has been developed with respect to the requirements of quantitative evaluations of UP and UE, whereby low-level perceptual and aesthetic investigations were under special focus. Forsell states that six steps are needed to conduct an evaluation: preparation of the study design, defining tasks and providing data, participant recruitment, study conduction and data collection, result analysis, and reporting [14]. While these steps are designed for conducting evaluations in a controlled environment, with some modifications they can also be applied to evaluations conducted using crowdsourcing platforms. Here, the participant recruitment is done through the crowdsourcing platform and the data collection step is automatically taken care of during the conduction step. Based on these steps, we have identified three main phases of a crowdsourced evaluation for scientific visualization as it is illustrated in Figure 1. In the first phase, the study is prepared by first creating the questionnaire designs to be used, before generating the visual stimuli, which are then combined with the questionnaire designs to generate trials. The second phase involves the conduction of the study on a crowdsourcing platform and the collection of the response data. The third phase is focused on analyzing the study data and reporting the results.

To enable the workflow depicted in Figure 1, we facilitate modern web technologies directly interfaced from within a versatile open source scientific visualization research framework, Inviwo [46]. Thus, the web-based parts of the system are realized in PHP, whereby a MySQL database is exploited for data storage. The choice for implementing the presented system as a web platform comes naturally, since crowdsourcing platforms recruit the participants over the web and require the participants to use their browser to complete tasks. The system comprises out of a front-end targeted towards study participants, and a back-end targeted towards the researchers planning and conducting a study. The front-end supports the actual study conduction, and collects the response data in the background. The back-end is the administrative entry point where visualization researchers perform study administration, design, and launch studies, before they perform result analysis and reporting. To enable an easy study initiation, we support interfacing with existing software frameworks through a RESTful API. Thus, we can generate stimuli, studies, and trials by transferring data over the HTTP protocol.

While in principle the concepts described in this paper can be integrated in any system used for developing new visualization algorithms, we have decided to realize them in close integration within Inviwo, a data-flow development environment for scientific visualization [46]. We see two main advantages arising from this deci-

sion. First, data-flow based development environments have a long tradition in visualization research as they enable a direct mapping of the visualization pipeline [36]. Second, as each data-flow network expresses its state through the set of parameters with their respective values, they are suitable for controlled user experiments, as one or several parameters can be varied in a controlled manner to generate a set of visualizations to be evaluated.

4 STUDY PREPARATION

The first phase of our system’s workflow consists of the questionnaire design and the creation of visual stimuli. These two tasks can be performed independently, and in case a similar study has been conducted earlier, components of that study can be reused. In the following subsections, we detail how a user of the proposed system can design questionnaires and create visual stimuli. For both tasks we will elaborate on the design considerations that have been taken into account when realizing the described functionality.

4.1 Questionnaire Design

Questionnaires are an easy to use and widely adopted mechanism to acquire user feedback, as they support objectivity, comparison and verification, as well as a quantification of the responses [15]. Shneiderman and Plaisant argue, that one way of assessing the efficacy of visualization tools is through documenting usage encoded as observations, interviews, surveys and logging [44]. Accordingly, questionnaires assessing this data are also used in the visualization community when users are involved in evaluations. In our workflow, selecting or designing a questionnaire is the initial step, as it defines the task to be conducted by the user. With respect to the user evaluation steps described by Forsell [14], this step combines the preparation of a study design together with the task definition.

In our system the questionnaire design is split into two steps. In the first step the user defines how many visual stimuli to be included, and which visual attachments to realize. Visual attachments, in the form of markers overlaid over the visual stimuli, allow for communication of points of interest or of widgets requiring location-based user input, a few examples of such attachments can be seen in Figure 2. In the second step, the user designs the layout of the questionnaire. For flexibility reasons, we have realized the designing of questionnaire layouts through an online questionnaire editor, which allows the user to flexibly build the questionnaire layout in PHP, HTML, CSS and JavaScript. Thus, the questionnaire editor enables researchers to realize their own questionnaire layouts which might be the result of a long research process [13]. Once the questionnaire design is complete it is stored in the database and can be used in a study. Thus, we allow for each design to be reused with or without modifications. We use PHP to fill the questionnaire with the trial content, that is, the stimuli and the visual attachments. JavaScript allows for advanced interaction, such as visual attachments realizing gauge figures, and extra data collection realized through logging of user interactions, such as for example mouse movements.

We believe that storing the questionnaire designs and allowing them to be reused and improved can serve as an important tool, when it comes to the standardization of questionnaires as it is called for by Forsell and Cooper [15]. As they expect the standardization process to involve a large number of studies and participants, the proposed questionnaire editor can help to realize this process, as varieties of questionnaires can be easily generated and tested with a large number of participants.

