DOI:
10.1007/978-3-319-66435-4

Document Version
Peer reviewed version

Crowdsourcing for Information Visualization: Promises and Pitfalls

Rita Borgo[1], Bongshin Lee[2], Benjamin Bach[3], Sara Fabrikant[4], Radu Jianu[5], Andreas Kerren[6], Stephen Kobourov[7], Fintan McGee[8], Luana Micallef[9], Tatiana von Landesberger[10], Katrin Ballweg[11], Stephan Diehl[12], Paolo Simonetto[1], Michelle Zhou[12]

[1] Swansea University, UK
[2] Microsoft Research, US
[3] Microsoft Research - Inria Joint Centre, FR
[4] University of Zurich, CH
[5] Florida International University, US
[6] Linnaeus University, SE
[7] University of Arizona, US
[8] Luxembourg Institute of Science and Technology, LU
[9] Helsinki Institute for Information Technology, Aalto University, FI
[10] Darmstadt University, DE
[11] University Trier, DE
[12] Juji, US

1 Introduction

The term *crowdsourcing*, coined in 2006 [52], describes a new labor market phenomenon where simple, often monotonous labor tasks are replaced by open self-managed recruitment of large groups of people from the general public. Online platforms such as Amazon Mechanical Turk and CrowdFlower have stimulated this trend, and made crowdsourcing attractive for user studies in visualization and human-computer interaction. The visualization community increasingly employs crowdsourcing mechanisms for conducting empirical visualization research with the goal to increase access to and take advantage of large and diverse participant groups for evaluation.

Crowdsourcing has the potential to overcome the limitations of controlled lab studies, such as small participant sample sizes and participant pools with narrow demographic backgrounds. These limitations can lead to empirical results that might be difficult to generalize or have low ecological validity. Through crowdsourcing, a large number of participants with a broad background can be recruited more easily and rapidly, often at a much lower cost compared to traditional lab studies. Within the visualization community, van Ham and Rogowitz [46] first set the scene for the use of online evaluations in the context of graph-layout aesthetics, clearly separating their game-inspired online study from a traditional laboratory setup.

However, the studies employing crowdsourcing pose additional conceptual and methodological challenges for rigorous empirical visualization research. Known

challenges to crowdsourcing-based studies relate to, but are not limited to: reduced control in the assessment of participants' background and training, use of evaluation criteria that go beyond classic performance measures (e.g., task completion time and accuracy), and need of additional testing mechanisms for complex evaluation tasks that require increased cognitive efforts over a prolonged period of time. The benefit of larger numbers of participants is contrasted by limited participant sampling and selection mechanisms, based on demographics or backgrounds of the participants required for the study. A large, potentially diverse but anonymous, and remote pool of participants can have undesired impacts on the internal validity of the empirical study, and thus can limit the quality of study results. Moreover, crowdsourcing-based experiments typically do not allow for direct interactions between experimenters and participants, and do not permit systematic control of the testing environment.

In this chapter, we review research that has attempted to take advantage of crowdsourcing for empirical evaluations of visualizations. With an aim to identifying best practices and potential pitfalls to guide future designs of crowdsourcing-based studies for visualization, we discuss core aspects for successful employment of crowdsourcing in empirical studies for visualization –participants (Section 2), study design (Section 3), study procedure (Section 4), data (Section 5), tasks (Section 6), and metrics & measures (Section 7). We also present case studies, discussing potential mechanisms to overcome the common pitfalls (Section 8). This chapter will help the visualization community understand how to effectively and efficiently take advantage of the exciting potential crowdsourcing might offer to empirical visualization research.

2 Participants

Scaling to a large number of participants and increasing their diversity (e.g., age, cultural background, or expertise), is the main objective in using crowdsourcing techniques. Typical lab studies in information visualization (InfoVis) involve a small number of participants. A larger and more diverse pool of participants can potentially provide the following advantages:

- **Large samples:** In most cases, participant sample sizes can be increased by simply running more Human Intelligence Tasks (HITs) [13]. A larger number of participants, first of all, result in larger samples (e.g., 480 participants in [74], 550 in [32]). Having more samples makes the data analysis more robust to outliers, since outliers can be removed while maintaining a large number of "good" samples. Larger samples can also provide more evidence with respect to distribution and significance between conditions.
- **Easier and faster data collection:** The time and effort that are dedicated to participant supervision in traditional studies are virtually eliminated in

[13] We adopt this terminology, which means a single self-contained task, from Amazon Mechanical Turk.

crowdsourcing studies. Crowdsourcing platforms make it convenient to recruit people automatically, while tasks are solved without direct interaction with the study experimenter. Moreover, multiple participants can perform their tasks in parallel, further speeding up the data collection process [50].
- **Diverse samples:** Accessing a larger pool of potential participants allows to search for participants with specific characteristics such as age, gender, educational background, familiarity with the visualization methods, visual abilities, profession, etc. These diverse criteria can be used to provide valuable insight which would be nearly impossible to find with typical lab studies.

To make the best use of these advantages, the experimenter needs to take into account a number of factors when including participants through crowdsourcing:

- **Anonymity:** The true identity and motivation of participants is unknown to the study experimenter. Thus, the experimenter should assess the level of expertise with explicit tests, and cannot entire trust the demographic data entered by the participant into the online system.
- **Reliability:** Participants in lab studies are typically in a more direct connection with the study experimenter, leading to a reasonable expectation of dedication of the participants and truthfulness of the answers. On the other hand, crowd-workers engage with the tasks without supervision, and there is no direct communication between the experimenter and the crowd-worker. The experimenter cannot check if they are working on multiple tasks at the same time [41], and needs to put extra efforts to check if a crowd-worker is paying attention to the task.
- **Confidentiality:** In having participants executing the study on a remote machine (in most cases their own machine), the study experimenter implicitly makes the study code and data available. Some studies might rely on confidential data or code that should not be made widely available.

The remainder of this section discusses visualization-related issues about *(2.1) how potential participants can vary, (2.2) how to find participants with a desired skill set, and (2.3) how to train the remote crowd-worker.*

2.1 Demographics and Expertise

There have been efforts to "measure the crowd," i.e., to analyze the demographics, characteristics, and habits of crowdworkers. Unfortunately, these statistics are extremely volatile, deeply influenced by the crowdsourcing platforms' policies, and easily biased by the population sampling method. They should therefore be interpreted more as a snapshot of a particular crowdsourcing platform at the time of the survey, rather than as demographics of general validity.

In 2010, Ross *et al.* [85] presented a demographic description of the workers in Amazon Mechanical Turk (AMT) based on surveys conducted in 2008 and 2009. The article suggests that in earlier years the population was mostly American, engaging in AMT typically for fun or some extra income, and with a distribution

across sex, income, and age that was fairly representative of the U.S. population. Around the time of the survey, there was however a gradual shifting toward an Indian-based population, which presents a strong bias toward young male individuals with a higher reliance on the AMT income for their sustenance. In 2015, Silberman *et al.*[14] remarked that the demographic presented is outdated for a number of reasons, including changes in Amazon policies, and provided further evidence suggesting the presence of a sampling bias in the previous study.

Fort *et al.* [40] further analyzed the above data and presented more details on the task distribution. According to the authors, about 80% of the tasks in AMT were carried out by less than 10,000 Turkers, which represented roughly one percent of the registered crowd-workers at that time. Moreover, considerations on the average wage obtainable in AMT, combined with the reasons provided by the Turkers for working on the tasks, made the authors raise ethical issues on the usage of AMT, apparently shared by the legal departments of some universities.

Hirth *et al.* [51], instead, attempted to provide a more general characterization by studying a platform with no explicit demographic restrictions, called Microworkers. At the time of the study, the majority of the workers on this platform were from Asia, typically from low wage countries. Employers were instead more likely from a western country, with the U.S. representing more than a quarter of the total number of employers. The distribution of reward suggests a polarization similar to the finding of Fort *et al.* for AMT, with a small number of employers and workers covering the vast majority of the tasks available. Their results also indicate major differences in preferences among the workers regarding accepted tasks, with some high-performing workers systematically accepting faster, less paid jobs and others mostly going for longer, better paid ones.

Martin *et al.* [71] employed a more qualitative approach to the characterization of some highly active Turkers. They detail the living and working conditions on these people, including the reasons why they work on the AMT platform, how they select the HITs to work on, and the possible disagreements between workers and employers. The authors also consider ethical considerations and opportunities for designing a better working platform.

A constantly updated summary of AMT can be found online[15] (gender, income, marital status, household size, etc.). Next we consider topics more specifically related to crowdsourced visualization work.

Visualization Literacy is a relatively new term, defined by Boy *et al.* as *"the ability to confidently use a given data visualization to translate questions specified in the data domain into visual queries in the visual domain, as well as interpreting visual patterns in the visual domain as properties in the data domain"*[19]. Related concepts include *graphicacy* [98] as the ability to understand simple bar charts and diagrams, and *visual literacy* [26] as the ability to understand signs. Studies on perception of visual variables (e.g., [36]) and how people associate

[14] https://medium.com/@silberman/stop-citing-ross-et-al-2010-who-are-the-crowdworkers-b3b9b1e8d300

[15] http://demographics.mturk-tracker.com/

values to visual variables, provide some general understanding of what to "expect" from a normal participant. However, these studies do not tell how to asses a person's visual understanding, or the ability to build up a methodology to correctly understand and interpret the meaning behind a picture. Assessing visualization literacy of potential crowd-workers can help define the type of studies possible with crowd-workers, design training conditions, and improve the overall experimental design.

According to Bertin [17], there are three levels of understanding visualizations. To understand a visualization on an *elementary level* means to be able to extract basic information from the data, such as to find a maximal value. Understanding on an *intermediate level* means to be able to extract trends and other higher-level structures. Finally, understanding on a *comprehensive level* means to be able to compare structures and make interpretations that involve domain knowledge. Based on Bertin's observations, Boy *et al.* define a methodology to measure visualization literacy, which involves a) stimuli (pictures, tables, text, etc.), b) tasks (e.g., *find maximum*), and c) a textual formulation (called a "question"). For questions, Boy *et al.* define the characteristic of *congruency*: a question with high congruency uses words related to the graphical elements (e.g., *"what is the highest bar?"*), while a question with low congruency uses domain language (e.g., *"which country spends the most on health care?"*). Questions with low congruency are expected to be harder to answer. Boy *et al.* also formulate a set of guidelines to test visual literacy, which include careful design and repetition of conditions.

However, the data gathered from such a visualization literacy test is rather complex to analyze, and the proposed visualization literacy tests require about 30 minutes, making it difficult to employ such tests in crowdsourcing experiments. Expressive and short tests are still missing. Two simpler examples of tests for visual literacy exist online, which embed simple multiple-choice tests in an HTML frame.[16] While the first test assesses the understanding of bar charts, the second asks questions about which of two representations is more readable, and attempts to determine whether people can spot deceptive charts [79].

Yet, every evaluation may want to define their own criteria of what participants' pre-knowledge is expected to be with respect to visualization literacy. Questions related to training Turkers are discussed further in Section 2.3. Study authors may want to carefully check the language and explanations of their tasks. It cannot generally be assumed that the average worker is able to translate a question from the domain space (low congruency) into the visual space (high congruency). Perhaps even simple graphics may benefit from explanation and a clarification of terminology. Similar problems may arise when requiring participants to interact with a visualization. Section 3.3 proposes a possible strategy on how to maximize the outcome of visualization literacy assessments while amortizing costs.

[16] http://www.quizrevolution.com/act101820/mini/go/
http://perceptualedge.com/files/GraphDesignIQ.html

Cultural codes define conventions about decoding information that is stored graphically. Some of these conventions are explicitly defined. For example, the *direction of reading* is different in many cultures: right-left, left-right, top-down. This can influence the order in which visual elements in a visualization are decoded (e.g., the orientation of a time axis [72, 3, 13]). The *formatting of number and date* affects labels and questions (e.g., 1.000 and 22/02/2016 in Europe vs. 1,000 and 02/22/2016 in North America). *Units and measures* as well as their abbreviations change from country to country (e.g., *MO* (MegaOctet) in France, *MB* in other countries). Time units can also be a source of confusion; day times should probably be indicated in both 12 and 24-hours notation (14:00/2pm), and fuzzy terms such as "semester" and "biweekly" should probably be avoided.

Other conventions such as colors and symbols can vary between sub-groups and with contexts. *Colors* can have, in many cases, more generally agreed upon meaning with respect to their effect [1], but very different symbolic meanings. For example, the colors white and green are associated with nature and well-being in western cultures, but they can be associated with death in Asia and South America, respectively. When using colors in textual descriptions, there may be discrepancies about the colors associated with a term [101], though generally color categorizations are consistent across cultures [16]. Finally, there may exist conventions about colors in the context of InfoVis: including the rainbow color-scale that, wrongly though, implies an order of colors, or dual scales ranging from blue (low or negative) via white (middle or zero), to red (high or positive), or vice versa. Such conventions need special explanation.

