VisLab: Enabling Visualization Designers to Gather Empirically Informed Design Feedback

Jinhan Choi
jinhan.choi@bc.edu
Boston College
Chestnut Hill, MA, USA

Changhoon Oh
changhoonoh@yonsei.ac.kr
Yonsei University
Seoul, South Korea

Yeaseul Kim
yeaseul.kim@cs.wisc.edu
University of Wisconsin-Madison
Madison, WI, USA

Nam Wook Kim
nam.wook.kim@bc.edu
Boston College
Chestnut Hill, MA, USA

ABSTRACT

When creating a visualization, designers face various conflicting design choices. They typically rely on their hunches to deal with intricate trade-offs or resort to feedback from their colleagues. On the other hand, researchers have long used empirical methods to derive useful quantitative insights into visualization designs. Taking inspiration from this research tradition, we developed VisLab, an open-source online system to complement the existing qualitative feedback practice and help visualization practitioners run experiments to gather empirically informed design feedback. We surveyed practitioners' perceptions of quantitative feedback and analyzed the research literature to inform VisLab's motivation and design. VisLab operationalizes the experiment process using templates and dashboards to make empirical methods amenable for practitioners while supporting sharing and remixing experiments to aid knowledge exchange and validation. We demonstrated the validity of experiments in VisLab and evaluated the usability and potential usefulness of VisLab in visualization design practice.

CCS CONCEPTS

• **Human-centered computing → Interactive systems and tools**.

KEYWORDS

design feedback, empirical feedback, data visualization, crowdsourcing, citizen science

ACM Reference Format:
Jinhan Choi, Changhoon Oh, Yeaseul Kim, and Nam Wook Kim. 2023. VisLab: Enabling Visualization Designers to Gather Empirically Informed Design Feedback. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3544548.3581132

CHI '23, April 23–28, 2023, Hamburg, Germany

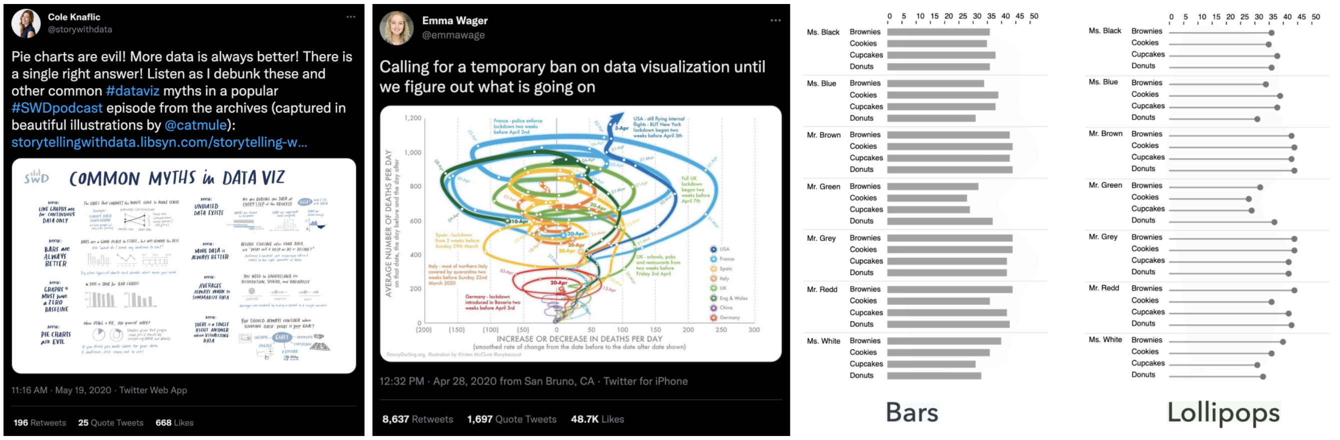
ACM ISBN 978-1-4503-9421-5/23/04.
https://doi.org/10.1145/3544548.3581132

1 INTRODUCTION

Data visualizations are now widely used across disciplines to understand and communicate complex data [73, 76, 94]. Visualization designers in the wild need to consider various factors, including perceptual effectiveness, aesthetics, memorability, and engagement. Although past empirical research provides fruitful visualization design knowledge, it is insufficient to cover the combinatorial space of visualization design in the wild [23]. Researchers often lack time and resources to investigate the vast variable space, leading to incomplete guidelines for the designers [35, 36]. On the other hand, designers commonly use their hunches to make nuanced design decisions involving heterogeneous data distributions, unconventional visuals, and narrational elements that are absent in typical research experiments. For instance, designers often employ unsubstantiated yet practical representations such as Bullet Graph [2], while scientists come up with new visuals to cope with domain-specific data such as Muller plots [11].

Practitioners often engage in active design discussions and experiments. For instance, they discuss common design myths (Figure 1 left) and alternative visual encodings [8, 10, 17, 46] and layouts [32]. Others debate unfamiliar designs on social media, such as Tornado plot [19] (Figure 1 middle), Circular Tube [3], and Sequence Logo [16], and Marimekko charts [25]. Moreover, several practitioners often run their own experiments, such as comparing charts for Likert scale data [1], visualizations for scientific results [21], bars vs. lollipops [63] (Figure 1 right) and bars vs. pies [15]. More often than not, practitioners seek qualitative design feedback from other experienced colleagues. While it is a quick and easy way to obtain rich design insights, this form of feedback can exhibit less specificity due to non-anonymity and the fear of criticism [64]. Since not everyone can afford such expert colleagues [80], existing research systems have investigated ways to gather *qualitative* design critiques from crowdworkers [79, 80, 108].

On the other hand, data visualization researchers have long been conducting empirical studies to produce visualization design knowledge [75]. Such empirical studies have typically remained within the strict realm of scientific investigation, requiring certain expertise in statistics and programming. Nevertheless, their underlying methodologies generally have structures and patterns [51] that can be streamlined to alleviate the complexity and have the potential to provide valuable insights into visualization designs in practice. Past crowdsourced studies have demonstrated the possibility of scalable



Cole Knaflic
@storywithdata

Pie charts are evil! More data is always better! There is a single right answer! Listen as I debunk these and other common #dataviz myths in a popular #SWDpodcast episode from the archives (captured in beautiful illustrations by @catmule): storytellingwithdata.libsyn.com/storytelling-w...

11:16 AM · May 19, 2020 · Twitter Web App

196 Retweets 25 Quote Tweets 668 Likes

Emma Wager
@emmawage

Calling for a temporary ban on data visualization until we figure out what is going on

12:32 PM · Apr 28, 2020 from San Bruno, CA · Twitter for iPhone

8,637 Retweets 1,697 Quote Tweets 48.7K Likes

SWD
COMMON MYTHS in DATA VIZ
MYTH: LINE GRAPHS are for CONTINUOUS DATA ONLY
MYTH: BARS are ALWAYS BETTER
MYTH: GRAPHS MUST have a ZERO BASELINE
MYTH: PIE CHARTS are EVIL
MYTH: UNBIASED DATA EXISTS
MYTH: MORE DATA is ALWAYS BETTER
MYTH: AVERAGES ALWAYS WORK to SUMMARIZE DATA
MYTH: THERE is a SINGLE RIGHT ANSWER when VISUALIZING DATA
AVERAGE NUMBER OF DEATHS PER DAY
on that date, the day before and the day after
INCREASE OR DECREASE IN DEATHS PER DAY
(smoothed rate of change from the date before to the date after date shown)
France - police enforce lockdown two weeks before April 2nd
USA - still flying internal flights - BUT New York lockdown began two weeks before April 5th
Full UK lockdown began two weeks before April 7th
Spain - lockdown from 2 weeks before Sunday 29th March
Italy - most of northern Italy covered by quarantine two weeks before Sunday 22nd March 2020
Germany - lockdown introduced in Bavaria two weeks before April 3rd
UK - schools, pubs and restaurants from two weeks before Friday 3rd April
USA
France
Spain
Italy
UK
Eng & Wales
China
Germany
DannyDorling.org. Illustration by Kirsten McClure @orpheuscat
0 5 10 15 20 25 30 35 40 45 50
Ms. Black
Ms. Blue
Mr. Brown
Mr. Green
Mr. Grey
Mr. Redd
Ms. White
Brownies
Cookies
Cupcakes
Donuts
Bars
Lollipops

Figure 1: Practitioners actively engage in online design discussions and experiments: (left) common myths in data visualizations [13], (middle) a controversial chart for pandemic data [19], a practitioner-run experiment comparing bars and lollipops [63]

experiment design and the viability of reaching a broader range of participants in online environments beyond restricted laboratory settings [26]. Notwithstanding, crowdsourced studies require the same research expertise that is not typically available to practitioners. Currently, no systems exist that can leverage the benefit of such *quantitative* experiments to meet practitioners' design needs.

This work seeks to enrich the existing feedback practice by supporting visualization practitioners to gather empirically informed, quantitative design feedback. First, we conducted a preliminary survey with 18 practitioners from the Slack workspace of Data Visualization Society [4] to understand the current feedback practices. The survey result showed that they gather feedback from others, look for examples to address the situation (83.33%), and rely on their gut instinct (50.00%) to address design choices. Moreover, many also found a quantitative experiment would be useful for evaluating their visualizations (4.38 on a 5-point Likert scale from 1: not very useful to 5: very useful). Next, we analyzed 140 experiments published in major data visualization venues to learn how to streamline the experiment process for quantitative feedback. Most experiments followed a standard procedure, including a pre-survey (e.g., qualification task), screening (e.g., testing color blindness), practice trials, main trials, and post-surveys (e.g., demographic survey). The analysis also revealed diverse evaluation topics and tasks & measures for practical visualization evaluation.

We developed VisLab, an open-source online system enabling visualization designers to run experiments to gather quantitative feedback. We derived an initial set of three experiment templates from the literature analysis, including graphical perception [41] GP, attention tracking AT [67], and memorability MB [29]. The templates were selected based on their scalability and generalizability in a practical context and designed to minimize the user's efforts (e.g., automatically suggesting perceptual task questions for GP). If needed, users can configure different experiment stages, such as the tutorial and post-survey. Moreover, VisLab provides a visual dashboard to aid the interpretation of experiment results using simple error bars and tables. The user can easily distribute the experiment to colleagues and potential readers through a shareable link. Individual experiment participants can see their performance results in light of the collective performance of all participants as a learning incentive. Lastly, to promote knowledge sharing, VisLab supports browsing and remixing finished experiments based on the chart, data, and task types.

We conducted two evaluation studies. First, we tested the feasibility of the templates by running replication experiments in VisLab based on the original studies [29, 41, 67]. We observed outcomes consistent with the originals in the perceptual comparison of bar vs. pie charts and the memorability rankings and click patterns of in-the-wild visualizations. Next, we conducted a two-phased user study with the practitioners from the preliminary survey. The first phase involved replication & reproduction tasks to evaluate the overall usability of VisLab, while the second phase involved creative tasks with three selected participants. The participants commented that the templates lowered barriers to creating an experiment and gathering quantitative feedback, while the dashboard facilitated interpreting the outcome and deriving design feedback. They also mentioned that VisLab would make it easier for other people to give objective feedback due to its quantitative form. The participants in the second phase also praised the value of sharing and remixing, such as reusing existing questions created by others and building a community-based visualization knowledge repository.

Our main contribution lies in the design, development, and evaluation of VisLab, consisting of the following sections:

- A practitioner survey illustrating how they collect design feedback to address conflicting design choices and perceive the value of empirically informed feedback.
- An analysis of existing empirical studies providing an overview of evaluation topics and common procedures and methods to run and report the studies.
- VisLab, a novel system that allows practitioners to run experiments and obtain empirically informed design feedback, share results with others, and extend existing experiments.

- User studies demonstrating the feasibility of the templates and the potential value of quantitative feedback in design practice.

We make the online system available in https://vislab.bc.edu/ and the source code in https://github.com/datawithinreach/vislab.

2 RELATED WORK

We review existing systems for gathering qualitative design feedback, empirical studies to gain quantitative design insights, past attempts to streamline the empirical processes, and inspirational knowledge-sharing platforms.

2.1 Gathering Design Feedback

Feedback is essential to assess a design and generate revision ideas in the design process [56, 97]. Gathering design critiques from colleagues is a common way to gather such feedback. However, the fear of criticism and non-anonymity can make people uncomfortable receiving feedback from colleagues [48, 64]. Moreover, self-employed people might struggle to find colleagues who can provide valuable feedback [107].

To address the problem of gathering quality feedback, past studies tapped into online crowdsourcing that can allow for easy access to large and diverse participants and faster feedback turnaround time with relatively low costs [31]. To overcome crowd workers' lack of appropriate expertise, existing approaches investigated various strategies to improve the quality of feedback from the workers, such as micro-tasking [79, 108], predefined rubrics [110], demonstrative examples [66], and feedback guidelines [74]. While these approaches make crowdsourced feedback more feasible and usable, the feedback still mostly takes the form of descriptive sentences that often contain differing or conflicting opinions and make the interpretation difficult as a result [109]. On the other hand, majority voting, ranking, and Likert scales can provide more interpretable numeric ratings and aggregate opinions and preferences [57].

Receiving design feedback is also an essential part of the visualization design process. For instance, Kosara [72] discusses the idea of visualization criticism, a critical thinking approach to discussing and assessing visualization designs. Viégas and Wattenberg similarly discuss redesigns as criticism [105]. Likewise, design study research [99, 102] and education curricula [47, 58, 95] have incorporated design critique and feedback; e.g., incorporating feedback from domain scientists and providing feedback to student projects, respectively. Researchers also adopted heuristic evaluation from the human-computer interaction literature [86] to help structure visualization design feedback [59, 112].