4.2 Visual Stimuli Generation

Visual stimuli are the backbone when evaluating visualizations, and therefore it is mandatory that the visual stimulus generation is powerful and flexible. To aid the result analysis, our system exploits textual label pairs, consisting of a group and a value, which are associated with the visual stimuli. These labels can for instance be used to describe the underlying visualization algorithm used to generate a stimulus, or alternatively encode the set of used parameters. In addition to these labels, researchers can attach annotation layers to the stimuli. These layers provide additional information for the stimuli, such as the surface normals or the depth values for all pixels in a 3D renderings. A set of example stimuli and the corresponding depth annotation layer can be seen in Figure 3 together with a label for the parameter the varies between the stimuli. The beauty of the concept of annotation layers in a visualization development system is, that specialized components can be easily developed which provide annotations based on the used visualization. By assigning these annotation layers to a visual stimulus, their content can be used to encode correctness. As we will discuss in Section 6, this encoding supports the automated analysis of the evaluation results during UP evaluations. We have identified three different approaches to generate visual stimuli together with associated keywords and annotation layers. We have implemented these three mechanisms in the data-flow based visualization framework underlying our system, whereby for each stimulus we also store the data-flow network used to generate the stimulus. Thus, we support researchers to trace back a stimulus to the underlying algorithms and parameters at any step in the study workflow. In the following paragraphs we describe the three stimulus creation mechanisms.

Manual stimulus creation. As the simplest form, the system supports to manually capture a single stimulus based on the current data-flow network using a widget (Figure 4). The widget allows the researcher to select which of the canvases to capture as stimulus, and if any annotation layers are to be captured from other canvases. Furthermore, the widget supports researchers to label the stimuli with keywords.

Automatic stimulus creation. In contrast to the manual stimulus creation, an automated approach is often beneficial to supply stimulus content. When dealing with quantitative evaluations, controlled parameter changes are often performed to simplify the analysis process [5]. These changes systematically analyze and conquer the parameter space to support the user when comparing visualizations. To our knowledge, the first application of parameter-sweep analysis in crowd sourcing visualizations has been presented by Ahmed et al., who successfully applied a stochastic approach to sample the parameter space [1]. Instead our system supports a linear parameter sweeping of the parameter space of several algorithm parameters of interest. We use this sampling for an automated visual stimulus creation based on these sweeps. Therefore, the user has to select one or more parameters of which the influence shall be investigated from within the connected visualization development environment. In Figure 3 we can see an example of a parameter sweep where a protein is rendered with depth of field and the focal distance is the swept parameter to generate a set of visual stimuli. This stimuli have been used in a study to evaluate optimal focal distances.

To realize the sweeping, we distinguish discrete and continuous parameters. For a discrete parameter p_d the user can directly select a subset of p_d 's allowed values to retrieve the n_{p_d} samples to be evaluated. To investigate the influence of a continuous parameter p_c , interpolation is required. Therefore, the user needs to specify p_c 's value domain, given as v_{min} and v_{max} as well as the number of

samples n_{p_c} to be evaluated. Based on these values we linearly generate n_{p_c} parameter samples $v_i = v_{min} + (i \cdot \frac{v_{max} - v_{min}}{n_{p_c}})$. Note, that this proceeding requires to implement a parameter interpolation function for all non-numerical parameters. Currently, we have such an implementation for most continuous parameters, whereby we have focused on linear interpolation only. In case the interdependency of multiple parameters shall be investigated, the user can select multiple discrete or continuous parameters and specify their value range as well as the number of samples. Thus, when having selected m parameters with n_p samples each, n_p^m possible parameter combinations exist. This vast growth of parameter combinations renders the usage of crowdsourcing technology an important ingredient, as a sufficient number of participants can be involved, which minimizes the risk of committing a type II error in the statistical analysis [8]. To give the user a hint on the number of possible parameter combinations, we prominently display it to the user. After all parameters have been swept, the actual parameter values are linked to the visual stimuli through automatic label generation. We see this parameter sweeping as one of the key features of our system, as many visualization algorithms require user set parameters [22], and the sweeping enables and effective evaluation of their influence.

Scripted stimuli creation. While the manual and automatic stimuli creation described above are quite flexible, they are still limited to the functionality exposed through the graphical user interface. To support less restricted stimulus creation we have realized a Python interface, in which the researcher can use the full potential of the Python programming language when generating visual stimuli. This can be very useful when it is desired to reproduce and import stimuli from evaluations conducted outside the proposed system, for example when information about the stimuli is encoded in the name of the file representing a stimulus. The Python interfaces allows for analyzing these file names and creating both system readable stimuli as well as the appropriate labels. The proceeding of parsing file name has been exploited for instance to replicate the multiclass scatter plot study discussed in Section 7.3.