Symbols are interpreted entirely by convention, according to the studies of semiotics; for example, in Poland a triangle indicates man's bathrooms while a circle indicates woman's. Simplified pictorial representations of an existing object (e.g., a man icon on bathroom doors) are termed *icons*. Icons are more universal than symbols, though they may still rely on cultural conventions (women wearing skirts, men trousers). Though the usage of symbols is generally discouraged in InfoVis, there may be intrinsic visual encodings, related to visualization literacy such as axis labels and scale tick-marks, and visual elements in visualizations (e.g., circles in node-link diagrams, contour lines in maps) are not generally self-explanatory per se, but learned by cultural convention.

Color blindness affects around 10% of the male population and 0.5% of female in the world [45], which amounts up to around 70 million affected people world-wide (these numbers are reported for 2016 and they vary across sources). *Color blindness* is a generic term that covers several types of color-perception deficiencies that involve almost every color hue [45], and appear on different scales (mild, moderate, high). As Ware notes [101], some people are not aware that they do not perceive color differences like the majority of the population.

The implications of color-blindness for crowdsourcing are three-fold: a) *Self assessment* for any type of color-blindness may be required to either categorize participants with different abilities or filter participants from the actual study; b) *Qualifying tests* may be required if color is an essential part of the evaluation

and adapted color schemes cannot be employed. An ad-hoc test could involve samples of actual study conditions from the experiment with an emphasis on color perception, or standardized tests and images (see [31] for a collection); and c) *Adapted color schemes* designed to work for most color-blind people [24, 63] can be employed to increase the soundness of a study.

Domain expertise of crowd-workers varies as they have different professions, each of which involves different activities and skills related to analysis, visualization, and domain knowledge. The specific domain expertise can influence a crowd-worker's interest in a task and the pre-knowledge he or she brings when decoding information (e.g., finance, biology, politics). More general analytical skills required in a certain field of daily work are related to visualization literacy (e.g., reading bar-charts, working with numbers and statistics). Both conditions may have an impact on tasks results and performances.

On the other hand, an evaluation of visualization may require participants with explicit knowledge in a certain domain. The problem at hand is to gain access to such experts. Domain experts may not participate in crowdsourcing platforms on their own and may not voluntarily spend time in evaluations. In working with domain experts, an appropriate compensation with respect to the expert's work or research may be a more promising approach than monetary rewards. For example, possible compensations of this type might include access to novel visualization tools, access to interesting datasets, etc.

2.2 Finding "The Right" Participants

As participants differ across a wide range of characteristics, a study author may want to find participants with certain characteristics, but exclude others from taking the HIT. As described in Section 2.1, visualization literacy tests are not yet generally applicable to crowdsourcing. Being aware of the problem should encourage study authors to include simple tests in their studies, and to focus on sufficient training. It is also important to deliver very precise task descriptions upfront, in order to discourage less motivated workers [36].

Some crowdsourcing platforms create participant profiles that allow a study author to directly contact participants after the HIT. This makes it possible to invite the participants for a post-study on the same topic (for example, when evaluating memorability), or to invite the participants to a new study that requires expertise and training obtained in earlier studies. However, platforms without such participant profiles make it difficult, if not impossible, to track workers who already have participated in a study.

2.3 Training

Most tasks in user studies require some sort of training a) to teach participants the goal of a task (*did the participants correctly understand the tasks and were they able to find the correct answers?*), b) to teach participants how to use

a specific visualization or interaction technique (*were the participants able to decode a visualization properly?, were the participants able to correctly interact?*), and c) to teach participants specific strategies on how to best solve a specific task (e.g., *first look at A, then adjust B, eventually interpret C*).

The main limitation with training in crowdsourcing is the quality assessment. In a lab study, the instructor can supervise the training, answer questions, and provide clarifications. Training represents a crucial aspects of any type of study design, we therefore address this issue in detail in Section 4.3.

3 Study Design

3.1 Types of Experiments and Associated Methodologies

The space of experimental designs for visualization studies can be described along multiple dimensions, several of which we describe here:

- **Study goal:** Studies may be employed to determine whether a visualization or visual technique is able to support the goals and tasks it was designed for (*usability*) and to quantify that ability (*quantification*), to understand how a visualization technique can support workflows in practice (*ethnographic*, to compare two visualization techniques in terms of their ability to support different tasks and workflows (*comparative*), and to understand and model mechanisms of human perception (*perceptual*).
- **Study target:** Studies may evaluate static visual encodings, non-interactive animations, visual encodings augmented by interaction, and visual analytic systems (i.e., multiple integrated and interactive visualizations).
- **Study duration:** Studies can be short or extended, and can be conducted over one or multiple experimental sessions. For example, perceptual studies often involve very short tasks [50], while studies that measure participants' ability to memorize visual information for an extended period of time may involve multiple sessions conducted several days apart [88].
- **Type of participants:** Studies may involve naive participants, or participants with a particular expertise or ability. Similarly, they may target either broad populations or populations with specific attributes (e.g., cultural background, visual impairment). A detailed discussion is provided in Section 2.
- **Type of methods and constraints:** Different types of studies typically pose unique challenges in the context of crowdsourcing. For example, ethnographic studies rely on participant observation in their environment, and thus need to capture context-data that may be difficult to acquire by a remote experimenter. Quantitative studies need to isolate the evaluated perceptual or data-reading tasks from other, non-related activities and processes. This can be difficult in crowdsourced environments, as unmonitored participants may engage in activities that experimenters are unaware of, and network, device, and browser variability can translate into recorded performance measures and significantly impact the study's outcome (also see Section 7). It is also only recently that complex interactive visualizations and visual analytics

systems can be distributed online, making them amenable to crowdsourcing. New research is required to understand the impact of such crowdsourcing particularities on different types of user studies, to create evaluation methods that can isolate the evaluated effects from the evaluation process, and to implement the tools to allow experimenters to conduct a wide range of study types with minimal overhead.

Lam et al. [58] provide a more comprehensive discussion on visualization evaluation, and we detail four examples of crowdsourcing-based studies in Section 8. Ideally, crowdsourcing technologies should eventually support the design and deployment of studies spanning this space with minimal overhead on the experimenter.

3.2 Study Design Considerations in Crowdsourced Environments

Study design. Visualization studies are typically designed as between-subjects, within-subjects, or a mixture of the two [84]. Traditionally, researchers gave preference to mixed or within-subjects designs as they were more robust to differences between individuals, and more amenable to the smaller number of participants that lab studies could attract. However, two characteristics of crowdsourcing lead an increasing number of online studies to recently opt for between-subjects designs [105, 5, 55]. First, unlike lab studies, crowdsourcing gives experimenters access to participant samples considered sufficiently large to offset participants' individual differences (Section 2). Second, between-subject studies are shorter, often significantly so, than within-subject ones, and thus fit better with the micro-task paradigm specific to crowdsourcing. Finally, as we will show in Section 3.3, because studies employing a between-subjects design are easily extendable (e.g., with new conditions, with additional tasks), they provide unique opportunities for incremental online experimentation.

Study duration. In line with the micro-task philosophy underlying crowdsourcing, online studies should be kept relatively short. This can be achieved in several ways. First, as mentioned above, between-subjects designs are shorter than within-subject designs. Depending on the study's goals, the evaluation work can be divided across multiple participants by one or a combination of its independent variables (e.g., by visualizations, by datasets, by tasks or groups of tasks). Second, piloting can more reliably inform the choice of reasonable time-limits for tasks, leading to shorter studies with less variance in duration. Finally, participant testing and training, typical components of a visualization study (Section 2), can significantly increase the duration of a study. As discussed in Section 3.3, allowing participants to save and reuse their demographic information, perceptual markers, and expertise information across multiple studies could significantly shorten the study duration.

Introductions and task descriptions. Introductions are perceived as overhead. Long, text-heavy, ambiguous study descriptions frustrate participants. Ex-

perimenters should use few words, avoid jargon, and exemplify encodings, interactions, and tasks using clear visual diagrams. Self-explanatory training sessions and task designs that can be picked up without excessive guidance are particularly effective in shortening introductions.

Study interfaces. Learning and interacting with the interface guiding participants through the study and collecting answers can introduce overhead. Experimenters should minimize this learning overhead by implementing GUI standards and affordances, and building on participants' pre-existing mental models to create study interfaces that can be learned and used without considerable effort. As a community, experimenters should strive to reuse and share study interfaces across their experiments to reduce the learning strain on participants.

Participant engagement. Unlike participants in lab studies, who typically are invited and participate in a limited number of studies, online participants often sift through many posted tasks before they choose one to participate in. While an important consideration in that choice is the amount of compensation, online participants also factor into a study's appeal and fun factor, intellectual reward, and significance of the study's expected results. In fact, there are online communities (e.g., Reddit) who participate in research studies voluntarily and whose members choose studies to participate in solely based on significance and appeal. Moreover, online participants often rate and discuss studies in online forums, building a collective memory and opinion about each study.

Studies that participants can link to their personal experiences can be more engaging. For example, finding paths in an abstract graph visualization is less likely engaging than finding the friends that connect two people in a social network. Micallef *et al.* [74] report on participants commenting on their interest and engagement in the study, and on things they learned while participating. Section 5 provides a few suggestions on how this could be achieved.

Experimenters can also consider using gamification (e.g., FoldIt [33]) and use one or both of two approaches. First, evaluated tasks could be gamified: participants would solve game-like tasks that are designed to translate or hide a meaningful research question. This is difficult to implement in practice as finding designs that hide meaningful research questions in appealing game-like setups can be challenging, and new creative effort would be necessary for each new studied task or measure. Alternatively, participants' performance on regular, un-gamified tasks could be used in a gaming scheme to motivate and engage participants. For example, based on their participation and performance on user studies, participants could earn points, reach and pass levels, or compete against each other. Since this approach is independent from the particularities of evaluated tasks, it could be integrated into reusable interfaces and platforms that service many diverse studies.

Malicious behavior. Workers may not take tasks seriously. Gadiraju *et al.* [41] define five categories of malicious behavior which all apply to evaluation in in-

formation visualization. *Ineligible workers* provide wrong pre-conditions about tasks, visualizations, domains, or other skills. *Fast deceivers* give random answers in order to finish a HIT as fast as possible, e.g., randomly selecting visual elements, or entering random numerical values. *Rule breakers* do not provide the required quality of the answer, e.g., giving 1 keyword, when the task requires at least 3 keywords, or by drawing a circle (or a cat) where a more complex drawing may be expected [99]. *Smart deceivers* conform to the rules but give semantically wrong answers. Finally, *Gold standard preys* can only be caught with repeated test questions during the evaluation.

Gadiraju *et al.* also provide a measure for the maliciousness of a worker and could report that several workers become malicious *during* the study. Fast deceivers can partially be excluded automatically by looking for consistently wrong or invalid answers. Detecting less salient malicious behavior can happen during training (Section 2.3) and by repeated tests for attention (Gold standard test) throughout the study. However, as Gadiraju *et al.* note, those techniques alone are not sufficient and suggest the need to carefully design the tasks to minimize the extent of cheating. Corresponding design guidelines can be found in the paper [41]. However, many tasks in information visualization are very open-ended and hence provide plenty of opportunities for malicious behavior and drawings of cats (Section 6).

3.3 Reusable Designs and Results

Controlled experiments often follow standardized procedures and materials. It is, for instance, typical to use entrance and exit questionnaires, to test participants' visual and cognitive abilities and to train them, to control for display or input factors, and to record performance data. In crowdsourced experiments, setting up each of these components involves web-development and requires programming expertise, can consume significant time, and is susceptible to implementation bugs. As such, reusable study components could be assembled into configurable frameworks, purposefully designed to support the crowdsourced evaluation of interactive visualizations in a plug-and-evaluate manner. Developers could connect interactive visualizations to evaluation engines, and specify tasks to be evaluated on those visualizations, data that should be collected, and the number and profiles of participants to be recruited. Creating studies interactively, by assembling existing building block components and workflows, would reduce the overhead of creating online content programmatically.

Such frameworks already exist to support the creation of computer based lab experiments (e.g., Touchstone [64], EvalBench [2]) and for very specific research domains (HVTE [9]). They have also started to emerge for web-based studies. For example, online interactive forms gained considerable popularity and have enabled a wide range of studies by simplifying the process of fielding a questionnaire and collecting data. Lightweight frameworks provide infrastructure for data collection (e.g., Experimentr [48]) or result visualization (e.g., VEEVVIE [80]). Much closer to our envisioned workflow is GraphUnit [77], a system which allows

even interactive web-content to be connected and evaluated online with minimal overhead. To fully realize the objective described here, additional work is necessary.

Shareable, reusable, and extendable online designs. Let us consider the following scenario. A visualization researcher or developer creates a new visualization design and evaluates it against its matching state of the art in a controlled experiment. The creators of a third design should be able to reuse as much as possible of this experiment's materials to compare their own solution to the previous two. Moreover, if the initial experiment was conducted using a between-subject methodology, and the second experiment can leverage the same or a similar crowd, then the possibility of simply extending the previous study with an additional condition, corresponding to the latest design, would be ideal.