Beyond qualitative evaluation, past empirical research in visualization takes a quantitative approach to evaluate visualizations. For instance, perception studies have been prevalent in quantifying the effectiveness of different visualization encoding designs [42, 60, 68]. Other recent studies investigate additional cognitive aspects of visualization design, including engagement [24, 28, 29] and memorability [54, 65, 81]. Similar studies (e.g., usability testing [84] and A/B testing [85]) are also often used in the industry to evaluate user interface designs. While visualization empirical studies have been primarily conducted in research contexts, their purpose is in a similar spirit to design feedback, as they both provide knowledge or insight into how a visualization would work.

While existing systems focus on providing *qualitative feedback*, we leverage empirical methods to help visualization practitioners gather *quantitative* feedback. The complementary use of qualitative and quantitative feedback can offer a broad range of design insights and perspectives.

2.2 Scaffolding Empirical Study Procedures

Empirical studies are typically conducted in controlled lab settings and often require diverse expertise such as experiment design, statistical analysis, and programming knowledge. They have been mainly used by specialists in particular contexts such as scientific investigation and user research. While these studies typically recruit people from the location of the studies and conduct study sessions in-person, recent studies leverage crowdsourcing to tap into more diverse and larger online labor markets [91]. These online markets also enable faster completion time with relatively low costs while also improving the ecological validity of the studies [31]. The online crowdsourcing approach has been employed in a variety of contexts, including software evaluation [70], behavioral studies [82, 89], and visualization experiments [27].

To adopt in-lab studies for the online environment, past research investigated ways to streamline the process of experiment tasks and procedures. For instance, microtasking is commonly used to break down a complex task into smaller tasks that can be done independently for a short amount of time without requiring specific skills [70]. Several other research also investigated quality control mechanisms, such as qualification tests and outlier removal [44], to address potential quality issues arising from anonymous participants with diverse skills and interests.

While many of the crowdsourced studies use paid platforms such as Amazon Mechanical Turk, CrowdFlower, and Prolific [91], recent research looks into volunteer-based approaches, including LabintheWild [93] and VolunteerScience [92]. Such volunteer-based platforms use personalized feedback as a learning incentive for participation rather than monetary rewards.

Crowdsourcing has become a popular alternative to visualization studies as well. Borgo et al. recently provided a comprehensive survey of visualization evaluation using crowdsourcing [27] and characterized existing crowdsourced experiments. For example, Heer & Bostock [60] crowdsourced graphical perception experiments that were originally run by Cleveland & McGill [41]. Many graphical perception studies have similarly been conducted online, comparing different error bars [43], correlation visualizations [55], deceptive designs [88], color differences [103], task effectiveness [68], and bar chart variations [104]. Other recent studies explored cognitive dimensions, such as memorability and recall [29, 54, 69], Bayesian reasoning [83], attention [67], attraction [49, 111], and persuasion [87]. All these studies leverage the crowdsourcing mechanism to enhance the scale and ease the deployment of experiments.

Current online empirical studies require the hands of expert researchers, statisticians, and programmers to craft the experiment procedures. Several existing commercial tools, such as LabVanced [52], Gorilla [6], and Qualtrics [7], provide complex graphical interfaces to build online experiments and surveys. However,

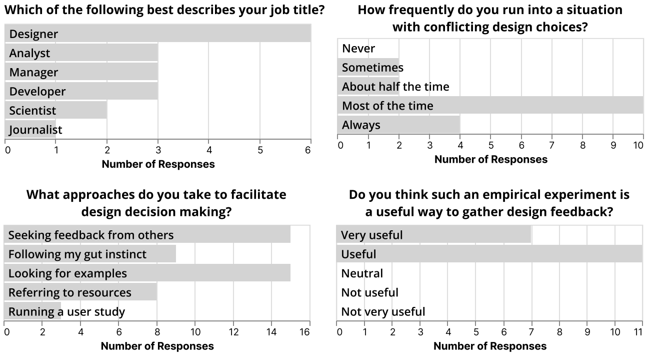


Figure 2: Overview of the preliminary survey with 18 participants. Many are designers, and they regularly run into design conflict situations. They typically rely on feedback from others, existing examples, or even their gut instinct to resolve the design conflict. Most said they found the empirical method useful for gathering design feedback.

they are still designed for researchers rather than practitioners. For instance, the templates provided in the existing tools are primarily for psychology and behavioral experiments. Moreover, they mainly serve as data collection tools, leaving the outcome analysis to researchers (e.g., downloading and analyzing in a separate tool), which is a critical step for deriving feedback.

In this work, we simplify the experiment process for practitioners to derive practical design feedback by leveraging templates derived from the empirical study literature. In contrast to the online research tools that focus on data collection, our system provides a visual dashboard to facilitate outcome interpretation.

2.3 Community-driven Knowledge Sharing

Fostering knowledge sharing can improve the collective knowledge of communities by alleviating trials and errors by individuals [61]. Many existing knowledge systems attempt to translate tacit knowledge from individuals into more explicit knowledge that can be easily accessed by other people [34, 100]. Examples include Scratch [96], Glitch [5], and Observable [12] in which people can build and share their programming projects (e.g., games, arts, visualizations) and also enable the cloning and extension of other projects. People on such knowledge-sharing platforms can have a variety of individual and social motivations, including altruism, reputation, and reciprocity [38]. Often, such collaborative efforts are vital to maintaining the integrity of large-scale online information, such as in Wikipedia [77].

We intend VisLab to facilitate sharing experiment results with the broader visualization community. VisLab provides a tag-based interface to annotate experiments and supports browsing & extending experiments. This sharing and remixing capability can provide additional channels to collaborate and provide feedback.

3 UNDERSTANDING DESIGN PRACTICE

We conducted a brief survey to gain insights into what design challenges visualization practitioners face, how they collect design feedback to address them, and how they perceive the value of empirical methods to derive design feedback.

3.1 Participants

We recruited 18 participants from the *Data Visualization Society* [4]'s *#help-general* channel. In the demographic part of the survey, six participants identified themselves as designers, three as developers, three as managers, three as analysts, three as (data) scientists, and one as a data journalist (Figure 2). In terms of experience in the visualization field, six participants reported they have two to four years of experience (33.33%), five have eight to ten years of experience (27.78%), four have five to seven years of experience (22.22%), two have less than two years of experience (11.11%), and one has more than 11 years of experience (5.56%). Almost all participants said their primary purpose in creating data visualizations is to communicate and present data to others (94.44%). In contrast, about half of them said they use visualizations to explore and analyze data (55.56%). When asked about the primary audience of their visualizations, stakeholders from a specific domain (e.g., policymakers and medical professionals) were the most common (66.67%), followed by internal members such as executives and managers (61.11%). Others also indicated that the general public (55.56%), as well as friends & family (33.33%), are their audiences, while six participants indicated "self" as one of the audiences (33.36%).

3.2 Results

3.2.1 *Design decision making.*

All participants indicated that they have been in a situation where they need to make decisions among conflicting design alternatives (100%). When asked about the frequency, four participants said they always run into such a situation (22.22%), ten for most of the time (55.56%), two for about half of the time (11.11%), and the rest two for sometimes (11.11%). We received varied yet consistent responses to the question asking the types of design conflicts/choices they run into. Making competing decisions between aesthetics and functions (77.78%) and selecting the right chart (72.22%) were the most common issues (72.22%), followed by interaction design to coordinate multiple visualizations (61.11%), designing chart elements such as picking appropriate color scales and selecting axes ranges (55.56%), choosing the right layout (55.56%), and addressing conflicts with user and business stakeholder needs (50.00%). When asked about what aspects they consider when making design decisions, clarity (77.78%), readability (77.78%), accuracy (72.22%), and understandability (72.22%) were generally higher than aesthetics (55.56%), storytelling (55.56%) and accessibility (50.00%).

3.2.2 *Design feedback practice.*

We asked what approaches they take to facilitate design decision-making. Seeking feedback from others (83.33%) and looking for existing examples were the most common responses (83.33%). Other responses included following their gut instinct (50.00%) and referring to resources such as guidelines and online courses (44.44%). Three participants said they ran a user study or experiment (16.67%). When asked about who they typically reach out to for feedback, colleagues were their top choice (88.89%), followed by actual potential users such as public audiences or people in the target domain (61.11%), experts (33.33%), friends or family (27.78%), and social networks such as Twitter or Slack community channels (22.22%). Responses were similar but slightly different in terms of “ideal” groups to seek feedback from. Actual potential users were the top choice (77.78%), followed by experts (66.67%), colleagues (55.56%), social networks (22.22%), and friends or family (11.11%). The appropriate number of people for seeking feedback was relatively on the low end: 2-3 people (50.00%), 4-10 people (33.33%), and 11 or more people (5.56%), which makes sense for qualitative feedback [22].

3.2.3 Attitude toward empirically-driven feedback. We were also interested in their perception of the utility of quantitative design feedback. As a representative empirical experiment, we requested them to participate in the pie vs. bar comparison study available in the Financial Times article titled *The science behind good charts* [15]. To ensure they participated in the experiment, we asked them to input their scores in the survey.

Subsequently, we asked about their awareness of such an empirical experiment. On a 5-point Likert scale from not well at all (1) to extremely well (5), the average awareness was 3.44 with $SD = 1.17$. When asked about where they have come across empirical studies, responses were diverse, ranging from books (55.56%), social media (50.00%), conferences and research paper archives (50.00%) to online websites such as blogs (38.89%), colleagues (33.33%), and online courses (11.11%). One participant has not seen any, while another has had experience running an experiment.

Regarding how approachable empirical research is, they were neutral ($M = 3.17, SD = 0.96$) on a 5-point Likert scale from not very accessible (1) to very accessible (5). Finally, we asked whether such an empirical experiment would be helpful for design feedback practice if it is easy to run such experiments to evaluate their visualizations. The overall response was positive; $M = 4.38, SD = 0.49$ on a 5-point Likert scale from not very useful (1) to very useful (5). In addition, we asked if they would use a tool that can facilitate the experiment process for gathering quantitative design feedback. The response was similarly positive ($M = 3.94, SD = 0.91$ on a 5-point Likert scale from 1: very unlikely to 5: very likely), although slightly less.

3.3 Takeaways

The results indicated that they commonly rely on qualitative feedback, such as opinions from colleagues and insights from examples, which aligns with what the existing feedback systems support (Section 2). On the other hand, their interest and willingness to incorporate empirical methods to evaluate visualizations demonstrate the potential of quantitative feedback in the design practice. In fact, their awareness of empirical research was not as low as we expected.

3.4 Limitations

The bar vs. pie perception experiment embedded in the survey was to introduce the idea of an empirical method for those unfamiliar with it. While it is the most representative and accessible one, a future investigation could employ a richer instrument to gauge practitioners’ interests in diverse empirical metrics such as engagement, aesthetic and affective responses.

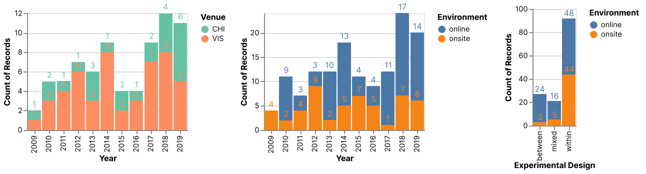


Figure 3: Our data collection consists of papers published in ACM CHI and IEEE VIS for the duration between 2009 and 2019. The number of papers is growing, especially for online studies, which typically have more between-subject studies due to easy recruitment.

4 UNDERSTANDING EMPIRICAL RESEARCH

To investigate practical ways to scaffold the process of deriving quantitative feedback, we collected and analyzed existing empirical studies that provide best practices for quantitative evaluation.

4.1 Data Collection

Our main inclusion criteria were empirical experiments evaluating user performance and experience in processing a data visualization, as they provide quantitative guidelines for visualization designs [33, 75]. To collect past empirical research in visualization, we used VisPerception, a public repository [20] of more than 200 visualization empirical studies published since 1926. We limited our analysis scope to articles published from 2009 to 2019 in ACM CHI and IEEE VIS, major data visualization venues, that provide enough representative samples to find patterns and trends in visualization experiments. The data collection was performed in early 2020 as part of our preliminary investigation of this work. The final data collection contains 74 papers, including 24 papers from ACM CHI and 50 papers from IEEE VIS. Finally, we divided each paper into multiple experiments, if applicable, resulting in a total of 140 experiments.

4.2 Analysis Method

Four researchers (one faculty, one graduate assistant, and two undergraduate assistants) went through the paper collection through an iterative, open coding process [101] with several high-level themes: evaluation aspects, procedures, and tasks & measures. We derived initial codes from the literature, such as task names, chart vocabularies, metrics, and environments [27, 30]. The researchers had weekly meetings to review the codes conducted by research assistants with guidance from the faculty researcher. They resolved conflicts and disagreements through consensus and discussion. In the end, the faculty researcher thoroughly reviewed the final codebook and resulting codes.

4.3 Results

4.3.1 Data collection overview. Figure 3 shows the overview of the data collection. Of the 140 experiments, 88 (62.9%) were online, and 52 (37.1%) were onsite experiments. The number of papers increased over time, showing the growth of empirical research in visualization, e.g., only two papers in 2009 vs. 11 papers in 2019. Online studies have been growing and are generally more common in recent years.

Regarding experiment design, single-level design, either within or between subjects, was the most common (85.0%), while we also observed 15.0% mixed designs. Between-subject designs were more common for online studies (27.3% vs. 5.8%), potentially due to their easy recruitment of larger participants.

4.3.2 Evaluation aspects, tasks, and measures. Perception experiments (77.1%) were more common than cognition experiments, such as memorability, comprehension, and engagement (20.7%). In terms of the types of visualizations considered, existing studies mostly focus on standard charts, including scatter plots (12.9%), bar charts (10.0%), and line charts (5.0%). At the same time, we also noticed custom charts (5.7%) and pictograms (3.6%). The frequency ranking was more or less similar across experiment environments, except all of the custom charts were tested onsite. Only 5% of the experiments tested animation, while 22.9% of them involved user interaction.