4.3 Trial Generation

The next phase, after the questionnaire has been designed and the stimuli have been created, is the trial generation phase, where the trials are generated which the participants will be asked to complete. Each trial has a related questionnaire design and a set of visual stimuli attached to it. While the number and type of stimuli is defined during the questionnaire design, our system supports various trial generation schemes based on standard experimental designs. To identify, which schemes an evaluation system should support, we have reviewed quantitative user evaluations earlier reported in the visualization literature. Based on this analysis we have decided to confine ourselves to evaluations which include static images only, as they have been described in Kosara's overview [27], and as they are facilitated in many recent publications about evaluations, for example, [31, 45, 18, 17, 3]. While not all publications describe the used questionnaires in all necessary detail, in most cases we were able to reconstruct the underlying scheme based on the result analysis. Thus, we were able to observe reoccurring schemes and patterns used in quantitative evaluations. Based on these observations, we have identified the four most important schemes reoccurring in evaluations in the area of scientific visualization. *Between-subjects design*, as for instance used by Gleicher et al. [17], confronts each user with only one level of the evaluated factor. In other words, N lists of trials, where N is the number of levels of the factor to be evaluated, have to be generated. Another important trial scheme is the *within-subjects design*, as used in several other studies [31, 18]. In this scheme, all participants will perform all available trials and therefore also be subject to all levels of the factor. For this type of studies, only one list of trials is needed and the trials are placed in a random order to reduce learning effects. If the user evaluation contains a large amount of trials, having all participants complete all trials might not be practical. For those cases *blockwise within-subjects design* needs to be supported, which is also widely used, for instance by Gleicher et al. [17]. The

Figure 3: An example sequence of stimuli generated using the automatic stimuli creation mechanism. By sweeping the focal distance parameter for a Depth of Field rendering we generate a set of stimuli to be used in our studies. Together with the color representation of the images we also store the depth buffer as annotation annotation layer.

user will be subject to two different levels divided into two blocks. The first block contains the set of trials for factor level A and the second block all trials for factor level B. To be able to test each factor against all other factors, it is necessary to generate $\sum_{i=1}^N i = 0.5(N^2 - N)$ lists of trials. The fourth scheme frequently used in scientific visualization evaluations, is the *counterbalanced measures design* [34]. To support this scheme N lists of trials need to be generated, where N is the number of levels of the factor of interest. Additionally, one second factor is needed for stimuli grouping. Thus, the i :th trial in each list of trials will have the same level of the additional factor and vary only with respect to the investigated factor. For example, in the replicated vector field visualization study described in Section 7.1, the investigated factor is *technique* while the additional factor is *dataset*. The ordering of the factors to use for each trial is usually defined by using a balanced Latin square. A Latin square is a $N \times N$ matrix in which each row and column contains non-repeated values on the interval $[1, N]$. In a balanced Latin square these numbers appear in an order such that each number is followed by all other numbers just once. Since the number of rows and columns in a Latin square is always equal, a set of Latin squares will be concatenated when we have more trials than the number of levels of the investigated factor. In our system, we perform this concatenation of two Latin squares by generating the second one with a row offset of one, that is, all numbers increase by one.

The selection of stimuli, questionnaire design and trial generation scheme to generate a study is done on the web platform aided by a wizard. Each step of the wizard aids the researcher in selecting data for that step and support visual preview of the outcome when possible. After the lists of trials has been generated we need to set parameters for the visual attachments. Our system provides editing of the attachments using a interface, which displays the stimuli alongside its annotation layers. When visual attachments with a position, the system will display the pixel color as well as the annotation layer information underlying that position. This information is useful in many cases, for instance, when defining markers for a depth comparison trial, the values from a depth annotation layer provides the depth at the markers to support trials of moderate difficulty. Furthermore, for each pixel color our system also calculates the complimentary color, which can be useful when deciding on a contrasting color for each marker. Lastly, the web-based interface supports population of the current visual attachment to all other trials fulfilling certain requirements. For example if the list of trials is generated using the counterbalanced measures design as described above, we can copy the attachment data to each trial using the same data set.

Validation trials. In a conventional lab-based user evaluation, participant recruitment plays a crucial role, as it ensures that the participants are motivated and have the required skills. In crowdsourcing the recruitment process is shifted towards the crowdsourcing platform, which gives researchers only relatively little control over participant selection. While *random clickers* were a serious problem during the advent of crowdsourcing, nowadays sophisticated techniques exist for detecting and eliminating these. As recommended by Heer and Bostock, we exploit verifiable questions in validation

trials to increase the likelihood of high quality results [21]. Therefore, the user can mark certain trials as validation trials in the trial list. These trials should contain stimuli and marker data, which make a correct answer possible for everybody paying a minimum of attention. During the analysis, this validation trials are then used to filter the results.

