How Do Researchers Manage Visualization Experiment Stimuli?

Hyeok Kim and Jeffrey Heer

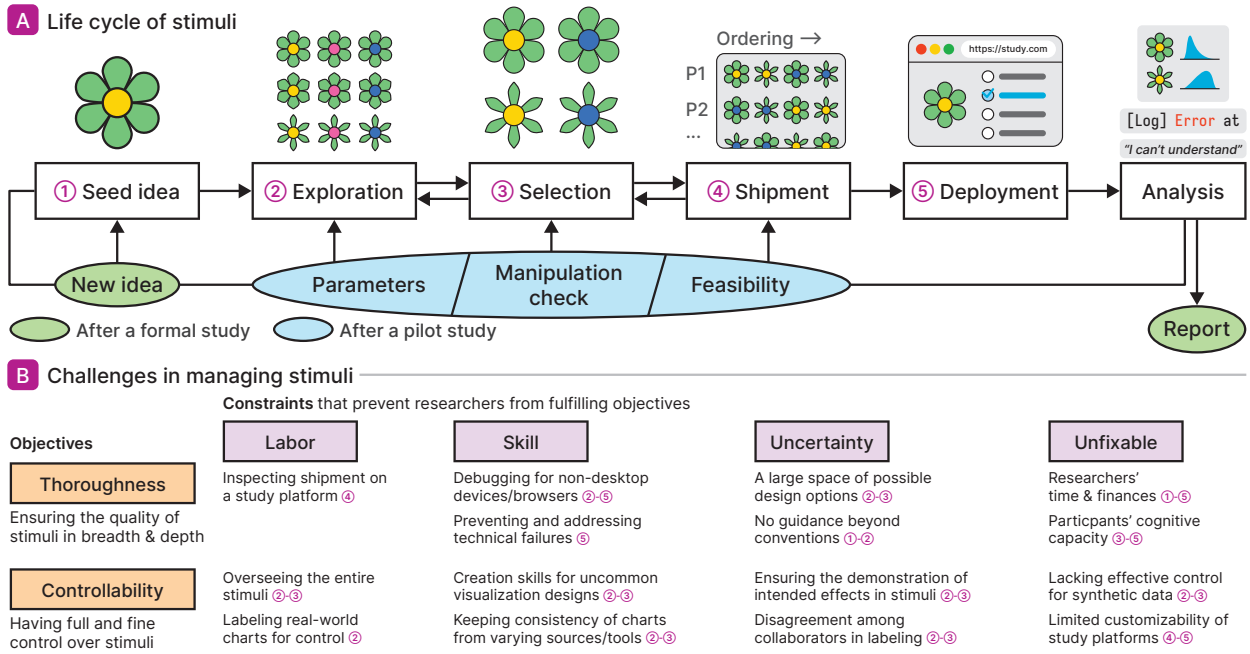

Figure 1: (A) Lifecycle of visualization experiment stimuli. Rectangles indicate current status, while circles indicate decisions made after running a pilot (lightblue) or formal (green) study. The locations of decisions can vary, but here indicate where our interviewees most commonly made those decisions. (B) We characterize researchers' challenges in creating and managing visualization experiment stimuli as *constraints* that prevent pursuing *objectives*. The listed challenges are examples.

Abstract—Visualization experiments need a set of "good" stimuli that effectively address research questions and hypotheses. Creating, managing, and deploying stimuli are often challenging, as these tasks require tremendous care. Inappropriate stimuli can make the outcome invalid or uninteresting, wasting both researchers' and participants' resources. As the speed of science increases, better support for stimuli-related tasks is essential, yet we lack a closer look at how visualization researchers deal with them. To understand the experiences of visualization experimenters and guide future improvements, we interviewed 19 visualization researchers with diverse backgrounds and experiences. Our findings describe practices and challenges across the life cycle of stimuli, from exploration and selection through shipment, deployment, and analysis. For example, stimuli management and deployment require tedious manual effort, which does not scale for experiments with many levels and complex conditioning. We also discuss both concerns and optimism around AI-assisted visualization experiment design. We conclude with future research opportunities in supporting stimuli creation, automated stimuli inspection, and experimental apparatus concerns.

Index Terms—Data visualization, experiment design, stimuli generation, interview

1 Introduction

Experiment studies help us to understand how people use visualizations to perceive data, update prior beliefs, and make decisions. Such findings provide guidelines on visualization design and also facilitate systems for design reasoning [36] and recommendation [27, 41, 47]. Experimenters usually design studies based on hypotheses, which are embodied as *stimuli* to elicit subjects' reactions. For example, to test the effectiveness of quantitative color scales [43] or build a related theoretical framework [8], the experimenters need to produce a set of relevant color scales. Given its essential role of bridging research questions to human perception, experimenters need to generate "good" stimuli to satisfy criteria like validity and generalizability. In the above example, the experimenter would want to test a "decently comprehensive" set of realistic color scales, including those commonly used. In addition, they may need to consider other relevant factors, such as data cardinality and distribution, that could influence the results.

However, decisions for stimuli generation are often nuanced and learned via apprenticeship, making it challenging for researchers without prior experience. For example, some studies use more than a few dozen charts, each of which needs careful inspection, potentially posing challenges in design iteration and version control. These challenges motivate better support for stimuli creation and management beyond those for web browser-based experiment development and deployment [14, 15, 51]. A first step is to closely understand researchers' challenges and strategies. Although prior work has documented practices in crowd-sourcing experiments [5], the curse of a large volume of stimuli [39], and sources of miscommunication in study stimuli [52],

• *Hyeok Kim is with the Korea Advanced Institute of Science & Technology, and the work was done at the University of Washington. Jeffrey Heer is with the University of Washington. E-mail: {hyeokk, jheer}@uw.edu.*

Manuscript received xx xxx. 202x; accepted xx xxx. 202x. Date of Publication xx xxx. 202x; date of current version xx xxx. 202x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.202x.xxxxxxx

little research has been done on the process of designing and deploying visualization experiment stimuli. Next, the research community anticipates and experiences substantial influences of AI in experiment design [4, 48] and visualization practices [76]. To better inform the design of appropriate AI and other automation methods for visualization stimuli, it is timely to contextualize them within researchers' practices. Therefore, we ask the following research questions.

- RQ1. What challenges and strategies do visualization experimenters have in designing, creating, and managing stimuli?
- RQ2. What are promising research opportunities to support how visualization experimenters work with stimuli?

To answer these questions, we interviewed 19 visualization experimenters from 14 academic organizations with varying degrees of experiences (from Ph.D. students to senior faculty). Our primary scope is online, crowd-sourcing experiments given its wide use in the field [5]. We report researchers' challenges in exploring, selecting, and shipping stimuli to study platforms by characterizing them as constraints in achieving *thoroughness* and *controllability* objectives, as summarized in Figure 1. Thoroughness objectives refer to researchers' goals to ensure the quality of stimuli in terms of both breadth and depth. For example, navigating a large space of possible designs and inspecting many stimuli are example challenges in pursuing thoroughness. Controllability objectives pertain to having full and fine control over stimuli. Limited customizability in stimuli design and an inability to oversee the entire stimuli set were common barriers in achieving controllability. We then present their strategies, such as the importance of "seeing" stimuli ideas and thinking through statistical analysis methods as guidelines for stimuli design. Regarding AI-assisted visualization experiments, interviewees had concerns about bias and maintaining authority, as well as optimism in automated creation and inspection. Based on our findings, we discuss future research opportunities in developing support for stimuli creation and management, response simulation for inspection, and cross-apparatus design.

2 Related Work

We refer to *visualization experiments* as controlled human-subject studies where data visualizations are primary stimuli. Based on the categorization by Lam et al. [39], our scope includes studies that evaluate *design*s (visual encodings and human perception & cognition) and *redesign*s (improvement for previously accepted designs). However, our scope does not cover experiments to evaluate *prototype systems* (i.e., newly proposed visual analytic or authoring tools), as these studies involve a different (although not mutually exclusive) set of challenges.

2.1 Visualization Experiment Design

Through a survey of empirical visualization studies, Lam et al. [39] characterize high-level objectives. First, experimenters want to understand design features that affect human visual data perception and cognition (*user performance*) and contribute to effective charts (*user experience*). Next, they want to test how different chart designs support different communication tasks, such as fact checking (e.g., [42]) and risk communication (e.g., [83]) (*communication through visualization*).

Carpendale [10] contextualizes literature on social science research methods to inform challenges in quantitative visualization studies. Among them, *internal*, *construct*, and *ecological* validity concerns are relevant to stimuli creation. An experiment achieves internal validity once stimuli test an effect with no confounders. If we broadly consider stimuli, including task instructions and solicitation questions, an experiment satisfies construct validity when the tasks and stimuli address the intended effect. An experiment holds ecological validity when stimuli can be translated to real world scenarios. Internal and ecological validity concerns inherently form a tradeoff, where the pursuit of ecological validity can sacrifice internal validity and vice versa. For example, using real-world visualizations as stimuli can cause confounders like color semantics, whereas participants from a general population are unlikely to see laboratory-generated charts in their everyday situations. Our findings illustrate challenges and strategies in navigating with this tradeoff.

Crowd-sourcing platforms, such as MTurk, Prolific, and Qualtrics, allow experimenters to run studies for a larger population beyond on-campus participants in a timely manner. Crowd-sourcing became a norm in contemporary visualization experiments. By analyzing 180 crowd-sourcing experiments from 82 papers, Borgo et al. [5] point out poor reporting quality and synthesize a checklist and template for reporting, which nevertheless lack considerations regarding stimuli. In addition, they pose a potential bias in favoring experiment types that are feasible for platforms or do not involve technical complexity in remote settings (e.g., static visualizations).

Elliot et al. [16] offer a design space for vision science methods for visualization experiments, in terms of participant tasks, stimuli adjustment, response solicitation techniques, and performance measurements. Among them, stimuli adjustment is most relevant to our scope. A given independent variable (e.g., the difference between two bar heights) can be adjusted gradually (increasing or decreasing unit by unit), interactively (participants indicate when they can tell a meaningful difference), randomly, and based on prior responses (making it more subtle after a correct response). Through a crowd-sourced study, Nobre et al. [52] identify literacy barriers for data visualization: human participants may misunderstand task instructions or objectives, misinterpret chart designs (e.g., visual encodings and references), or read wrong values (e.g., maximum instead of average; reading y instead of x). To mitigate such barriers, experimenters may need to choose simpler visual designs and tasks, despite some criticism against simplicity [59]. Alternatively, Burns et al. [9] propose a framework for accounting for different levels of understanding in a crowd-sourcing setting.