We observed a variety of user tasks in the experiments. The top three tasks include *compare* (25.7% of 140 experiments), *identify* (21.4%), *derive* (11.4%), *select* (5%), and *recall* (3.6%), while the *derive* task also often involved other lower-level tasks including *identify* and *compare* tasks. Example user tasks include identifying the extremum and deriving the ratio of two numbers. In terms of measures, accuracy (40.0%), time (38.6%), error rate (28.6%), and confidence (7.9%) were the most common. While not prevalent, other experiment-specific metrics include hit rate & false alarm rate, perceived effectiveness, fixation locations, engagement, and discriminability rate.

4.3.3 Experiment procedures, participants, and reports. Most experiments followed a standard procedure, including pre-survey (28.6%), screening (41.4%), training/practice (60.0%), and post-survey (32.1%), although they often do not explicitly mention whether they had each component or not.

Out of 40 experiments that had pre-surveys, common pre-survey questions include gender (60.0%), age (50.0%), chart literacy (20.0%), education (17.5%), and academic major (15%). Among 45 experiments that had post-surveys, common post-survey questions also frequently involved demographics (20.0%), followed by free-form comment (15.6%), preference (11.1%), task strategy (6.7%), and familiarity (6.7%). The rate of having pre-surveys was higher for in-lab studies (46.2%) compared to online studies (18.2%), while the rate for post-surveys was about even (51.1% onsite vs. 48.9% online).

Out of 58 experiments that had screening tasks, many of them involved color blindness (48.3%), while others included vision tests (13.8%) and the acceptance rate on Amazon Mechanical Turk (10.3%). In contrast to pre-surveys, online studies had more screenings (43.2%) than in-lab studies (38.5%). Among 84 experiments that had training and practice, the frequency of online studies was higher than that of onsite studies (57.1% vs. 42.9%). The type of practice tasks depended on the task types in the experiments.

For online studies, the average number of participants was 216 ($median = 96, SD = 287$), while the average was 24 in in-lab studies ($median = 20, SD = 14$). Although we did not quantify the gender balance due to a lack of reports, the experiments that reported gender distribution had sensible balance except in a few cases (e.g., 14 female vs. 59 male) if it matters for the result [62].

In terms of reporting experimental results, they employed diverse charts. The most common forms of reporting methods were error bars (37.1%) and tables (17.1%), while often simple bar charts (7.9%) and box plots (8.6%) were used as well. Often, the task time was constrained (33.6%). The ratio for time constraints was higher for onsite studies (42.3%) than online studies (28.4%).

4.4 Lessons Learned

The shared components across the experiment procedures hinted at ways to devise a standardized experiment process that would be amenable for practitioners. Although the experiments had varied purposes and settings, several had relatively simple goals and setups that could be templatized to different visualizations, datasets, and tasks. Likewise, common graphical and table reporting methods suggested ways to support easy-to-understand interpretations without resorting to complex statistical analysis. The current emphasis on a few standard charts may point to a potential knowledge gap for a wide variety of custom visualizations used in the wild.

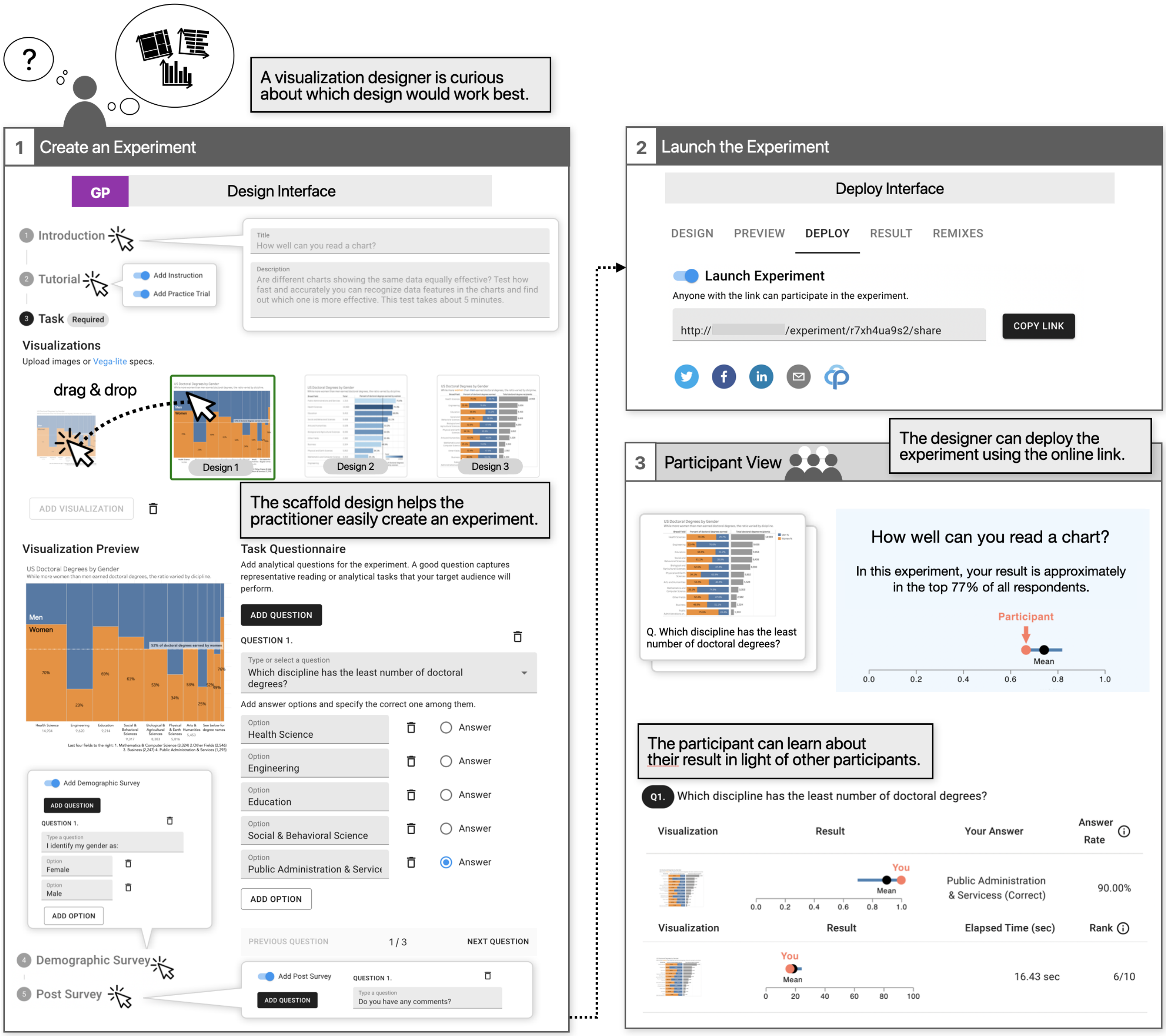
5 DESIGN GOALS

Based on the practitioner perception survey and empirical research analysis, we derived the following high-level design goals to guide the development of VisLab.

D1. Support gathering quantitative design feedback. Qualitative feedback provides a holistic diagnosis, while quantitative feedback can help funnel successes or failures in a particular context. While the preliminary survey indicated empirical methods would be helpful for visualization evaluation, existing feedback systems currently do not support gathering such quantitative feedback. We aim to adopt empirical processes currently reserved for scientific investigation to supplement the current design practice, enabling visualization practitioners to gather empirically-informed, quantitative design feedback. On the other hand, the empirical study analysis suggests a potential lack of design knowledge for extensive and custom visualization types available in the wild. Being able to investigate a personal design space on their own might help alleviate the knowledge gap.

D2. Streamline the process of running an evaluation experiment. The survey participants indicated varied experiences from design to management and analytics in data visualization. To run an experiment without expertise in experiment design, statistical analysis, and computer programming, they would need a scaffold on which they can easily create their own experiments, receive responses from participants, and interpret the feedback outcome. We do not intend practitioners to invent new experimental methods or initiate a novel scientific investigation. Instead, we can leverage already established procedures that can be used to evaluate particular aspects of visualization designs. Providing off-the-shelf templates can reduce the complexity of the empirical process in which they only supply their own visualizations and datasets. The experiment should be easy to deploy and provide a meaningful motivation for external participants to ease recruitment.

D3. Foster visualization design knowledge sharing and extension. Existing experiments could provide valuable insights for others



A visualization designer is curious about which design would work best.

1 Create an Experiment

GP Design Interface

1 Introduction

2 Tutorial

3 Task Required

Title: How well can you read a chart?

Description: Are different charts showing the same data equally effective? Test how fast and accurately you can recognize data features in the charts and find out which one is more effective. This test takes about 5 minutes.

Add Instruction

Add Practice Trial

Visualizations

Upload images or Vega-lite specs.

ADD VISUALIZATION

The scaffold design helps the practitioner easily create an experiment.

Visualization Preview

US Doctoral Degrees by Gender

While more women than men earned doctoral degrees, the ratio varied by discipline.

Task Questionnaire

Add analytical questions for the experiment. A good question captures representative reading or analytical tasks that your target audience will perform.

ADD QUESTION

QUESTION 1.

Type or select a question: Which discipline has the least number of doctoral degrees?

Add answer options and specify the correct one among them.

Option: Health Science ○ Answer

Option: Engineering ○ Answer

Option: Education ○ Answer

Option: Social & Behavioral Science ○ Answer

Option: Public Administration & Servic ● Answer

ADD OPTION

PREVIOUS QUESTION 1 / 3 NEXT QUESTION

Add Demographic Survey

ADD QUESTION

QUESTION 1.

Type a question: I identify my gender as:

Option: Female

Option: Male

ADD OPTION

4 Demographic Survey

5 Post Survey

Add Post Survey

ADD QUESTION

QUESTION 1.

Type a question: Do you have any comments?

2 Launch the Experiment

Deploy Interface

DESIGN PREVIEW DEPLOY RESULT REMIXES

Launch Experiment

Anyone with the link can participate in the experiment.

http:// /experiment/r7xh4ua9s2/share

COPY LINK

The designer can deploy the experiment using the online link.

3 Participant View

Q. Which discipline has the least number of doctoral degrees?

How well can you read a chart?

In this experiment, your result is approximately in the top 77% of all respondents.

The participant can learn about their result in light of other participants.

Q1. Which discipline has the least number of doctoral degrees?

Visualization	Result	Your Answer	Answer Rate
		Public Administration & Servicess (Correct)	90.00%

Visualization	Result	Elapsed Time (sec)	Rank
		16.43 sec	6/10

Figure 4: VisLab’s workflow from design to deployment: (1) A designer creates an experiment using their visualizations based on the graphical perception template GP and (2) deploys the experiment using an online link. (3) Participants can see their result relative to the collective outcome.

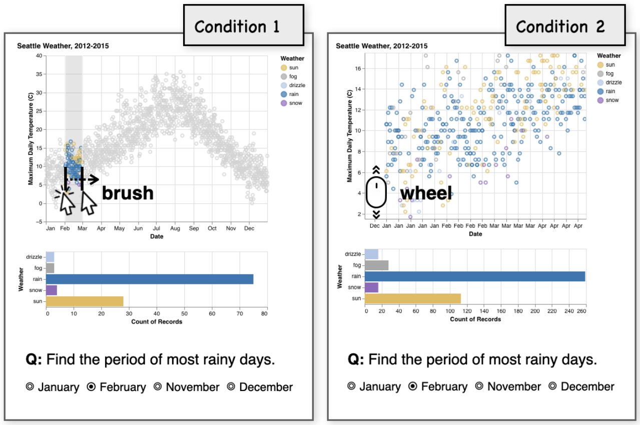
even if used to derive feedback for specific visualizations in particular contexts. In addition, since quantitative feedback is more explicit than tacit [34], it would be more amenable to sharing and extension. Drawing inspiration from existing community-oriented knowledge-sharing platforms [5, 12, 96], we seek to facilitate sharing and browsing experiment outcomes based on familiar design criteria such as charts, data types, and task types. Furthermore, we aim to support the remixing of existing experiments [96], which will help communities validate and extend the experiments [37].

6 VISLAB

We developed VisLab guided by the aforementioned design goals. Below, we describe the main components of VisLab based on a potential user workflow.

6.1 Deciding Which Experiment Template to Use (D1)

First, the user needs to decide which design aspect they are interested in, determining what template to use. VisLab currently



Q: Find the period of most rainy days.

○ January ● February ○ November ○ December

Q: Find the period of most rainy days.

○ January ● February ○ November ○ December

Figure 5: An example experiment using interactive charts, comparing two different interaction modalities for a search task: (left) *brush* and (right) *wheel* conditions.

supports three templates drawn from the existing literature: 1) graphical perception GP [41], 2) attention tracking AT [67], and 3) memorability MB [29]. Each template is designed to evaluate a specific design aspect. GP is based on the seminal experiment by Cleveland & McGill, where study participants perform elementary tasks such as perceiving different bar lengths or pie angles to extract and compare values. GP can be used to evaluate which design variation is perceptually effective on user-formulated reading questions (e.g., How much debt does the US have than Canada?). AT is based on the simulated eye-tracking experiment where participants were presented with a series of blurred images and asked to click to reveal small, circular areas of the image at original resolution, similar to having a confined area of focus like the eye fovea. AT can provide feedback on where the reader attends to based on the click patterns. Lastly, MB is based on the at-a-glance image recognition game where participants were presented with a sequence of images and had to press a key if an image appeared the second time in the sequence. MB can inform how well the reader can recognize a visualization among many filler images.