Similar scenarios include extending the range of tasks evaluated by existing studies, increasing their sample size or diversity, or replicating the studies with modified conditions. By and large, supporting such workflows would allow researchers to incrementally build on top of their own and their colleagues' findings in an unprecedented way.

Two significant technological advancements are necessary to lead towards this goal. First, storing studies online, both in terms of their designs and in terms of their data, in public or shareable repositories, would provide direct access for researchers and developers interested in understanding the design and results of studies or in replicating and extending them. Second, a standardization of the technologies and procedures used to create and deploy online user studies would allow a more seamless integration of new conditions or study components into existing ones.

Reusable participant profiles and qualifications. When evaluating visual and interactive content, it is often imperative to test participants on their perceptual, motor, and cognitive abilities (e.g., testing for color-blindness) and on their general visual literacy, and train them to understand or use specific visual encodings or interactive visualizations. But, it can also be prohibitively time consuming. An ability to allow habitual study participants to create profiles in which to input demographic information and store results and certifications of their testing and training, and an ability to allow them to reuse this data in subsequent studies, would allow researchers to capture more data and make better use of their participants' time. Some existing crowdsourcing platforms (e.g., Amazon Mechanical Turk) provide such features, but additional support needs to be researched and implemented to support the evaluation of visual and interactive content.

4 Study Procedure

Following the principles for standard laboratory experiments, the study procedure in a crowdsourced context involves four stages: experiment setup, pre-

experiment activities, experiment activities, and post-experiment activities. Similar to brick-and-mortar experiments (e.g., [70]), the study procedure in crowdsourced settings should be carefully planned and systematically executed. The study procedure follows directly from a concrete research question, and is the result of the operationalization of an experimental study design. Experimental procedures also need to be adapted to the selected crowdsourcing platform (i.e., technical requirements, invitation and task assignment of registered crowd workers, selection of desired participant sample, etc.). This needs to be carefully tested before the actual study is executed. For this reason, pilot experiments are especially critical in crowdsourced contexts, so as to achieve high internal validity of the study, despite of the limited experimental control compared to traditional lab studies. Information visualization empirical studies are characterized by tasks relying on both perceptual and cognitive abilities of participants. The nature of such tasks demand care in validation of aspects that might hinder the soundness of the collected results. These aspect include not only design but also study deployment (i.e. platform type vs. architecture used) and participant selection (i.e. spatial and visual abilities).

4.1 Experiment Setup

The limitation of experimental control in online studies that are executed without the presence of an experimenter can of course affect participants' responses irrespective of their actual ability, and thus influence the quality of collected performance data. As mentioned in Section 2 one of the main differences with online studies in general, and crowdsourced experiments in particular, is that the experimenter cannot ensure that the intended experiment procedure is followed, and thus is identical for each participant, as intended for a laboratory study. The still open research question for crowdsourced studies is thus how procedural control can be achieved in crowdsourced studies. For example, one solution could be the development of detailed standardized instructions, and the inclusion of automatic setup and procedure checks for crowdsourced experiments.

The experimental design, experiment setup, and its deployment on a crowdsourcing platform should be equally well documented, as is the norm for traditional experiments, as to ensure transparency, and reproducibility for given crowdsourcing platforms. For instance, items to report relate to full disclosure of specification details of the computing environment, such as the type and location of the server used, the type of crowdsourcing platform, and any technical details of the apparatus used to run the study. This might include the specifications of the employed video camera, eye-movement, mouse tracking or other equipment, or any other remote participant behavior tracking technology. Another important procedural control includes the recording and reporting of how and when micro-tasks were uploaded on the crowdsourcing platform, and to whom. Procedural information to report would have to further include whether offered micro-tasks were presented in batch mode or in a particular sequence, at which exact date and time of the day, and whether participants had to satisfy certain prerequisites for participation (e.g., response quality record, geolocation, cultural

background, specific work environment, language, etc.). The exact duration of the micro-task, the specific mechanism adopted to engage with participants before the study, or motivate participants to stay focused on the tasks during the study, and what type of reward was offered to them, also needs to be reported.

4.2 Pre-experiment Activities

Similar to laboratory experiments, pilot experiments should be conducted with the target population of crowd workers. These participants need to be recruited in identical ways as for the main experiment, including the same reward type. This is especially important for crowdsourced experiments, as the availability of crowd workers is more volatile, and the crowd workers backgrounds are more diverse. More importantly, crowd workers' motivations for participating in an online experiment are likely to be different from those in typical laboratory experiments, for example, carried out with students at universities. Professional crowd workers might engage in a crowdsourced experiment as part of their job, and thus might wish to finish as many micro-tasks as quickly as possible, even in parallel, so as to increase their income. Study participants recruited for laboratory studies typically do not depend on participation rewards as their sole source of income. For many traditional experiments, especially those carried out at universities, participation is either required for degree completion (i.e., psychology), or for small rewards such as course credits, or some such. Moreover, participants in controlled experiment settings are closely monitored to stay focused on one experiment task at a time.

Collected data from pilot experiments should be analyzed as thoroughly as for laboratory experiments, as to ensure that the planned procedure is appropriate for targeted participants, task formulations are comprehensible, enough time is allocated for study completion, and that the reward for study participation is fair. Additionally, for pilot studies in crowdsourced contexts it is particularly important to ensure that:

- task instructions are clear and understood for the diverse set of online participants;
- participant attention checks are robust, to ensure that participants stay focused on tasks;
- apparatus checks are robust, to ensure the experiment setup works as planned on the crowdsourcing platform, and crowd workers' devices;
- participants are able to run the study apparatus as intended and instructed;
- online micro-tasks work as expected on different display types and web interfaces;
- anticipated target group is reached, i.e., language, geo-location, and other sorts of study requirements are met.

As for controlled laboratory studies, full disclosure of pilot study details and respective sample analysis is necessary in study publications and reports, including when, how and with whom pilots were conducted, the reward offered

and given, and why and how the experiment procedures were modified due to pilot experiments.

4.3 Experiment Activities

As the experimenter or study supervisor is not physically present during a crowdsourced study, the experiment introduction and respective instructions need to be carefully designed, complete, and unambiguous for the diverse set of potential crowd workers. Compared to laboratory studies, the following expectations need particular consideration and communication to participants, and respective mechanisms for removal of participants when study expectations are not met:

- screening for repeated study participation by crowd workers;
- information about attentional demands (i.e., lighting conditions, noise levels, interruptions, etc.);
- required skills and abilities (i.e., language, expertise);
- technology configuration requirements (e.g., speed of CPU, plugins, browser type and versions, screen size, resolution, and color depth);
- anticipated response time limits (i.e., entire study, sections, and micro tasks);
- expected reward structure including minimal response standards;
- consent for participation in the study.

The crowdsourced study should always include warm-up trials and/or a training session with analysis of the response quality before the actual experiment can be run. Training could be complemented with tests that have to be passed before the actual data is recorded. Training could then, theoretically, be repeated until a certain test is passed, provided that proper feedback is given to the participant, explaining mistakes and pointing out how to arrive at the correct answer. Eventually, more training could be provided on demand. This can be useful to assure that participants understand the expected type of questions and experiment tasks, and that the expected experiment procedures stated in the recruitment phase of the study are met (e.g., check of display type, device type, browser configuration, etc.) and thus are identical across participants and repeatable for future studies. However, long training leads to participant fatigue and crowd-workers may complain and discredit the campaign among their peers.

The response procedures should be well explained and amply practiced before the actual experiment, i.e., whether the response type is active (i.e., participants need to complete a task and the answer is displayed later) or passive (i.e., questions and answers are provided jointly). It should also be communicated ahead of time and documented whether participants are allowed to revise answers by going backwards in the study, skip trials, or whether and when they are allowed to take a break.

The experiment trials portion in a crowdsourced study basically follows the standards for laboratory experiments. However, participants' response behaviors must also be carefully monitored and compared to the planned procedures. Hence, a full account of what happened, when, how, and by whom needs to be

documented automatically, and digitally recorded such as, an anonymized identifier for the participant, the number, order, and type of trials; how, when, and where the response was recorded; and possibly any other user interaction logs with the system during the entire experiment (i.e., whether a participant revised answers of previous trials, or moved on to the next trials without completing a prior trial, idle times, etc.), as to be able to trace what exactly happened during the experiment. In crowdsourcing studies, participants' task-relevant attention needs to be monitored remotely. Procedures can include forced breaks and distractor tasks to monitor participant attentional demands throughout the study.

4.4 Post-experiment Activities

As with traditional experiments, post-test questionnaires might include a series of recruitment checks or tests of control variables, such as the assessment of individual differences (e.g., spatial abilities, numeracy abilities, visual literacy, color-blindness, etc.), group differences (i.e., gender, age group, expertise levels, etc.), and/or any other user background or demographic assessments and self reports. Also, experiment-related questionnaires for study monitoring purposes might be very useful (e.g., whether participants used additional tools beyond instruction to solve a task; whether participants were confident in their answers; and/or self-reports on strategies used to complete the task). Other aspects such as debriefings, thank you, and free-form comments, simply follow traditional experiment procedures, but need to be built-in in the online experiment.

In lab experiments, participants are compensated at the start of the study to ensure that they could stop anytime they want. However, in crowdsourced studies, the participants are typically compensated after completion of the experiment, as a means to assure response quality.

Processing and Filtering of Collected Data A special aspect of crowdsourcing experiments is to systematically validate the collected data, as to assure that anticipated procedures based on a specific experimental design were indeed followed. For that, the quality of the response data needs to be carefully assessed before it can be statistically analyzed. Data from inattentive participants, participants who did not follow the stated instructions, did not meet experiment requirements such as a specific language or similar need be removed from the analysis, and possibly replaced. Such participants could be identified through mechanisms proposed and discussed in Section 2. Data from participants who did not complete the entire experiment as stated (i.e., repeating participants, participants who took long breaks between sections, or similar) need be removed too. In specific task types, such as with image tagging, outlier analysis could be performed on the response data, and responses that are beyond 2-3 standard errors above and below the response mean for the sample could be removed. Besides filtering the response data, one could also perform other kinds of validation assessments such as, response error pattern analysis, response time pattern

analysis, etc. Any post-test filtering or data validation analysis need to be additionally reported together with the rationale for adopting such approaches, and a description of the final data used for the actual statistical analysis.

Participant Compensation and Experiment Completion Researchers wishing to run crowdsourced studies especially with well-established crowdsourcing platforms should strive for a high online reputation with crowdsourced workers, such that their profile with the workers is enhanced, so as to attract reliable crowd workers. A good reputation can be built first and foremost by being honest with study participants, and by compensating them rapidly after completion of the experiment with the promised award or bonus, stated in the experiment instructions. Experimenters need to clearly indicate in the experiment instructions what the expectations for compensations are (i.e., following task instructions, satisfaction of study prerequisites and requirements, etc.). In a crowdsourced setting, where experiment participation might be considered as "employment for a micro task," one might debate about the ethical basis for compensation conditions based on task quality. This is because in regular employment settings, once a person is employed for a given job, the prior agreed pay might not be as easily revoked due to low quality of delivered work. It would simply be at the employer's discretion not to employ that person again for future tasks.

Conversely, crowd workers also have an incentive to keep up their reputation on crowdsourcing platforms with micro-task providers. Various crowdsourcing platforms provide assessment measures of a crowd worker's reliability, e.g., based on the percentage of the completed tasks that were approved by a task provider, including any comments or feedback on the quality of the performed tasks by task providers. Crowd workers' reliability data are then perused by other task providers to decide upon selection of study participants. This also means that task providers should carefully consider quality of their assessments of crowd workers, as this could have a great impact on crowd workers' profiles, and thus might in extreme cases lead to the blocking of a crowed worker's account.

A simple strategy for researchers to avoid the collection of poor response data for future crowdsourced studies is to keep a log of all the participants that did not complete the task appropriately, or not with the desired focus of attention, and thus disallowing these crowd workers to participate in future crowdsourced experiments. It is generally good practice to keep a log of the study participant IDs in case these crowd workers need to be contacted again for follow-up questionnaires or tasks, or who might actually be interested in receiving the publication of study results. A log of IDs might also be useful for cross-checks to exclude crowd workers from participation in studies that are too similar, as to avoid potential learning or knowledge transfer effects.

Once the study participants are compensated and the necessary information is logged, the micro-task should be removed from the crowdsourcing platform.

5 Study Data

Study data forms an important part of the study design as the tasks are performed on the visualized data. Data specifics substantially influence the visualization and interaction techniques used as well as the tasks to be performed. In addition, the meaning and size of the data can influence the incentives for task completion. Thus, an appropriate choice of data, with respect to study tasks and research questions, is crucial to the success of the entire crowdsourcing experiment and of gaining new findings.