While the literature documents various concerns in visualization experiment design, these are limited in their ability to provide insights about the nuanced decisions that experimenters make while creating stimuli. Ziemkiewicz [84] only point out related gaps in our understanding, such as the number of cases to test and the lack of characterization of common tradeoffs. Our work aims to understand what factors form challenges in creating stimuli and how researchers deal with them in their lived experience as empirical visualization researchers.

2.2 Technical Support for Visualization Experiments

As surveyed by Cutler et al. [14], researchers build and deploy crowd-sourcing visualization experiments using survey platforms (e.g., Google Forms, Qualtrics), psychology experiment toolkits (e.g., psytoolkit [69], PsychoPy [57, 58], jsPsych [15]), and visualization-specific toolkits (e.g., VisUnit [29], GraphUnit [53]). Recently, ReVISit [14, 51] helps researchers build experiment websites with customizable components, support for external libraries (e.g., Vega-Lite, custom JavaScript), diverse response methods (e.g., survey forms, audio and screen recording, drawing), and domain-specific language-based creation. While ReVISit [14] supports browsing individual responses, other research has focused on more general data analysis toolkits. To highlight some recent examples, Jun et al. [30–32] offer a range of systems to automate coding statistical analyses based on conceptual hypotheses, and Lam et al. [40] propose a Large Language Model (LLM)-based tool for qualitative analysis. Currently, we have little understanding regarding whether we need support for stimuli creation, and if so, what kind of support the research community wants. To better inform future tool development, we aim to uncover what forms of technical support researchers desire regarding stimuli, as well as how their practices and concerns translate to future tool development.

2.3 AI in Experiments

Adoption of AI for scientific research is an active topic across diverse domains. Given the breadth of this discourse, we only point to topics relevant to our scope: literature review, hypothesis generation, stimuli creation, experiment design, and response simulation. Readers can refer to review papers [4, 48] for further reading. Wu et al. [76] also provide a detailed review of the use of AI in data visualization.

First, AI shows promise to support reviewing a wide range of prior literature, with emphasis on querying related papers [1, 49] and generating and/or assisting writing of reviews [1, 3, 11]. Second, generating hypotheses requires a well-structured, domain-specific corpus, such as

molecular structures for material discovery [46]. In data visualization, Zeng et al. [80, 81] present a corpus for graphical perception studies in a principled methods. Kim & Heer [34] propose a method for finding design features (e.g., the use of log scale on color) that are not covered by such corpora, which can potentially be turned into hypotheses. Next, hypotheses about data and visualization need to be reified as stimuli. Prior work on automatic visualization creation tends to suppose design recommendation scenarios (e.g., [27, 35, 47, 55, 56, 74, 75, 77, 79, 82]) and data-driven Q&A settings (e.g., [33, 62]).

Fourth, independent, moderation, and control variables in hypotheses need to be assigned as study conditions (e.g., within, between, and randomization). While ReVISit [14] supports enumerating condition assignments with different randomization options (i.e., the order of stimuli per participant), deciding how to assign variables remains the researchers' job given that it involves thorough, iterative reasoning about validity and study feasibility. Lastly, simulated responses (using AI in addition to, or even instead of, human participants) have been proposed to reduce resource constraints in human-subject studies, with some successful adoptions in some settings like marketing [6].

Researchers have both hopes and concerns in using AI for scientific knowledge discovery. Through a survey of 1,600 scientists, Van Noorden and Perkel [70] find that scientists are optimistic that AI can help with lowering language barriers, reducing time and financial costs, and making coding and presentation easier. However, they also point out shared concerns, such as false information and bias in overall writing and literature review, not being able to understand underlying mechanisms, and widespread plagiarism and fraud (e.g., via data fabrication). Indeed, Hao et al. [23] find that AI-assisted research papers exhibit higher publication and citation rates, but constricted topical coverage. In social science contexts, Grossman et al. [20] identifies issues with the use of LLM for response simulation, such as biases due to reusing the same underlying distribution and reproducibility concerns of commercial models. Taking a more formal lens, Hullman et al. [28] note that AI response simulation violates the assumption of independence and identical distribution (IID), may fail to capture detailed distributional characteristics (e.g., variance, skewness), and provides a "modest gain in precision" of findings. In visualization, computer vision-based methods [22], human-AI alignment [33, 72], and/or LLM-based design judgment [71] could be used for simulated user responses. Based on our findings, we discuss how these concerns are contextualized in visualization experiment settings.

3 Methods

To understand challenges, strategies, and desired tools in dealing with visualization experiment stimuli, we interviewed 19 visualization researchers with experience in running experimental studies. Our study was approved by the IRB of the University of Washington (No. STUDY00024973).

3.1 Interviewees

We recruited 19 visualization researchers (Table 1) through professional networks (research Slack groups, LinkedIn, and Bluesky) and by directly contacting those who recently published experiment research. To solicit diverse ideas, we carefully recruited them by diversifying their academic organizations and their experience levels. Interviewees were from 14 different organizations, and the maximum number of interviewees from the same institute was three. Our interviewees' experience levels were reasonably distributed: nine junior (Ph.D. students), five early career (postdocs, assistant professors), and five senior (tenured professors) researchers. Their research areas collectively covered perception & cognition, decision making, communication, data/visualization literacy, equity & accessibility. Our recruitment criteria were: (1) having completed visualization experiment studies and (2) being able to communicate in English (our study language).

3.2 Procedures

We conducted semi-structured interviews. Each session took about 45 minutes and was conducted in English through a remote meeting platform (Teams). We prepared our interview questions after three pilot sessions (one junior, one early career, and one senior researcher). After obtaining verbal consent, we asked questions about their years of experience running visualization experiments and elicited keywords for their research background. To gain context on their research practices, we asked about their most recent experiment study experience. For faculty interviewees, we asked them to describe the study they could best remember. Then, we inquired about their challenges in designing experiment stimuli, the impact of stimuli design on data analysis, and how they managed such challenges. We asked follow-up questions regarding aspects that either commonly or rarely appeared in prior sessions to reason about consistency and diversity across their experiences. Finally, we asked them to brainstorm technological support for experiment stimuli design, as well as tasks that they did not want to replace with automation technologies. Interviewees were compensated with a USD $20 gift card. Supplementary Material includes the detailed interview protocol.

Table 1: The demographics of interviewees. Exp. Levels (Experience levels): Junior (Ph.D. students), Early career (postdocs, assistant professors), and Senior (tenured professors). Years. (Years of experience in visualization experiments, not entire research career). P&C (perception & cognition), DM (decision making), E&A (equity & accessibility), Comm. (communication). The creation modes of our interviewees reflect only the experiences mentioned in their sessions.

Participant	Exp. Levels	Years	Broad areas	Creation modes
P1	Junior	3-5	P&C	Synthetic
P2	Junior	0-2	DM, Comm.	Synthetic
P3	Junior	3-5	P&C	Synthetic
P4	Early Career	5-10	E&A, Comm.	Synthetic
P5	Early Career	5-10	P&C	Real-world
P6	Early Career	10-15	P&C	Both
P7	Junior	3-5	P&C, DM	Synthetic
P8	Junior	3-5	P&C	Synthetic
P9	Senior	16-20	P&C	Both
P10	Senior	10-15	DM, literacy	Both
P11	Senior	16-20	P&C	Synthetic
P12	Early Career	10-15	P&C, Comm.	Both
P13	Junior	3-5	P&C, Comm.	Both
P14	Early Career	3-5	DM, Comm.	Real-world
P15	Junior	3-5	Literacy	Synthetic
P16	Junior	0-2	DM	Real-world
P17	Junior	0-2	P&C, E&A	Real-world
P18	Senior	16-20	P&C	Synthetic
P19	Senior	10-15	Literacy, design	Real-world

3.3 Analysis

We collected audio recordings from the sessions. We analyzed the recording transcripts through thematic analysis [7]. Initially we performed inductive open coding of the interview transcripts, identifying codes like 'language-driven confounders,' 'using custom tools for inspection,' and 'novelty in task design.' Then, guided by our research questions, we deductively categorized the open codes into themes regarding challenges, strategies, and desired tools in stimuli creation and management in visualization experiments. This step resulted in initial themes, such as 'making decisions for using real-world or synthetic stimuli' and 'expanding and pruning stimuli.' To build a more cohesive structure between those themes and better illuminate future research opportunities, we reorganized our themes in terms of objectives and constraints in dealing with stimuli. We include the evolution of our coding process in Supplementary Material.

4 Findings

We first describe challenges and strategies occurring while creating and managing visualization experiment stimuli. We then report concerns and opportunities in adopting AI for stimuli. Below, "interviewee" strictly refers to those who participated in our study; "participant" and "user" refer to those who took part in our interviewees' experiments.

Privacy statement. Describing detailed information about our interviewees' experiments—which often cover niche topics studied by a limited number of people—can inadvertently reveal their identities especially when this information is combined with interviewees' demographics. To protect the privacy of our interviewees, we only provide background information that is minimally necessary for understanding their challenges and answering our research questions. We also do not identify whether two or more interviewees were from the same institute.

4.1 Life Cycle of Stimuli

To contextualize our findings, we start with an overview of decision making steps that our interviewees made while dealing with stimuli, as summarized in Figure 1A. First, there is a **seed idea** about the stimuli, which could be from real-world experiences, examples, or demands (P2, P4, P9, P10, P16, P18), a specific prior paper on similar problems (P3, P6, P9, P14, P15), or what is absent in a broader theoretical framework (P8). Second, experimenters go through an **exploration** phase to populate more examples. Initial examples (typically one or two) may not be sufficient to test intended effects with a reasonable degree of generalizability. In this phase, interviewees add or remove independent variables (P2, P11) or increase or decrease levels of each variable (P1, P15). When using real-world examples, they often felt the need to include diverse examples or data patterns (P5, P10, P13, P17). Third, they make a **selection** of final stimuli to use in experiments. This selection step can provide a reasonable set of constraints and hence facilitate further exploration. For example, P11 initially considered using various chart types, which exploded the number of stimuli with no relationship between them. Once fixing it to bar charts, they were able to explore possible variations within that scope with no explosion.

