Towards Designing Unbiased Replication Studies in Information Visualization

Poorna Talkad Sukumar *   Ronald Metoyer†

Department of Computer Science and Engineering
University of Notre Dame

ABSTRACT

Experimenter bias and expectancy effects have been well studied in the social sciences and even in human-computer interaction. They refer to the nonideal study-design choices made by experimenters which can unfairly influence the outcomes of their studies. While these biases need to be considered when designing any empirical study, they can be particularly significant in the context of replication studies which can stray from the studies being replicated in only a few admissible ways. Although there are general guidelines for making valid, unbiased choices in each of the several steps in experimental design, making such choices when conducting replication studies has not been well explored.

We reviewed 16 replication studies in information visualization published in four top venues between 2008 to present to characterize how the study designs of the replication studies differed from those of the studies they replicated. We present our characterization categories which include the prevalence of crowdsourcing, and the commonly-found replication types and study-design differences. We draw guidelines based on these categories towards helping researchers make meaningful and unbiased decisions when designing replication studies. Our paper presents the first steps in gaining a larger understanding of this topic and contributes to the ongoing efforts of encouraging researchers to conduct and publish more replication studies in information visualization.

Index Terms: Human-centered computing—Visualization—Empirical studies in visualization; Human-centered computing—Visualization—Visualization design and evaluation methods; Replication; Information Visualization

1 INTRODUCTION

Experimenter biases refer to the nonideal choices made, intentionally or unintentionally, when designing experiments which can hinder their meaningfulness and validity. For example, experimenters can select weak control conditions for evaluating their systems, i.e. "Straw Man Comparisons" [43], which can produce biased results. While these biases need to be kept in mind when designing any empirical study, their consideration is vital to replication studies where the goal is to not only design a meaningful study but to also design a meaningful replication. These biases can be thought of as having a bearing on and a precursor to failed replications and replication crises [40].

Guidelines for making unbiased choices in each of the several steps involved in experimental design are generally known. However, making such choices in replication studies, which are constrained by the studies being replicated to begin with, is less known. We address this shortcoming by characterizing the study-design differences in the replication studies in information visualization and the studies they replicated. We sampled and reviewed 16 replication studies published in four venues – IEEE TVCG, CHI, EuroVis, and PacificVis, between 2008 and 2018.

Our characterization discusses the prevalent use of crowdsourcing and its suitability for replication studies, the types of replications found ranging from near-accurate to high-level conceptual replications, and the commonly-found differences in the experimental designs. We draw guidelines based on this characterization towards the goal of helping researchers design valid, unbiased replication studies.

Previous work by Hornbæk et al. [25] on characterizing replication studies in human-computer interaction (HCI) has mostly focused on the interpretation and reporting of findings in the studies. To our knowledge, no prior work has characterized replication studies in information visualization alone or from an experimental-design perspective and addressed how they can be designed in a more meaningful and unbiased manner, as done in this paper. Additionally, we go beyond the conventional definition of replication and present a broader view including even the replication of only conditions or concepts, which can all contribute to the development of theories and further the maturity of the information visualization field.

We begin by outlining the significance of replication both in general and specifically in information visualization in Section 2. Section 3 is intended to give an overview of the evaluation methods in information visualization and indicate that they are largely comprised of quantitative empirical studies with human subjects. These studies are inherently replicable and the experimenter biases described in Section 4 can be applied to them. Our search, sampling, and coding methods are described in Section 5. We then present our main contributions–the characterization of the study-design differences embedded with the formulated guidelines in Section 6 followed by a brief discussion and conclusion.

2 SIGNIFICANCE OF REPLICATION

A replication is generally defined as a re-evaluation of a previous study, by different researchers and employing different methods, in order to confirm, extend, or generalize its findings [25, 49]. It is considered an essential component of scientific research and should ideally follow breakthroughs in any domain to enable the development of sustaining rules and theories which further the maturity of the domain [15]. It is especially important in domains concerned with studying human factors, such as the social sciences and HCI, because measurements of human behavior normally fluctuate [37].

There is not only a dearth of replication studies generally published in the social sciences and HCI, but most of the replication studies that are indeed published also appear only to be replications of controversial and counterattitudinal findings, modified or "imprecise" replications, or replications of highly-cited work [5, 19, 25, 29, 57].