Rationale for the currently supported templates. We had several inclusion criteria for template candidates. First, the underlying empirical methods need to be useful for practical evaluation. Thus, we chose the initial three templates to cover a holistic visualization reading experience; a reader first perceives a chart GP, browses to extract its meaning AT, and stores the message in a brain MB. We also considered whether they are amenable to templatization; do they follow a standard procedure? And are they generally applicable to a wide variety of visualizations? Lastly, we favored empirical methods proven possible in online studies. These online studies also tend to have simpler setups for easy recruitment and participation (e.g., microtasking for crowdsourcing). Although we currently support three experiment scenarios, VisLab is extensible to support other types of evaluation experiments as long as they can fit the common standard procedure. We discuss possible extensions in the discussion section (Section 8).

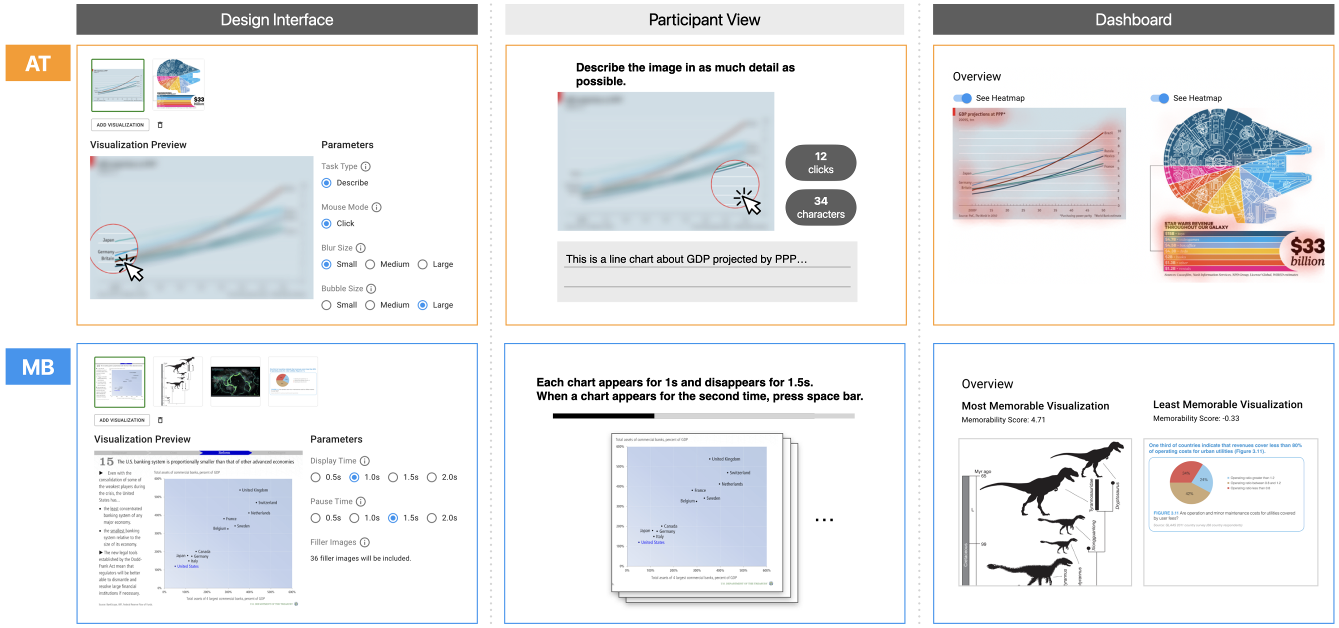
6.2 Creating an Experiment using the Selected Template (D1)

Once the user chooses a specific template, it creates a new experiment with default settings. The overall procedure is shared across all templates and is divided into clear steps so the user can easily understand and follow the experiment design process. There are currently five steps: introduction, tutorial, task, demographic survey, and post-survey (Figure 4). At the minimum, the user only needs to supply their visualizations in the task step, except for GP that requires additional task questions (e.g., "which discipline has the least number of doctoral degrees?").

Since each template generally determines what actions the participant will perform, we can pre-fill many default settings. Examples include the introductory text, tutorial description, practice trials, demographic survey questions (e.g., age, ethnicity, and gender), and post-survey questions (e.g., user preference). For instance, we use the default title *"How well can you read a chart?"* for GP, *"Where do you pay attention when reading a chart?"* for AT, and *"How well can you recognize a chart?"* for MB. For GP, we suggest generic analytical questions derived from the visualization literature, such as classification, recognition, localization, visual search, discrimination, identification, and estimation [50]. If desirable, the user can customize the default settings or skip certain steps, such as the practice trial and surveys; e.g., the post-survey can be customized to gather qualitative opinions.

In GP, the user can drag and drop visualizations into the thumbnail containers in the task *design* interface (Figure 4). VisLab supports an image file, as well as Vega-lite specification [98] that can support interactive visualizations. Figure 5 shows an example experiment using an interactive Vega-lite chart in which we ask participants to find the period of most rainy days using two different interaction modalities, including brushing and wheel-scrolling. For simplicity, we use a single-level between-subject design so that each participant sees a different visualization while the questions are the same across all participants. We do not expose these experimental design details to users. While we originally implemented ways to enable a full factorial experimental design, we decided to remove it at the moment for its complexity and potential impracticality. The GP template currently supports three types of questions—multiple-choice, multiple answers, and short-answer (number)—in order to support common tasks we found in the empirical study analysis, such as *identify*, *compare*, and *derive*.

Similarly, in MB, the user can upload target visualizations through drag-and-drop and configure image display & pause time (Figure 6); i.e., how long will each image be visible for a participant and how long will there be a gap time between consecutive images? The user can also customize filler (non-target) images, while we provide default filler images from the original study consisting of visualization images sampled from various sectors (e.g., charts from infographics, news media, government websites, and science journals). In AT, the user also adds stimuli in the same way, while configuring the bubble and blur sizes (Figure 6). They can preview how the visualization will be interactive upon changing the parameters. We also provide detailed descriptions of what the parameters mean in the information tooltips in all templates.



AT — Design Interface: Task Type ● Describe; Mouse Mode ● Click; Blur Size ● Small ○ Medium ○ Large; Bubble Size ○ Small ○ Medium ● Large

AT — Participant View: Describe the image in as much detail as possible.

This is a line chart about GDP projected by PPP...

MB — Design Interface: Display Time ○ 0.5s ● 1.0s ○ 1.5s ○ 2.0s; Pause Time ○ 0.5s ○ 1.0s ● 1.5s ○ 2.0s; Filler Images: 36 filler images will be included.

MB — Participant View: Each chart appears for 1s and disappears for 1.5s. When a chart appears for the second time, press space bar.

MB — Dashboard: Overview. Most Memorable Visualization, Memorability Score: 4.71. Least Memorable Visualization, Memorability Score: -0.33

Figure 6: Design, participant, and dashboard interfaces showing example AT and MB experiments using images from the original studies [29, 67].

6.3 Deploying the Experiment to the Public (D2)

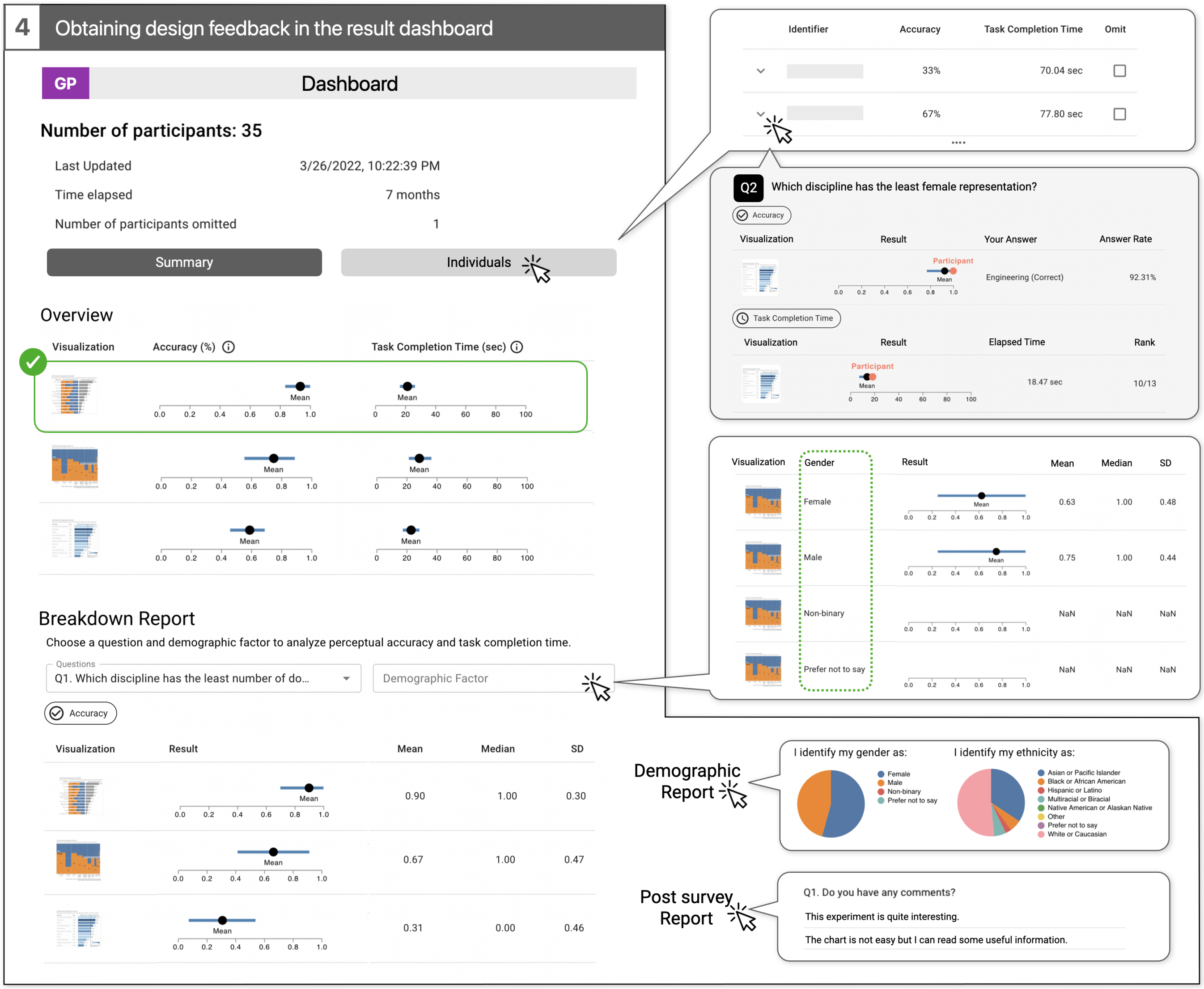
Once the user is satisfied with the experiment configuration, they can preview it by submitting a pilot response as necessary; they can omit it from the final result (Figure 7). Once they launch it to the public, VisLab generates an online link (Figure 4.2). The user can share the link with their colleagues or online communities, seeking voluntary participation. Alternatively, they can deploy the link on paid platforms such as Prolific and Amazon Mechanical Turk; the designer can embed the completion code in VisLab. VisLab does not require sign-up to lower the barrier for participation while using IP addresses to prevent duplicate participation.

To further protect against malicious participation, the user can add an attention question using the questionnaire editor in GP (e.g., obvious questions identifying the x-axis title). For MB, similar to the original study [29], we randomly choose vigilant images among fillers and check if participants failed to recognize a significant portion of them. In AT, the user can require participants to describe at least 150 characters and generate a minimum number of clicks as in the original study [67]. To encourage participation, VisLab provides a personalized result at the end of each experiment as a learning incentive (Figure 4.3), similar to the LabintheWild platform [93]. The participant can compare their performance against the aggregate performance of others.

6.4 Obtaining Design Feedback in the Result Dashboard (D2)

The user can monitor the current participation status of the experiment to see if anything goes wrong and inspect the final result in the analytic dashboard (Figure 7). The dashboard summarizes the results and shows individual responses for the designer to inspect and filter malicious participants. The user can explore the result based on the survey responses if there is a demographic survey or post-survey. VisLab provides basic descriptive statistics using plots and tables that can help derive design feedback. The dashboard presents the experiment result in a plain language, such as *“half of the participants took less than 7.8 seconds”* and *“around 19 out of 20 scored between 5% and 40%”*. These explanations appear when the user places a mouse cursor over the resulting charts. Also, we provide informational tooltips explaining how to read the charts (e.g., *“the circle is the average while the horizontal bar indicates a confidence interval”*). Likewise, the tooltips also convey how we derived specific metrics, such as the answer rate and rank in the participant result view.

The GP template measures the duration and accuracy of answering each question. For instance, when a designer uses two different visualizations with the same questionnaire, they can gain insights into which visual encoding might convey their intended information more accurately and potentially faster (Figure 7). AT aggregates all user clicks to generate a heatmap, providing feedback on where people would look. They can gauge the importance of design elements in their visualization (Figure 6), and compare attention patterns in multiple design alternatives. The MB template computes the same memorability score as the original paper [29], which is roughly defined by successful hit rate minus false alarm rate (Figure 6). We also present plain successful and false recognition rates for easy interpretation. The memorability score can hint at how well people would recall information after perceiving and



4 Obtaining design feedback in the result dashboard

GP Dashboard

Number of participants: 35

Last Updated	3/26/2022, 10:22:39 PM
Time elapsed	7 months
Number of participants omitted	1

Summary Individuals

Overview

Visualization	Accuracy (%)	Task Completion Time (sec)

Breakdown Report

Choose a question and demographic factor to analyze perceptual accuracy and task completion time.

Questions: Q1. Which discipline has the least number of do...