Next, the selected stimuli are **shipped** to a crowd-sourcing platforms like Prolific or Qualtrics (P8, P12, P17), a study website using ReVISit [14, 51] (P3, P9, P14, P16), or custom websites (P9, P10, P16). While our interviewees generally run experiments via crowd-sourcing, a few of them also mentioned prior or on-going experiences in running in-person studies (P9, P18) using websites or instruments. Here, they need to check if stimuli are correctly assigned to corresponding conditions, if the randomization is working as intended, and if there are any technical bugs, for example. Once ready, experimenters **deploy** their experiment and wait for responses. To avoid confusion with visualized data within stimuli, we use *response* to refer to collected user reactions. Finally, experimenters **analyze** the collected responses. In this step, stimuli are revisited to match the findings. Note that these steps are not necessarily linear, as prior steps may be revisited back and forth.

This cycle applies not only to final formal studies, but also pilot studies. Our interviewees ran multiple pilot studies for various purposes, including parameterization of stimuli (e.g., correlation, difficulty), manipulation checks (e.g., natural language consistency, the existence of effects), feasibility (e.g., time and load on participants), and correctness (e.g., technical bugs). After running pilots, experimenters analyze the log data or open-ended feedback along with quantitative measures, and then revisit related steps to fix (P4, P5, P14). Sometimes, they realize missing controls or unexpected patterns after a formal study, which becomes inspiration for a subsequent study (P15, P18).

4.2 Challenges: Unmet Objectives

We characterize challenges in visualization experiment stimuli as researchers' *unmet objectives* due to *constraints*, as summarized in Figure 1B. Our interviewees shared high-level motivations, or internal virtues, in designing stimuli and other parts of experiments. Their motivations include, but are not limited to, generalizability (within a given scope), novelty (in stimuli design, analytic methods, study variables, theoretical frameworks), experimental validity, and research ethics (e.g., avoiding p-hacking). These motivations tended to manifest as objectives regarding stimuli design. We categorize objectives into two groups: *thoroughness* and *controllability*, as illustrated in Figure 2. First, thoroughness means researchers' objectives to ensure the quality of an experiment in breadth and depth. Thoroughness objectives include exploring a wide variety of possible stimuli ideas, having a generalizable set of stimuli, running technically sound experiment

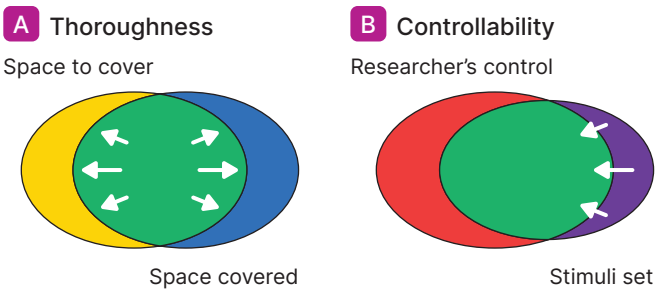

Figure 2: (A) The *thoroughness* objective is to have stimuli that sufficiently cover the breadth and depth of the research space. (B) The *controllability* objective is to fully govern and manage stimuli and their design details.

websites, and involving a broader audience by adjusting difficulty or apparatus concerns. Next, controllability means researchers' objectives to produce and manage stimuli as they want, while having full oversight. For example, researchers may want to manipulate stimuli in detail to show and isolate intended effects, design custom stimuli to test novel ideas, and oversee stimuli and study platforms.

However, there are constraints that prevent researchers from fulfilling their objectives. We classify those constraints in terms of amount of labor, required skills, inherent uncertainty in solutions, and (un)fixability. First, researchers can fix some constraints, but doing so requires a lot of labor. For example, researchers can create and update lots of stimuli (e.g., 100+), but it is time-consuming to do so. Second, some problems require advanced skills from different areas, such as programming and statistics. Third, researchers may eventually end up with some solutions, but they may find it difficult to justify them. For example, they may not have clear ideas about the range of parameters for stimuli. To address these, they may have to run multiple pilots or iterative discussions. Lastly, researchers just cannot solve some constraints. For example, the number of pilots may be constrained by financial reasons, and a researcher cannot test everything within a single study due to the limited attention of participants.

Below, we elaborate these challenges with detailed experiences of our interviewees in pursuing their objectives. While some challenges may involve different constraints, we focus on the primary causes that interviewees identified. For example, shipment inspection is challenging primarily because researchers often need to manually go through multiple stimuli, which better tools can support. On the other hand, creating sophisticated stimuli is fundamentally difficult due to limited technical skills, while also requiring some labor.

4.2.1 Thoroughness Objective

We describe challenges in ensuring quality in creating and managing stimuli.

Labor constraint. Shipping stimuli was a highly labor-intensive task of our interviewees. While some checks like stratification of a random parameter or counterbalancing of conditions over time can be done in a closed form (i.e., writing a test case or checking distributions), there are other sanity checks that need manual attention, such as data collection and assignment. First, experimenters need to check whether stimuli are assigned to right conditions; in other words, making sure that participants are seeing what they are supposed to be seeing. This is different from closed form checks that are done at algorithm levels; it is more about whether a stimulus (e.g., an image file, a design specification) is correctly mapped to the study conditions. Our interviewees tended to do it manually. For example, P18's lab practice was basically to go through every possible condition and see if assignments are correct. Similarly, P1 said, "I need to grab a paper and I need to write these conditions down and I record every single value of the data I want to collect." Using Qualtrics, P7 manually had to "set up a link to an image for each stimuli corresponding to each condition" and then "[go] through the survey with the preview function and actually [check] if it's showing the right image." However, when there were too many

possible conditions due to randomization, staircase design, or on-the-fly generation, interviewees could not just do 100% but a small portion of it (P2, P5, P15, P16, P17, P19). P16 said they only go through "the first 10" iterations where there were more than several thousands of possible orderings of stimuli.

This challenge persists when recycling prior stimuli that were successful due to the changes to platform. For example, P19 had to convert prior experiments done in different survey platforms to ReVISit for better data management. This recreating process took a few weeks for their team consisting of technically skilled researchers.

Skill constraint. To create complete visualization stimuli beyond static images or specification-based charts, researchers may need to have advanced technical skills. For example, P9 said that due to "packagers and minifiers," it became more difficult to find source codes for real-world interactive visualizations. P19 wanted to run a study with touch screen-specific interaction techniques on tablet devices. However, the lack of developer tools for non-desktop-facing browsers made it highly difficult to debug those techniques, such as the live log of every user movement.

Sometimes, researchers faced unexpected technical failures, which resulted in financial and time-related consequences. For example, P5 and P14 shared experiences where their study websites failed to collect the data. In P5's case, it was after adding an attention check, and they eventually had to re-run the study. In P14's case, it was only demographic information, so they had to contact the participants with extra compensation and about 90% of them responded back.

Uncertainty constraint. Regardless of knowing technical methods for designing stimuli, interviewees had difficulties in figuring out the scope of stimuli. This challenge occurred across different types of stimuli creation: synthetic, real-world, and replicated designs. First, interviewees often expanded their anecdotal real-world observations to a broader set of synthetic visualizations by looking to prior theoretical discussion to complement them (P10, P18) or generating comparative cases (P4, P14). In doing so, they often adopted theoretical frameworks from outside of the visualization literature (e.g., vision science, fact-checking, affordance, etc.)[1], which required further translation jobs. P4 and P14 noted that their chosen framework was intended for text-based communication, so it was challenging to create visualizations in a way that is "very closely equivalent to the text-only condition" (P14).

When using synthetic data for visualizations, interviewees adopted conventional approaches, when available, to avoid unnecessary reviewer concerns (P6, P8, P10, P17), for convenience (P8, P15), or for relationship to prior work (P9). Yet, P8 said that the frequent (almost only) use of the Gaussian distribution for a certain task made them unsure of whether the same treatment would work with other distribution types, which was often a concern of reviewers (including themselves). If no convention is available or suitable for their study, interviewees relied on heuristics and tested later (P2, P15).

Next, when using real-world visualizations, interviewees struggled with collecting qualified, representative cases (P5, P14, P17), while guidelines for comprehensiveness were not straightforward. For example, P14 struggled with "collect[ing] [N] clean examples," which resulted in quite a small set of stimuli (less than 10) out of a few thousands. But, this invited reviewer criticism regarding comprehensiveness. To achieve better comprehensiveness (e.g., in terms of chart types), P5 and P17 had to expand their collections by adding novel designs while ensuring consistency with the existing ones.

Lastly, other interviewees wanted to compare their work with previous experiments or had questions in how prior work analyzed and interpreted responses, so they decided to run replication studies. P3 and P9 (separately) wanted to find better ways to explain individual differences instead of reporting averages only. Their goal was to replicate related prior studies as much as possible, but the prior work did not provide full replication details. For instance, P9 mentioned that studies based on Cleveland & McGill [12, 13] were highly replicated in visualization experiments, yet no exact implementation of stimuli were available, which gave them uncertainty in precision.

Across stimuli types, an experiment variable can have continuous parameters (e.g., correlation, latency) or discrete levels (e.g., color hue; chart layout). Some junior interviewees noted that finding parameters and levels often felt like "art and not science" (P2) because there is no standardized method or concrete guidelines. Even if there is a guideline, how to divide or stratify a continuous parameter was quite tricky (P2, P15). When using real-world charts, they had to label a large volume of examples to ensure even assignment (P5, P17) and find a way to quantify certain variables like visual complexity (P6). Because these processes often involved different, at times contradicting, opinions of collaborators, they had to eventually run a pilot study to confirm.