Efforts have been made to both encourage researchers to conduct more replications and reviewers to attribute more value to replications. Researchers have continually stressed their importance in the social sciences [5, 38, 44, 48, 57] and more recently in the fields

*e-mail: ptalkads@nd.edu
†e-mail: rmetoyer@nd.edu

2018 IEEE Evaluation and Beyond -
Methodological Approaches for Visualization (BELIV)
21 October 2018, Berlin, Germany
978-1-5386-6884-9/18/$31.00 ©2018 IEEE

of HCI and information visualization [2, 6, 6, 19, 20, 25, 30, 33, 51]. Authors provide additional information in their papers regarding their research methods and data to facilitate replications of their work [34, 54, 68]. The *RepliCHI* workshop [70] has been organized as part of the *CHI* conference to spur a "culture shift" towards viewing the merit of replications more favorably within the HCI community.

In the fields of HCI and information visualization, researchers are generally driven to present radical, novel solutions to problems in keeping with the advances in technology and available tools [25, 45, 70]. Hence publication venues don't value replications as highly as cutting-edge work and novel results [19, 70]. The irony, however, is that, while these fields place much emphasis on including sound evaluations in the papers, they are less receptive to replication studies which further solidify the findings of these usually formative evaluations [19].

In information visualization, as in HCI, one of the main challenges in replication is the rebuilding of the visualization interfaces used in earlier studies [2]. Additionally, given that evaluating visualizations is generally hard, the evaluation methodologies in this field are constantly evolving and the use of more qualitative [7] and exploratory techniques [12] are being advocated. While these make replicating more difficult, the proposed methodologies also need to be put into practice and generalized for us to better understand their effectiveness. As Kosara posits, this field is "an empire built on sand" [33] with somewhat loose foundations and in dire need of replications to strengthen the underpinnings of the field.

3 Empirical Research Methods in Information Visualization

Information visualization is full of borrowed methods from HCI and the behavioral and social sciences [7, 41, 67]. Evaluation in information visualization particularly pertains to how visualizations support the perception, cognition, and specific tasks and goals of users [67]. More recent work also include studying the context of visualization-use and how visualizations support communication and collaboration processes [35].

It is well acknowledged within the information visualization community that evaluating visualizations is difficult [7, 12, 50]. The evaluation challenges not only include those commonly found in HCI, such as finding a representative sample of participants or formulating suitable tasks, but also include challenges specific to visualizations. For example, visualization techniques are *generative* in nature, i.e. they can be used for various data sets, tasks, and users and may be effective for only some of these [12]. While we can validate a technique for particular instances, it may not be possible to evaluate the technique itself.

Quantitative empirical evaluations measuring user performance and experience to make statistical inferences are still commonplace in visualization evaluation [35]. However, to address the needs specific to visualization research as well as to support the constantly growing field with new avenues and target users, new and alternative methods and measures for evaluation have also been employed.

To address the complexity inherent in visualization evaluation, such as user interactions occurring on both coarse- and fine-grained levels and difficulty in ascertaining what factors contributed to a finding, the use of more holistic or qualitative [7, 61] and explorative [12] approaches have been proposed. These methods generally result in detailed descriptions of the processes rather than statistical reports and also in uncovering new concepts. However, the main goal in these methods is *transferability* and not *reproducibility* [7, 59], unlike quantitative empirical methods which are more replicable and generalizable.

Information visualization caters to a wide range of applications and appropriate evaluation methods should be used for each. For example, in addition to data analysis, finding insights, and decision making in the presence of large amounts of data, visualizations are increasingly used for light reading in the form of narrative visualizations [60]. In such applications where users' limited time and interest are key factors, it becomes important to not only evaluate the *usability* of the visualizations but to also evaluate the *user experience* afforded by them, often measured by their memorability, engagement, and enjoyment aspects [58].

The experimenter biases that we are concerned with in this paper, which are mainly drawn from the social-science literature [56], don't pertain only to quantitative empirical studies. They can potentially occur even in the context of qualitative methods, such as observations and interviews, and are, in general, applicable even to the recently added methods and measures in visualization evaluation.

However, replicability is ill-defined for purely qualitative methods given that they are subjective, their recording and analysis processes are more variable, and their quality largely depends on the experience of the researchers [7]. Hence replication studies can be expected to only include quantitative empirical studies and qualitative measures that can be numerically recorded using Likert scales. This is evidenced by the *type of evaluation* of the papers sampled for our characterization and presented in Table 1.

4 Experimenter Bias in Empirical Studies with Human Subjects

Experimenter bias refers to the biased or nonideal choices made by experimenters intentionally or unintentionally in designing and running their experiments. These biased choices are often linked to the expectations of the experimenters and can unduly influence the effects observed in the process. We present the potential sources of experimenter bias in quantitative empirical study designs based on Rosenthal's work in the social sciences [56] as well as the research methods in HCI and information visualization [7, 14, 24, 37, 43].