Demographic Factor

Accuracy

Visualization	Result	Mean	Median	SD
		0.90	1.00	0.30
		0.67	1.00	0.47
		0.31	0.00	0.46

Identifier	Accuracy	Task Completion Time	Omit
	33%	70.04 sec	□
	67%	77.80 sec	□

Q2 Which discipline has the least female representation?

Accuracy

Visualization	Result	Your Answer	Answer Rate
		Engineering (Correct)	92.31%

Task Completion Time

Visualization	Result	Elapsed Time	Rank
		18.47 sec	10/13

Visualization	Gender	Result	Mean	Median	SD
	Female		0.63	1.00	0.48
	Male		0.75	1.00	0.44
	Non-binary		NaN	NaN	NaN
	Prefer not to say		NaN	NaN	NaN

Demographic Report

I identify my gender as: I identify my ethnicity as:

Post survey Report

Q1. Do you have any comments?

This experiment is quite interesting.

The chart is not easy but I can read some useful information.

Figure 7: An overview of the dashboard interface for the GP template. The user can examine the overall comparison results and dive into the details broken down by individual questions and demographic factors. In addition, they can review the participants' demographic information and their qualitative comments in the post-survey. The experimenter may inspect individual responses to remove potential outliers and malicious participation. The dashboard interface and layout are consistent with other templates.

comprehending it [28]. Thus, the designer might use the three templates in the above order to have a holistic view of the experience of potential readers of their visualizations.

6.5 Exploring and Remixing Existing Experiments (D3)

Once the user finishes the experiment, they may share the outcome with other users. The user can decide whether to share only the result, the content, or both (Figure 8). Other users can inspect the experiment's visualizations and other configuration parameters and settings if the content is shared (similar to sharing underlying code in addition to the website, such as in Glitch [5]). Otherwise, they can only examine the final result in the dashboard. The user can also assign useful tags to categorize the experiment (Figure 8). VisLab provides predefined tags in three high-level themes, including chart type, data type, and task type. We leverage the vocabularies found in the empirical study analysis to suggest potential relevant tags such as *"scatter plot"*, *"time-series"*, and *"compare"*. The users can then browse all shared experiments based on the tags, e.g., finding all experiments relevant to *"bar chart"*.

In addition, users can remix existing experiments to create new experiments for reproduction or extension if the experiment content is shared. For instance, another user might be interested in using the same visualizations but with different data distributions and embellishments or asking additional questions to see how the outcome might change (Figure 8). The remix activity can promote user collaborations, producing valuable design insights as a community. VisLab keeps track of the remixes of each experiment so that the users can see the genealogy of an experiment if it was remixed, providing another lens to examine relevant knowledge in addition to the tag-based navigation.

7 EVALUATION

We conducted evaluation studies to assess the viability of experiment templates and also the usability and usefulness of VisLab in designers' visualization design practice.

7.1 Feasibility Study

First, we wanted to verify the feasibility of the templates in VisLab. In other words, although our templates are directly derived from the original studies, we wanted to double-check if we could get similar results in the context of VisLab's use case. Our experiments are simplified from the perspective of practical evaluation rather than trying to fully replicate the original studies, e.g., comparing a few competing designs rather than tens of thousands of repeated visualizations with randomly generated datasets. We provide the stimuli image in the supplement.

7.1.1 Stimuli, Tasks & Procedures. For GP, we used a bar chart drawn from the Guardian [18, 71] since Cleveland & McGill's original study used synthetic data. Given the bar chart, we created a pie chart alternative for comparison. For MB and AT, we downsampled images from the original studies [29, 67], available in the MassVis dataset [9]. For MB, we used two most memorable and two least memorable visualizations. For AT, we used two visualizations from the news media (horizontal, diverging stacked bar chart) and the infographic category respectively.

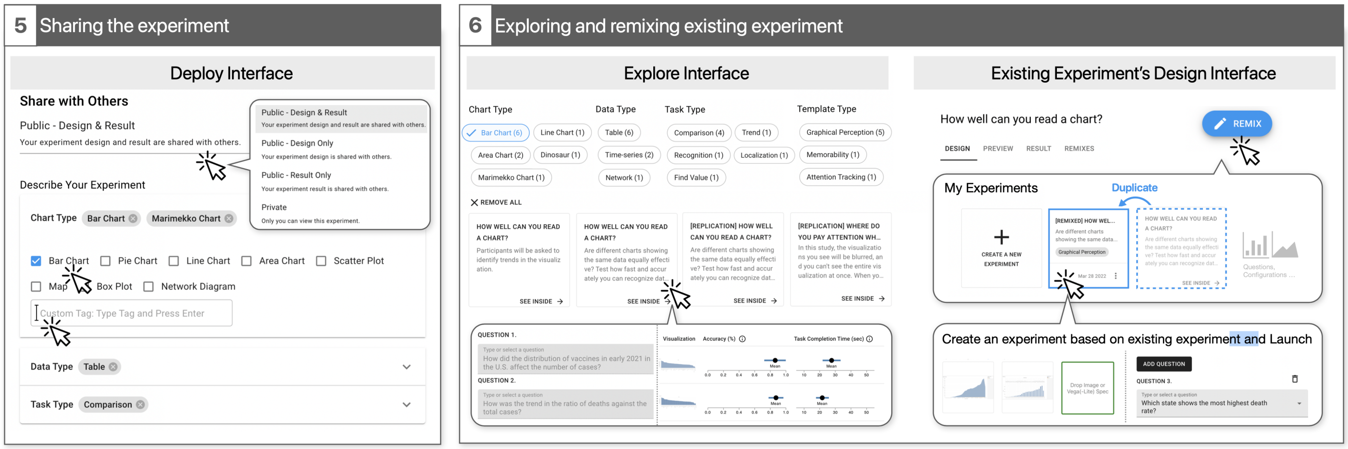
The tasks are also similar to the original studies. For GP, we asked two questions: one finding a specific target (*Which country has the sixth most donation amount?*) and another deriving the ratio of larger to smaller (*What is the percentage of donations from the European Commission compared to that of the U.S.?*). The first question was multiple-choice, while the second question was in a numeric text format. For AT, participants were asked to browse a blurred visualization image and describe it, as in the original study [67]. We used the bubble radius of 32 pixels and 40 blur sigma, which was found to be appropriate in the original study [67]. For MB, they were presented with a sequence of images and asked to press a space bar when they saw an image a second time. We generated the image sequence, following the same mechanism in the original study [29]; i.e., maintaining reasonable space among fillers and repeated targets. The size of the image sequence is variable based on the target size; in our case, it had 120 images, including 15 random vigilant images and four targets. We did not use images when their aspect ratio was greater than 3:1. The attention tracking and memorability experiments were within-subject design, while the graphical perception experiment was between-subjects. All experiments were deployed in Prolific [14].

7.1.2 Analysis Methods. For GP, we measured the accuracy and time taken for the questions. For the multiple-choice question, we used Fisher's exact test for bar chart and pie chart groups. For the numeric answer question, we computed absolute log error (distance from the participant's answer to the correct answer). We then performed Levene's test for equality of variances, followed by the independent t-test. For AT, we aggregated click-maps across all participants for each image. We then measured the cross-correlation (CC) between the aggregate click-map of the image and the ground truth fixation map of the same image. We qualitatively compared the CC score to the CC score in the original paper. For MB, we compute an image's hit rate, false alarm rate, and memorability score. The hit rate is defined as *hits/(hits+misses)*, while the false alarm rate is defined as *false alarms / (false alarms + correct rejections)*. The memorability score is *d' = Z(hit rate) - Z(false alarm rate)* where *Z* is the inverse cumulative Gaussian distribution. A higher memorability score requires a high hit rate and a low false alarm rate.

7.1.3 Results. We collected 25 responses for GP, 12 for the bar chart and 13 for the pie chart. We used the interquartile (IQR) range-based outer removal, filtering items out of the bound *[first quartile - 1.5 x IQR, third quartile + 1.5 x IQR]*. We filtered four outliers in the second question, three for the bar chart and one for the pie chart. We did not detect any outliers in the first question. Our result was largely consistent with the position-angle experiment in Cleveland & McGill. For the target finding task (first question), the bar chart had a lower accuracy ($M = 0.67, SD = 0.50$) and shorter task time ($M = 18.14sec, SD = 7.67$) compared to the accuracy ($M = 0.91, SD = 0.30$) and time ($M = 42.85sec, SD = 25.33$) for the pie chart. The accuracy was not significantly different according to Fisher's exact test ($p = 0.285$), while the time difference was significant ($t(20) = -2.811, p = 0.012$). For the ratio derivation task (second question), the bar chart was better in accuracy ($M = 0.60, SD = 2.80$) and lower in time ($M = 36.61sec, SD = 14.13$) in contrast to the accuracy ($M = 3.89, SD = 1.19$) and time ($M = 61.19sec, SD = 34.04$) of the pie chart. Unlike the first question, the accuracy (or error rate) had a significant difference ($t(20) = -3.547, p = 0.002$), although the time did not ($t(20) = -2.021, p = 0.058$).

The AT experiment had eleven participants. We removed one response whose number of clicks was less than 10. All other responses' clicks were within the interquartile outlier bound. Overall, our CC scores were substantially close to the original CC scores. The news media chart had an average number of clicks of 62.5 ($SD = 34.19$) and an average completion time of 171.6 sec ($SD = 38.61$). The CC score of our click map to the ground truth fixation map was 0.94, while the CC score of the original click map was 0.96. On the other hand, the infographic chart had an average click count of 33.4 ($SD = 13.02$) and an average task time of 183.9 ($SD = 78.37$). The CC score of our click map was 0.62, close to the CC score of the original click map, 0.60.

For MB, we collected 12 responses. We filtered one response whose false alarm rate was more than 50%. Overall, hit and false



Public - Design & Result
Your experiment design and result are shared with others.

Public - Design Only
Your experiment design is shared with others.

Public - Result Only
Your experiment result is shared with others.

Private
Only you can view this experiment.

HOW WELL CAN YOU READ A CHART?
Participants will be asked to identify trends in the visualization.

HOW WELL CAN YOU READ A CHART?
Are different charts showing the same data equally effective? Test how fast and accurately you can recognize dat...

[REPLICATION] HOW WELL CAN YOU READ A CHART?
Are different charts showing the same data equally effective? Test how fast and accurately you can recognize dat...

[REPLICATION] WHERE DO YOU PAY ATTENTION WH...
In this study, the visualizations you see will be blurred, and you can't see the entire visualization at once. When yo...

How did the distribution of vaccines in early 2021 in the U.S. affect the number of cases?

How was the trend in the ratio of deaths against the total cases?

How well can you read a chart?

[REMIXED] HOW WEL...
Are different charts showing the same data...

HOW WELL CAN YOU READ A CHART?
Are different charts showing the same data equally effective? Test how fast and accurately you can recognize dat...

Create an experiment based on existing experiment and Launch

Which state shows the most highest death rate?

Figure 8: Other users can inspect visualizations and configuration parameters within existing experiments and remix them for reproduction or extension as needed.

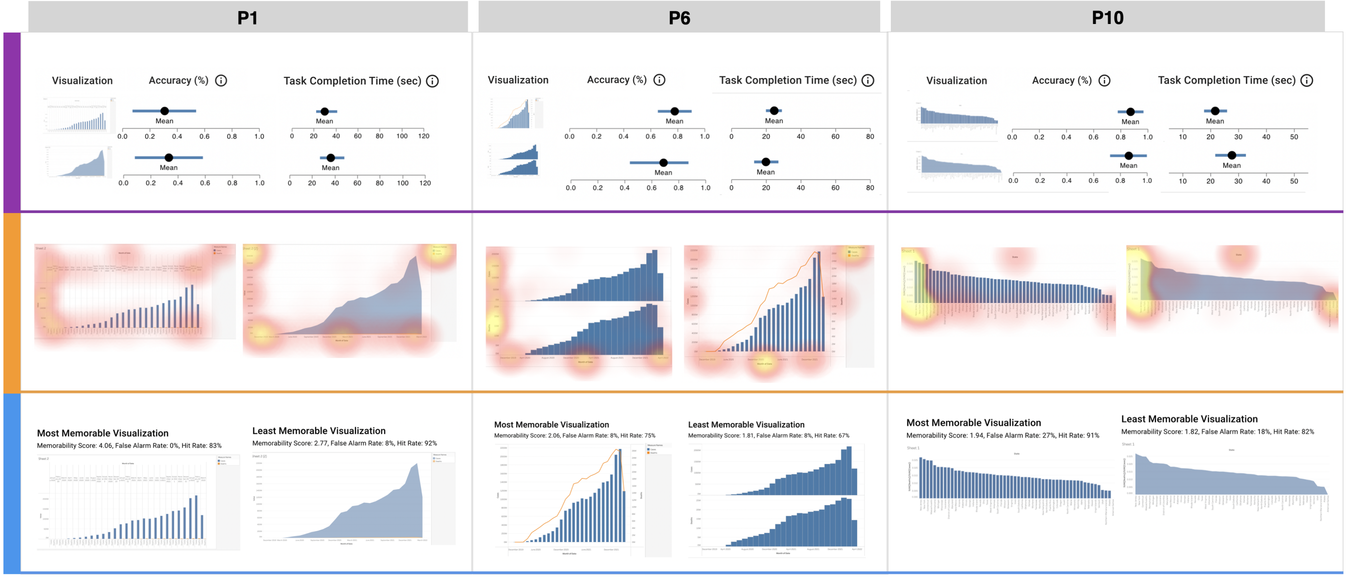


Figure 9: The experiments results of the visualizations created by the participants.

alarm rates were comparable to the original study to a great extent, as did the memorability scores. The most memorable visualization had a hit rate of 0.727 and a false alarm rate of 0.091, while the original study had 0.783 and 0, respectively. The second memorable visualization's hit rate and false alarm rate were 0.727 and 0.182, comparable to 0.810 and 0.010 from the original study. Likewise, the second least memorable visualization had a hit rate of 0.636 and a false alarm rate of 0.182 compared to 6.333 and 0.452 in the original study. The least memorable visualization had a 0.364 hit rate and 0.091 false alarm rate compared to 0.238 and 0.144. The ranking of the memorability scores was also consistent (ours: 1.94 → 1.51 → 1.26 → 0.99 vs. original: 3.87 → 3.20 → 0.46 → 0.35).