5 STUDY CONDUCTION

While in principle any crowdsourcing platform can be steered from within our system, in the current implementation we have chosen to use the CrowdFlower platform¹, a meta-platform which enables distribution of trials to various other crowdsourcing platforms. Our system communicates with and CrowdFlower using their RESTful API, where data is sent using URL-encoded key-value pairs and through the JSON format. When submitting studies, the user can tune how many participants should be involved, how much to pay for each trial list and the maximum number of trial lists each participant is allowed to work on. Based on these parameters, we calculate and display the resulting costs, in order to enable the user to optimize these. Upon submission we create a CrowdFlower job with one data row for each trial list in our study. When a participants start the CrowdFlower job they are presented with a link that points them directly to a trial list in our system.

6 RESPONSE DATA PROCESSING

After an evaluation has been successfully conducted, the user feedback needs to be acquired and the results need to be analyzed and reported. In this section, we describe how these steps are realized in our system.

Encoding correctness. While UP evaluations measure the user’s performance in terms of correctness and used time, UE evaluations enable an estimation of the user’s satisfaction when using the tested visualizations. Both evaluation scenarios, UP and UE evaluations can be conducted with our proposed system. However, to support analysis when dealing with UP evaluations, it is important to have a metric to measure the correctness or error for each judgment. As visual stimuli are the primary entities of a questionnaire for visualization evaluation, the correctness of an answer can often be calculated using the information stored in the stimuli’s annotation layers. Therefore, for each question asked in a UP questionnaire, we support the specification of correctness rules. These rules take the trial data and the user feedback as input to evaluate a participant’s estimation of a judgment task. In many cases, this correctness might be encoded in a pixel’s color, either in the stimuli itself but mostly in one of the annotation layers. We found this especially useful for evaluations exploiting markers, where the distance to the camera [18], surface normals [45], or object size [31] have to be judged. In some occasions the correctness cannot be specified a priori or annotation layers are not available, for example when reusing stimuli from old user evaluations, then the correct value can be stored in the stimulus as a label. Listing 1 shows an example rule which uses

¹www.crowdflower.com

Figure 4: The manual stimulus creation widget enables the selection of one canvas to serve as stimulus source, additional canvases for annotation layers and textual labels to be associated with the stimulus in order to support the subsequent analysis. The automatic stimuli creation widget supports to capture several stimuli by sweeping over parameters, it supports canvas selection and textual labels like the manual stimulus creation widget but also shows parameters selected for sweeping and allows for previewing of stimuli before uploading them to the system.

a depth image as an annotation layer to evaluate a user’s depth estimation. The feedback is given as a floating point value representing the estimated depth, and is compared to the actual depth stored in the annotation layer of the visual stimulus.

Listing 1: An example implementation of a correctness rule which calculates the given correctness for a depth judgment trial.

```
function depthEstimate(Trial $t, Judgment $j) {
  $depthLayer = $t->getAnnotationLayer('depth');
  $pos = $t->markers[0]->pos;
  $depth = $depthLayer->pixelAt($pos);
  $diff = $depth - $j->getFeedbackAsFloat();
  return 1 - abs($diff);
}
```

The correctness rules are programmed in PHP, and they can be created and edited using a code editor integrated in our web-based platform. Once created they can be run on the entire user feedback of an evaluation to process the response data.

Analysis and Reporting. Once all participants have completed the evaluation and the users feedback has been evaluated using the correctness rules, statistical analysis can be performed. Our system currently has support for repeated measurement ANOVA (rANOVA) and post-hoc analysis using Tukey’s tests of Honest Significant Difference (HSD). To perform the analysis, the researcher selects a factor through the labels keyword groups associated with the visual stimuli. When the factor has been selected the system builds a CSV file which will be passed as input to an R-script. The R-script is executed on the web-server and its output is parsed using PHP. The result is sent to the researchers browser where the results are presented using HTML tables and plots using D3 [4]. The built in analysis capability is currently limited, therefore we allow the researcher to export the data as a CSV file, which can be used in any statistical analysis tool to perform more in depth analysis.