These biases are often alluded to in HCI and information visualization literature presenting guidelines for how to design and run experiments and to avoid pitfalls during the process [14, 24, 43], even though they are not labeled as "biases" as we do in this paper. We make this distinction for two reasons:

1. To distinguish the experimenter biases, which are essentially a type of *systematic* errors, from *random* errors which occur by chance

   Random errors are unpredictable and uncontrollable errors that occur in the observed values in both (high and low) directions and can be offset by averaging. Systematic errors, on the other hand, cause errors in one direction and result in observations that are consistently too low or too high [37, 56]. While both types of errors reduce the reliability of experiments, experimenter biases can be avoided by becoming aware of them and by following appropriate guidelines.

2. To allow for the fact that replication studies, unlike other general studies, may not always be motivated to find significant results and can sometimes intend to do the opposite

   Experimenter bias has been studied to be particularly consequential in replication studies in the social sciences where experimenters are often driven to *disconfirm* an earlier study's (significant) findings [5, 57]. It then only becomes trivial to achieve this goal through a flawed experiment design. For example, noise can be introduced by not controlling for various factors or inconsistent procedures can be followed, which can all result in failed replications (i.e. type 2 errors) and these are aspects that are seldom reported or reported in detail in the papers.

This section is not intended to present an exhaustive list of all the possible experimenter biases that can occur but is intended to give an overview to help the readers better understand how the biases can come into play and how they differ from random errors. We only focus on the processes of designing and running experiments and don't discuss the steps of interpreting and reporting the results, although experimenter biases can also influence these steps. The listed biases are well known, not specific to replication studies, and are generally applicable to any domain employing empirical studies with human subjects.

4.1 Devising hypotheses

Experiments, except purely exploratory studies, normally begin with hypotheses construction which are based on theories and prior work on the topic. However, emphasis on a hypothesis can bias experimenters into expecting particular outcomes. A suggested solution is often to construct multiple, alternative hypotheses which can provide explanations when the experiment results don't conform to one hypothesis [24].

4.2 Independent variables

Biases can occur when the conditions compared in an experiment are incomplete instances and/or not representative of the constructs specified in the hypotheses [24]. They also occur when outdated or weak baseline conditions are chosen for comparison, i.e. "win-lose setups" [24] or "Straw Man Comparisons" [43] and the conditions being compared differ significantly resulting in confounding factors. Lam and Munzner discuss such nonoptimal comparisons between conditions found in information visualization studies where the conditions differed with respect to aspects including their basic visual elements, information content, and interaction complexity [36]. Additionally, one-off studies, i.e. without comparisons, can also be biased in that more favorable ratings may be obtained for the condition being tested [66].

4.3 Dependent variables and measures

The use of measures that don't adequately reflect the qualities of the conditions compared and specified in the hypothesis can cause the study outcomes to be biased. Additionally, in information visualization, measures should be able to sufficiently capture the granularity of the user interactions so that more introspection is afforded in the absence of differences [36].

4.4 Tasks

Common biases occurring with respect to tasks include selecting tasks that are not representative of real-world activities and tasks that can be better performed using the experimenter's system [43].

4.5 Experiment procedure

Many biases can occur when structuring experiments including deciding what variables to control or randomize and whether to use a within- or between-subjects design. These seemingly subtle decisions have the potential to substantially alter the study results. A suggested solution to identify such biases and improve the study design is to run *pilot* studies with real participants [37].

4.6 Sampling

The importance of sampling in empirical studies with human subjects has been widely discussed [6, 7, 24, 40, 43]. The rule of thumb is that the sample should be large enough which can help in offsetting the random errors. However, the composition of the sample also plays a major role in the experiment outcome and validity. Common biases that occur include the selection of a sample that is not representative of the target population and failing to account for the individual differences of the participants which can act as confounding factors in the study.

4.7 Experimenter behavior

In experimenter-participant interactions, the experimenter's body language, delivery of instructions, and inconsistency among different experimenters in conducting the experiment can influence the participant responses in the study [37, 56]. Additionally, participants' positive attitudes towards the experimenter can motivate them to behave as "good subjects" and respond in a way that confirms the experimenter's hypothesis and vice versa [46]. These biases are generally absent in crowdsourced studies where there are no direct interactions between the experimenters and participants.

4.8 Experiment setting

Regardless of whether the experiment is run in a laboratory or in a real-world setting, *physical* environmental aspects, such as the lighting and temperature, and *social* environmental aspects, such as the people in the vicinity and experimenter behaviors, can influence participant responses in the study [37].

It should be noted that while controlling these variables, i.e. keeping them constant throughout the study, may help in alleviating random errors, they may still cause systematic errors to occur during the study by consistently lowering or increasing participant responses.