7.2 Usability Study

Our usability study had two phases. The first phase involved active exploration of the core features of VisLab using provided visualizations. In the second phase, selected participants created their own visualization experiments and experienced the browsing and remixing interfaces.

7.2.1 Participants. We recruited 10 participants (three female and seven male, self-identified) for the user study. Five (one female) were from the design practice survey (Figure 3) who agreed to participate in the user study, and the other five (two female) were from the authors' university alumni network. Six participants were visualization designers, among which three of them have 8-10 years of experience, and the other three have 4-6 years of experience. Two participants were data analysts, among which one had 4-6 years of experience, and the other had less than two years of experience. We also had one journalist with 8-10 years of experience and one technical writer with 2-4 years of experience. We paid participants a $50 gift card as compensation for the first phase and an extra $100 gift card to each of the three participants engaged in the second phase.

7.2.2 Procedure & Tasks. The first phase consisted of five steps: (1) tutorial, (2) replication, (3) reproduction, (4) survey, and (5) interview. The tutorial explained the three templates and demonstrated how to use them by creating example experiments. For GP, we adopted a use case scenario comparing three different bar charts: a grouped bar chart, a vertical stacked bar chart, and a horizontal stacked bar chart. For AT and MB, we borrowed sample images from original studies [29, 67] (Figure 6). The image stimuli can be found in the supplement. In the replication task, participants were asked to replicate the three example experiments in the tutorial to demonstrate their understanding of VisLab. In the reproduction task, participants created an experiment without any guidance from a moderator, based on a provided scenario comparing a bar and pie chart. Once they completed the tasks, they answered usability survey questions on a 5-point Likert scale, followed by a semi-structured interview discussing their overall experience and the potential use of VisLab in their work practice. Each participant engaged in a one-time remote session that took about 60 minutes.

In the second phase, we carried out a follow-up study with three selected participants (two females and one male) based on availability. The participants were a visualization designer with 8-10 years of experience, a data journalist with 8-10 years of experience, and a data analyst with 4-6 years of experience. We wanted to have a more realistic creation task and evaluate the result-sharing and exploration capability we added. We provided a simplified COVID-19 dataset [6], and participants had an asynchronous self-moderated session to create their two visualization variants and formulate questions to ask. We deployed their experiments in Prolific on their behalf. Once the experiments received enough responses, we had a synchronous remote session in which the participants went through the feedback result, publicized it, and explored other experiments. We also asked them to remix one of the existing experiments as well. We had an interview discussing the experience of the additional tasks. Each participant spent roughly two hours in the second phase.

7.2.3 Results. All participants successfully completed all the tasks in both phases. Each participant in the second phase created two charts and two questions to gather design feedback (Figure 9). P1 created a bar chart and an area chart with the following questions: *how did the distribution of vaccines in early 2021 in the U.S. affect the number of cases?* and *how was the trend in the ratio of deaths against the total cases?*. P6 made a dual-axis chart and a small-multiple chart to ask the following questions: *how has the death count changed across time by each state?* and *is this trend different from the cases count?*. P10 also created a bar chart and an area chart, along with the accompanying questions: *which state shows the highest death rate?* and *did Connecticut's death rate show higher than Maryland's?*. Below, we discuss our observations, survey results, and interview insights.

Templates lower barriers for a design feedback experiment. Participants highly rated VisLab's overall usability (M=4.33, SD=0.71), usefulness (M=4.50, SD=0.53), and learnability (M=4.50, SD=0.71) on a 5-point Likert scale (1 – strongly disagree, 5 – strongly agree). Among all templates, the highest-rated template in usability was AT (M=4.60, SD=0.52), followed by GP (M=4.30, SD=0.67), and MB (M=4.00, SD=0.94). On the other hand, the participants rated GP most useful (M=4.50, SD=0.71) compared to AT (M=4.00, SD=0.94) and MB (M=3.40, SD=1.26). For learnability, AT was also the highest (M=4.60, SD=0.52), while the mean scores for GP (M=4.40, SD=0.70) and MB (M=4.40, SD=0.97) were still comparable. A few participants liked the step-wise *design* interface. They said *"I liked that I can build the experiment easily according to the experiment flow."*—P1 and *"I liked the five steps of the design interface."*—P9. Ready-to-use templates with clear goals also seemed to help participants learn and use them easily. For instance, P7 commented that *"It was so intuitive that I could only drag and drop."* and P8 said *"I think it was easier and faster to learn because the functions with a clear purpose were provided as templates."*.

Dashboard eases the interpretation of design feedback. Participants said the resulting dashboard is useful for interpreting the outcome (M=4.90, SD=0.32). They commented *"It was easy to understand because it showed visual stats rather than just text stats."*—P6 and *"I think this is so good because otherwise, I have to analyze all the statistics one by one."*—P10. P6 also said that *"It is good that the order of information coincides with the order of thoughts. The conclusion is presented at the top, and detailed analysis results to support the conclusion are presented at the bottom."* P10 said *"The tooltip description of the graph is helpful for those with limited statistical knowledge."* They also commented that the breakdown report (e.g., by demographic factors) in the dashboard is valuable; as P6 said, *"I can analyze all the results here. I think I save my time and effort a lot."*. In the second phase, the participants also derived design improvement ideas from the dashboard. For instance, P1 said, *"[In the bar vs. area chart experiment with COVID-19 dataset], the bar chart might be better than the area chart, and I want to compare it once more with the visualization with a guide bar indicating the timing of vaccination on March 2021."* P6 said, *"[In attention tracking experiment of the dual axis chart vs. small-multiple chart] I see that people looked around titles, axes, and legends a lot. I might have to pay more attention to their legibility, like font size."*

Quantitative format can help gather feedback and enhance current design practice. Participants indicated that VisLab's quantitative format would make it easier for people to give design feedback. For example, P6 said *"people usually don't know what they want, so it's hard to say which part should be better. I think the way that answers the questions makes it easier to provide feedback."* P10 commented, *"It is useful to obtain quantitative indicators of which charts were good and which specific parts of the chart were important."* The participants contrasted quantitative feedback from anonymous users with qualitative feedback from known colleagues. P1 said *"It's quick and easy to ask my colleagues, but I think there's a high risk of biased*

[6] https://github.com/nytimes/covid-19-data/blob/master/us-states.csv

feedback because of the limited pool of background or chart literacy. On the other hand, this kind of experiment can easily get feedback from the general public." P2 mentioned, *"It might be more objective if I ask someone I don't know than ask directly."*

They also discussed how VisLab would fit in their current design process. P4 commented *"I have never seen such a tool where this process has been implemented. [...] I will use it right before preparing some key presentations. It might be very helpful."* Similarly, P6 said, *"I need to evaluate and select charts from an objective perspective to communicate effectively with my colleagues. I think I can use this tool [to decide which chart is appropriate] when I present the results of my data analysis."* On the other hand, P10 commented on using it to test ideas in the prototyping stage, saying *"I initially preferred the area chart as I believed its representation would more vividly convey the significance of how many people died [because of COVID-19]. Looking at the design feedback result, the bar chart seems better and faster to read."*

Result sharing and experiment remixing promote knowledge sharing. Participants in the second phase commented on the potential benefits of sharing and remixing in generating future experiment ideas and expanding their visualization design knowledge. P1 said, *"It can be difficult to compose multiple-choice questions for visualizations, but the remixing can provide great starting points."* P10 said *"Even comparing within the same chart may yield different results depending on the tasks. So I could expand my knowledge by looking at other similar experiments."* P6 commented that the result-sharing feature also contributes to building a visualization literacy resource by saying, *"I could say to the junior designer that it will be helpful to see what I made on VisLab. This archive would help to spread knowledge easily. When new employees come in, they can start learning from the archive."* Nevertheless, participants also indicated they might hesitate to share publicly as their experiments might involve private internal data. But, they might still want to share some results for intrinsic rewards such as recognition; as P10 commented, *"[I will share] when I want to show off my best outcomes, like uploading music to Soundcloud, or I'm curious what people think of my experiment results."*

8 DISCUSSION

The user study suggests that VisLab is promising to help practitioners gather useful, quantitative feedback to improve their visualization design. Below, we reflect on the lessons learned from building and evaluating VisLab and discuss limitations and opportunities for future work.

8.1 Limitations and Opportunities

Experiment Templates. Our templates are currently limited by the scope of the three original studies. While they cover various general evaluation use cases from perception to cognition, there are many possibilities to bring other useful templates. VisLab's internal architecture is designed to adopt an additional template as a plug-in component. It only requires the task and dashboard interfaces while reusing other elements such as tutorial and post-survey. Potential additional templates, which might be helpful in practice, include evaluating discriminability (e.g., correlation judgement [55], color difference [53]) and assessing working memory [26].

As P5 noted that *"I use aesthetics to make it more playful and user friendly and to draw people in."*, the post-survey could include a default questionnaire to gather "qualitative" feedback and to evaluate subjective design dimensions, including aesthetics and engagement [65]. Likewise, VisLab could further extend existing templates to support other visualization forms (e.g., embedding external charts such as Tableau and D3) and tasks. For instance, P1 commented on a potentially additional task, *"I think there will be situations where attention tracking tasks also require multiple-choice questions to get more specific feedback than where participants paid attention to the visualization."*. While our templates ease the overall process, we noted in the second phase that it could still be difficult for users to develop meaningful analytical questions. Investigating the literature in graphical perception can help expand our current limited set of default questions in GP.

Dashboard. Our current dashboard design was inspired by how experiment results are reported in the literature. In the second phase of the user study, our dashboard helped participants derive design feedback from the results of their experiments. Overall, the participants were satisfied with the scores and distribution on the *dashboard*, though some offered new suggestions. For example, they mentioned explicitly presenting clear messages written in natural language to directly instruct design feedback. Presenting a likelihood percentage in addition to descriptive statistics might be more understandable for users [45]. In addition, a few participants noted that they wanted to do further analysis beyond what we provide in the *dashboard*. P3 stated *"I wanted to perform a multivariate analysis as granular as possible using the multiple demographic factors."* P7 said *"I'm always going to have some weird questions, and so that's where the raw data is going to be useful. I can go and visualize it myself."* After the study, we added data exports to VisLab to meet these needs.

Recruitment. While VisLab provides various strategies to assist recruitment, including the shareable experiment link and learning benefits for participants, the recruitment procedure is mainly left up to the users. While affluent and experienced users might use private dedicated channels or paid platforms to recruit appropriate participants, other lay users might still find the recruitment difficult. In particular, P3 also indicated a potential need for reaching multi-cultural audiences, *"...how you choose to represent those colors are going to have cultural meanings that the author's intent and the audience interpretation and may not align."*. Similar to existing volunteer-based platforms such as LabintheWild [93], VisLab could maintain its own recruitment channels, such as a social media page and mailing list. Showing other experiments on the participant result page (Figure 4.3) might nourish additional curiosity [78] for participants to engage in the additional experiments [40]. Furthermore, fostering social relationships around VisLab can lead to reproductive benefits by encouraging users to participate in each other's experiments [106]. While heterogeneous participation can be beneficial for external validity, it would be helpful to incorporate advanced outlier detection and attention-checking mechanisms in the future and manage multiple target audiences.

Process & Sharing. An experiment in VisLab is currently a one-off, while the user might go through a series of design variations and multiple stages of design feedback. It would be beneficial to provide ways to connect related experiments in the iterative process. Being able to categorize associated experiments can be helpful for sharing the outcomes with coworkers or the public, as P6 suggested the need for curating all experiment results as an organizational effort for training (e.g., minimizing repeating mistakes) and collaboration.

Participants also suggested expanding categories for navigating publicized experiments, such as based on topics and target audiences. While VisLab supports custom tags, providing such predefined tags can help curate the knowledge base and support more structured navigation. P10 even indicated the possibility of VisLab as a forum or community where people share visualization design ideas.

The current remixing is in its simplest form, and potential modifications include changing questions and parameters. P10 suggested the ability to modify existing visualizations in order to make the remixing more meaningful. Since VisLab supports Vega-lite specifications [98], a future version could allow modifying the specs directly in the experiment design interface.

Qualitative Feedback & Practical Experimentation. VisLab focuses on quantitative feedback as it is currently a missed opportunity for practitioners. That said, combining both types of feedback would benefit practitioners the most. Currently, the post-survey questionnaire can provide a way to collect qualitative feedback. However, adopting ideas from past qualitative feedback systems [66, 74, 79, 108, 110] can offer more structured methods and improve the holistic quality of the feedback.

Our development goal for VisLab was to find a minimum viable product, considering the practicality and complexity of the system for practitioners. For instance, we dropped the feature supporting multi-level mixed design for GP. We wanted to make VisLab accessible for those unfamiliar with empirical methods and certainly did not intend VisLab to be a research tool but rather a quick turnaround feedback tool. While out of scope for this work, it would be interesting to investigate further the level of user control & freedom best suited for practical experimentation and explore the tension and trade-off between feedback and science.