Selecting suitable data for crowdsourcing experiments is a challenging core step in the experiment design. For a specific data type, the study designers need to consider several factors when choosing suitable datasets. For instance, they need to decide upon usage of real or controlled data; they also need to consider data suitability for the crowdsourcing studies. This may include data size, data confidentiality, or privacy issues. Moreover, they should take into consideration data attractiveness which influences the participant's engagement and willingness to conduct the study properly.

In the following, we discuss main factors influencing data selection for crowdsourcing studies noting that many factors influencing data choice in general apply to crowdsourcing data and there are some specific factors that play role solely in crowdsourcing studies. We summarize all factors (both general and specific).

5.1 Data Source: Real versus Controlled Data

The dataset should suit the goal of the study and the tasks to be performed. Depending on the study goal, the designers may decide among the following main data sources: *real-world data*, *controlled data*.

Real Data Real world datasets are gained from domain-specific applications. Therefore, this type of data reflects real user problems. Their closeness to real-world situations may raise the attractiveness of the data for the participants (see below for details). At the same time, real world datasets may be very domain specific. They may require domain expertise, too. This may reduce the suitability of real data for crowdsourcing.

The datasets offer interpretability and thus also are appropriate for testing insight-focused tasks. Often, real data does not include a "ground truth" and hence is not usable for crowdsourced perception studies. Moreover, real datasets are often very limited with respect to variability. Often, only one dataset of a kind is available. This limits the tasks and designs to open-end questions.

Real datasets may be difficult to obtain and to use in crowdsourcing studies. Frequently, they also have confidentiality and privacy constraints making them unusable in crowdsourcing studies, where the participants can freely access and possibly also share the data without access control. Real datasets are often very large and complex. The data size may increase loading times and hardware requirements. This may be problematic in crowdsourcing studies, where the participants have only limited internet access or only simple hardware available (e.g.,

crowdsourcing participants in India). Data complexity and size may also lead to long task completion times, thus distracting and frustrating many crowdsourcing participants (see Section 5.2 for a more detailed discussion of such issues).

Controlled Data Controlled datasets differ from real datasets in one main feature: they have specific "controlled" properties and often provide a variability of these features. These are often more suitable for crowdsourcing studies (e.g., can be created such that they are small and simple and non-confidential). However, there are only limited ways of obtaining controlled data, such that they are suitable for the study design at hand. In the following, we present the advantages and disadvantages of controlled data for crowdsourcing studies. We also provide several pointers to sources of these datasets. Here, we focus on three types of controlled data sources: benchmarks, synthetic data creation, and curated real data.

- **Benchmark datasets:** Benchmark data repositories offer public datasets that have specific properties. They are often used in both laboratory and crowdsourcing studies, thus they support comparability across studies.

  Main advantages of benchmark datasets are their public availability and re-usability in research. In contrast to real-world datasets, they have well-known properties that can be tested for task accuracy and completion time. Many benchmark datasets have small sizes and often have real-world interpretation. This may make them favorable for crowdsourcing studies.

  The main drawbacks of benchmark datasets are their limited number and often specialized focus, of which the UC Irvine Machine Learning Repository[17] is an example. Nevertheless, data from benchmarks are often used in various studies. This brings along an additional problem for crowdsourcing: datasets reuse may pitentially lead to repeated participation. The participation in various studies using the same dataset may lead to learning effects and thus skew the collected results. As a general issue to bear in mind for both crowdsourcing and laboratory studies is that benchmark datasets have a limited set of specific properties. While they are suitable for comparability across approaches using standardized tasks, they may not be suitable for novel tasks or for testing novel visualizations (esp. visualizations of complex data types). Such datasets may not be available in benchmarks, or may be very difficult to find. For example, the analysis of dynamic geo-located networks requires specific properties, while many benchmark network datasets are static or do not have geo-location at all.
- **Synthetic data creation:** As an alternative, the study designers can develop proprietary datasets specifically for the study at hand. A clear advantage would be that the individual creation of datasets can consider all requirements of the crowdsourcing study. This, however, requires the careful consideration of all criteria including study tasks, dataset specifics, possible

[17] http://archive.ics.uci.edu/ml/

target participants, attractiveness, as well as statistically-significant variability (see sections on Tasks, Design, Metrics and Requirements).

Creating such datasets manually can be cumbersome and time consuming. In many cases, study designers can use automatic or visual-interactive data generation tools. For example, the PCDC System [23], SketchPadN-D [100] or the system developed by Albuquerque et al. [4] allow for visual-interactive creation of multivariate data with specified properties. Random data generators, such as graph generators, can automatically create data with special properties (e.g., [25, 106, 7, 6]). This can be joined with visual-interactive means, for instance, Bach et al. [14] developed an evolutionary graph generation algorithm. Another data type – geographic data can be generated using spatio-temporal patterns [95, 90–92].

– **Curating real datasets:** Pre-processing real data for crowdsourcing experiments can bring advantages of both real and synthetic datasets. The resulting datasets are close to reality and, at the same time, have properties needed in a specific crowdsourcing study. For example, the study designers may select a suitably small subset that can be tackled also in crowdsourcing study on small screens or with slow internet connection. Moreover, the study designers may encode ground truth into it, which is then suitable for measuring accuracy with large number of participants. This is an advantage for crowdsourcing studies, where other assessment methods such as think aloud protocols are not feasible. However, the data curation process can be tedious. In addition to the above-mentioned requirements, in order to be able to real datasets in crowdsourcing studies, often also data anonymization is needed. Anonymization needs to ensure that the study participants cannot reveal private or confidential information in the original data. This may be difficult to ensure. But, visualization may help to check anonymity and privacy issues in the data. For example, anonymity in multivariate data can be analyzed with the tool by Dagsputa [35], and spatio-temporal data privacy issues can be revealed by data mining and visualization approaches [44, 75, 10].

5.2 Data Specifics

In this subsection, we highlight specific issues and characteristics of the data used in crowdsourcing experiments. Our assumption is that the data specifics have a great influence on the design and results of the planned experiments. Thus, the dataset should be carefully chosen. For example, the larger and more complex the data, the less likely it is to be suitable for crowdsourcing due to the fact that the participants need to invest considerably more time and effort to understand the data itself. In addition, domain-specific knowledge often plays an important role when using complex datasets, for instance, consider data coming from biochemistry (e.g., biological networks with experimental data attached to the network elements [57]). We briefly describe the most important data specifics in the following.

Data type and complexity. In context of information visualization, people usually differentiate between several data types: univariate data (1D), bivariate data (2D), trivariate data (3D), multidimensional or multivariate data (nD), temporal data, tree or hierarchical data, and network or graph data [56]. The data values themselves can be classified according to diverse scales: nominal, ordinal, and quantitative. The different data types lead to various data complexities, e.g., univariate data is surely easier to understand and visualize compared to network data. When using real datasets (see the previous Section 5.1), their structure is mostly more demanding because those data is often a mixture between the above mentioned data types. All these properties have a great influence on which visualization (visual encoding) and interaction technique should be chosen. They also have an effect on the tasks that the participants have to cope with (cf. the next Section 6).

Data size. So-called *data scalability*, i.e., the capability of a visualization to handle an increasing amount of data is a well-known challenge in information visualization [59]. This applies also to crowdsourcing experiments. On the one hand side, the chosen visual encoding and visualization in general must be able to efficiently deal with a large dataset. On the other hand, large datasets may be a problem for the crowdsourcing infrastructure and the technical equipment of the participants as they may be difficult for small screens, slow internet connections, or small computing power.

Data familiarity. Depending on the tasks to solve in an experiment, familiarity with the data is of crucial importance. Participants should normally know the data domain (there might be study designs where this is not the case), i.e., they should be able to understand and interpret it. If that is the case, we can also assume that we have selected the "right" people for the experiment who have a relationship to the data. But, there are even more perhaps unexpected effects that the data may have. For instance, if the data contains information which may be problematic for participants due to cultural differences and similar reasons.

Another factor not to be ignored is the language used in textual or audio-visual data sources, such as extracted text parts from newspapers. Designers of crowdsourcing experiments should either take care to carefully select participants who are able to understand such data or translate the data.

Data attractiveness. Generally, it is believed that suitable data can improve participant's motivation and engagement in studies. Data attractiveness can be raised by familiarity and by including a "fun" or "game" factor in the data and the tasks. For example, a task of finding a shortest path in an abstract graph may be less engaging, than finding a shortest way to a home of a friend. Another factor that can improve attractiveness is reward from solving a task, especially, educational reward. When the participants see that they also learn by solving the task with special data, this can improve their motivation to participate in

a study. A great challenge is to provide attractive datasets and tasks. This may involve long data curation or synthetic data creation processes. Yet, whether a dataset is attractive depends on a particular participant. So, a right match of participants and the data is crucial.

Data confidentiality. A special case of a data characteristics are privacy and confidentiality issues, for instance, when real data from medical records are used. Generally, confidential data and/or private data are not suitable for crowdsourcing, because we cannot protect them. A natural way of dealing with private or confidential data is anonymization. However, a full anonymization is a difficult challenge. Therefore, crowdsourcing platforms should also provide additional technical support for dealing with confidential data. For example, they should hinder data download and its subsequent distribution. Moreover, they should enable access only to selected participants.

6 Study Tasks

Many taxonomies have been suggested to organize tasks performed by participants, working with visualizations (e.g., [93, 8, 61, 96, 21]). The purpose of these taxonomies is to support visualization experts in creating the visualization design and to support the evaluation of visualizations.

The purpose of this section is to help researchers determine *if* their tasks are suited to a crowdsourcing-based evaluation, and *how* they may instead construct their evaluation tasks. Note that this does not mean that every task *can* be made suitable to crowd-workers. Some tasks may be simply too difficult, even in lab studies. We provide a list of considerations to help determine whether or not a task is suitable for a crowd sourcing based evaluation. This is not a new taxonomy, but rather a new dimension of categorization to be considered in addition to those offered by existing taxonomies.

6.1 Tasks in existing studies

In this section, we briefly describe some of the visualization tasks that have been successfully used in a crowdsourcing-based evaluation. The type of evaluation that can be performed using crowdsourcing is usually referred to as a participant performance evaluation. Lam et al. [58] identify two question types for these studies:

- *What are the limits of human perception for a technique?*
- *How does one technique compare to another, in terms of human performance?*

A perfect example of the first question is provided by Harrison *et al.* [49]. The authors use a Mechanical Turk study to determine the perception of correlation in commonly used visualizations. They use a staircase methodology to infer

the Just Noticeable Differences (JNDs) for perception of correlation in each visualization type. For each trial, the participants were shown two visualization of the same type with different datasets and asked which one was most correlated. As part of the staircase methodology, the data displayed for a new trial depends on the result of the previous trial. If a trial is answered correctly the next trial is more difficult, if it is answered incorrectly the next trial is easier.

Jianu *et al.* [55] perform the type of study suggested in the the second question in their study on displaying community information on node link diagrams. The authors performed ten different experiments each with a different task. Their tasks are inspired by the graph task taxonomy of Lee *et al.* [61] which is in turn inspired by the information visualization task taxonomy of Amar *et al.* [8]. Neither of these taxonomies has any consideration about the impact of using crowdsourcing for an evaluation.

The two questions are not mutually exclusive. Heer and Bostock [50], in their pioneering mechanical turk studies (also discussed in Section 8.1) perform perceptual experiments, replicating earlier studies. They applied both types of the above questions in their study, quantifying perceptual distortion of area estimates, and providing information about which area representation was superior in terms of accuracy of human perception.

6.2 A crowdsourcing dimension for task taxonomies

Task complexity and task effort Similar to Bertin [17], several other taxonomies distinguish between simpler and more complex tasks. For example, Amar et al. [8] name *low-level* and *high-level*, where low-level tasks being smaller units related to unique actions in analytic activity: *Retrieve Value*, *Filter*, *Compute Derived Value*, *Find Extremum*, *Sort*, *Determine Range*, *Characterize Distribution*, *Find Anomalies*, *Cluster*, and *Correlate*. These low level tasks are very concrete and cover a wide range of tasks which people try to solve with information visualizations. They focus on identifying specific entities or finding clear correlations. High level tasks, on the other hand, are more general and may involve complex decision making, uncertainty, identifying trends and outliers, and domain knowledge.

While it may be tempting to use the notion of task complexity as a category for determining suitability for a crowdsourcing evaluation, there are many different interpretations of complexity. The notion of complexity can refer to the perceptual complexity of the task and visualization itself, or the cognitive complexity of the task. These are very much participant based considerations, which may have a different impact on different participants in an experiment. Therefore, rather than task complexity, we suggest task effort as a consideration for crowdsourcing. Effort can be used to not just characterize the task, but also how that task is performed in a crowdsourcing evaluation.

Consider the task of path-tracing as an example. Path tracing is a frequently used task for graph evaluation [83, 102, 73, 55], and would be considered a connectivity task in the graph task taxonomy of Lee at al. [61]. If the task is to determine if the shortest path between two nodes is 1,2, or 3 hops between a

pair of nodes, as done in [73], the participant may find the shortest path, but have to continue searching to verify that it is indeed the shortest. This may lead to a longer experiment time and frustration if they have to spend a long time verifying that an initial answer was correct. The approach to path tracing taken by Jianu et al. [55] provides the participant with a series of node titles and asks if these titles form a path. This format of the question allows the participant to quickly see if the path is invalid, and does not result them in searching for potential alternatives. In a crowdsourced approach, which is usually a between subjects evaluation, this allows for quicker answers and more trials.