5 METHOD

Our goal was not to study the aforementioned biases per se in the replication studies but to deduce how unbiased and meaningful replication studies can be designed in information visualization. Towards this goal, we began by searching and sampling replication studies published in four top venues, namely, IEEE TVCG, CHI, EuroVis, and PacificVis, between 2008 to 2018. Our methodology is inspired by and draws from that of Hornbæk et al. [25].

5.1 Paper Search

We faced the same difficulty in searching for relevant papers as reported by Hornbæk et al. [25]. There are no definite keywords for finding replications and they may be called by various other terms, such as reproduction, duplication, or revisiting. Searching for papers including one or more of these terms returned a few hundred papers in each of the digital libraries. Hence we restricted our search to include both the terms "information visualization" (or "infovis") and "replication" (or "replicates", "replicated", "replicating") anywhere in the full text of the papers.

The search returned 80, 86, 6, and 3 results in the IEEE TVCG, CHI, EuroVis, and PacificVis proceedings, respectively. We briefly scanned these articles to identify if they contained replications of previous work and topics pertaining to information visualization. This further reduced our sample to 20, 5, 0, and 1 papers in each of the above proceedings respectively. We then reviewed these papers more carefully considering the sampling criteria described below and finally collected 16 papers overall, 13 published in IEEE TVCG and 3 in CHI. An overview of the collected studies is presented in Table 1.

We are aware that our sample may not include replication studies that are described using different terms or pertain to information visualization without its mention in the papers. However, given that our purpose was to study how the experiments are designed in these studies and not to produce a meta-analysis, our search and sampling methodology suffices our purpose. Additionally, we were also particularly interested in learning what the authors mean when they say they have "replicated" prior work in their studies.

5.2 Sampling Criteria

Similar to Hornbæk et al. [25], we required the papers to contain *quantitative empirical studies with human subjects* and *report an experiment* to be eligible. Furthermore, to ascertain that the studies fell within the precincts of information visualization, we required them to be representative of one or more of the *scenarios* of empirical

Table 1: An overview of the replication studies in information visualization sampled and used in our characterization.

Replication Study		Study being replicated		Type of evaluation according to [35]	Characterization categories (Section 6) applicable to replication study
Publication	Publication venue and year	Publication	Publication venue and year		
Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design [23]	CHI 2010	1. Graphical perception: Theory, experimentation, and application to the development of graphical methods [8] 2. Alpha, contrast and the perception of visual metadata [62]	1. J. Am. Statistical Assoc., 1984 2. Color Imaging Conf. 2009	User Performance	6.1, 6.2, 6.7
Perceptual Guidelines for Creating Rectangular Treemaps [32]	IEEE TVCG 2010	Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design [23]	CHI 2010	User Performance	6.2
The Impact of Social Information on Visual Judgments [26]	CHI 2011	1. Graphical perception: Theory, experimentation, and application to the development of graphical methods [8] 2. Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design [23]	1. J. Am. Statistical Assoc., 1984 2. CHI 2010	User Performance	6.1, 6.4
Assessing the effect of visualizations on bayesian reasoning through crowdsourcing [42]	IEEE TVCG 2012	1. How to improve Bayesian reasoning without instruction: frequency formats [16] 2. Pictorial representations in statistical reasoning [4] and others	1. Psychological Review 1995 2. Applied Cognitive Psychology 2009	User Performance	6.1, 6.4, 6.5, 6.7
How Visualization Layout Relates to Locus of Control and Other Personality Factors [72]	IEEE TVCG 2012	1. Towards the personal equation of interaction: The impact of personality factors on visual analytics interface interaction [17] 2. Using personality factors to predict interface learning performance [18]	1. IEEE VAST 2010 2. HICSS 2010	User Performance, User Experience	6.1, 6.5, 6.7
Does an eye tracker tell the truth about visualizations?: findings while investigating visualizations for decision making [31]	IEEE TVCG 2012	A comparative study of three sorting techniques in performing cognitive tasks on a tabular representation [27]	IJHCI 2013	User Performance	6.7
Influencing Visual Judgment through Affective Priming [21]	CHI 2013	1. Graphical perception: Theory, experimentation, and application to the development of graphical methods [8] 2. Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design [23]	1. J. Am. Statistical Assoc., 1984 2. CHI 2010	User Performance	6.1, 6.4, 6.6
Interactive visualizations on large and small displays: The interrelation of display size, information space, and scale [28]	IEEE TVCG 2013	Sizing up visualizations: effects of display size in focus+context, overview+detail, and zooming interfaces [55]	CHI 2011	User Performance, User Experience	6.2, 6.7
Ranking Visualizations of Correlation Using Webers Law [22]	IEEE TVCG 2014	The perception of correlation in scatterplots [53]	Computer Graphics Forum 2010	User Performance	6.1, 6.2, 6.6, 6.7, 6.8
Four Experiments on the Perception of Bar Charts [65]	IEEE TVCG 2014	Graphical perception: Theory, experimentation, and application to the development of graphical methods [8]	J. Am. Statistical Assoc., 1984	User Performance	6.1, 6.3, 6.7
Improving Bayesian reasoning: The effects of phrasing, visualization, and spatial ability [47]	IEEE TVCG 2016	1. Assessing the effect of visualizations on bayesian reasoning through crowdsourcing [42] and others	1. IEEE TVCG 2012	User Performance	6.1, 6.4, 6.5, 6.7
HindSight: encouraging exploration through direct encoding of personal interaction history [13]	IEEE TVCG 2017	Storytelling in information visualizations: Does it engage users to explore data? [3]	CHI 2015	User Performance, User Experience	6.1, 6.4
The attraction effect in information visualization [10]	IEEE TVCG 2017	1. Between a rock and a hard place: The failure of the attraction effect among unattractive alternatives [39] 2. Distinguishing among models of contextually induced preference reversals [69]	1. Journal of Consumer Psychology 2013 2. Journal of Experimental Psychology: Learning, Memory, and Cognition 1991	User Performance	6.1, 6.4, 6.5, 6.7
Correlation Judgment and Visualization Features: A Comparative Study [71]	IEEE TVCG 2018	1. The perception of correlation in scatterplots [53] 2. Ranking visualizations of correlation using Webers law [22]	1. Computer Graphics Forum 2010 2. IEEE TVCG 2014	User Performance	6.1, 6.3, 6.8
Blinded with Science or Informed by Charts? A Replication Study [11]	IEEE TVCG 2018	Blinded with science: Trivial graphs and formulas increase ad persuasiveness and belief in product efficacy [64]	Public Understanding of Science 2016	User Performance	6.1, 6.3, 6.7, 6.8
Modeling Color Difference for Visualization Design [63]	IEEE TVCG 2018	Enabling designers to foresee which colors users cannot see [52] and others	CHI 2016	User Performance	6.1, 6.5