Isenberg et al. emphasize, that many researchers fail when reporting the results of user evaluations, as the reports lack important methodological details [23]. As a similar discovery has been made by Ellis and Dix [12], we have decided to integrate reporting into our system. Currently, in addition to tables and plots mentioned above, we also allow for exporting various parts of the results as LaTeX. The exported LaTeX can be directly integrated into papers and other documents. The output consists of a textual reporting template and accompanying tables showing the distribution of achieved results. The LaTeX output is meant as a first draft which should be modified by the reporting researchers to match their needs. However, for demonstration purposes, we have included this output unmodified for the first application case (*Study 1: 2D Vector Fields Visualization*), which is further discussed in the appendix.

7 APPLICATION CASES

To demonstrate the capabilities of the presented system, we have used the system to replicate three studies which have been previ-

ously conducted and reported in the literature. The first study was originally conducted by Laidlaw et al. and compares different techniques for 2D vector field visualization [28]. The second study is evaluating the perceptual impact of seven volume illumination techniques, a study originally conducted by Lindemann and Ropinski [31]. While our system is targeting scientific visualization it is still possible to use it for studies in other fields, which we demonstrate by replicating the study originally conducted by Gleicher et al., which compares participants ability to estimate averages in various multiclass scatterplots [17].

In the following subsections we will discuss the results from our studies and compare them to the results reported in the original publications. For more details on study setup, participants and analysis we refer the reader to the appendix.

7.1 Study 1: 2D Vector Fields Visualization

This study is a replication of Laidlaw’s et al. study comparing techniques for visualization of two dimensional vector fields [28]. The study is divided into three parts. In the first part, classifying critical points, the rANOVA shows that there is a significant difference between the means. Enhanced LIC performed significantly better than all technique except against regular LIC. In the original study [28], Laidlaw’s et al. fail to find any significant results, except between their best technique, GSTR and worst technique, LIC. This may be due to the fact that their participant pool was much smaller than ours, (17 vs 73).

In the second part, locating critical points, the rANOVA shows that there is a significant difference between the means. Enhanced LIC outperformed all other techniques followed by LIC. This is consistent with the results from Laidlaw’s et al. original study [28] where LIC were better than GRID, JIT and LIT. It was expected that Enhanced LIC and LIC would perform well in this study since they are both dense visualization techniques, meaning they can depict the flow in every pixels, while the other techniques are sparse and values between glyphs has to be visually interpolated. In our study GRID and JIT does not perform significantly different, this is expected since they are quite similar, both uses arrows to show the direction of flow at a single point. Furthermore, LIT performed significantly worse than the other techniques, this is slightly different from Laidlaw’s et al. results, where they found no significant difference between LIT, GRID and JIT. This might be due to the parameters used when rendering our images. It was not completely clear what parameters was used for the original study, which may lead to slightly different results. Regarding measured time there were no significant difference between the various techniques.

For the third part of the study, tracing a particle, the rANOVA fails to show any significant difference in user performance. In Laidlaw’s et al. original study the analysis for this task shows two groupings, with OSTR and GSTR in one group performing significantly better than GRID, JIT, LIT and LIC. Since we only included GRID, JIT, LIT and LIC in our study our results agree with the results of Laidlaw’s et al.

7.2 Study 2: Volume Illumination

This study compares seven illumination techniques for volume rendering, originally published by Lindemann and Ropinski's [31]. The study consist of three parts, the first parts evaluates absolute depth perception. Both our study and the original study has significant strength according to the rANOVA, the distribution of the results are similar, with just a slight difference in the ordering. Their results showed that Half Angle Slicing and Shadow Volume Propagation where the top two techniques for absolute depth with 18.5% and 20.3% discrepancy, in our study they are on second and third place with 20.9% and 22.1% discrepancy and Directional Occlusion Shading performed best with 19.4% discrepancy. Similarly in the ordinal depth part we have similar results, our top three techniques was placed in their top four, after Directional Occlusion Shading. While our results is similar there are still some differences, this may be because of having too few validation trials, Lindemann and Ropinski's had one validation trials per task type which might not be enough when crowdsourcing is used for recruiting participants. For the beauty comparison study, the subjective preference on which method was the most beautiful we have very similar results. The only place were the results do not completely agree is place 3 an 4, were the order has been swapped.

7.3 Study 3: Multiclass Scatter Plots

This study is a replication of Gleicher et al. study evaluating perception of average values in multiclass scatter plots [17]. Our analysis of how the distance between the means affect the difficulty of the task shows similar finding as Gleicher et al., both with p-values very close to zero. In our study we had to discarded close to 50% of the participants due to the fact they failed 50% or more of the validation trials. Gleicher et al. also had to discard some users, though not as many. After discarding participants, Gleicher et al. recruited new participants, which made their participant count remain high. While they used the Amazon's Mechanical Turk platform to recruit participants, they did not have to pay the rejected users, we used the CrowdFlower platform, while they allow for rejecting participants, the cost of the rejected participants will still be charged and if new participants are to be recruited more money has to be spent. Furthermore, CrowdFlower uses various platforms on the web to distribute the tasks, while they assume all participant speaks English, it is a high probability the many participants of the participants does not have English as native language. CrowdFlower gives us a report on what country the participants who performed our study was located and only around 11% of the participants who completed the study were from a country were English is the native language. We believe that this might affect the ability to fully understand the instructions and therefore some participants may have completed the study incorrectly. We like to redo the study in the future, where we will limit the recruitment to include only countries where English is the native language.