Unfixable constraint. Our interviewees found challenges that raised important concerns, but they could not fix, such as losing comprehensiveness, participants' device types, and participants' cognitive load. First, real-world stimuli often convey multiple factors together, causing generalizability and internal validity concerns. P14's experiment had to use data already widely adopted in daily lives (e.g., those from government agencies or news media), and P19 wanted to involve human-made data fallacies in visualizations (e.g., improper log scale use). Because they needed to use entire charts as is, they had to justify the lack of diverse data patterns as a limitation (P5, P14). P5 said that expert designers tended to use bar and line charts more often than pie charts and scatter plots, so they "dropped them and focus[ed] on line and bar," sacrificing comprehensiveness. Next, interviewees were concerned with potential influences of device types other than desktops/laptops, so they had to exclude participants using mobile devices (P3, P5, P12, P13), yet P3 and P12 hoped for a method to account for device factors as an increasing number people may use crowd-sourcing platforms via smartphones. Lastly, when a research problem itself was quite complicated (e.g., uncertainty communication), it was not possible to use basic charts, increasing participants' cognitive load. To balance such complexity, for example, P16 omitted some stimuli to reduce the number of trials, and P12 had to add detailed instructions and training sessions.

4.2.2 Controllability Objective

We report challenges in controlling and overseeing experiment stimuli.

Labor constraint. Because exploration and selection stages often involve a lot of candidate stimuli, managing and overseeing them caused labor-intensive challenges. First, management of synthetic stimuli has a version control challenge. For example, P2 actively used LLMs to create stimuli but found it difficult to manage AI-generated stimuli, describing the process was "totally duct tape and glued by me in a folder and I'm constantly making copies of different code." They also said, "I'm not certain which direction I'm going. So I'm constantly saving lots of different drafts and I'm constantly taking screenshots." P1 and P2 said that updating stimuli became even more complicated when a study condition was added or removed. P4 desired an easier way to propagate changes made to one stimuli to others.

In contrast to synthetic charts created by a set of parameters and metadata, experimenters need to extract such parameters from real-world charts to make sure some degree of controllability and variability in stimuli and refer back to them during analysis. This labeling may involve a lot of manual effort (P5, P17). Although P17 used a repository with some metadata, for example, they were not enough for operationalization. Their team had to go through several rounds of labeling and calibration of conflicting opinions, requiring further user testing.

For internal communication and reporting purposes, interviewees wanted a way to overview the study design (e.g., stimuli laid out by variables) and export it as a figure for a paper. For example, P15 stored 24 stimuli in a folder, which they manually opened and viewed to sanity-check; they found that this tactic was not scalable for larger experiments with 100+ stimuli. P14 said, "figures for papers [are] both labor intensive and a lot of designing when it's a complicated study. So, if something did that for me automatically, that would be great."

[1] As visualization is interdisciplinary, it is hard to demarcate boundaries between the visualization literature and others. Here, we simply differentiate frameworks that were *originated* within visualization venues like VIS, TVCG, and EuroVis. They may still have been *used* in the visualization literature.

While describing themselves as an "AI Luddite," P18 noted that such automation could be "fundamental" to experiment studies.

Skill constraint. Interviewees had skill-related difficulties in keeping design consistency, creating unconventional designs, controlling parameterization, and accounting for cross-browser experiences. First, interviewees found it challenging to keep design consistency when they were using bespoke charts (P12), switching among tools (P7), or including non-visual stimuli like text (P2, P14, P15). For example, P7 initially created stimuli using Excel, and later needed to switch to D3, which made it difficult to replicate some of Excel's design choices. Second, interviewees had challenges creating "edge-cases" (P3) that were not widely adopted but seemed to open up interesting opportunities (P3, P12, P16). P16 was trying out map designs that were "not ... available [via] traditional visualization libraries" like Vega-Lite [65] or ggplot2 [73], so had to author new JavaScript code.

Lastly, even if experimenters exclude non-desktop devices, different monitors have different levels of resolutions and sizes, different browsers may vary in rendering the same code, and people may have personalization settings like dark mode, which interviewees learned from pilot studies (P3, P13, P15, P19). Solving cross-browser compatibility was often beyond our interviewees' skill stacks, threatening their control over stimuli. In response, they decided to use static images instead of on-the-fly generation, such as using Vega-Lite [65]. P3 further included a screen resolution check in their study. In one of P13's experiments, pixel-level differences were highly important, so they had to find some workarounds to prevent zooming on browsers. A related issue is handling of floating point numbers. P3 and P15 used NumPy and SciPy (on Python) to generate synthetic data points with long decimal values. However, browsers did not render these in the desired way, which was another reason they stuck to static images.

While platforms like Qualtrics or Prolific can mitigate such technical issues, they can severely limit customizability. For example, they do not require setting up a new server and their data collection is quite stable. P8, P10, and P12 found those platforms to be sufficient enough for experiments with static visualizations. For this convenience, P10 said that they had recommended Qualtrics to masters or undergraduate students' projects. However, such platforms were not enough to run more customized studies. For instance, P12 needed to solicit some responses and show them in later tasks. To do so, they had to inject custom JavaScript codes, which was challenging and added further inspection needs. P17 also mentioned the difficulty of customizing output formats, which added unnecessary data cleaning procedures.

Uncertainty constraint. During the exploration and selection steps, a major concern, particularly for perception studies, was to find stimuli that can "demonstrate that [intended] effects exist" (P11), which required fine control. P3 discussed the excitement of "testing edge cases" that "[do not have] abundant literature" but also anxiety that they "will perhaps give you interesting results or no results," a sentiment that was shared by P2 as well. More technically, our interviewees had challenges in achieving contrast and subtlety in stimuli for showing intended effects. Some effects can be achieved via contrastive stimuli with clear differences. For example, P4 was interested in how emojis (as in icon arrays) can invoke different affects while making decisions. Because some emojis have shared shapes (e.g., smiley faces), their team had to carefully find emojis that have different shapes as well as imply different emotions. It could also mean having a clear baseline case without any treatment like text descriptions (P14). On the other hand, there are cases where the difference among stimuli needs to be vague. When a stimulus has an easy-to-detect data pattern (e.g., correlation of 0.95), people are unlikely to make wrong answers, making it difficult to compare the effectiveness of different visual design choices (P1, P15, P18). Instead, experimenters needed to show subtle cases where other factors can lead to individual variance. Yet, there is no simple answer for the appropriate degree of subtlety in stimuli. For example, to see the influence of people's attitude toward certain topics on graphical perception, P15 needed to run pilots only for topic sensitivity as highly sensitive topics may overwhelm effects due to chart design. Furthermore, P13 randomly pulled out data from a distribution, and later found that a pattern where users performed well under certain numbers (those divisible by 5), which was a confounder for task difficulty.

When using real-world stimuli, interviewees had to account for internal validity to a degree by fixing elements that interrupt with isolating main effects, such as annotations explicitly describing the chart (P9, P14), excessive visual embellishment (P5), and unfamiliar or irrelevant topics (P19). When tasks are at a higher level than visual perception, experimenters also need to select a topic of the data (P1, P10), which may introduce additional confounders.

Unfixable constraints around controllability included the difficulty to synthesize data, real-world chart quality, and external constraints. First, using synthetic data distributions does not mean that you can fully control them. For example, there may not be a way to produce synthetic data that exhibits exact data parameters such as correlation, skewness, or entropy. So, P15 had to repeatedly create a bivariate distribution and test if it achieves a certain correlation until they filled out all the needed correlation values. This makes it nearly impossible to produce synthetic data on the fly, as P2 did. Furthermore, a targeted distribution may involve many underlying parameters with correlations or causal relationship (e.g., temporal, spatial data). After a few months of iterations, for example, P16 learned that it was impossible to synthesize their targeted distribution for geospatial data. At the same time, synthetic data may have additional graphical constraints like no overlapping data points (P3, P12), requiring either algorithmic or manual verification.

Next, using real-world visualizations or data may confine the room for control. For example, P14 noted that they had too few high-quality options once filtering out illegible charts from their corpus of real-world charts, which limited sophisticated control over stimuli. P10 also had to carefully choose a real-world data distribution that exhibits a uniform distribution of intended parameters (pairwise differences) while not provoking any strong emotions like political topics.

Lastly, some interviewees had external constraints in choosing a certain platform or toolkit. For example, P17 had to use a specific platform that their institution had a contract with, and P13 said that ReVISit is not compliant with local data privacy regulations (e.g., GDPR). P10 and P4 mentioned that sometimes they needed to carefully think about the total byte count of their study materials when they actually run studies with populations with limited internet access.

4.2.3 Merits and Challenges in using Off-the-shelf Toolkits

Our interviewees tended to use various off-the-shelf toolkits, such as Vega-Lite [65], Qualtrics, and Prolific. Among them, ReVISit [14, 51] was a popular keyword across our sessions. While only a few interviewees had a complete experience of using ReVISit to run studies, others were trying it out or had strong hopes for it. Overall, ReVISit had benefits including customizability (P3), more types of collectible responses (P9, P19), reusability of prior stimuli (P12), and in-situ support for stimuli inspection (P3, P13); yet some interviewees with deep skills in building study websites (P10, P12, P13) found some limitations in its capability. For example, P19 said that its screen-recording (a response option) allowed for seeing "where [people] click," which helped them to debug stimuli. On the other hand, P10 said, "I'm accustomed to having low-level control over my experiment design and ReVISit is this mix of expecting you to work [at a] high level, but it also expects you to have technical skills." P10 further shared concerns where adding a custom function or fixing some codes in ReVISit's source code could cause downstream problems, such as failing to properly show stimuli and collect responses.

This duality of attitudes is not unique to ReVISit. For example, P4 found Vega-Lite [65] to be useful to test out different ideas by just changing some properties in a design specification, whereas P3 sometimes had difficulty using Vega-Lite to add custom interactions or control low-level visual properties like padding between elements. P12 shared a failure experience when adding custom JavaScript code to Qualtrics platform. Due to the limited customizability, their team decided to stick to custom React websites whenever their study would involve interactions.

Table 2: Low-level tactics used by our interviewees. Note that these tactics are highly contextualized. The stages are: Exp(loration), Sel(lection), Ship(ment), Depl(loyment), and Anal(ysis).