studies in information visualization described by Lam et al. [35] with the exception of the algorithm-evaluation scenario which typically does not involve human subjects.

Our paper differs from Hornbæk et al. [25] in that we did not set particular criteria for a study to be regarded as a replication. There is ambiguity in how the term "replication" is used and differing definitions of replications can be found [25, 49]. Hence we also wanted to learn how replications can differ and included papers that claimed to replicate previous work irrespective of how the replications were conducted.

5.3 Coding Process

We reviewed the experimental designs described both in the replication studies and the studies they replicated and identified the *differences* between them. These differences pertained to their hypotheses, tasks, conditions, choice of within- or between-subjects, and other aspects of study design. We labeled/coded these differences and then found labels that recurred among the papers. We then identified related labels to form higher-level categories.

Our focus was solely on the differences in the study designs described in the papers and what they imply for the design of replication studies. We were not concerned with the results of the replication studies, i.e what was found in comparison to the study that was replicated and how they were reported.

6 Characterization and Guidelines

The resulting categories of our coding process are presented below along with the associated guidelines within each category.

6.1 Crowdsourced studies

Beginning with Heer and Bostock's replication [23] of the seminal Cleveland and McGill's graphical perception study [8] using Amazon Mechanical Turk (AMT), many replication studies have followed suit. A majority of the studies reviewed in this paper have used crowdsourcing platforms, such as AMT or Crowdflower, to run replication studies (see Table 1).

Crowdsourced studies offer many advantages including large sample sizes and diverse populations with varying backgrounds, education, and gender [23, 42], which can be especially beneficial to replication studies. While most of the studies include strategies that were used to screen suitable participants or to remove the responses of insincere participants, these are still a small price to pay for the advantages provided by crowdsourcing studies.

With the exception of certain types of studies, for example, those requiring eye-tracking [31] or specific testing devices [28], crowdsourcing has been used for replicating visualization evaluations involving visual judgments [21, 23, 26, 32, 65, 71], reasoning [10, 42, 47, 64], and even color perception [63].

Guidelines: Crowdsourcing can help in alleviating biases related to sampling, experimenter behavior, and experiment settings. **Researchers should consider crowdsourcing for replication studies given that it provides opportunities for reevaluating an earlier study's findings with larger, more diverse populations and also to leverage existing experimental data increasingly made available by authors, of their (crowdsourced) studies.**

6.2 Different conditions

Each of the studies in this category have dutifully replicated the methods of an earlier study but also with *different conditions* either to compare an alternative means to test the findings [28] or to validate additional conditions [22, 23, 32].