8 LIMITATIONS

While the conducted studies described in the previous section demonstrate that the proposed system can be used to perform user evaluations quickly and easily, there are several limitations which we would like to address in the future.

Technical limitations. Currently our system interfaces the meta crowdsourcing platform CrowdFlower, which is used to submit studies. This has been done to enable an international usage, as Amazon's Mechanical Turk is only available in the US. However, this has the downsides that study prices increase, as CrowdFlower adds ~33% to the price for administration. Furthermore, rejecting a participant who fails the validation trials will not be refunded and an additional cost to recruit new participants will be added. Therefore, in the future we would like to integrate other platforms as well such as Amazon's Mechanical Turk, in order to reduce study costs. While the current system have only been used to do evaluation of static images, the availability of JavaScript and WebGL also supports the exposition to dynamic and possibly interactive stimuli. As we currently use JavaScript to overlay markers on the visual stimuli,

we are positive that we can exploit a similar proceeding to enable such trials. One other shortcoming regarding the correctness checking is, that each rule works only with the given trial and judgment, for example it is not possible in the current implementation to compare a judgment with aggregated data from other participants which could be useful to find patterns in the participants behavior.

Conceptual limitations. The system provides a set of tools which aids the developer in the process of creating and launching studies, analyzing study data and reporting results. We have developed these tools with a focus on the system being flexible and expendable. This comes at the cost, when a user with little or no knowledge about user evaluations tries to set up a new study, currently there is nothing to prevent this user from designing and launching a sub-standard study or misinterpret the statistical analysis. Though, we believe that when the system has launched and several studies have successfully been conducted, the design of the used questionnaires can be reused with new data which will hopefully decrease the risk of using fault questionnaires. Another issue where inexperienced users may do mistakes is when selecting values for the stimuli attachments. For example, when selecting locations for markers used in a depth comparison task it is important to select locations such that the difference in depth is not too small nor not too large in order to make a task with moderate difficulty. This is something we would like to explore in the future, by investigating methods to automatically generate marker locations. To overcome these conceptual limitations, more user guidance principles would be necessary, which are currently not supported by our system.

9 CONCLUSIONS AND FUTURE WORK

In this paper, we have described an interactive system, which enables the easy conduction and analysis of crowdsourced user evaluations in scientific visualization. Combining crowdsourcing techniques with an appropriate user interface design, enables the initiation of large scale studies with only a few mouse clicks, and support initial analysis of the study results. To our knowledge, this is the first system of its kind, which provides this functionality to support quantitative evaluations in scientific visualization. We believe that this has an impact in the visualization community, as it allows researchers for the first time to conduct quantitative evaluations with a minimal effort. However, we see the system's future benefits also beyond study administration, as combining conduction and analysis allows for adaptive studies and considering cross-experiment validation. This can lead to new knowledge and a reinvention of the visualization design process, as the feedback of a large set of participants can be directly considered during the design. While the system described in the paper is designed to utilize crowdsourcing for participant recruitment, there is nothing that prevents researchers to use the system without crowdsourcing, for example to perform the evaluation in a controlled environment.

In this paper, we have been focusing on the conduction and analysis of quantitative evaluation approaches. While these evaluations are considered an important tool in science, one evaluation type is never enough when investigating a visualization [5]. Therefore, by allowing to share and reuse previously used questionnaire designs, we see the described system as one step towards more reliable evaluation in scientific visualization. We would like to investigate in the future how other evaluation scenarios can benefit from a similar setup. In its current state, our system only enables the manual submission of trials, while in the future we would like to support automatic trial submission based on the results of previous trials. Thus, more dynamic experiments become possible, and the role of the crowd as a fitness function is emphasized. Finally, we see several opportunities for exploiting the data collected using our studies. As all data is collected in a centralized database, once a critical mass is reached, it will be possible to analyze the data and derive models for selected aspects of visualization design. We believe, that this will on the long run contribute to a better understanding of visualization design with respect to the perceptual and cognitive capabilities of the user. While these are all interesting research goals we would like to focus on in the future, our greatest

interest is to release the proposed system open source within the next few months to make it available to other researchers, and thus contribute to future evaluations in visualization.