Stage	Context	Tactic	Concerns	Participants
Overall	Real-world stimuli	Label the stimuli	↑Labor	5/17
	Synthetic stimuli	Use conventions	↓Design space	6/8/9/10 /15/17
Exp.	No prior conventions	Use heuristics and pilot them	Rationalizing decisions	2/15
	Complex task/ chart designs	Reduce trials	↓Statistical power	16
		Add instructions & training sessions	↓Time for trials	12
	Uncontrollable stimuli parameters	Generate & choose	↑Labor ↑Verification skills	12/13/15
		Use real-world data	↓Stimuli control	16
	Confounders (data patterns, topics)	Run pilots	↑Cost	13/15
	Stimuli applicability	Mix real-world and synthetic	Rationalizing decisions	10/14
	Uncertainty in stimuli exploration	Implement stimuli	↑Skill burdens	7/9/11/13 /15
Sel.	Deciding stimuli scopes	Run a pilot or main study early	↑Cost	15/16/19
	Too many possible stimuli	Fix to a certain chart type	↓Design space	4/5/11/12 /14
	Limited coverage of real-world stimuli	Create novel designs	↓Ecological validity ↓Consistency	15/17
	Real-world stimuli with confounders	Fix the stimuli designs	Rationalizing decisions	5/9/14/19
Ship.	Too many stimuli to inspect	Inspect a small portion	↓Ensuring overall quality	2/5/15/16 /17/19
	Not too many stimuli to inspect	Manually check every stimulus	↑Labor	1/7/15
Depl.	Limited skills for infrastructure	Use off-the-shelf platforms	↑Labor (switching) + ↓Customizability	8/10/12 /17
	Legal constraints	Use off-the-shelf platforms	↓Customizability	13/17
	Device-related concerns	Use static stimuli	↓Stimuli control	3/15
		Stick to desktop	↓Participant pools	3/5/12/13
Anal.	Wrong stimuli selection	Consider analytic choices early (e.g., preregistration)	-	6/9/14

4.3 Strategies

Interviewees shared strategies that they gained from their experiences and now use for mentoring. We introduce some recurring, high-level patterns below, along with a list of low-level tactics in Table 2.

Testing everything is a "fool's errand" (P10). Emphasizing the importance of admitting there is no perfect study, interviewees desired comprehensively testing in a well-scoped space rather than "false narrative of generalizability" for every possible case (P10). Specifically, P6 said such behavior would encourage P-hacking while analyzing responses, and P11 said adding too many factors would only increase more noise in responses. P11 also mentioned that after fixing a chart type, they were able to enumerate and reason about other possible conditions in a more explainable way. Studies comparing complete chart designs (e.g., quantile dot plot, error bars) as constructs can have fewer examples often based on popularity in literature (P4, P12) or in society (P14). On the other hand, experiments for isolating certain perceptual variables (P2, P18) would still need to create a lot of stimuli with a couple of variables as those variables form continuous parameter spaces or have more than a few dozens of levels.

Internal vs. ecological: Either path over middle ground. When it comes to internal and ecological validity concerns, our interviewees tended to choose either path and mitigate remaining concerns rather than find a middle ground. Some used synthetic, highly controlled charts to isolate a variable of interest, admitting loss of realism (P2, P7, P14, P15, P19). Others used real-world visualizations and data sets because they had related application domains (P5, P10, P13), wanted to convey the historical trajectories with some charts (P7, P14, P19), or were not able to synthesize one (P16). In addition, interviewees often mixed synthetic and real-world sources. For example, P10 used real-world data but created synthetic charts; while P14 used real-world charts but added synthetic text elements to them.

Seeing vs. thinking. Many interviewees reported that they had gone through long periods of exploring designs. These processes commonly include brainstorming different ideas, getting feedback from mentors and collaborators, and fixing them. In doing so, interviewees experienced technical burdens, such as version control and programming (P1, P2, P8, P12). Technical complexity made P12 hesitant to explore more stimuli. However, implementing stimuli is an important step because there is "a big difference between thinking ... versus seeing ... variations" in stimuli (P11). Implementation can function as a first pilot with themselves (P13, P15), produce reusable material for later stages (P7, P9), and reduce ambiguity among collaborators (P9).

Analysis as guidelines. After running several experiments, P14 said, "now I think more deeply about ... how to design an experiment in a way to make the stats later work better." With the same realization, P6 and P9 mentioned that they had used thinking about analysis ahead of the time as a key principle in their mentorship. For example, P9 said, "when you're designing the stimuli, you might control factors better, you might structure the stimuli itself or the instrumentation around the stimuli, so that you can more readily do the types of analysis that you anticipate." P6 had a rule that made their students complete pre-registration and IRB, which was in turn a forcing function to reason about data analysis while designing experiments. A challenge here is that many Ph.D. students with computer science backgrounds did not have prior statistics experiences, so they had to find related resources on their own (P2, P12) or take courses from another department (P6, P10).

"Close the loop ASAP". Junior interviewees (P15 and P16) said that their advisor often instructed them to run a study based on their best understanding at the moment "rather than planning everything" (P15). Once they analyzed the first study, they reasoned about what were missing or further interesting questions, which "pave[d] the path" (P16) for the next steps. P9 observed that they had "one pattern" of "closing the loop as soon as possible with sort of a minimal, viable experiment. ... within the lab or with a few participants at most," which not only made their students move on but also get some "muscle memory" of how to run an experiment. P1 also mentioned that running a small pilot study tended to be more useful because just speculating about stimuli designs often wasted their time without any confirmation.

4.4 Concerns and Opportunities for AI

In our interview sessions, we asked interviewees to brainstorm how automation can help their work and what boundaries they might draw for AI-assisted visualization experiments. At a high level, participants voiced concerns with losing control, but they expected reduced labor and enhanced thoroughness of the stimuli scope.

4.4.1 Concerns

"I want to be in the driver's seat." Our interviewees unanimously wanted to take the control in AI-assisted experiment design, with varying degrees of openness to AI assistance. At the one extreme, P18 did not want any kind of automation in experiment design but coding, whereas P14, P2, and P11 had more optimism in using AI for various tasks. Still, they "want[ed] to be in the driver's seat" (P11) at every step. P11 was worried that, if AI does things on behalf of them, then they get "this black box that [they] have to decode" regarding important decisions and rationales, which could be an additional, nebulous job. As "someone who's very open to AI," P14 said, "I wouldn't have [AI] write a section of the paper for me" because "it can't take accountability." They also pointed out that AI suggestions for confounding factors in stimuli could encourage "fishing for hypotheses."

Interviewees also expressed the importance of human care as researchers. While noting that "AI is going to continue to get better," P2 said that "the care to want to solve the problem and the interest to put the time and energy" would be something that they wanted to keep

holding on. Plus, some interviewees did not want automation because the very process of experiment design was enjoyable. For example, P16 said they would not use LLMs to ask about research ideas because they "find a lot of joy in figuring out what a good study is."

Bias and homogeneity. Interviewees shared about their concerns regarding biases of AI assistance. For example, P10 was concerned with the poor quality of AI-based literature review and summarization, which further caused distrust in automated literature solicitation. More importantly, some were worried about the homogeneity of AI-suggested topics, which is already being observed in science [23], and hence resulting in similar stimuli designs. For instance, P3 pointed out historical bias of AI which could suggest stimuli designs that "you haven't considered but other people have already [done]," which would in turn strip the novelty of their work. In a similar vein, P9 noted that AI-suggested experiments would be "more constrained than what people do now, which is almost anything." P14 was also worried that AI suggestions would bias them, making it difficult to "independently verify [their] work." P18 was similarly concerned that their advisees would not be able to develop an ability to independently think—not only about stimuli designs, but also overall experiment designs.

4.4.2 Opportunities

Below, we summarize ideas from our interviewees. In the following section, we will discuss their potential impacts and relationship to the literature.

Stimuli creation and management was a frequently mentioned avenue for future tool support, as noted above. This tendency was because they often found technical creation of stimuli (e.g., using D3) or setting up an experiment platform to be relatively less essential parts than experiment design and data analysis. P12 and P16 noted that they could explore a broader set of possible designs if they were able to promptly generate visualizations that are not well-supported by toolkits. P12 envisioned further use cases, such as creating charts on the fly so that participants can provide more contextual qualitative feedback. P2 and P4 emphasized the need to propagate design changes, as making consistent changes across different stimuli was challenging. Because LLM prompts tend to be natural language, P4 suggested a tool that generates semantically equivalent or varying icons for icon arrays. For example, if they want to compare "concrete" (e.g., faces, materials) and "abstract" shapes (e.g., arrows), then an AI model should allow manipulations based on such concepts. P9 suggested that automated stimuli creation could benefit "closing the loop" early. Yet, such support should come with a caution. P19 shared an experience where the stimuli and tasks were too easy for LLMs to do but slightly troublesome for people to do, resulting in LLM-generated, invalid, homogeneous responses with no error.

Automated shipment verification. Interviewees often imported stimuli images to study platforms or websites by manually matching image URLs to study conditions, which they found unscalable (P15). This inconvenience is further problematic when experimenters are making changes to stimuli design or conditioning after shipping it to a platform, which could mean starting from scratch (P10). Plus, experimenters may want to reuse their stimuli from pilots partially or entirely after pilot studies, which could require another from-scratch shipment (P9). Thus, a "unified" ecosystem (P15) could integrate design iteration, shipment to study platforms, verification of assignments, and reuse of stimuli (P9, P15, P16). As a verification method, some participants brainstormed around simulated participants, which we further discuss below.

Simulated responses. ***While all of our interviewees noted that they would never replace human responses with synthetic ones*** (formal study or parameterization pilots), they pointed to places where response simulation could be useful. On the one hand, interviewees wanted sanity checks of their stimuli (e.g., shipment, bugs) using AI. For example, a computer vision agent could go through the study for each possible condition and check if a stimuli is visually correct (not just by the file name) for a condition or if there are any bugs (P15, P8). On the other hand, interviewees wished for methods to check downstream effects of their stimuli to statistical analysis (P5, P16). For example, to assess the suitability of stimuli for different literacy levels, P5 wished for simulated responses for different participant demographics (e.g., education level). P16 also mentioned that some realistically simulated responses could help them envision final statistical modeling and assess if their study design is internally valid.