Rensink and Baldridge's popular study [53] demonstrating that the perception of correlation in scatterplots can be modeled using the psychophysical Weber's law has spurred much interest in the visualization community. Harrison et al. [22] closely replicated their study in a crowdsourced experiment to confirm their results and further extended the study to eight other commonly-used visualizations.

In the two studies presented by Jakobsen and Hornbæk [28, 55], they studied how three different display sizes (small, medium, and large) affect the three visualization interactions of focus+context, overview+detail, and zooming. While they kept the information space constant in the earlier study, i.e. information was cropped in the smaller displays, they scaled the information space in the latter study to fit the display size.

In replicating Cleveland and McGill's graphical perception study [8] for ranking the effectiveness of visual encodings such as length, position, and angle, Heer and Bostock [23] also included additional encoding types such as circular area (e.g., bubble chart) and rectangular area (e.g., treemaps). This latter part of their study involving rectangular area judgments was in turn replicated by Kong et al. [32] in their paper on studying various design attributes associated with rectangular treemaps. They also tested additional conditions including rectangles with more extreme aspect ratios and with different dominant orientations (horizontal or vertical).

Guidelines: The above examples can be considered as conventional replication studies wherein an earlier study is closely replicated to confirm, extend, or generalize its findings [25]. **When the goal is to extend the findings of an earlier study to new or different conditions, it is important to first closely replicate the study for the original conditions.** This process will determine if the replication is successful or not and if it is not, it may be possible to ascertain the causes and potential biases in the study design by comparing it with the earlier study. This comparison will be more difficult if the new conditions are directly validated. **One should also ensure that the new conditions can be meaningfully validated using the same experiment methodology, tasks, and measures.**

6.3 More introspection

We found introspective replication studies where the main goal was to find out *why* an earlier study found what they did. In particular, the studies specified certain attributes that they were more interested in investigating with respect to an earlier study.

Unconvinced by Tal and Wansink's study reports [64] that simple charts can *persuade* people to have a greater belief in a drug's efficacy, Dragicevic and Jansen [11] replicated their study. They added a comprehension test which could provide additional explanations since, to persuade or bias user judgments, the charts must not also increase user understanding [11].

Talbot et al. [65] have built on the aforementioned study by Cleveland and McGill [8], later replicated by Heer and Bostock [23] (Section 6.2), to further explore perceptual tasks associated with bar charts alone. They introduced new conditions to find answers to their open-ended questions including why it's easier to compare adjacent than separated bars and aligned than unaligned bars, and how *distractors*, i.e. intermediate bars between bars being compared, affect these comparisons [65]. They also studied if the inconsistent placement of the *marking dot* in the original study (to denote which bars were to be compared) affected user accuracy.

Yang et al. [71] replicated the aforementioned study by Rensink and Baldridge [53] (Section 6.2) to further investigate if perceiving correlation in a scatterplot is indeed synonymous with perceiving its *visual features* and quite unrelated to one's statistical training.

Guidelines: This theme demonstrates that replication studies can also be more investigative or exploratory in nature. Introspection can be achieved both by adding additional attributes or by simplification, i.e. adopting more controlled designs to study the factors in isolation. **When the goal is to delve deeper into the whys and hows of an earlier study's findings, it is important to (1) formulate high-level (as opposed to well-defined) hypotheses or research questions supported by theory or prior work to guide the study and**

(2) ensure that the specified attributes of interest are adequately justified and their exploration in the context of the earlier study is meaningful and unbiased.

6.4 Replicating conditions rather than study

A common theme found in our paper sample was the replication of a *condition* of an earlier study and not the study itself. The conditions (and associated tasks) of prior work have been leveraged to evaluate newly-proposed ideas in these studies. There are also examples where the replicated conditions are subjected to the same method and measures as in the earlier study making their corresponding results comparable [10].

Feng et al. [13] evaluated their proposed *interaction-history* visual indicators by adding them to existing visualizations. In one of their experiments, a visualization developed by Boy et al. [3] was replicated to be used as the control condition and augmented with their interaction-history aids for comparison.

Hullman et al. [26] and Harrison et al. [21] studied how social information and affective priming can influence visual judgments, respectively. The chart conditions and associated tasks of Cleveland and McGill's graphical perception study [8] (also of Heer and Bostock's study [23]) served as the visual-judgment tasks in their studies to which social information [26] or priming effects [21] were added.

Micallef et al. [42] and Ottley et al. [47] conducted crowdsourced studies to study how Bayesian reasoning is affected by various textual and visualization designs, many of which were adapted from earlier studies including those of Brase and Gary [4], Comsides and Tooby [9] and others.