There may be cases where a researcher desires the participant to search many possible alternatives, but in evaluations where this is not a goal of the task, the shorter validation approach, in which the participant has to determine whether or not the given information is true, is a more desirable approach. In summary, to reduce task effort for crowd workers, and avoid fatigue and distraction, low-level tasks may be preferred. Studies that involve higher-level tasks may benefit from a reduced number of trials, careful explanations, training (Section 2.3) and additional motivation (Section 2.5).

Task expertise. Related to task complexity and visualization literacy is the level of expertise required to perform a task. However, complexity and expertise are not mutual. For example, sometimes complex tasks can be explained in a simple way, by breaking down the description into low-level and high-congruent [19] (Section 2.1) task and formulation. Often this happens by explaining a specific strategy to the participant, such as *"To compare the two datasets, you could first look at X, then filter Y until you find diverging values, and finally report how often you found different values"*. However, other high-level tasks are much harder to break down in to low-level tasks and the study instructor cannot or does not want to reveal specific strategies.

As discussed in Section 2.2 finding the right participants can be a challenge in performing a crowdsourced based evaluation. Even if a pre-qualified set of participants is available care must be taken in the selection and definition of tasks. In use cases where specific tasks are used to determine if participants are qualified, as discussed in Section 2.2, care must be taken to ensure the qualification task guarantees the correct minimum level of expertise. Eventually, study authors may want to carefully train participants to perform specific tasks or instead report on the different strategies participants invent and apply.

Technical Task Feasibility. The heterogeneous nature of computer hardware and software means that researchers cannot assume that all experiment participants will have a similar environment to perform the experiment tasks. This may affect the task performance and may influence the results. Especially, screen size, input devices and calibration, and hardware performance influence task performance. For example, perception or interaction studies may result in different accuracy depending on the used screen size and hardware. Other technical issues related to display capability can also be a factor. An experiment that involves

human perception of color requires careful consideration of the fact that different display devices have different color gamuts. Depending on the level of accuracy required it may be possible to calibrate response based on some initial questions, asked for this purpose.

In addition, internet connection speed, affecting page loading times, can interfere with reported measurement times. For some experiments these delays may not interfere with results; however for others the reporting functionality of a crowdsourcing platform may not offer enough accuracy. Heer and Bostock [50] recommend that if researchers require fine grained timing, that they use their own technical implementation of a task interface which participants can access through mechanical turk.

7 Study Measures and Metrics

In visualization, quantitative evaluation usually entails measuring the participants performance of tasks in terms of accuracy (how many tasks were solved correctly) and time (how long did it take to complete the tasks). More recently, there has been a concerted effort to take into account aspects beyond time and error. For example, the BELIV workshop series is a well-known venue created to encourage the study of novel evaluation methods, such as memorability of visualizations, memorability of the underlying data, subjective preferences, engagement and enjoyment. Visualization researchers have also started to include psycho-physiological and neuro-biological measures to study the effectiveness and efficiency in their visualization evaluations. Measures include eye tracking, galvanic skin response measures (GSR) [67], and Electroencephalogram (EEG) recordings [66]. Video cameras help further assess the process with which participants arrive at a certain response. Facial expressions can reveal the emotional state of the participants, and these can be further analyzed using standardized questionnaires [65].

In the context of crowdsourcing, most types of measurements pose interesting challenges, mostly due to the lack of experimental control. For example, response time can be affected by participants' use of different hardware configurations (e.g., desktop computer, laptop, or mobile device). When measuring long-term memorability (e. g., days after initial visualization interaction), ensuring that the same participants are again available for a second assessment can be difficult. Similarly, measuring enjoyment and engagement by observing behavior, or via think-aloud protocols, poses additional challenges in a crowdsourced setting.

Self-reporting methods are possible alternatives and there is good evidence that people are capable of giving numerical or graphical indication of their emotions [78]. Similarly, interaction logging and basic eye-tracking (e. g., via laptop cameras) might be possible.

Attention to the correctness of the experimental procedure in a crowdsourced context, as described in section refsec:procedure, is crucial especially when attempting non conventional types of measurements and novel metrics as described in section 7.4.

7.1 Methodological Background

Borrowing from usability studies in human-computer interaction (HCI) research, visualization designers typically employ one or a combination of two evaluation approaches, broadly categorized in formative and summative evaluation methods. Formative evaluation approaches involve human participants early in the design cycle and are often of a more qualitative or quantitative, but subjective, nature such as Likert-style self-reports, preference ratings, and response questionnaires. Summative evaluation methods include more typically controlled, lab-based methods, borrowed from empirical research in psychology such as response time (e. g., efficiency) and response accuracy (e. g., effectiveness). Formative and summative methods can help guide what decision has been made with a visualization, and then validate the effectiveness of resulting visualizations.

The evaluation approaches, measures and metrics above are similar to those used for crowdsourcing studies, yet there are certain differences; we next address the topic of measures and metrics in the crowdsourcing context.

7.2 Measure Types

As mentioned above, there are two major types of measures: quantitative (e.g., time and error) and qualitative (e.g., think-aloud protocols, self-reports, focus group discussions, interviews, observations). Each respective measure type has its own advantages and disadvantages in a typical lab setting. In the context of crowdsourced visualization studies, there are additional considerations due to the greater lack of experimental control.

Quantitative measurements consist of counts, frequencies, rates, and percentages that document the actual existence or absence of occurrences and participants' behavior, beliefs, preferences, or attitudes. These methods are considered objective, although they require standardization in order to fit answers into a response scale and/or a number of predetermined response categories. Examples for such standardized measures are standardized questionnaires, psychophysiological measures, or success and error rate. Quantitative methods are often used to evaluate new visualization methods in lab studies, as well as in crowdsourced evaluations, as they are typically easy to administer, may include many questions, may yield a large amount of clearly structured responses that can be easily summarized and statistically evaluated. But clearly, the actual choice of a particular quantitative measure also includes a subjective component and is dependent on the expertise of the experimenter. Furthermore, sometimes difficulties arise, for instance, if the "correct response" cannot sufficiently be specified [27, 103].

Qualitative measurements consist of descriptions or lists of recorded visualization use events, unstructured text from questionnaires, interviews, or transcribed focus group discussions, video and audio tapes, or observed behaviors.

Qualitative measures can provide rich information about thought processes, as well as opinions, experiences, feelings, and attitudes. Qualitative methods can be very valuable for understanding how and why visualizations are used in realistic and meaningful contexts [82]. Qualitative methods, such as those that apply grounded theory, can provide a useful and holistic analysis of visual analytics applications [53]. Since qualitative data are typically collected through direct observation, interviews, and talk-aloud protocols, they are well-suited to lab studies. However, qualitative measures have certain drawbacks in lab studies. Beyond the difficulty of systematic evaluation, due to the individuality inherent in such data, there is the time consuming nature of qualitative questions for the participants. In addition to the already mentioned challenges, qualitative measures pose further significant challenges in crowdsourced evaluations. For example, in crowdsourced settings it is more difficult to follow talk-aloud protocols, to collect audio and video of the experiment, or to conduct 1-on-1 interviews with the participants.

In a nutshell, qualitative methods can provide deeper insights and can help clarify quantitative data by providing missing explanatory details and semantic nuances which are not inherent in quantitative data [28]. This, however, makes the systematic evaluation of qualitative data distinctly harder. Vice versa, it is easier to evaluate quantitative data systematically. The combination of quantitative and qualitative data regarding study participants can help to tackle the qualitative systematization issues and give the quantitative data an enriched context. However, since even in controlled lab studies, qualitative data is considered to be more subjective and may be difficult to summarize and compare systematically. The challenges increase in the crowdsourced environment, where the choice of qualitative measures that are easily deployed and analyzed is limited to response questionnaries, while quantitative, yet subjective, measures allow for the use of Likert-style self reports and preference ratings.

7.3 Standard Measures

The most commonly collected performance data in visualization evaluations is task performance data (efficiency (e. g., response time), effectiveness (e. g., response accuracy)), measured in terms of time to complete the tasks and errors made. Most often the time to complete a task is measured directly in seconds or minutes. Alternatively, tasks can be given a fixed amount of time and then analyzed for completion within the given time limits (e. g., count of the completed tasks, percentage of completed tasks, ratios of success to failure). Errors can similarly be measured via counts of (in)correctly completed tasks, percentage of (in)correctly completed tasks, and ratios of success to failure [84].

7.4 Measures Beyond the Standard

While controlled lab studies using standard evaluation measures are typical in InfoVis, over the last decade there has been the desire to design and implement

new methods of evaluation, from longitudinal field studies, insight based evaluation and other metrics adapted to the perceptual aspects of visualization as well as the exploratory nature of discovery. This desire is embodied in the BELIV workshop, which began in 2006, and which aims to collect and discuss innovative ideas about InfoVis evaluation methods, including new ways of conducting user studies, definition and assessment of InfoVis effectiveness through the formal characterization of perceptual and cognitive tasks and insights, and definition of quality criteria and metrics. Several of the proposed measures can be applied in crowdsourced settings.

Recently there has been an increased interest in measuring recognizability and memorability. A number of studies investigate the effect of embellishments on visualization memorability and comprehension. Bateman et al. [15] conducted a study to test the comprehension and recall of charts using an embellished version and a plain version. Bateman's study has been somewhat controversial, and Li et al. [62] recently reported a replication, limiting their selection to those charts that consisted of datasets with 10 or more observations. They found that the presence of a time limit affected comprehension and short-term recall performance, while the type of chart significantly affected short-term recall. Borgo et al. [18] showed that visual embellishment improves information retention in terms of both accuracy of and time required for memory recall. Since their focus was on "visual perception and cognitive speed-focused tasks" that leverage cognitive abilities, they used analytical tasks, where they enforced attention to switch from one task to another. Another study by Vande Moere et al. [97] showed that visual metaphors do not have a significant impact on perception and comprehension. Short term recall can be measured just as well in crowdsourced studies as in lab studies.

Ghani and Elmqvist [42] studied the effect of visual landmarking in node-link diagrams and found that landmarking is generally promising for *graph revisitation*, i.e., the "task of remembering where nodes in the graph are and how they can be reached." Marriott et al. [69] investigated the cognitive impact of various layout features, such as symmetry and alignment, on the recall of graphs. They asked participants to look at drawings and redraw them. Perceptual characteristics and memorability in dynamic graphs have also been studied [11, 12, 39, 43]. As a part of an experiment measuring the effectiveness of four visualizations (BubbleSets, Node-link, LineSets, GMap) Jianu et al. [54] asked participants to perform 10 different tasks, including one task related to the memorability of the data. Graph revisitation tasks and simple memorability tests can be performed in crowdsourced settings, although drawing tasks will likely be much more difficult. Saket et al. [89] present evidence that different visual designs can significantly impact the recall accuracy of the data being visualized, specifically, comparing *node-link* visualizations to *map-based* visualizations. This was measured by asking participants to perform certain tasks with both types of visualizations and later on asking them again to perform a subset of the tasks without the visualization. This type of data recall experiment can be performed in crowdsourced settings.

Other aspects, such as enjoyment and engagement, are not as well explored, even though enjoyment is often given as a reason to consume visualizations [22]. Enjoyment has been carefully studied in psychology. One of the most well-known models for understanding and measuring enjoyment in psychology is the flow model of Csikszentmihalyi [34]. Elmqvist et al. [37] define fluid interaction in the context of information visualization. In a recent study, Haroz et al. [47] assessed user engagement with ISOTYPES by measuring the total amount of time participants spent looking at different visualizations. Boy et al. [20] investigated the effects of initial narrative visualization techniques and storytelling on user engagement by examining interaction logs (e. g., amount of time spent on exploration, number of meaningful interactions). Recently, Mahyar et al. [68], Tanahashi et al. [94], and Saket et al. [86] proposed models of enjoyment in visualization. In particular, Saket et al. considered different elements of flow (challenge, focus, clarity, feedback, control, immersion) and argued that these elements correspond to specific levels of Munzner's nested model [76]. Later Saket et al. [87] used the flow-based evaluation in a study of the enjoyment of two different visualization methods of the same relational data: node-link and node-link-group visualizations. The results indicated that the participants in this study found node-link-group visualizations more enjoyable than node-link visualizations. Measuring time spent looking at different visualizations might be difficult in crowdsourced setting when there is a financial incentive to complete the job as fast as possible. However, in crowdsourced setting it should be possible to measure flow elements via Likert-style self reports, preference ratings, and response questionnaires.