Scaffolding methods. Given that stimuli designs can have downstream effects (Section 4.3), interviewees often had difficulty reasoning about possible next steps in stimuli design, or saw students who did. While acknowledging that many decisions do not have clear solutions, they brainstormed some potential opportunities to help with their reasoning. For example, P2 often experienced challenges in choosing statistical methods. Based on experiment conditions and stimuli parameter settings, they wanted an AI tool that suggests "standard guidelines" along with their pros and cons, so that they could reason about them and later discuss with their advisor. Similarly, P6 imagined a linter-like tool that inspects their experiment design and helps them to understand the limitations. Their intention was not to fix that limitations but to have confidence in what they could generalize through their experiments. Lastly, with P9, we brainstormed about AI-based educational tools that can instruct possible next steps based on previous decisions, acknowledging potential biases. For example, if a study design is tentatively decided, a simulation tool can suggest potential stimuli options and help experimenters to think about their applicability, consequences to statistical methods, etc.

5 Discussion: Future Research Opportunities

Based on our findings and prior literature, we outline several promising research opportunities in terms of stimuli creation and management, response simulation, cross-apparatus support, and scaffolding for experiment design.

5.1 Stimuli Creation and Management

Our findings confirm that one-shot generation of possible stimuli given a hypothesis will not fulfill researchers' needs at various stages of experiment design. Instead, stimuli creation and management systems must integrate support for creating, exploring, selecting, verifying, and shipping stimuli. Suppose that a researcher has some seed idea about stimuli and related hypotheses, and a system can enumerate a set of initial designs. For synthetic visualizations, prior techniques in partial specification support [47, 74] and design augmentation [34] could be useful. For real-world visualizations, style transfer [17, 24, 25] and classification [66] techniques could further support filling missing chart designs based on existing ones or reduce confounders between stimuli. Replicating interactive visualizations from real-world cases could adopt automated methods for adding interactions to static visualizations [68]. Natural language-based visualization creation [50, 62] will also play an important role for semantic consistency. Interesting future research is to understand how AI-assistance for visualization stimuli supports or disrupts replication studies in terms of code quality and readability.

When exploring and selecting stimuli design, we point out four important requirements: propagating design changes, version control, interactions for adding/removing conditions, and addressing biases. First, updating stimuli consistently may not necessarily mean repeating the same changes. For example, making changes to mark type in Vega-Lite (e.g., `point` → `rect`) can also change the color palettes (white-to-blue → Viridis). Plus, a common exploration trajectory in any design task is trying out an idea and going back to an original, which can happen much later in the process. This nature requires effective version control methods (e.g., selectively canceling design changes) as well as an overview dashboard. Third, such an overview dashboard could interactively support adding or removing conditions to adjust the study scope or match participants' cognitive capacity. Future work can refer to responsive visualization literature [26, 35, 67] as they share similar concerns. Lastly, automated stimuli creation systems should be able to assess and inform homogeneity and bias in suggested designs. A potential method is a platform for sharing experiments without requiring technical burdens, such as refactoring their study designs. For example, a generative AI-based method can produce a preliminary study specification from a pre-registration text (as a starting

point), and the researchers can clarify it with minimum effort. Such a structured dataset of study designs could allow for assessing the frequency of stimuli designs and illuminating gaps that a proposed study might be able to fill.

To correctly ship multiple stimuli to a study platform, our interviewees tended to rely on tedious manual tasks. Thus, a stimuli management system must consider relevant automation techniques. A potential approach is to produce stimuli in ReVISit-supported formats (e.g., React components, Vega-Lite specifications) and their assignment as a ReVISit config spec [14]. For those with limited tool options, systems can generate equivalent configuration files for platforms like Prolific [60] and Qualtrics [61] given that they offer programmatic methods for creating a study form. Shipment automation will require high accuracy or inspection methods, which we discuss in the following section.

5.2 Guaranteed Large-scale Inspection

Inspecting stimuli does not mean checking them one by one, but seeing them in full sequences of a given study. For example, bugs can happen in specific orderings of stimuli, and there can be unexpected priming effects. Furthermore, an update to stimuli after that inspection can mean additional inspection. However, when there are more than a few dozens of stimuli, and they are ordered based on complex logic, it becomes hard to inspect them with 100% coverage.

We suggest two approaches—machine learning-based automated checks and finding a representative subset—for hybrid, guaranteed, large-scale inspection. First, automated inspection systems can go through study websites, perform tasks, and check certain concerns like repeated visual similarity. They can utilize ML-based visualization perception techniques [22,72] and AI agents like Claude Cowork. Next, manual inspection is still essential to ensure stimuli quality. To reduce manual efforts, algorithmic methods to find a minimum, guaranteed subset of stimuli assignments (e.g., random ordered stimuli) could be useful. Systems could adaptively modify a guaranteed subset as inspection proceeds, based on initial automated checks.

5.3 Response Simulation

We clarify again that simulating experiment responses can be harmful in conducting experiments ethically. Our discussion about response simulation is bounded to support pre- or post-study tasks that could inform stimuli design, such as choosing statistical methods, inspecting lack of effects of stimuli before pilots, and simulating various normative scenarios given stimuli designs. With this in mind, an interesting application scenario is mimicking different population demographics, such as literacy or education levels, as those factors can have influence on how participants interpret stimuli (e.g., P5). Researchers may take a closer look for cases where a performance measure drops or surges in particular demographic groups. Inspired by a *rational agent framework* [78] and *causal quartets* [19], responses can be simulated to represent normative statistical scenarios (e.g., no effects, too strong effects) to help researchers interpret human response data.

There are important considerations in terms of the performance, behavior, and safeguards of automated response simulation. First, performance of automated techniques could be problematic in simulating responses. Prior work on ML-based visualization perception and evaluation techniques [22,71,72] tend to exhibit low alignment with human counterparts, which is also prevalent in LLM-based response generation [20,28]. Next, ML-based methods may have different objectives than human-subject experiments. Many of them tend to focus on how accurately an automated technique can process visualization images, while visualization experiments essentially want to collect how humans make errors. A possible concern is that automated agents make no errors with stimuli designed to provoke specific human errors. It will be an important research question whether such models should exhibit human-like errors. For instance, while human-like behaviors could simulate patterns that are better aligned with human cognition [38], they might reinforce historical biases and harms [45]. Lastly, we can imagine malicious researchers using response simulation techniques for data fabrication, as science history shows [18]. Thus, response simulation techniques would ideally come with safeguards, such as watermarking (c.f. [2,54]) that is robust against data removal or permutation.

5.4 Accounting for Apparatus

As described in Section 4.2.2, our interviewees had controllability challenges regarding apparatus, stemming from the fact that they cannot fully anticipate and control participant devices in crowd-sourcing settings. In practice, browsers may behave differently on desktop and smartphones, and mobile devices have a wide range of screen sizes, which makes it nearly impossible for even professional designers to inspect them all [26]. Due to discrepancies in browser behaviors, interviewees could not test stimuli design on all possible platforms even if they wanted to recruit participants on diverse devices. Furthermore, it is extremely difficult to obtain real-time logs for interactions with stimuli (as we do with browser developer tools) on smartphones and tablets, as P19 indicated, so debugging capacity is limited. To mitigate this technical difficulty, responsive visualization-based approaches could be useful with careful coordination of perceptual differences with respect to physical devices (e.g., same pixels may result in different millimeters). While tools like Dupo [35] propose some physical size simulation methods, every device has different pixel resolutions, so a more granular control is necessary, ideally using perceptual units like a degree of visual angle. At the same time, future work is warranted to extend off-the-shelf toolkits and study platforms to support robust cross-device experiences.

5.5 Scaffolding

Stimuli creation and management cannot be separated from the overall experiment design processes. For learning purposes, some interviewees brainstormed scaffolding support in experiment design. For example, given a hypothesis, an educational system could provide possible applicable next steps, such as nudging how many stimuli are needed in what study conditions, instead of suggesting the results directly. Potentially, an early analysis of participants' time and cognitive load would help assess the study scope. Based on stimuli design, their assignment, and simulated responses, such a system could further inform analytic choices, adopting modeling support systems [21,30–32] and multiverse analysis systems [44,63,64], for example. With P9 we brainstormed visual demonstrations of influences of previous decisions to next steps, inspired by *Seeing Theory* [37] that illustrates how statistical distributions and testing methods work in a detailed step-by-step animation, as a method to amplify the educational value of scaffolding systems for experiment design.

6 Conclusion

Stimuli play an essential role in visualization experiments and need tremendous care of researchers. Yet, researchers strategies, challenges, and needs regarding stimuli management have remained untold. To understand practices and illuminate future paths, we conducted a semi-structured interview study with 19 visualization experimenters with diverse levels of experience from 14 different institutions. Our findings inform how researchers' motivations in creating and managing stimuli thoroughly with controllability are constrained due to labor, skill, uncertainty, and (un)fixability, as well as their strategies. We also report optimism in AI for facilitating such processes as well as concerns and criticism regarding the use of AI for visualization science. Based on our findings, we discuss future research opportunities in terms of stimuli creation and management, guaranteed shipment verification, response simulation, and apparatus concerns.

Limitations. Our work has multiple limitations with respect to participant demographics. Our participants tended to have experiences in online, crowd-sourcing experiments, reflecting its wide use in the field. A large portion of participants were junior and early-career researchers as we prioritized capturing recent, hands-on experiences. Lastly, except for a few participants who were interested in individual variances, most of them designed their experiments based on a positivist paradigm. Future work can extend our work regarding the above aspects.

Supplementary Materials

We provide our study protocol and qualitative codes.

Acknowledgments

The authors thank Cindy Xiong Bearfield, Danielle Szafir, and Paula Kayongo for initial feedback. This research was supported in part by NSF IIS Award (2402718).