8.2 Fostering a Community around Visualization Design Knowledge

Practitioners, researchers, and educators constitute one holistic visualization community. While there are constant efforts to meet and communicate with each other, it is known that there exist knowledge gaps among them [39, 90]. VisLab could be used by practitioners to help fill in missing knowledge in practice when there are no other resources available. For instance, one participant (P7) indicated evaluating the applicability of existing guidelines in their own context and visualization by saying *"There's a lot of guidelines in there that I think is correct, but I do not have a good way to test them, so this would be an excellent tool for testing those things."*

In addition, experiments created and curated in VisLab can provide insights into what practitioners might be interested in, offering useful initial hypotheses for researchers to test in rigorous lab environments and produce generalizable knowledge. Moreover, VisLab can provide educational opportunities for visualization learners by gaining hands-on experience in visualization experiments [15] as a starting point to understand the science behind visualization design. In this manner, exploring ways to bring practitioners, researchers, and educators together to promote visualization design knowledge as collaborative efforts would be valuable in the future.

9 CONCLUSION AND FUTURE WORK

In this work, we presented VisLab, an interactive system that enables designers to run visualization experiments and derive empirically-driven design feedback. Our perception survey revealed their genuine interest in empirical methods for obtaining quantitative feedback despite a lack of awareness and experience. We also derived a standardized procedure and representative experiment templates based on the literature review of existing empirical studies in the visualization field. Our user study demonstrates that VisLab's templates and dashboard make it easy to get informative insights for their visualization designs. For future work, we plan to deploy VisLab in the wild, along with detailed tutorials and various example experiments.

REFERENCES

[1] Accessed: September 9, 2022. Approaches to displaying Likert-scale data. https://www.surveymonkey.com/r/STLFN23.

[2] Accessed: September 9, 2022. Bullet graph. https://en.wikipedia.org/wiki/Bullet_graph.

[3] Accessed: September 9, 2022. Circular Tube Chart Twitter Discussion. https://twitter.com/questionsinDV/status/1480532267084754952.

[4] Accessed: September 9, 2022. Data Visualization Society. https://www.datavisualizationsociety.com/.

[5] Accessed: September 9, 2022. Glitch. https://glitch.com/.

[6] Accessed: September 9, 2022. Gorilla. https://gorilla.sc/.

[7] Accessed: September 9, 2022. Gorilla. https://www.qualtrics.com/.

[8] Accessed: September 9, 2022. Heatmap Twitter Discussion. https://twitter.com/FILWD/status/1303112447163879426.

[9] Accessed: September 9, 2022. MASSVIS Dataset. http://massvis.mit.edu/.

[10] Accessed: September 9, 2022. Misleading Color Schemes Twitter Discussion. https://twitter.com/NicholasDanfort/status/1484587593182420992.

[11] Accessed: September 9, 2022. Muller plot. https://en.wikipedia.org/wiki/Muller_plot.

[12] Accessed: September 9, 2022. Observable. https://observablehq.com/.

[13] Accessed: September 9, 2022. Pie charts are evil. https://twitter.com/storywithdata/status/1262764077690105856.

[14] Accessed: September 9, 2022. Prolific. https://www.prolific.co/.

[15] Accessed: September 9, 2022. The science behind good charts. https://ig.ft.com/science-of-charts/.

[16] Accessed: September 9, 2022. Sequence Logo Twitter Discussion. https://twitter.com/FILWD/status/1213520691489857538.

[17] Accessed: September 9, 2022. Stack Bar Chart Twitter Discussion. https://twitter.com/VictimOfMaths/status/1514220648524046340.

[18] Accessed: September 9, 2022. Syrian refugees: how many are there and where are they? http://www.guardian.co.uk/news/datablog/2013/mar/06/syrian-refugee-crisis-in-numbers.

[19] Accessed: September 9, 2022. Tornado Chart Twitter Discussion. https://twitter.com/emmawage/status/1255172980788785152.

[20] Accessed: September 9, 2022. VisPerception. http://visperception.com/.

[21] Accessed: September 9, 2022. What do you see? http://graphs.labinthewild.org/.

[22] Accessed: September 9, 2022. Why You Only Need to Test with 5 Users. https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/.

[23] Alfie Abdul-Rahman, Min Chen, and David H. Laidlaw. 2020. *A Survey of Variables Used in Empirical Studies for Visualization.* Springer International Publishing, Cham, 161–179. https://doi.org/10.1007/978-3-030-34444-3_7

[24] Scott Bateman, Regan L. Mandryk, Carl Gutwin, Aaron Genest, David McDine, and Christopher Brooks. 2010. Useful Junk? The Effects of Visual Embellishment on Comprehension and Memorability of Charts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 2573–2582. https://doi.org/10.1145/1753326.1753716

[25] Lindsay Betzendahl. 2019. Don't Mekko with My Marimekko. https://vizzendata.com/2019/10/18/dont-mekko-with-my-marimekko/. [Online; posted 9-September-2022].

[26] Rita Borgo, Alfie Abdul-Rahman, Farhan Mohamed, Philip W. Grant, Irene Reppa, Luciano Floridi, and Min Chen. 2012. An Empirical Study on Using Visual Embellishments in Visualization. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (dec 2012), 2759–2768. https://doi.org/10.1109/TVCG.2012.197

[27] R. Borgo, L. Micallef, B. Bach, F. McGee, and B. Lee. 2018. Information Visualization Evaluation Using Crowdsourcing. *Computer Graphics Forum* 37, 3 (2018), 573–595. https://doi.org/10.1111/cgf.13444

[28] Michelle A. Borkin, Zoya Bylinskii, Nam Wook Kim, Constance May Bainbridge, Chelsea S. Yeh, Daniel Borkin, Hanspeter Pfister, and Aude Oliva. 2016. Beyond Memorability: Visualization Recognition and Recall. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (2016), 519–528. https://doi.org/10.1109/TVCG.2015.2467732

[29] Michelle A. Borkin, Azalea A. Vo, Zoya Bylinskii, Phillip Isola, Shashank Sunkavalli, Aude Oliva, and Hanspeter Pfister. 2013. What Makes a Visualization Memorable? *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 2306–2315. https://doi.org/10.1109/TVCG.2013.234

[30] Matthew Brehmer and Tamara Munzner. 2013. A Multi-Level Typology of Abstract Visualization Tasks. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 2376–2385. https://doi.org/10.1109/TVCG.2013.124

[31] Michael Buhrmester, Tracy Kwang, and Samuel D. Gosling. 2011. Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science* 6, 1 (2011), 3–5. https://doi.org/10.1177/1745691610393980 PMID: 26162106.

[32] Alberto Cairo. 2015. Redesigning a circular timeline. http://www.thefunctionalart.com/2015/02/redesigning-circular-timeline.html. [Online; posted 20-June-2022].

[33] Sheelagh Carpendale. 2008. *Evaluating Information Visualizations.* Springer Berlin Heidelberg, Berlin, Heidelberg, 19–45. https://doi.org/10.1007/978-3-540-70956-5_2

[34] Murillo V. H. B. Castro, Monalessa P. Barcellos, Ricardo de A. Falbo, and Simone D. Costa. 2021. Using Ontologies to Aid Knowledge Sharing in HCI Design. In *Proceedings of the XX Brazilian Symposium on Human Factors in Computing Systems* (Virtual Event, Brazil) *(IHC '21)*. Association for Computing Machinery, New York, NY, USA, Article 50, 7 pages. https://doi.org/10.1145/3472301.3484327

[35] Min Chen, Alfie Abdul-Rahman, and David H. Laidlaw. 2020. The Huge Variable Space in Empirical Studies for Visualization – A Challenge as well as an opportunity for Visualization Psychology. https://doi.org/10.48550/ARXIV.2009.13194

[36] Min Chen, Georges Grinstein, Chris R. Johnson, Jessie Kennedy, and Melanie Tory. 2017. Pathways for Theoretical Advances in Visualization. *IEEE Computer Graphics and Applications* 37, 4 (2017), 103–112. https://doi.org/10.1109/MCG.2017.3271463

[37] David S. Chester and Emily N. Lasko. 2021. Construct Validation of Experimental Manipulations in Social Psychology: Current Practices and Recommendations for the Future. *Perspectives on Psychological Science* 16, 2 (2021), 377–395. https://doi.org/10.1177/1745691620950684 PMID: 32975479.

[38] Hichang Cho, MeiHui Chen, and Siyoung Chung. 2010. Testing an integrative theoretical model of knowledge-sharing behavior in the context of Wikipedia. *Journal of the American Society for Information Science and Technology* 61, 6 (2010), 1198–1212. https://doi.org/10.1002/asi.21316

[39] Jinhan Choi, Changhoon Oh, Bongwon Suh, and Nam Wook Kim. 2021. Toward a Unified Framework for Visualization Design Guidelines. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–7.

[40] Robert B. Cialdini and Noah J. Goldstein. 2004. Social Influence: Compliance and Conformity. *Annual Review of Psychology* 55, 1 (2004), 591–621. https://doi.org/10.1146/annurev.psych.55.090902.142015 PMID: 14744228.

[41] William S Cleveland and Robert McGill. 1984. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American statistical association* 79, 387 (1984), 531–554. https://doi.org/10.2307/2288400

[42] William S. Cleveland and Robert McGill. 1985. Graphical Perception and Graphical Methods for Analyzing Scientific Data. *Science* 229, 4716 (1985), 828–833. https://doi.org/10.1126/science.229.4716.828

[43] Michael Correll and Michael Gleicher. 2014. Error Bars Considered Harmful: Exploring Alternate Encodings for Mean and Error. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (2014), 2142–2151. https://doi.org/10.1109/TVCG.2014.2346298

[44] Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. 2018. Quality Control in Crowdsourcing: A Survey of Quality Attributes, Assessment Techniques, and Assurance Actions. *ACM Comput. Surv.* 51, 1, Article 7 (jan 2018), 40 pages. https://doi.org/10.1145/3148148

[45] Nediyana Daskalova, Jina Yoon, Yibing Wang, Cintia Araujo, Guillermo Beltran, Nicole Nugent, John McGeary, Joseph Jay Williams, and Jeff Huang. 2020. SleepBandits: Guided Flexible Self-Experiments for Sleep. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376584

[46] Nick Desbarats. 2021. I've Stopped Using Box Plots. Should You? https://nightingaledvs.com/ive-stopped-using-box-plots-should-you/. [Online; posted 9-September-2022].

[47] Alexandra Diehl, Elif E. Firat, Thomas Torsney-Weir, Alfie Abdul-Rahman, Benjamin Bach, Robert Laramee, Renato Pajarola, and Min Chen. 2021. VisGuided: A Community-driven Approach for Education in Visualization. In *Eurographics 2021 - Education Papers*, Beatriz Sousa Santos and Gitta Domik (Eds.). The Eurographics Association. https://doi.org/10.2312/eged.20211003

[48] Michael Diehl and Wolfgang Stroebe. 1987. Productivity loss in brainstorming groups: Toward the solution of a riddle. *Journal of personality and social psychology* 53, 3 (1987), 497. https://doi.org/10.1037/0022-3514.53.3.497

[49] Evanthia Dimara, Anastasia Bezerianos, and Pierre Dragicevic. 2017. The Attraction Effect in Information Visualization. *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (2017), 471–480. https://doi.org/10.1109/TVCG.2016.2598594

[50] Madison A. Elliott, Christine Nothelfer, Cindy Xiong, and Danielle Albers Szafir. 2021. A Design Space of Vision Science Methods for Visualization Research. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2021), 1117–1127. https://doi.org/10.1109/TVCG.2020.3029413

[51] Niklas Elmqvist and Ji Soo Yi. 2012. Patterns for Visualization Evaluation. In *Proceedings of the 2012 BELIV Workshop: Beyond Time and Errors - Novel Evaluation Methods for Visualization* (Seattle, Washington, USA) *(BELIV '12)*. Association for Computing Machinery, New York, NY, USA, Article 12, 8 pages. https://doi.org/10.1145/2442576.2442588

[52] Holger Finger, Caspar Goeke, Dorena Diekamp, Kai Standvoß, and Peter König. 2017. LabVanced: a unified JavaScript framework for online studies. In *International Conference on Computational Social Science (Cologne)*. https://www.labvanced.com/static/2017_IC2S2_LabVanced.pdf

[53] Connor C. Gramazio, David H. Laidlaw, and Karen B. Schloss. 2017. Colorgorical: Creating discriminable and preferable color palettes for information visualization. *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (2017), 521–530. https://doi.org/10.1109/TVCG.2016.2598918

[54] Steve Haroz, Robert Kosara, and Steven L. Franconeri. 2015. ISOTYPE Visualization: Working Memory, Performance, and Engagement with Pictographs. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 1191–1200. https://doi.org/10.1145/2702123.2702275

[55] Lane Harrison, Fumeng Yang, Steven Franconeri, and Remco Chang. 2014. Ranking Visualizations of Correlation Using Weber's Law. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (2014), 1943–1952. https://doi.org/10.1109/TVCG.2014.2346979

[56] John Hattie and Helen Timperley. 2007. The Power of Feedback. *Review of Educational Research* 77, 1 (2007), 81–112. https://doi.org/10.3102/003465430298487

[57] Saskia Haug and Alexander Mädche. 2021. Crowd-Feedback in Information Systems Development: A State-of-the-Art Review. In *ICIS 2021 Proceedings*. Association for Information Systems (AIS). https://doi.org/10.5445/IR/1000139669

[58] Shiqing He and Eytan Adar. 2017. VizItCards: A Card-Based Toolkit for Infovis Design Education. *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (2017), 561–570. https://doi.org/10.1109/TVCG.2016.2599338

[59] Marti A. Hearst, Paul Laskowski, and Luis Silva. 2016. Evaluating Information Visualization via the Interplay of Heuristic Evaluation and Question-Based Scoring. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5028–5033. https://doi.org/10.1145/2858036.2858280

[60] 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* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 203–212. https://doi.org/10.1145/1753326.1753357

[61] Paul Hendriks. 1999. Why share knowledge? The influence of ICT on the motivation for knowledge sharing. *Knowledge and Process Management* 6, 2 (1999), 91–100. https://doi.org/10.1002/(SICI)1099-1441(199906)6:2<91::AID-KPM54>3.0.CO;2-M

[62] Anita Holdcroft. 2007. Gender bias in research: how does it affect evidence based medicine? *Journal of the Royal Society of Medicine* 100, 1 (2007), 2–3. https://doi.org/10.1177/014107680710000102 PMID: 17197669.