Alternative methods for measuring enjoyment and engagement in visualizations have also been considered. Cernea et al. [29, 30] employed a mobile electroencephalographic headset for detecting emotional responses, when working with a visualization [65]. Peck et al. [81] argue that functional, near-infrared spectroscopy is a viable technology for understanding the effect of visual design on a person's cognition processes. Fabrikant et al. [38, 67, 66, 65] measured the emotional responses of participants in a cartographic experiment about interactions with maps, using sensors that monitor psycho-physiological responses and eye movement data. Novel approaches to include eye tracking methodologies [60, 104] also in crowdsourcing contexts provide interesting future possibilities in the assessment toolbox of the empirical visualization researcher.

7.5 Challenges for Study Measures and Metrics in Crowdsourcing Studies

There are increased difficulties in performing both quantitative and qualitative evaluations via crowdsourcing. From differences in hardware (desktop, laptop, mobile device) to differences in viewing capabilities (screen size and resolution), and the availability of camera and microphone, such variations can dramatically affect most measurements. Variations in the crowdsourcing platform (which place different restrictions on the experimenter and the participants), as well as environmental conditions (e. g., light conditions, distractions such as noise level,

help from another person), that cannot be controlled, pose additional challenges. Consequently, the validity of such experiments and the associated experimental conclusions can be widely open to debate and challenges.

Various strategies to address some of these issues can be employed. If a minimum hardware standard is needed, qualifications tasks can be used to select participants with devices that meet the standard (e. g., spoken responses to establish access to microphone, video responses to establish access to camera, etc.). The standard timing of tasks, which can be affected by many factors (e.g., device type, internet connection quality, screen size, etc.) can be replaced by timed tasks, where each task has a time limit (and if the answer is not given within that time limit the result is recorded as incorrect). Crowdsourcing platforms that allow the experimenter to identify and contact the participants can be used for evaluations that require repeated sessions (e.g., memorability).

8 Case Studies

This section discusses four case studies that demonstrate diverse ways how crowdsourcing can be used for visualization research. Not only can crowdsourcing be used to perform simple visual perceptual micro-tasks as described in case study 1, but it can also be used to understand user's complex visual comprehension of composite visualizations as summarized in case study 2. While the first two studies demonstrate the use of crowdsourcing for evaluating and comparing static visualizations, the third case study shows how different user individual traits can be assessed and included in the visualization study, and the last case study indicates the use of crowdsourcing for interactive visualizations, including data collection for informing the design of visualization techniques and algorithms. Below, we summarize the four case studies by their participants, procedures, data, tasks, and measures. For each case study, we also discuss their take-away points and limitations.

8.1 Case Study 1: Assessing Graphical Perception

Jeffrey Heer and Michael Bostock (2010). *Crowdsourcing graphical perception: using mechanical turk to assess visualization design.* In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 203-212). ACM.

- **Crowdsourcing Usage**: Used Amazon Mechanical Turk to replicate previous laboratory studies in spatial coding and luminance contrast and then compare the results of the two.
- **Design**: The study design follows the design of previous laboratory studies where a user is asked to accomplish various visual perception tasks, ranging from ranking visual variables by their effectiveness for conveying quantitative values to judging how a chart size may affect visual comparison accuracy.
- **Participants**: For task 1's sub-task 1, there are 70 trials and 50 turkers per trial; a total of 3481 responses are received and each trial is paid $0.05. For

task 1's sub-task 2, there are 108 trials and 24 turkers per trial; each trial is paid $0.02 (10 second per trial). For task 2, there are 60 trials and 24 turkers per trial; each trial is paid $0.02 per trial. Task 3 includes 48 trials and 24 turkers per trial; each trial is paid $0.04.
- **Procedure**: For each task, a turker first performs a qualifying task and then the perceptual task.
- **Data**: The data used to create the visualization used in the experiments are gathered from the previous laboratory studies.
- **Tasks**: The study includes three main tasks. The first task is to replicate Cleveland and McGill's studies on spatial coding. The task includes two sub-tasks: Proportional Judgment and Rectangular Area Judgment. The second task is on another perceptual task: separation and layering via luminance contrast. It replicates an alpha contrast experiment by Stone and Bartram. The third task is on the effects of chart size and gridline spacing on the accuracy of visual comparison.
- **Measures and Metrics**: The study collects the turkers' judgment from a set of visual perception tasks and compares the crowdsourced judgment with that obtained from previous laboratory studies. Depending on the task, the metrics/measure are different. For example, for the alpha contrast task, the alpha value, time to completion, and the turker's screen resolution, color depth, and browser type are recorded.
- **Take-Away Points**: (1) Since this is an early crowdsourced study for visual perception tasks, it demonstrates the viability of such studies, since the study successfully replicated prior experiments in three visual perceptual tasks. (2) The study also demonstrates the use of crowd to gain new insights into visualization design. (3) It also characterizes the use of Mechanical Turk for conducting web-based experiments. (4) It shows certain advantages of using crowdsourced studies over lab studies, including its low cost, speed, as well as participant diversity.
- **Limitations**: The main limitations lie in the type of visual perceptual tasks being investigated. When such tasks become more complex and require more visual literacy, it is unknown how the crowd would perform.

8.2 Case Study 2: Understanding Users' Comprehension and Preferences for Complex Information Visualization

Huahai Yang, Yunyao Li, and Michelle X. Zhou (2014). *Understand Users' Comprehension and Preferences for Composing Information Visualization.* ACM Transactions on Computer-Human Interaction (TOCHI), 21(1), 6.

- **Crowdsourcing Usage**: Used Amazon Mechanical Turk to crowdsource participants' insights and preferences for using complex information visualization to accomplish real-world visual analytic tasks.
- **Design**: The paper presents two crowd-sourced between participant-design studies. The first study aims at soliciting the participant's comprehension of a typical information visualization by asking the participant to articulate

the insights s/he has derived from the visualization given a specific, realistic analytic task with real datasets. This study contains a total 10 sets of visualization, each of which contains three visualizations, two simple visualizations that present the same dataset in different ways and one composite visualization that is supposed to provide additional insight compared to the two simple ones. The second study aims at soliciting the user's preferences when using a composite information visualization to accomplish an analytic task. This study consists of 8 groups of visualizations, each of which includes 5 different composite visualizations of a dataset. And each participant is asked to assess the 5 composite visualizations by accomplishing an analytic task, as well as rank his/her preference amongst the 5 composite designs.

- **Participants**: In the first study, 50 turkers were recruited for each of 10 sets of visualization; a total of 524 responses were received and each response was paid $1.50 (about 20-minute per response). In the second study, 30 turkers were recruited for each of eight groups of visualization; a total of 240 responses received.
- **Procedure**: In the first study, each turker is asked to articulate the insights that they derive from each visualization in free text. In the second study, each turker is asked to rank his/her preferences for each composite visualization that s/he uses to accomplish an analytic task.
- **Data**: Six real-world datasets from SPSS associated with real-world analysis tasks were used in both studies.
- **Tasks**: In both studies, turkers are asked to perform visual analytic tasks by deriving certain types of insights from a given visualization. The turkers were also asked to describe any derived insights in free text.
- **Measures and Metrics**: Both studies collected rich data, ranging from free text to ranked user preferences. A set of measures and metrics is also derived from extensive data analysis. In study 1, from user-articulated visual insights in free text, a taxonomy of user-perceived visual insights is derived. A set of metrics is also derived to measure the taxonomy, including the quality of insights (accuracy + depth), easiness of comprehension, usefulness of insights, and distribution of insights. In Study 2, user preferences of composite visualization are derived from the collected data.
- **Take-Away Points**: (a) This is an early study that crowdsources users' complex, high-level comprehension of information visualization beyond simple visual perception experiments. It thus provides methods to systematically instrument such crowdsourced studies for complex visual cognitive tasks and rigorously analyzes the quality and reliability of free-text-based crowdsourced results beyond structured, multi-choice survey answers. (b) The study also presents a systematic content analysis method for other researchers to harvest insights from such crowdsourced rich content in free text. The collected raw text as well as the derived visual insight taxonomy establishes the connections between one's verbal expressions and information visualization, which lays a foundation to develop more advanced in-

formation visualization systems, e.g., natural-language-based visualization retrieval and generation.
– **Limitations**: The main limitations lie in the type of visualizations and analytic insights being investigated. When interactive visualization is involved, deriving insights from such an interactive visualization may introduce unknown challenges (e.g., a wide diversity of actions amongst participants) that this study has not addressed.

8.3 Case Study 3: Analyzing Deceptive Visualizations

Anshul Vikram Pandey, Katharina Rall, Margaret L. Satterthwaite, Oded Nov, and Enrico Bertini (2015). *How deceptive are deceptive visualizations?: An empirical analysis of common distortion techniques.* In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 1469-1478). ACM.

– **Crowdsourcing Usage:** Used Amazon Mechanical Turk (AMT) to study deceptive visualizations, assessing in particular: (a) the deceptiveness of different distortion techniques in visualization; (b) the type of questions for which such visualizations are mostly deceptive; (c) the effect of users' various individual traits on the deceptive effect.
– **Design:** Four within-group experiments were conducted to assess four different types of deception caused by different distortion techniques ("truncated axis", "aspect ratio", "area", "inverted axis") and evaluate the deception effect on the users' responses. The deception type was the independent variable, while the user response was the dependent variable.
– **Participants:** Recruited 330 unique AMT workers who reported to be located in the Unites States and who had a task approval rate of at least 99%. Participants were paid $0.30 for a 5- to 10-minute experiment.
– **Procedure:** An experiment website was hosted on a server external to AMT, and a link to the webpage was provided in AMT's task description. The experiment stages, shown as different webpage pages, included: (a) consent form; (b) personal information form; (c) chart familiarity test; (d) visual abilities test; (e) deception test, including a chart overview, the chart, the deception test question and an attention check question; (f) need for cognition scale.
– **Data:** The context and the axes of the charts were made up for the study but non-abstract. Example of a chart title, 'Access to safe drinking water by minority ethnic group over time'. The type of data used is not clearly explained in the paper.
– **Tasks:** Two types of tasks in accordance with the type of deception: (a) "how much" questions ("how much better is A compared to B"), when the visualization message is exaggerated or understated; (b) "what" questions ("what does chart A show?") with multiple-choice answers, when the visualization message is reversed.
– **Measures and Metrics:** (a) user response, including response accuracy (percentage of correct answers) and mean user response; (b) measures of the

deceptive effect occurring when the visualization message is exaggerated or understated (results of the correctly and incorrectly represented charts were compared in a between-participants analysis) and when the message is reversed (response accuracy was compared in a between-participants analysis); (c) measures of individual traits, including their familiarity with basic charts, their visual literacy, their need for cognition, age, gender and education, used to regulate the user response.
- **Take-Away Points:** (a) To our knowledge, this is the only crowdsourced visualization study which takes into account the effect of various user individual traits on the collected responses. As shown in previous laboratory user studies (e.g., [107], user individual traits influence the effectiveness of visualizations, yet such traits are often not tested in crowdsourced experience due to the need to keep online experiments short. (b) This study indicated that good quality results could be achieved by employing attentive check questions, which are then used to filter out the data before analysis, and by testing various individual traits such as visual literacy which can then be used to regulate the user response accordingly.
- **Limitations:** The data collected for the user individual traits did not provide any statistically significant results, possibly indicating that these trait tests should be redesigned and adapted for crowdsourced experiments.

8.4 Case Study 4: Identifying Graph Layout Aesthetics

Steve Kieffer, Tim Dwyer, Kim Marriott, and Michael Wybrow (2016). *Hola: Human-like orthogonal network layout.* IEEE Transactions on Visualization and Computer Graphics, 22(1), 349-358.

- **Crowdsourcing Usage:** Built an online system (using HTML5/Javascript) named Orthowontist to conduct two studies: (a) collect data about the aesthetic criteria a graph layout algorithm should optimize to ensure the generation of human-readable network layouts with a comparable quality to manual layouts produced by hand; (b) evaluate the effectiveness of the layouts generated by the proposed automatic orthogonal network layout algorithm, HOLA, that took into account the aesthetic criteria collected in the first phase.
- **Design:** Both studies adopted a within-group design.
- **Participants:** Both studies were advertised on a university-wide bulletin. For study 1, part 1 had 17 participants who could have won one of three $50 gift cards if their layouts were ranked high in part 2, and part 2 had 66 participants who could have won a $50 gift card only if their answers were the closest to the aggregated answers of other participants. For study 2, part 1 was completed by 89 participants, part 2 by 84, and up to parts 3 and 4 by 83.
- **Procedure:** The overall procedure for both studies involved: (a) consent form and instructions; (b) questionnaire about their experience in using node-link diagrams; (c) technical training on how to use Orthowontist along with training tasks; (d) the study tasks; and (e) comments about the study.