References

[1] S. Agarwal, G. Sahu, A. Puri, I. H. Laradji, K. D. Dvijotham, J. Stanley et al. LitLLMs, LLMs for literature review: Are we there yet? *Transactions on Machine Learning Research*, 2025. 2

[2] S. Al-Marzouki, S. Evans, T. Marshall, and I. Roberts. Are these data real? statistical methods for the detection of data fabrication in clinical trials. *BMJ*, 331(7511):267–270, 2005. doi: 10/btnnrv 9

[3] H. An, A. Narechania, E. Wall, and K. Xu. vitaLITy 2: Reviewing academic literature using large language models, 2024. NLVIZ Workshop at IEEE VIS '24. https://arxiv.org/abs/2408.13450. 2

[4] S. Bianchini, M. Müller, and P. Pelletier. Artificial intelligence in science: An emerging general method of invention. *Research Policy*, 51(10):104604:1–15, 2022. doi: 10/gqm63c 2

[5] R. Borgo, L. Micallef, B. Bach, F. McGee, and B. Lee. Information visualization evaluation using crowdsourcing. *Comput. Graphics Forum*, 37(3):573–595, 2018. doi: 10/gdxdk8 1, 2

[6] J. Brand, A. Israeli, and D. Ngwe. Using LLMs for market research. *Harvard Business School Marketing Unit Working Paper No. 23-062*, pp. 61–71, 2023. doi: 10/qw6j 3

[7] V. Braun and V. Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006. doi: 10/fswdcx 3

[8] R. Bujack, T. L. Turton, F. Samsel, C. Ware, D. H. Rogers, and J. Ahrens. The good, the bad, and the ugly: A theoretical framework for the assessment of continuous colormaps. *IEEE Trans. Vis. Comput. Graphics*, 24(1):923–933, 2018. doi: 10/gcqb9w 1

[9] A. Burns, C. Xiong, S. Franconeri, A. Cairo, and N. Mahyar. How to evaluate data visualizations across different levels of understanding. In *2020 IEEE Workshop on Evaluation and Beyond - Methodological Approaches to Visualization (BELIV)*, pp. 19–28, 2020. doi: 10/kz34 2

[10] S. Carpendale. *Evaluating Information Visualizations*, pp. 19–45. Springer-Verlag, Berlin, Heidelberg, 2008. 2

[11] K. Choe, S. Park, S. Jung, H. Kim, J. W. Yang, H. Hong et al. Supporting novice researchers to write literature review using language models. In *ACM Human Factors in Computing Systems*, CHI EA '24, pp. 307:1–9. ACM, 2024. doi: 10/qw5n 2

[12] W. S. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *J. Am. Stat. Assoc*, 79(387):531–554, 1984. doi: 10/gdvmwd 5

[13] W. S. Cleveland and R. McGill. An experiment in graphical perception. *International Journal of Man-Machine Studies*, 25(5):491–500, 1986. doi: 10/b6vt7v 5

[14] Z. Cutler, J. Wilburn, H. Shrestha, Y. Ding, B. Bollen, K. A. Nadib et al. ReVISit 2: A full experiment life cycle user study framework. *IEEE Trans. Vis. Comput. Graphics*, 32(1):13–23, 2026. doi: 10/hbkxwp 1, 2, 3, 4, 6, 9

[15] J. R. de Leeuw, R. A. Gilbert, and B. Luchterhandt. jsPsych: Enabling an open-source collaborative ecosystem of behavioral experiments. *Journal of Open Source Software*, 8(85):5351:1–4, 2023. doi: 10/gtzcxz 1, 2

[16] M. A. Elliott, C. Nothelfer, C. Xiong, and D. A. Szafir. A design space of vision science methods for visualization research. *IEEE Trans. Vis. Comput. Graphics*, 27(2):1117–1127, 2021. doi: 10/ghgt5q 2

[17] M. H. Everts, H. Bekker, J. B. T. M. Roerdink, and T. Isenberg. Interactive illustrative line styles and line style transfer functions for flow visualization, 2015. 8

[18] D. Fanelli. How many scientists fabricate and falsify research? a systematic review and meta-analysis of survey data. *PLOS ONE*, 4(5):1–11, 05 2009. doi: 10/bn5pnj 9

[19] A. Gelman, J. Hullman, and L. Kennedy. Causal quartets: Different ways to attain the same average treatment effect. *Am. Stat.*, 78(3):267–272, 2024. doi: 10/qw5z 9

[20] I. Grossmann, M. Feinberg, D. C. Parker, N. A. Christakis, P. E. Tetlock, and W. A. Cunningham. AI and the transformation of social science research. *Science*, 380(6650):1108–1109, 2023. doi: 10/gscnp5 3, 9

[21] Z. Guo, A. Kale, M. Kay, and J. Hullman. VMC: A grammar for visualizing statistical model checks. *IEEE Trans. Vis. Comput. Graphics*, 31(1):798–808, 2025. doi: 10/qw57 9

[22] D. Haehn, J. Tompkin, and H. Pfister. Evaluating 'graphical perception' with CNNs. *IEEE Trans. Vis. Comput. Graphics*, 25(1):641–650, 2019. doi: 10/gd52dt 3, 9

[23] Q. Hao, F. Xu, Y. Li, and J. Evans. Artificial intelligence tools expand scientists'impact but contract science's focus. *Nature*, 649(8099):1237–1243, 2026. doi: 10/hbj27r 3, 8

[24] J. Harper and M. Agrawala. Deconstructing and restyling D3 visualizations. In *ACM Symposium on User Interface Software and Technology*, UIST '14, pp. 253–262. ACM, 2014. doi: 10/ghvkfk 8

[25] J. Harper and M. Agrawala. Converting basic D3 charts into reusable style templates. *IEEE Trans. Vis. Comput. Graphics*, 24(1):1274–1286, 2017. doi: 10/gcxtgt 8

[26] J. Hoffswell, W. Li, and Z. Liu. Techniques for flexible responsive visualization design. In *ACM Human Factors in Computing Systems*, CHI '20, pp. 1–13. ACM, 2020. doi: 10/gk9mqp 8, 9

[27] K. Hu, M. A. Bakker, S. Li, T. Kraska, and C. Hidalgo. VizML: A machine learning approach to visualization recommendation. In *ACM Human Factors in Computing Systems*, CHI '19, pp. 128:1–12. ACM, 2019. doi: 10/gf2b86 1, 3

[28] J. Hullman, D. Broska, S. Huaman, and S. Aaron. This human study did not involve human subjects: Validating LLM simulations as behavioral evidence, 2025. Working manuscript. doi: 10/qw5p 3, 9

[29] R. Jianu, A. Slingsby, D. Laksono, and M. Okoe. VisUnit: Literate visualisation studies assembled from reusable test-suites. In *ACM Human Factors in Computing Systems*, CHI '25, pp. 917:1–15. ACM, 2025. doi: 10/qw5j 2

[30] E. Jun, M. Daum, J. Roesch, S. Chasins, E. Berger, R. Just et al. Tea: A high-level language and runtime system for automating statistical analysis. In *ACM Symposium on User Interface Software and Technology*, UIST '19, pp. 591–603. ACM, 2019. doi: 10/gmbdfj 2, 9

[31] E. Jun, E. Misback, J. Heer, and R. Just. rTisane: Externalizing conceptual models for data analysis prompts reconsideration of domain assumptions and facilitates statistical modeling. In *ACM Human Factors in Computing Systems*, CHI '24, pp. 1037:1–16. ACM, 2024. doi: 10/qw5k 2, 9

[32] E. Jun, A. Seo, J. Heer, and R. Just. Tisane: Authoring statistical models via formal reasoning from conceptual and data relationships. In *ACM Human Factors in Computing Systems*, CHI '22, pp. 490:1–16. ACM, 2022. doi: 10/hbsvwh 2, 9

[33] K. Kafle, B. Price, S. Cohen, and C. Kanan. DVQA: Understanding data visualizations via question answering. In *IEEE CVPR*, June 2018. 3

[34] H. Kim and J. Heer. Data augmentation for visualization design knowledge bases. *IEEE Trans. Vis. Comput. Graphics*, 32(1):538–548, 2026. doi: 10/qw5x 3, 8

[35] H. Kim, R. Rossi, J. Hullman, and J. Hoffswell. Dupo: A mixed-initiative authoring tool for responsive visualization. *IEEE Trans. Vis. Comput. Graph.*, 30(1):934–943, 2024. doi: 10/qw5s 3, 8, 9

[36] Y. Kim, K. Wongsuphasawat, J. Hullman, and J. Heer. GraphScape: A model for automated reasoning about visualization similarity and sequencing. In *ACM Human Factors in Computing Systems*, CHI '17, pp. 2628–2638, 2017. doi: 10/gd7hg5 1

[37] D. Kunin, J. Guo, T. Dae Devlin, and D. Xiang. Seeing theory, 2016. Last accessed Mar. 18, 2026. https://seeing-theory.brown.edu/probability-distributions/index.html. 9

[38] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman. Building machines that learn and think like people. *Behav. Brain Sci.*, 40:e253:1–72, 2017. doi: 10/gcw3n8 9

[39] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale. Empirical studies in information visualization: Seven scenarios. *IEEE Trans. Vis. Comput. Graphics*, 18(9):1520–1536, 2012. doi: 10/drrh6j 1, 2

[40] M. S. Lam, J. Teoh, J. A. Landay, J. Heer, and M. S. Bernstein. Concept induction: Analyzing unstructured text with high-level concepts using LLooM. In *ACM Human Factors in Computing Systems*, CHI '24, p. 766. ACM, 2024. doi: 10/qw5m 2

[41] H. Li, Y. Wang, S. Zhang, Y. Song, and H. Qu. KG4Vis: A knowledge graph-based approach for visualization recommendation. *IEEE Trans. Vis. Comput. Graph.*, 28(1):195–205, 2022. doi: 10/gndq6h 1

[42] M. Lisnic, A. Lex, and M. Kogan. "yeah, this graph doesn't show that": Analysis of online engagement with misleading data visualizations. In *ACM Human Factors in Computing Systems*, CHI '24, pp. 199:1–14. ACM, 2024. doi: 10/qhrc 2

[43] Y. Liu and J. Heer. Somewhere over the rainbow: An empirical assessment of quantitative colormaps. In *ACM Human Factors in Computing Systems*,

CHI '18, pp. 598:1–12. ACM, 2018. doi: 10/gfxbcb 1

[44] Y. Liu, A. Kale, T. Althoff, and J. Heer. Boba: Authoring and visualizing multiverse analyses. *IEEE Trans. Vis. Comput. Graphics*, 27(2):1753–1763, 2021. doi: 10/ghgt7j 9