[63] Eli Holder. 2020. Settling the Debate: Bars vs Lollipops (vs Dot Plots). https://3iap.com/bar-graphs-vs-lollipop-charts-vs-dot-plots-experiment-PP8-qapwQe2fRBJu1-ADfA/. [Online; posted 9-September-2022].

[64] Julie Hui, Amos Glenn, Rachel Jue, Elizabeth Gerber, and Steven Dow. 2015. Using Anonymity and Communal Efforts to Improve Quality of Crowdsourced Feedback. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 3, 1 (Sep. 2015), 72–82. https://ojs.aaai.org/index.php/HCOMP/article/view/13229

[65] Ya-Hsin Hung and Paul Parsons. 2017. Assessing User Engagement in Information Visualization. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI EA '17)*. Association for Computing Machinery, New York, NY, USA, 1708–1717.

https://doi.org/10.1145/3027063.3053113

[66] Hyeonsu B. Kang, Gabriel Amoako, Neil Sengupta, and Steven P. Dow. 2018. Paragon: An Online Gallery for Enhancing Design Feedback with Visual Examples. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3174180

[67] Nam Wook Kim, Zoya Bylinskii, Michelle A. Borkin, Krzysztof Z. Gajos, Aude Oliva, Fredo Durand, and Hanspeter Pfister. 2017. BubbleView: An Interface for Crowdsourcing Image Importance Maps and Tracking Visual Attention. *ACM Trans. Comput.-Hum. Interact.* 24, 5, Article 36 (nov 2017), 40 pages. https://doi.org/10.1145/3131275

[68] Younghoon Kim and Jeffrey Heer. 2018. Assessing Effects of Task and Data Distribution on the Effectiveness of Visual Encodings. *Computer Graphics Forum* 37, 3 (2018), 157–167. https://doi.org/10.1111/cgf.13409

[69] Yea-Seul Kim, Katharina Reinecke, and Jessica Hullman. 2017. Explaining the Gap: Visualizing One's Predictions Improves Recall and Comprehension of Data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 1375–1386. https://doi.org/10.1145/3025453.3025592

[70] Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing User Studies with Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08)*. Association for Computing Machinery, New York, NY, USA, 453–456. https://doi.org/10.1145/1357054.1357127

[71] Nicholas Kong, Marti A. Hearst, and Maneesh Agrawala. 2014. Extracting References between Text and Charts via Crowdsourcing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 31–40. https://doi.org/10.1145/2556288.2557241

[72] R. Kosara. 2007. Visualization Criticism - The Missing Link Between Information Visualization and Art. In *2013 17th International Conference on Information Visualisation*, Vol. 1. IEEE Computer Society, Los Alamitos, CA, USA, 631–636. https://doi.ieeecomputersociety.org/10.1109/IV.2007.130

[73] Robert Kosara and Jock Mackinlay. 2013. Storytelling: The Next Step for Visualization. *Computer* 46, 5 (2013), 44–50. https://doi.org/10.1109/MC.2013.36

[74] Markus Krause, Tom Garncarz, JiaoJiao Song, Elizabeth M. Gerber, Brian P. Bailey, and Steven P. Dow. 2017. Critique Style Guide: Improving Crowdsourced Design Feedback with a Natural Language Model. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 4627–4639. https://doi.org/10.1145/3025453.3025883

[75] Heidi Lam, Enrico Bertini, Petra Isenberg, Catherine Plaisant, and Sheelagh Carpendale. 2012. Empirical Studies in Information Visualization: Seven Scenarios. *IEEE Transactions on Visualization and Computer Graphics* 18, 9 (2012), 1520–1536. https://doi.org/10.1109/TVCG.2011.279

[76] Bongshin Lee, Eun Kyoung Choe, Petra Isenberg, Kim Marriott, and John Stasko. 2020. Reaching Broader Audiences With Data Visualization. *IEEE Computer Graphics and Applications* 40, 2 (2020), 82–90. https://doi.org/10.1109/MCG.2020.2968244

[77] Jun Liu and Sudha Ram. 2011. Who Does What: Collaboration Patterns in the Wikipedia and Their Impact on Article Quality. *ACM Trans. Manage. Inf. Syst.* 2, 2, Article 11 (jul 2011), 23 pages. https://doi.org/10.1145/1985347.1985352

[78] George Loewenstein. 1994. The psychology of curiosity: A review and reinterpretation. *Psychological bulletin* 116, 1 (1994), 75. https://doi.org/10.1037/0033-2909.116.1.75

[79] Kurt Luther, Amy Pavel, Wei Wu, Jari-lee Tolentino, Maneesh Agrawala, Björn Hartmann, and Steven P. Dow. 2014. CrowdCrit: Crowdsourcing and Aggregating Visual Design Critique. In *Proceedings of the Companion Publication of the 17th ACM Conference on Computer Supported Cooperative Work &amp; Social Computing* (Baltimore, Maryland, USA) *(CSCW Companion '14)*. Association for Computing Machinery, New York, NY, USA, 21–24. https://doi.org/10.1145/2556420.2556788

[80] Kurt Luther, Jari-Lee Tolentino, Wei Wu, Amy Pavel, Brian P. Bailey, Maneesh Agrawala, Björn Hartmann, and Steven P. Dow. 2015. Structuring, Aggregating, and Evaluating Crowdsourced Design Critique. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work &amp; Social Computing* (Vancouver, BC, Canada) *(CSCW '15)*. Association for Computing Machinery, New York, NY, USA, 473–485. https://doi.org/10.1145/2675133.2675283

[81] Narges Mahyar, Sung-Hee Kim, and Bum Chul Kwon. 2015. Towards a taxonomy for evaluating user engagement in information visualization. In *Workshop on Personal Visualization: Exploring Everyday Life*, Vol. 3. 2. https://groups.cs.umass.edu/wp-content/uploads/sites/8/2018/08/IEEEVIS_engagement-taxonomy.pdf

[82] Winter Mason and Siddharth Suri. 2012. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior research methods* 44, 1 (2012), 1–23. https://doi.org/10.3758/s13428-011-0124-6

[83] Luana Micallef, Pierre Dragicevic, and Jean-Daniel Fekete. 2012. Assessing the Effect of Visualizations on Bayesian Reasoning through Crowdsourcing. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2536–2545. https://doi.org/10.1109/TVCG.2012.199

[84] Kate Moran. 2019. Usability Testing 101. https://www.nngroup.com/articles/usability-testing-101/. [Online; posted 9-September-2022].

[85] Jakob Nielsen. 2005. Putting A/B Testing in Its Place. https://www.nngroup.com/articles/putting-ab-testing-in-its-place/?lm=ab-testing-101&pt=youtubevideo. [Online; posted 9-September-2022].

[86] Jakob Nielsen and Rolf Molich. 1990. Heuristic Evaluation of User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Seattle, Washington, USA) *(CHI '90)*. Association for Computing Machinery, New York, NY, USA, 249–256. https://doi.org/10.1145/97243.97281

[87] Anshul Vikram Pandey, Anjali Manivannan, Oded Nov, Margaret Satterthwaite, and Enrico Bertini. 2014. The Persuasive Power of Data Visualization. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (2014), 2211–2220. https://doi.org/10.1109/TVCG.2014.2346419

[88] 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 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 1469–1478. https://doi.org/10.1145/2702123.2702608

[89] Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. 2010. Running experiments on amazon mechanical turk. *Judgment and Decision making* 5, 5 (2010), 411–419. https://doi.org/10.1037/t69659-000

[90] Paul Parsons. 2022. Understanding Data Visualization Design Practice. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2022), 665–675. https://doi.org/10.1109/TVCG.2021.3114959

[91] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. 2017. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology* 70 (2017), 153 – 163. https://doi.org/10.1016/j.jesp.2017.01.006

[92] Jason Radford, Andy Pilny, Ashley Reichelmann, Brian Keegan, Brooke Foucault Welles, Jefferson Hoye, Katherine Ognyanova, Waleed Meleis, and David Lazer. 2016. Volunteer Science: An Online Laboratory for Experiments in Social Psychology. *Social Psychology Quarterly* 79, 4 (2016), 376–396. https://doi.org/10.1177/0190272516675866

[93] Katharina Reinecke and Krzysztof Z. Gajos. 2015. LabintheWild: Conducting Large-Scale Online Experiments With Uncompensated Samples. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Vancouver, BC, Canada) *(CSCW '15)*. Association for Computing Machinery, New York, NY, USA, 1364–1378. https://doi.org/10.1145/2675133.2675246

[94] Nathalie Henry Riche, Christophe Hurter, Nicholas Diakopoulos, and Sheelagh Carpendale. 2018. *Data-driven storytelling*. CRC Press. https://doi.org/10.1201/9781315281575

[95] Jonathan C. Roberts, Chris Headleand, and Panagiotis D. Ritsos. 2016. Sketching Designs Using the Five Design-Sheet Methodology. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (2016), 419–428. https://doi.org/10.1109/TVCG.2015.2467271

[96] Ricarose Roque, Natalie Rusk, and Mitchel Resnick. 2016. *Supporting Diverse and Creative Collaboration in the Scratch Online Community*. Springer International Publishing, Cham, 241–256. https://doi.org/10.1007/978-3-319-13536-6_12

[97] D Royce Sadler. 1989. Formative assessment and the design of instructional systems. *Instructional science* 18, 2 (1989), 119–144. https://doi.org/10.1007/BF00117714

[98] Arvind Satyanarayan, Dominik Moritz, Kanit Wongsuphasawat, and Jeffrey Heer. 2017. Vega-Lite: A Grammar of Interactive Graphics. *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (2017), 341–350. https://doi.org/10.1109/TVCG.2016.2599030

[99] Michael Sedlmair, Miriah Meyer, and Tamara Munzner. 2012. Design Study Methodology: Reflections from the Trenches and the Stacks. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2431–2440. https://doi.org/10.1109/TVCG.2012.213

[100] Mark Stover. 2004. Making tacit knowledge explicit: The ready reference database as codified knowledge. *Reference services review* (2004). https://doi.org/10.1108/00907320410537685

[101] Anselm Strauss and Juliet M Corbin. 1997. *Grounded theory in practice*. Sage. https://us.sagepub.com/en-us/nam/grounded-theory-in-practice/book6165

[102] Uzma Haque Syeda, Prasanth Murali, Lisa Roe, Becca Berkey, and Michelle A. Borkin. 2020. Design Study "Lite" Methodology: Expediting Design Studies and Enabling the Synergy of Visualization Pedagogy and Social Good. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376829

[103] Danielle Albers Szafir. 2018. Modeling Color Difference for Visualization Design. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2018), 392–401. https://doi.org/10.1109/TVCG.2017.2744359

[104] Justin Talbot, Vidya Setlur, and Anushka Anand. 2014. Four Experiments on the Perception of Bar Charts. *IEEE Transactions on Visualization and Computer*

Graphics 20, 12 (2014), 2152–2160. https://doi.org/10.1109/TVCG.2014.2346320

[105] Fernanda Viégas and Martin Wattenberg. 2015. Design and Redesign in Data Visualization. https://medium.com/@hint_fm/design-and-redesign-4ab77206cf9. [Online; posted 9-September-2022].

[106] Edward O. Wilson. 2000. *Sociobiology: The New Synthesis, Twenty-Fifth Anniversary Edition*. Harvard University Press. http://www.jstor.org/stable/j.ctvjnrttd

[107] Anbang Xu and Brian Bailey. 2012. What Do You Think? A Case Study of Benefit, Expectation, and Interaction in a Large Online Critique Community. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work* (Seattle, Washington, USA) *(CSCW '12)*. Association for Computing Machinery, New York, NY, USA, 295–304. https://doi.org/10.1145/2145204.2145252

[108] Anbang Xu, Shih-Wen Huang, and Brian Bailey. 2014. Voyant: Generating Structured Feedback on Visual Designs Using a Crowd of Non-Experts. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work &amp; Social Computing* (Baltimore, Maryland, USA) *(CSCW '14)*. Association for Computing Machinery, New York, NY, USA, 1433–1444. https://doi.org/10.1145/2531602.2531604

[109] Yu-Chun Grace Yen, Joy O. Kim, and Brian P. Bailey. 2020. Decipher: An Interactive Visualization Tool for Interpreting Unstructured Design Feedback from Multiple Providers. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376380

[110] Alvin Yuan, Kurt Luther, Markus Krause, Sophie Isabel Vennix, Steven P Dow, and Bjorn Hartmann. 2016. Almost an Expert: The Effects of Rubrics and Expertise on Perceived Value of Crowdsourced Design Critiques. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work &amp; Social Computing* (San Francisco, California, USA) *(CSCW '16)*. Association for Computing Machinery, New York, NY, USA, 1005–1017. https://doi.org/10.1145/2818048.2819953

[111] Caroline Ziemkiewicz and Robert Kosara. 2010. Laws of Attraction: From Perceptual Forces to Conceptual Similarity. *IEEE Transactions on Visualization and Computer Graphics* 16, 6 (2010), 1009–1016. https://doi.org/10.1109/TVCG.2010.174

[112] Torre Zuk, Lothar Schlesier, Petra Neumann, Mark S. Hancock, and Sheelagh Carpendale. 2006. Heuristics for Information Visualization Evaluation. In *Proceedings of the 2006 AVI Workshop on BEyond Time and Errors: Novel Evaluation Methods for Information Visualization* (Venice, Italy) *(BELIV '06)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/1168149.1168162