Similarly, in studying the attraction effect in visualizations, Dimara et al. [10] replicated the numerical-table stimuli used in previous experiments [39, 69] as the control conditions with which their scatterplot-counterparts were compared.

Guidelines: The discussed examples have illustrated two main goals of replicating conditions of previous studies – (1) to augment the conditions with their proposed ideas, and (2) to use them as the control to evaluate other conditions. **Towards these goals, it is important to ensure that (1) the conditions provide a meaningful context for augmenting and testing the newly-proposed ideas, and (2) the replicated conditions are comparable to the other conditions being evaluated to avoid task-comparison biases (see Section 4.4)**.

6.5 High-level, conceptual replications

We found studies [42, 47, 63, 72] that replicated the findings of previous studies on a very high level, with different study designs, e.g., the replication of "prior results that indicate color perceptions can be measured in crowdsourced environments" [63] and failed replication of previous findings in that "visualizations exhibited no measurable benefit" in facilitating Bayesian reasoning [42].

We also found an overlap between some of these study designs and those of the prior studies, such as the aforementioned replicated conditions [42, 47] and dependent variables [72].

Guidelines: High level, conceptual replications can be considered as independent studies each attempting to test a high-level concept in one of potentially several ways. While they can borrow methods and variables from previous studies, they have relatively more freedom and are less constrained by the study-design choices of previous studies. **Hence in designing such studies, as in designing any study, the general experimenter biases such as those discussed in Section 4 will have to considered.**

6.6 Changing within-subjects to between-subjects

While most of the replication studies retained the same within- or between-subject design as in the studies that were replicated, we found two instances where a within-subjects study design was changed to a between-subjects study in the replication studies [21, 22].

However, this change was adequately justified by the authors. A between-subjects design was used to compensate for the variability in the participant responses collected from the crowdsourcing platform in Harrison et al. [22]. Additionally, a larger sample size (~30) was used for each condition compared to the sample size (20) used in the earlier within-subjects study [53].

Although the graphical perception studies [8, 23] adapted by Harrison et al. [21] are typically within-subjects, they used a between-subjects design so that the participants were subjected to only one priming condition, which is typical in priming studies.

Guidelines: While it's best to adhere to the design choices of the study being replicated, additional constraints may arise depending on the type of replication, experiment settings, or participant demographics and individual differences. **When the study design is changed from within-subjects to between-subjects or vice versa in a replication study, adequate justification must be provided ensuring that the study is still meaningful (and also a meaningful replication if a strict replication is the goal) despite the change.**

Although certain types of studies mandate the use of either within-subjects or between-subjects, most other study types don't have this constraint. In such studies, within-subjects (with counterbalancing techniques) is generally used since it requires fewer participants and also minimizes random noise [1]. **When between-subjects is used instead when replicating these studies, it is important to ensure that the study is adequately powered.**

6.7 Larger samples

We observed that bigger samples were generally used in the replication studies, especially those using crowdsourcing, compared to those of the studies being replicated (see Table 1.)

Sampling is a driving factor for replications and it is commonly expressed that low-powered experiments and experiments with small sample-sizes should be replicated [6]. Additionally, failures to replicate and replication crises have often been attributed to smaller sample-sizes in replication studies [40]. Hence, while studies, in general, should use large-enough samples and there can be valid reasons to reject papers for not including enough participants [6], it is more imperative for replication studies to include bigger sample-sizes.

Guideline: Whatever the aims of the replication – to buttress the claims made in a prior study, demonstrate the utility of crowdsourcing, or find alternative explanations for prior studies' findings, it is important to use bigger sample-sizes for the replication to be meaningful and to avoid replication crisis.

6.8 Disparity in the amount of study detail

We found a disparity in the amount of study-design details presented between the replication study papers and the corresponding studies that were replicated. There were both instances where certain aspects were better explained in the former than the latter and vice versa.

For example, although the study by Rensink and Baldridge [53] was replicated by both Harrison et al. [22] and Yang et al. [71], we found the methodology details in the former replication to be more descriptive than those in the latter where it is probably implied that the omitted details were exactly the same as in the earlier study. We also, however, did not find mentions of specific instructions given to participants in the former replication, such as "... it was mentioned that accuracy was important" found in the original study.

While such differences in reporting may seem trivial, even such disparity in the actual study designs can potentially introduce noise and cause biases to occur. On the other hand, the lack of details in the studies being replicated can hinder accurate replications. For example, Dragicevic and Jansen [11] mention that they had difficulty replicating certain aspects due to the lack of details in the original experiment description.