- **Data:** Study 1 used small random abstract graphs with an incomprehensible layout for part 1, and the layouts used in part 1 together with the participants' improved layouts in part 1 for part 2. Study 2 used graphs with diverse number of nodes and edges. A few graphs depicted real-data (e.g., Sydney's metro map, the Glycolysis-Gluconeogenesis pathway), others were random graphs. None of the nodes and edges were labeled and no context was provided for any of the graphs.
- **Tasks:** In study 1, the participants were asked to (1a) manually edit the layout of graphs to make them more human-readable, and (1b) choose the best layout with respect to their aesthetic preference. In study 2, the participants were asked to (2a) rank graph layouts based on their aesthetic preference, (2b) find the shortest path between two nodes in a graph, (2c) identify all neighbouring nodes of a highlighted node by clicking on the nodes, and (2d) choose the best of two layouts for the same graph and explain why.
- **Measures and Metrics:** For both studies, user preference, response accuracy and response time were recorded. For some tasks, the participants were asked to explain in writing their response.
- **Take-Away Points:** (a) Crowdsourcing is not only useful to evaluate visualization, but also to collect data to inform the design of novel visualization algorithms. (b) It is possible to use crowdsourcing for interactive tasks, such as manually editing the layout of graphs (e.g., moving nodes or edges, adding or deleting edges) or interacting (e.g., clicking) with parts of the visualization (e.g., the nodes of a graph) to provide an answer. (c) It is also possible to log complex interactions when crowdsourcing user studies.
- **Limitations:** It is unclear whether the study included participant attentive checks and how the experimenter ensured the data reliability.

8.5 Summary

As described above, the four case studies have used crowdsourcing for different aspects of visualization research. Case study 1 demonstrates that crowdsourcing can be used to replicate previous laboratory studies on understanding people's visual perceptions. Moreover, a crowdsourced approach greatly reduces the cost and time required to perform such studies, let alone having access to the large, diverse participant population. Case study 2 goes further to demonstrate that crowdsoucing can also be used to understand participants' comprehension in composite visualizations. It also indicates how to crowdsource rich participant input in free text and harvest insights from such input beyond crowdsourcing and analyzing just simple micro-task data. Case study 3 further shows the power of crowdsourcing in understanding participants' perception of complex and potentially deceptive visualizations. It also shows how various individual traits can be measured and assessed in crowdsourced studies. Case study 4 solicits the crowd's aesthetic criteria for network layout and then incorporates the crowdsourced results into layout algorithms. It demonstrates the effectiveness of harvesting the crowd's creativity to inform new visual designs beyond studying participants' visual perceptions.

While the four case studies demonstrate the effectiveness of crowdsourcing in visualization research, they also point out the challenges and limitations in such studies. In particular, the difficulty in instrumenting interactive visualizations for a diverse crowd as well as acquiring comprehensive participant behavioral data during the study (e.g., a participant's attentiveness and experimental condition) that might be easier to control or observe in a traditional laboratory condition.

9 Conclusion

In this chapter, we have highlighted what can be considered the most relevant dimensions in the use of crowdsourcing for Information Visualization research and application development, to which its use brings some genuine advantages and challenges.

Strengths and Opportunities. Literature shows how access to a larger and diverse cohort enriches the amount of information that can be collected as well as the types of data analysis that can be conducted. Financial effectiveness is one of the most mentioned features especially for research on a budget, and crowdsourcing supports easy scaling to large samples that would otherwise be prohibitive, greatly expanding the space of feasible study designs. Crowdsourcing provides opportunities beyond simple cost-cutting, support from crowdsourcing platforms considerably reduces recruiting effort, which is an extremely time consuming task.

Crowdsourcing as a concept is still evolving, the diversity of approaches deployed on existing platforms and interpretations of the concept itself, which transcends off the shelf environments like Amazon Mechanical Turk, provide the opportunity for the research community to tap into dimensions not yet explored. First and foremost is the development of platforms capable of supporting InfoVis type of experiments. The literature shows how community requirements go beyond simple data collection typical of marketing research, for which most of the existing platforms have been initially developed. Literature also shows how the ability to scale to a large cohort and to increase user community diversity can lead to new analytical methods which might strengthen existing or lead to new findings. Comparison of traditional laboratory based studies and crowdsourcing based studies is a powerful mean to replicate and compare results which can lead to consolidate or question field knowledge foundations. Challenges posed by crowdsourcing environment also represent an opportunity to re-think study settings and propose novel designs.

Weaknesses and Threats. Scalability to large cohorts comes at a loss in ability to control several aspects of a study execution: recruitment and filtering of participants, monitoring of task execution from both experimental setting and participants level of involvement. These aspects imply a considerable increase in the complexity of designing a study, more factors need to be taken into account to avoid confounding effects and guarantee reliability of the collected data.

Cost-effectiveness carries also non negligible ethical issues when monetary transactions are involved. Work ethics is not only an ethical issue but a fundamental aspect in research, therefore the book devotes an entire chapter to its role in crowdsourcing (see Chapter **??**).

Crowdsourcing provides access to the power of the crowd which is a fascinating phenomenon. The crowd itself is, however, a very complex entity and as such not suited for each and every task. Threats that might be looming at the horizon include the fallacious perception that quantity implies quality. Crowdsourcing-based studies should not be interpreted as a replacement for traditional laboratory studies and neither a requirement to support research findings. It is also easy to overestimate a crowd knowledge basis, when tasks demand specific skills the chance of overestimation is a highly dangerous threat to the soundness of a study results.

References

1. Adams, F.M., Osgood, C.E.: A cross-cultural study of the affective meanings of color. Journal of Cross-Cultural Psychology 4(2), 135–156 (1973), http://jcc.sagepub.com/content/4/2/135.abstract
2. Aigner, W., Hoffmann, S., Rind, A.: Evalbench: a software library for visualization evaluation. In: Computer Graphics Forum. vol. 32:3:1, pp. 41–50. Wiley Online Library (2013)
3. Aigner, W., Miksch, S., Schumann, H., Tominski, C.: Visualization of Time-Oriented Data. Human-Computer Interaction, Springer, London, UK (2011)
4. Albuquerque, G., Lowe, T., Magnor, M.: Synthetic generation of high-dimensional datasets. Visualization and Computer Graphics, IEEE Transactions on 17(12), 2317–2324 (Dec 2011)
5. Alsallakh, B., Micallef, L., Aigner, W., Hauser, H., Miksch, S., Rodgers, P.: Visualizing sets and set-typed data: State-of-the-art and future challenges. In: Eurographics conference on Visualization (EuroVis)–State of The Art Reports. pp. 1–21 (2014)
6. Alvarez-Garcia, S., Baeza-Yates, R., Brisaboa, N.R., Larriba-Pey, J., Pedreira, O.: Graphgen: A tool for automatic generation of multipartite graphs from arbitrary data. In: Web Congress (LA-WEB), 2012 Eighth Latin American. pp. 87–94. IEEE (2012)
7. Álvarez-García, S., Baeza-Yates, R., Brisaboa, N.R., Larriba-Pey, J.L., Pedreira, O.: Automatic multi-partite graph generation from arbitrary data. Journal of Systems and Software 94, 72–86 (2014)
8. Amar, R., Eagan, J., Stasko, J.: Low-level components of analytic activity in information visualization. In: Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on. pp. 111–117. IEEE (2005)
9. 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. pp. 81–86. IEEE (2007)
10. Andrienko, G., Andrienko, N.: Privacy issues in geospatial visual analytics. Springer (2012)
11. Archambault, D., Purchase, H.C.: The mental map and memorability in dynamic graphs. In: Pacific Visualization Symposium (PacificVis). pp. 89–96 (2012)

12. Archambault, D., Purchase, H.C.: Mental map preservation helps user orientation in dynamic graphs. In: Graph Drawing. pp. 475–486 (2013)
13. Bach, B., Dragicevic, P., Archambault, D., Hurter, C., Carpendale, S.: A descriptive framework for temporal data visualizations based on generalized space-time cubes. In: Computer Graphics Forum. Wiley Online Library (2016)
14. Bach, B., Spritzer, A., Lutton, E., Fekete, J.D.: Interactive random graph generation with evolutionary algorithms. In: Graph Drawing. pp. 541–552. Springer (2013)
15. Bateman, S., Mandryk, R.L., Gutwin, C., Genest, A., McDine, D., Brooks, C.: Useful junk? the effects of visual embellishment on comprehension and memorability of charts. In: CHI '10 (2010)
16. Berlin, B., Kay, P.: Basic color terms. University of California Press, Berkeley, CA (1969)
17. Bertin, J.: Sémiologie graphique: Les diagrammes-Les réseaux-Les cartes. Gauthier-VillarsMouton & Cie (1973)
18. Borgo, R., Adul-Rahman, A., Mohamed, F., Grant, W.p., Reppa, I., Floridi, L., Chen, M.: An empirical study on using visual embellishments in visualization. In: IEEE Transactions on Visualization and Computer Graphics (InfoVis '12) (2012)
19. Boy, J., Rensink, R.A., Bertini, E., Fekete, J.D.: A principled way of assessing visualization literacy. IEEE Transactions on Visualization and Computer Graphics 20(12), 1963–1972 (Dec 2014)
20. Boy, J., Detienne, F., Fekete, J.D.: Storytelling in information visualizations: Does it engage users to explore data? In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. pp. 1449–1458. CHI '15, ACM, New York, NY, USA (2015)
21. Brehmer, M., Munzner, T.: A multi-level typology of abstract visualization tasks. IEEE Transactions on Visualization and Computer Graphics 19(12), 2376–2385 (2013), http://dx.doi.org/10.1109/TVCG.2013.124
22. Brehmer, M., Munzner, T.: A multi-level typology of abstract visualization tasks. IEEE Transactions on Visualization and Computer Graphics 19(12), 2376–2385 (2013)
23. Bremm, S., Von Landesberger, T., Heß, M., Fellner, D.: Pcdc-on the highway to data-a tool for the fast generation of large synthetic data sets. In: EuroVis Workshop on Visual Analytics. pp. 7–11 (2012)
24. Brewer, C.A., MacEachren, A.M., Pickle, L.W., Herrmann, D.: Mapping mortality: Evaluating color schemes for choropleth maps. Annals of the Association of American Geographers 87(3), 411–438 (1997)
25. Brinkmann, G., McKay, B.D.: Fast generation of planar graphs. MATCH Commun. Math. Comput. Chem 58(2), 323–357 (2007)
26. Bristor, V.J., Drake, S.V.: Linking the language arts and content areas through visual technology. T.H.E. Journal 22(2), 74–77 (1994)
27. Çöltekin, A., Fabrikant, S.I., Lacayo, M.: Exploring the efficiency of users' visual analytics strategies based on sequence analysis of eye movement recordings. International Journal of Geographical Information Science 24(10), 1559–1575 (2010)
28. Çöltekin, A., Heil, B., Garlandini, S., Fabrikant, S.I.: Evaluating the effectiveness of interactive map interface designs: A case study integrating usability metrics with eye-movement analysis. Cartography and Geographic Information Science 36(1), 5–17 (2009)
29. Cernea, D., Kerren, A., Ebert, A.: Detecting insight and emotion in visualization applications with a commercial EEG headset. In: SIGRAD 2011 Conference on

Evaluations of Graphics and Visualization-Efficiency, Usefulness, Accessibility, Usability,(Stockholm, Sweden). pp. 53–60 (2011)
30. Cernea, D., Weber, C., Ebert, A., Kerren, A.: Emotion scents – a method of representing user emotions on gui widgets. In: Proceedings of the SPIE 2013 Conference on Visualization and Data Analysis (VDA 2013). IS&T/SPIE (2013)
31. Colblindor: Colblindor. http://www.color-blindness.com/color-blindness-tests/
32. Cole, F., Sanik, K., DeCarlo, D., Finkelstein, A., Funkhouser, T., Rusinkiewicz, S., Singh, M.: How well do line drawings depict shape? In: ACM Transactions on Graphics (ToG). vol. 28:3, p. 28. ACM (2009)
33. Cooper, S., Treuille, A., Barbero, J., Popović, Z., Baker, D., Salesin, D.: Foldit. online:https://fold.it/portal/ (2008)
34. Csikszentmihalyi, M.: Flow: The Psychology of Optimal Experience. Harper Perennia, New York (1990)
35. Dasgupta, A., Kosara, R.: Privacy-preserving data visualization using parallel coordinates. In: IS&T/SPIE Electronic Imaging. pp. 78680O–78680O. International Society for Optics and Photonics (2011)
36. Demiralp, Ç., Bernstein, M.S., Heer, J.: Learning perceptual kernels for visualization design. IEEE Trans. Vis. Comput. Graph. 20(12), 1933–1942 (2014), http://dx.doi.org/10.1109/TVCG.2014.2346978
37. Elmqvist, N., Vande Moere, A., Jetter, H.C., Cernea, D., Reiterer, H., Jankun-Kelly, T.J.: Fluid interaction for information visualization. Information Visualization 10(4), 327–340 (Oct 2011)
38. Fabrikant, S.I., Christophe, S., Papastefanou, G., Maggi, S.: Emotional response to map design aesthetics. In: 7th International Conference on Geographical Information Science, Columbus, Ohio. pp. 18–21 (2012)
39. Farrugia, M., Quigley, A.: Effective temporal graph layout: a comparative study of animation versus static display methods. Information Visualization 10(1), 47–64 (2011)
40. Fort, K., Adda, G., Cohen, K.B.: Amazon mechanical turk: Gold mine or coal mine? Computational Linguistics 37(2), 413–420 (2011)
41. Gadiraju, U., Kawase, R., Dietze, S., Demartini, G.: Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. pp. 1631–1640. CHI '15, ACM, New York, NY, USA (2015), http://doi.acm.org/10.1145/2702123.2702443
42. Ghani, S., Elmqvist, N.: Improving Revisitation in Graphs Through Static Spatial Features. In: Graphic Interface (GI '11). pp. 737–743 (2011)
43. Ghani, S., Elmqvist, N., Yi, J.S.: Perception of Animated Node-link diagrams for dynamic graphs. Computer Graphics Forum 31(1), 1205–1214 (2012)
44. Giannotti, F., Pedreschi, D.: Mobility, data mining and privacy: Geographic knowledge discovery. Springer Science & Business Media (2008)
45. Group, C.B.A.: Color blindness. http://www.colourblindawareness.org/colour-blindness/types-of-colour-blindness/
46. van Ham, F., Rogowitz, B.: Perceptual organization in user-generated graph layouts. Visualization and Computer Graphics, IEEE Transactions on 14(6), 1333–1339 (Nov 2008)
47. Haroz, S., Kosara, R., Franconeri, S.L.: Isotype visualization–working memory, performance, and engagement with pictographs. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. pp. 1191–1200. ACM (2015)