[45] E. J. Livesey, M. B. Goldwater, and B. Colagiuri. Will human-like machines make human-like mistakes? *Behav. Brain Sci.*, 40:e270:41, 2017. doi: 10/g537kf 9

[46] A. Merchant, S. Batzner, S. S. Schoenholz, M. Aykol, G. Cheon, and E. D. Cubuk. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85, 2023. doi: 10/gs7sbc 3

[47] D. Moritz, C. Wang, G. L. Nelson, H. Lin, A. M. Smith, B. Howe et al. Formalizing visualization design knowledge as constraints: Actionable and extensible models in Draco. *IEEE Trans. Vis. Comput. Graph.*, 25(1):438–448, 2018. doi: 10/cs68 1, 3, 8

[48] S. Musslick, L. K. Bartlett, S. H. Chandramouli, M. Dubova, F. Gobet, T. L. Griffiths et al. Automating the practice of science: Opportunities, challenges, and implications. *Proceedings of the National Academy of Sciences*, 122(5):e2401238121:1–11, 2025. doi: 10/g84s7d 2

[49] A. Narechania, A. Karduni, R. Wesslen, and E. Wall. VITALITY: Promoting serendipitous discovery of academic literature with transformers & visual analytics. *IEEE Trans. Vis. Comput. Graphics*, 28(1):486–496, 2022. doi: 10/g6sc9c 2

[50] A. Narechania, A. Srinivasan, and J. Stasko. NL4DV: A toolkit for generating analytic specifications for data visualization from natural language queries. *IEEE Trans. Vis. Comput. Graphics*, 27(2):369–379, 2021. doi: 10/ghgthp 8

[51] C. Nobre, D. Wootton, Z. Cutler, L. Harrison, H. Pfister, and A. Lex. ReVISit: Looking under the hood of interactive visualization studies. In *ACM Human Factors in Computing Systems*, CHI '21, pp. 25:1–13. ACM, 2021. doi: 10/gksk5b 1, 2, 4, 6

[52] C. Nobre, K. Zhu, E. Mörth, H. Pfister, and J. Beyer. Reading between the pixels: Investigating the barriers to visualization literacy. In *ACM Human Factors in Computing Systems*, CHI '24, pp. 197:1–17. ACM, 2024. doi: 10/n98g 1, 2

[53] M. Okoe and R. Jianu. GraphUnit: Evaluating interactive graph visualizations using crowdsourcing. *Comput. Graphics Forum*, 34(3):451–460, 2015. doi: 10/f7kr8h 2

[54] L. Pan, A. Liu, Z. He, Z. Gao, X. Zhao, Y. Lu et al. MarkLLM: An open-source toolkit for LLM watermarking. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 61–71. ACL, 2024. doi: 10/qw6f 9

[55] A. Pandey, S. L'Yi, Q. Wang, M. A. Borkin, and N. Gehlenborg. GenoREC: A recommendation system for interactive genomics data visualization. *IEEE Trans. Vis. Comput. Graph.*, 29(1):570–580, 2023. doi: 10/qw5t 3

[56] A. Pandey, A. Srinivasan, and V. Setlur. MEDLEY: Intent-based recommendations to support dashboard composition. *IEEE Trans. Vis. Comput. Graph.*, 29(1):1135–1145, 2023. doi: 10/g9g7qk 3

[57] J. Peirce, J. R. Gray, S. Simpson, M. MacAskill, R. Höchenberger, H. Sogo et al. PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, 51(1):195–203, 2019. doi: 10/gft89w 2

[58] J. W. Peirce. PsychoPy—psychophysics software in python. *J. Neurosci. Methods*, 162(1):8–13, 2007. doi: 10/b5dqq8 2

[59] C. Plaisant. The challenge of information visualization evaluation. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, AVI '04, pp. 109–116. ACM, 2004. doi: 10/c7g3x4 2

[60] Prolific. Your first data collection, 2026. Last visited Mar. 18, 2026. https://docs.prolific.com/documentation/get-started/your-first-data-collection. 9

[61] Qualtrics. Survey API Introduction, 2026. Last visited Mar. 18, 2026. https://api.qualtrics.com/f8fdcc4586e1f-survey-api-introduction. 9

[62] S. Sah, R. Mitra, A. Narechania, A. Endert, J. Stasko, and W. Dou. Generating analytic specifications for data visualization from natural language queries using large language models, 2024. NLVIZ Workshop at IEEE VIS '24. https://arxiv.org/abs/2408.13391. 3, 8

[63] A. Sarma, K. Hwang, J. Hullman, and M. Kay. Milliways: Taming multiverses through principled evaluation of data analysis paths. In *ACM Human Factors in Computing Systems*, CHI '24, pp. 607:1–15. ACM, 2024. doi: 10/qw56 9

[64] A. Sarma, A. Kale, M. J. Moon, N. Taback, F. Chevalier, J. Hullman et al. multiverse: Multiplexing alternative data analyses in r notebooks. In *ACM Human Factors in Computing Systems*, CHI '23, pp. 148:1–15. ACM, 2023. doi: 10/qw55 9

[65] A. Satyanarayan, D. Moritz, K. Wongsuphasawat, and J. Heer. Vega-Lite: A grammar of interactive graphics. *IEEE Trans. Visualization & Comp. Graphics*, 23(1):341–350, 2017. doi: 10/f92f32 6

[66] M. Savva, N. Kong, A. Chhajta, L. Fei-Fei, M. Agrawala, and J. Heer. ReVision: automated classification, analysis and redesign of chart images. In *ACM Symposium on User Interface Software and Technology*, UIST '11, pp. 393–402. ACM, 2011. doi: 10/bcbnfw 8

[67] S. Schöttler, J. Dykes, J. Wood, U. Hinrichs, and B. Bach. Constraint-based breakpoints for responsive visualization design and development. *IEEE Trans. Vis. Comput. Graphics*, pp. 4593–4604, 2024. doi: 10/m3f3 8

[68] L. S. Snyder and J. Heer. DIVI: Dynamically interactive visualization. *IEEE Trans. Vis. Comput. Graphics*, 30(1):403–413, 2024. doi: 10/qw54 8

[69] G. Stoet. PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology*, 44(1):24–31, 2017. doi: 10/gftft5 2

[70] R. Van Noorden and J. M. Perkel. AI and science: what 1,600 researchers think. *Nature*, 621:672–675, 2023. doi: 10/gtw8f2 3

[71] H. W. Wang, M. Gordon, L. Battle, and J. Heer. DracoGPT: Extracting visualization design preferences from large language models. *IEEE Trans. Vis. Comput. Graph.*, 31(1):710–720, 2025. doi: 10/qw5w 3, 9

[72] H. W. Wang, J. Hoffswell, S. M. Thazin Thane, V. S. Bursztyn, and C. X. Bearfield. How aligned are human chart takeaways and llm predictions? a case study on bar charts with varying layouts. *IEEE Trans. Vis. Comput. Graphics*, 31(1):536–546, 2025. doi: 10/n82w 3, 9

[73] H. Wickham. A layered grammar of graphics. *J. Comput. Graphical Stat.*, 19(1):3–28, 2010. doi: 10/cttwkf 6

[74] K. Wongsuphasawat, D. Moritz, A. Anand, J. Mackinlay, B. Howe, and J. Heer. Voyager: Exploratory analysis via faceted browsing of visualization recommendations. *IEEE Trans. Vis. Comput. Graph.*, 22(1):649–658, 2016. doi: 10/bdsz 3, 8

[75] K. Wongsuphasawat, Z. Qu, D. Moritz, R. Chang, F. Ouk, A. Anand et al. Voyager 2: Augmenting visual analysis with partial view specifications. In *Proc. CHI*, CHI '17, pp. 2648–2659. ACM, 2017. doi: 10/b7jv 3

[76] A. Wu, Y. Wang, X. Shu, D. Moritz, W. Cui, H. Zhang et al. AI4VIS: Survey on artificial intelligence approaches for data visualization. *IEEE Trans. Vis. Comput. Graphics*, 28(12):5049–5070, 2022. doi: 10/grxn3d 2

[77] A. Wu, Y. Wang, M. Zhou, X. He, H. Zhang, H. Qu et al. MultiVision: Designing analytical dashboards with deep learning based recommendation. *IEEE Trans. Vis. Comput. Graph.*, 28(1):162–172, 2022. doi: 10/gm393p 3

[78] Y. Wu, Z. Guo, M. Mamakos, J. Hartline, and J. Hullman. The rational agent benchmark for data visualization. *IEEE Trans. Vis. Comput. Graphics*, 30(1):338–347, 2024. doi: 10/hbjqpb 9

[79] J. Yang, P. F. Gyarmati, Z. Zeng, and D. Moritz. Draco 2: An extensible platform to model visualization design. In *IEEE Visualization and Visual Analytics*, VIS '23, pp. 166–170. IEEE, 2023. doi: 10/qw5r 3

[80] Z. Zeng and L. Battle. A review and collation of graphical perception knowledge for visualization recommendation. In *ACM Human Factors in Computing Systems*, CHI '23, pp. 820:1–16. ACM, 2023. doi: 10/kz36 3

[81] Z. Zeng, J. Yang, D. Moritz, J. Heer, and L. Battle. Too many cooks: Exploring how graphical perception studies influence visualization recommendations in draco. *IEEE Trans. Vis. Comput. Graph.*, 30(1):1063–1073, 2024. doi: 10/qw5q 3

[82] S. Zhang, H. Li, H. Qu, and Y. Wang. AdaVis: Adaptive and explainable visualization recommendation for tabular data. *IEEE Trans. Vis. Comput. Graph.*, 30(9):5923–5938, 2024. doi: 10/g94bsg 3

[83] J. Zhu, J. Zhang, Q. Zhu, W. Li, J. Wu, and Y. Guo. A knowledge-guided visualization framework of disaster scenes for helping the public cognize risk information. *International Journal of Geographical Information Science*, 38(4):626–653, 2024. doi: 10/g4jdj9 2

[84] C. Ziemkiewicz, M. Chen, D. H. Laidlaw, B. Preim, and D. Weiskopf. *Open Challenges in Empirical Visualization Research*, pp. 243–252. Springer International Publishing, Cham, 2020. doi: 10/qw5h 2