REFERENCES

- [1] N. Ahmed, Z. Zheng, and K. Mueller. Human computation in visualization: Using purpose driven games for robust evaluation of visualization algorithms. *IEEE TVCG*, 18(12):2104–2113, 2012.
- [2] W. Aigner, S. Hoffmann, and A. Rind. EvalBench: A software library for visualization evaluation. *Computer Graphics Forum*, 32(3):41–50, 2013.
- [3] M. A. Borkin, A. A. Vo, Z. Bylinskii, P. Isola, S. Sunkavalli, A. Oliva, and H. Pfister. What makes a visualization memorable? *IEEE TVCG*, pages 2306–2315, 2013.
- [4] M. Bostock, V. Ogievetsky, and J. Heer. D³ data-driven documents. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2301–2309, 2011.
- [5] S. Carpendale. Evaluating information visualizations. In A. Kerren, J. T. Stasko, J.-D. Fekete, and C. North, editors, *Information Visualization*, pages 19–45. Springer, 2008.
- [6] M. Chen, D. Ebert, B. Fisher, R. S. Laramée, and T. Munzner. Evaluation: How much evaluation is enough? In *IEEE VisWeek Panels*, 2013.
- [7] M. Chen, K. Gaither, E. Groeller, P. Rheingans, and M. Ward. Quality of visualization: the bake off. In *IEEE VisWeek Panels*, 2012.
- [8] J. Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, 1988.
- [9] M. Correll and M. Gleicher. Error bars considered harmful: Exploring alternate encodings for mean and error. *Visualization and Computer Graphics, IEEE Transactions on*, 20(12):2142–2151, 2014.
- [10] C. D. Demiralp, M. S. Bernstein, and J. Heer. Learning perceptual kernels for visualization design. *Visualization and Computer Graphics, IEEE Transactions on*, 20(12):1933–1942, 2014.
- [11] D. L. Donoho. An invitation to reproducible computational research. *Biostatistics*, 11(3):385–388, 2010.
- [12] G. Ellis and A. Dix. An explorative analysis of user evaluation studies in information visualisation. In *Beyond Time and Errors - Novel Evaluation Methods for Visualization (BELIV)*, pages 1–7, 2006.
- [13] N. Elmqvist and J. S. Yi. Patterns for visualization evaluation. In *Beyond Time and Errors - Novel Evaluation Methods for Visualization (BELIV)*, 2012.
- [14] C. Forsell. A guide to scientific evaluation in information visualization. In *Conf. Information Visualisation (IV)*, pages 162–169, 2010.
- [15] C. Forsell and M. Cooper. Questionnaires for evaluation in information visualization. In *Beyond Time and Errors - Novel Evaluation Methods for Visualization (BELIV)*, 2012.
- [16] J. Fuchs, P. Isenberg, A. Bezerianos, F. Fischer, and E. Bertini. The influence of contour on similarity perception of star glyphs. *Visualization and Computer Graphics, IEEE Transactions on*, 20(12):2251–2260, 2014.
- [17] M. Gleicher, M. Correll, C. Nothelfer, and S. Franconeri. Perception of average value in multiclass scatterplots. *IEEE TVCG*, 19(12):2316–2325, 2013.
- [18] A. Grosset, M. Schott, G.-P. Bonneau, and C. D. Hansen. Evaluation of depth of field for depth perception in DVR. In *IEEE Pacific Visualization Symposium*, pages 81–88, 2013.
- [19] L. Harrison, F. Yang, S. Franconeri, and R. Chang. Ranking visualizations of correlation using weber’s law. *Visualization and Computer Graphics, IEEE Transactions on*, 20(12):1943–1952, 2014.
- [20] C. G. Healey. On the use of perceptual cues and data mining for effective visualization of scientific datasets. In *Graphics Interfaces (GI)*, pages 177–184, 1998.
- [21] J. Heer and M. Bostock. Crowdsourcing graphical perception: Using mechanical turk to assess visualization design. In *ACM Conf. Human Factors in Computing Systems (CHI)*, pages 203–212, 2010.
- [22] C. Heinzl, S. Bruckner, M. E. Grller, A. Pang, H.-C. Hege, K. Potter, R. Westermann, T. Pfaffelmoser, and T. Möller. Uncertainty and parameter space analysis in visualization. In *IEEE VisWeek Tutorials*, 2012.
- [23] T. Isenberg, P. Isenberg, J. Chen, M. Sedlmair, and T. Möller. A systematic review on the practice of evaluating visualization. *IEEE TVCG*, 19(12), 2013.
- [24] A. Kachkaev, J. Wood, and J. Dykes. Glyphs for exploring crowd-sourced subjective survey classification. In *Computer Graphics Forum*, volume 33, pages 311–320. Wiley Online Library, 2014.