48. Harrison, L.: Experimentr. https://github.com/codementum/experimentr (2014)
49. Harrison, L., Yang, F., Franconeri, S., Chang, R.: Ranking visualizations of correlation using weber's law. Visualization and Computer Graphics, IEEE Transactions on 20(12), 1943–1952 (2014)
50. 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. pp. 203–212. ACM (2010)
51. Hirth, M., Hoßfeld, T., Tran-Gia, P.: Anatomy of a crowdsourcing platform-using the example of microworkers. com. In: Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2011 Fifth International Conference on. pp. 322–329. IEEE (2011)
52. Howe, J.: The rise of crowdsourcing. WIRED (June 2006), http://www.wired.com/2006/06/crowds/
53. Isenberg, P., Zuk, T., Collins, C., Carpendale, S.: Grounded evaluation of information visualizations. In: Proceedings of the 2008 Workshop on BEyond Time and Errors: Novel evaLuation Methods for Information Visualization. pp. 6:1–6:8. BELIV '08, ACM, New York, NY, USA (2008), http://doi.acm.org/10.1145/1377966.1377974
54. Jianu, R., Rusu, A., Hu, Y., Taggart, D.: How to Display Group Information on Node–Link Diagrams: an Evaluation. IEEE Trans. Visualization and Computer Graphics (TVCG) 20, 1530–1541 (2014)
55. Jianu, R., Rusu, A., Hu, Y., Taggart, D.: How to display group information on node-link diagrams: an evaluation. Visualization and Computer Graphics, IEEE Transactions on 20(11), 1530–1541 (2014)
56. Kerren, A., Ebert, A., Meyer, J. (eds.): Human-Centered Visualization Environments, LNCS Tutorial, vol. 4417. Springer (2007)
57. Kerren, A., Schreiber, F.: Network visualization for integrative bioinformatics. In: Chen, M., Hofestädt, R. (eds.) Approaches in Integrative Bioinformatics, pp. 173–202. Springer, Heidelberg (2014)
58. Lam, H., Bertini, E., Isenberg, P., Plaisant, C., Carpendale, S.: Empirical studies in information visualization: Seven scenarios. Visualization and Computer Graphics, IEEE Transactions on 18(9), 1520–1536 (2012)
59. Laramee, R.S., Kosara, R.: Challenges and Unsolved Problems. In: Kerren et al. [56], pp. 231–254
60. Lebreton, P., M/"aki, T., Skodras, E., Hupont, I., Hirth, M.: Bridging the gap between eye tracking and crowdsourcing. In: Proc. SPIE. vol. 9394, pp. 93940W–93940W–14 (2015), http://dx.doi.org/10.1117/12.2076745
61. Lee, B., Plaisant, C., Parr, C.S., Fekete, J.D., Henry, N.: Task taxonomy for graph visualization. In: Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization. pp. 1–5. ACM (2006)
62. Li, H., Moacdieh, N.: Is "chart junk" useful? An extended examination of visual embellishment. Proceedings of the Human Factors and Ergonomics Society Annual Meeting 58(1), 1516–1520 (2014)
63. Light, A., Bartlein, P.J.: The end of the rainbow? color schemes for improved data graphics. Eos 85(40), 385–391 (2004)
64. 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. pp. 1425–1434. ACM (2007)

65. Maggi, S., Fabrikant, S.I.: Embodied decision making with animations. 8th International Conference on Geographic Information Science (2014)
66. Maggi, S., Fabrikant, S.I.: Triangulating eye movement data of animated displays. 2nd International Workshop on Eye Tracking for Spatial Research 1241, 27–31 (2014)
67. Maggi, S., Fabrikant, S.I., Imbert, J.P., Hurter, C.: How do display design and user characteristics matter in animations? an empirical study with air traffic control displays. Cartographica (2016)
68. Mahyar, N., Kim, S.H., Kwon, Bum, C.: Towards a taxonomy for evaluating user engagement in information visualization. Workshop on Personal Visualization: Exploring Everyday Life (2015)
69. Marriott, K., Purchase, H., Wybrow, M., Goncu, C.: Memorability of visual features in network diagrams. Visualization and Computer Graphics, IEEE Transactions on 18(12), 2477–2485 (Dec 2012)
70. Martin, D.: Doing psychology experiments, 7th Edition. Thomson Wadsworth, Belmont, CA (2008)
71. Martin, D., Hanrahan, B.V., O'Neill, J., Gupta, N.: Being a turker. In: Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing. pp. 224–235. ACM (2014)
72. McCloud, S.: Understanding Comics: The Invisible Art. HarperPerennial, New York (1994)
73. McGee, F., Dingliana, J.: An empirical study on the impact of edge bundling on user comprehension of graphs. In: Proceedings of the International Working Conference on Advanced Visual Interfaces. pp. 620–627. ACM (2012)
74. Micallef, L., Dragicevic, P., Fekete, J.D.: Assessing the effect of visualizations on bayesian reasoning through crowdsourcing. IEEE Transactions on Visualization and Computer Graphics 18(12), 2536–2545 (2012)
75. Monreale, A., Andrienko, G.L., Andrienko, N.V., Giannotti, F., Pedreschi, D., Rinzivillo, S., Wrobel, S.: Movement data anonymity through generalization. Transactions on Data Privacy 3(2), 91–121 (2010)
76. Munzner, T.: A nested model for visualization design and validation. IEEE Transactions on Visualization and Computer Graphics 15(6), 921–928 (2009)
77. Okoe, M., Jianu, R.: Graphunit: Evaluating interactive graph visualizations using crowdsourcing. In: Computer Graphics Forum. vol. 34:3, pp. 451–460. Wiley Online Library (2015)
78. Paas, F., Tuovinen, J.E., Tabbers, H., Van Gerven, P.W.: Cognitive load measurement as a means to advance cognitive load theory. Educational psychologist 38(1), 63–71 (2003)
79. Pandey, A.V., Rall, K., Satterthwaite, M.L., Nov, O., Bertini, E.: How deceptive are deceptive visualizations?: An empirical analysis of common distortion techniques. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. pp. 1469–1478. ACM (2015)
80. Papadopoulos, C., Gutenko, I., Kaufman, A.: Veevvie: Visual explorer for empirical visualization, vr and interaction experiments. Visualization and Computer Graphics, IEEE Transactions on 22(1), 111–120 (2016)
81. Peck, E.M.M., Yuksel, B.F., Ottley, A., Jacob, R.J., Chang, R.: Using fNIRS brain sensing to evaluate information visualization interfaces. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 473–482. ACM (2013)

82. Plaisant, C.: The challenge of information visualization evaluation. In: Proceedings of the Working Conference on Advanced Visual Interfaces. pp. 109–116. AVI '04, ACM, New York, NY, USA (2004), http://doi.acm.org/10.1145/989863.989880
83. Purchase, H.: Which aesthetic has the greatest effect on human understanding? In: Graph Drawing. pp. 248–261. Springer (1997)
84. Purchase, H.C.: Experimental human-computer interaction: a practical guide with visual examples. Cambridge University Press (2012)
85. Ross, J., Irani, L., Silberman, M., Zaldivar, A., Tomlinson, B.: Who are the crowdworkers?: shifting demographics in mechanical turk. In: CHI'10 extended abstracts on Human factors in computing systems. pp. 2863–2872. ACM (2010)
86. Saket, B., Scheidegger, C., Kobourov, S.: Towards understanding enjoyment and flow in information visualization. In: EuroVis. The Eurographics Association (Short Paper) (2015)
87. Saket, B., Scheidegger, C., Kobourov, S.: Comparing node-link and node-link-group visualizations from an enjoyment perspective. In: Computer Graphics Forum. vol. 35:3 (2016)
88. Saket, B., Scheidegger, C., Kobourov, S.G., Börner, K.: Map-based visualizations increase recall accuracy of data. In: Computer Graphics Forum. vol. 34:3, pp. 441–450. Wiley Online Library (2015)
89. Saket, B., Scheidegger, C., Kobourov, S.G., Börner, K.: Map-based visualizations increase recall accuracy of data. Comput. Graph. Forum 34(3), 441–450 (2015), http://dx.doi.org/10.1111/cgf.12656
90. Sakshaug, J.W., Raghunathan, T.E.: Synthetic data for small area estimation. In: Privacy in Statistical Databases. pp. 162–173. Springer (2010)
91. Sakshaug, J.W., Raghunathan, T.E.: Generating synthetic data to produce public-use microdata for small geographic areas based on complex sample survey data with application to the national health interview survey. Journal of Applied Statistics 41(10), 2103–2122 (2014)
92. Sakshaug, J.W., Raghunathan, T.E.: Nonparametric generation of synthetic data for small geographic areas. In: Privacy in Statistical Databases. pp. 213–231. Springer (2014)
93. Shneiderman, B.: The eyes have it: A task by data type taxonomy for information visualizations. In: Proceedings of the 1996 IEEE Symposium on Visual Languages, Boulder, Colorado, USA, September 3-6, 1996. pp. 336–343. IEEE Computer Society (1996)
94. Tanahashi, Y., Ma, K.L.: Stock lamp: An engagement-versatile visualization design. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. pp. 595–604. ACM (2015)
95. Theodoridis, Y., Silva, J.R., Nascimento, M.A.: On the generation of spatiotemporal datasets. In: Advances in Spatial Databases. pp. 147–164. Springer (1999)
96. Valiati, E.R., Pimenta, M.S., Freitas, C.M.: A taxonomy of tasks for guiding the evaluation of multidimensional visualizations. In: Proceedings of the 2006 AVI workshop on Beyond time and errors: novel evaluation methods for information visualization. pp. 1–6. ACM (2006)
97. Vande Moere, A., Tomitsch, M., Wimmer, C., Christoph, B., Grechenig, T.: Evaluating the effect of style in information visualization. Visualization and Computer Graphics, IEEE Transactions on 18(12), 2739–2748 (Dec 2012)
98. Wainer, H.: A test of graphicacy in children. Applied Psychological Measurement 4(3), 331–340 (1980)

99. Walny, J., Huron, S., Carpendale, S.: An Exploratory Study of Data Sketching for Visual Representation. Computer Graphics Forum (2015)
100. Wang, B., Ruchikachorn, P., Mueller, K.: Sketchpadn-d: Wydiwyg sculpting and editing in high-dimensional space. Visualization and Computer Graphics, IEEE Transactions on 19(12), 2060–2069 (2013)
101. Ware, C.: Information Visualization: Preception for Design, Third Edition. Elsevir (2013)
102. Ware, C., Mitchell, P.: Visualizing graphs in three dimensions. ACM Transactions on Applied Perception (TAP) 5(1), 2 (2008)
103. Wilkening, J., Fabrikant, S.I.: How users interact with a 3d geo-browser under time pressure. Cartography and Geographic Information Science 40(1), 40–52 (2013)
104. Xu, P., Ehinger, K.A., Zhang, Y., Finkelstein, A., Kulkarni, S.R., Xiao, J.: Turkergaze: Crowdsourcing saliency with webcam based eye tracking. CoRR abs/1504.06755 (2015), http://arxiv.org/abs/1504.06755
105. Yang, H., Li, Y., Zhou, M.X.: Understand users' comprehension and preferences for composing information visualizations. ACM Transactions on Computer-Human Interaction (TOCHI) 21(1), 6 (2014)
106. Ying, X., Wu, X.: Graph generation with prescribed feature constraints. In: SDM. vol. 9, pp. 966–977. SIAM (2009)
107. Ziemkiewicz, C., Kosara, R.: Preconceptions and individual differences in understanding visual metaphors. Computer Graphics Forum 28(3), 911–918 (2009)