- [25] A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with mechanical turk. In *ACM Conf. Human Factors in Computing Systems (CHI)*, pages 453–456, 2008.
- [26] M. Kleiner, D. Brainard, and D. Pelli. What’s new in Psychtoolbox-3? In *Perception ECVF Abstract Supplement*, 2007.
- [27] R. Kosara, C. G. Healey, V. Interrante, D. H. Laidlaw, and C. Ware. Thoughts on user studies: Why, how, and when. *IEEE Computer Graphics and Applications (CGA)*, 23(4):20–25, 2003.
- [28] D. H. Laidlaw, R. M. Kirby, C. D. Jackson, J. S. Davidson, T. S. Miller, M. Da Silva, W. H. Warren, and M. J. Tarr. Comparing 2d vector field visualization methods: A user study. *Visualization and Computer Graphics, IEEE Transactions on*, 11(1):59–70, 2005.
- [29] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale. Empirical studies in information visualization: Seven scenarios. *IEEE TVCG*, 18(9):1520–1536, 2012.
- [30] R. S. Laramée and R. Kosara. Challenges and unsolved problems. In *Human-Centered Visualization Environments*, pages 231–254. Springer, 2007.
- [31] F. Lindemann and T. Ropinski. About the influence of illumination models on image comprehension in direct volume rendering. *IEEE TVCG*, 17(12):1922–1931, 2011.
- [32] G. Little, L. B. Chilton, M. Goldman, and R. C. Miller. TurkIt: human computation algorithms on mechanical turk. In *ACM Symp. User interface software and technology (UIST)*, pages 57–66, 2010.
- [33] W. E. Mackay, C. Appert, M. Beaudouin-Lafon, O. Chapuis, Y. Du, J.-D. Fekete, and Y. Guiard. Touchstone: exploratory design of experiments. In *ACM Conf. Human Factors in Computing Systems (CHI)*, pages 1425–1434, 2007.
- [34] I. S. MacKenzie. *Human-computer interaction: An empirical research perspective*. Newnes, 2012.
- [35] S. Matht, D. Schreij, and J. Theeuwes. Opensesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, 44(2):314–324, 2012.
- [36] K. Moreland. A survey of visualization pipelines. *IEEE TVCG*, 19(3):367–378, 2013.
- [37] T. Munzner. A nested model for visualization design and validation. *IEEE TVCG*, 15(6):921–928, 2009.
- [38] M. Okoe and R. Jianu. Graphunit: Evaluating interactive graph visualizations using crowdsourcing. In *Computer Graphics Forum*, volume 34, pages 451–460. Wiley Online Library, 2015.
- [39] H. Pashler and E. Wagenmakers. Special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, 6(7):645–654, 2012.
- [40] C. Plaisant. The challenge of information visualization evaluation. In *Conf. Advanced Visual Interfaces (AVI)*, pages 109–116, 2004.
- [41] H. Purchase. Effective information visualisation: a study of graph drawing aesthetics and algorithms. *Interacting with Computers*, 13(2):147–162, 2000.
- [42] A. J. Quinn and B. B. Bederson. Human computation: A survey and taxonomy of a growing field. In *ACM Conf. Human Factors in Computing Systems (CHI)*, pages 1403–1412, 2011.
- [43] M. Sedlmair, M. Meyer, and T. Munzner. Design study methodology: Reflections from the trenches and the stacks. *IEEE TVCG*, 18(12):2431–2440, 2012.
- [44] B. Shneiderman and C. Plaisant. Strategies for evaluating information visualization tools: Multi-dimensional in-depth long-term case studies. In *Beyond Time and Errors - Novel Evaluation Methods for Visualization (BELIV)*, 2006.
- [45] V. Solteszova, C. Turkay, M. Price, and I. Viola. A perceptual-statistics shading model. *IEEE TVCG*, 18(12):2265–2274, 2012.
- [46] E. Sundén, P. Steneteg, S. Kottavel, D. Jönsson, R. Englund, M. Falk, and T. Ropinski. Inviwo - An Extensible, Multi-Purpose Visualization Framework. Poster at IEEE Vis, 2015.
- [47] J. Talbot, V. Setlur, and A. Anand. Four experiments on the perception of bar charts. *Visualization and Computer Graphics, IEEE Transactions on*, 20(12):2152–2160, 2014.
- [48] M.-C. Yuen, I. King, and K.-S. Leung. A survey of crowdsourcing systems. In *ASE Conf. Information Privacy, Security, Risk and Trust*, pages 766–773, 2011.