Guidelines: It is necessary for all studies, replication and those that can potentially be replicated, to include detailed study-design descriptions in their publications. Authors should elaborate on all the key aspects of their experiment design (Section 4), and any different choices made as well as the same choices retained in the case of replication studies. While publication page limits can hinder the amount of details reported, these details can be also be included as supplementary material as suggested by Lam and Munzner [36]. While these details can contribute to more faithful replications of the studies, they are also mainly needed for reviewers to ascertain that valid, unbiased experimental designs were employed in the replication studies, which can especially be crucial in the event that the replication results don't conform with those of the study replicated.

Table 2: A summary of the formulated guidelines for designing unbiased replication studies in information visualization based on our characterization.

1. **Consider crowdsourcing** for replication studies which can mitigate biases related to sampling, experimenter behavior, and experiment settings.
2. **Replicate original study closely** before extending to new conditions to facilitate comparison of the study designs. **Ensure the new conditions can be meaningfully validated** using the same approach when extending prior findings.
3. **Apply high-level hypotheses and meaningful exploration of valid attributes** in introspective replications, which delve deeper into the whys and hows of a prior study's findings either by adding additional attributes or by isolating attributes.
4. In studies replicating only a condition(s) of an earlier study, (i) **ensure that the replicated conditions (and tasks) provide a meaningful context** when embedding newly-proposed ideas and (ii) **ensure that the replicated conditions (and tasks) are comparable** when using them as the control to evaluate other conditions.
5. **Consider general experimenter biases** in conceptual replications with independent study designs.
6. **Provide justification** when changing a within-subjects study to between subjects and vice versa. **Ensure the study is adequately powered** when changing a within-subjects study to between subjects.
7. **Use larger samples** in replication studies for the replication to be meaningful and to avoid replication crisis.
8. **Include detailed study-design descriptions** in publications both to facilitate more faithful replications of the studies and for reviewers to ascertain that the replications are valid and unbiased.

7 Discussion and Conclusion

Our guidelines, summarized in Table 2, present the first steps towards developing experimental-design requirements for replication studies. While they are drawn from existing replication studies, there are numerous other possibilities for designing replication studies which are yet to be explored. To expand these guidelines and obtain a more complete understanding of how to perform replications in a meaningful and unbiased manner, it is vital that more replication studies as well as replications of other evaluation types are attempted and published in information visualization.

Our characterization reflects the critical observations made in the social sciences [5] that replication studies are required to "add" to earlier studies' findings in order to be published. This often results in an inclination to further replicate only those studies that have proved generalizable. We observed this trend even in our characterization where studies, such as those of Cleveland and McGill [8] and Rensink and Baldridge [53], are replicated more because they've been proved generalizable. The different replication types covered in our characterization can all contribute to the knowledge associated with the findings of prior work. However, Rosenthal [57] states that, to capitalize on the theory-building potential of replications, it is necessary to also attempt and publish replication studies without extension.

Replications of other evaluation scenarios should also be attempted. All of the sampled studies presented in Table 1 only correspond to the *User Performance* and/or *User Experience* scenarios presented by Lam et al. [35]. There are also many other evaluation types which are replicable, such as the controlled studies described under the *visual data analysis and reasoning* and *communication through visualization* scenarios [35].

We can see from Table 1 that the studies that were replicated are predominantly from non-visualization publication venues. This further demonstrates that information visualization is theoretically built on and continues to derive from various fields such as cognitive psychology. Hence we reiterate what we mentioned in the beginning of this paper that more replication studies are needed to strengthen the underpinnings of this field.

References

[1] Between-Subjects vs. Within-Subjects Study Design. `https://www.nngroup.com/articles/between-within-subjects/`. Accessed: 06-07-2018.

[2] J. Alexander, A. Roudaut, J. Steimle, K. Hornbæk, M. Bruns Alonso, S. Follmer, and T. Merritt. Grand Challenges in Shape-Changing Interface Research. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, p. 299. ACM, 2018.

[3] J. Boy, F. Detienne, and J.-D. Fekete. Storytelling in information visualizations: Does it engage users to explore data? In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 1449–1458. ACM, 2015.

[4] G. L. Brase. Pictorial representations in statistical reasoning. *Applied Cognitive Psychology*, 23(3):369–381, 2009.

[5] R. F. Bronstein. Publication politics, experimenter bias and the replication process in social science research. *Journal of Social Behavior and Personality*, 5(4):71, 1990.

[6] K. Caine. Local standards for sample size at CHI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 981–992. ACM, 2016.

[7] S. Carpendale. Evaluating information visualizations. In *Information visualization*, pp. 19–45. Springer, 2008.

[8] W. S. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American statistical association*, 79(387):531–554, 1984.

[9] L. Cosmides and J. Tooby. Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *cognition*, 58(1):1–73, 1996.

[10] E. Dimara, A. Bezerianos, and P. Dragicevic. The attraction effect in information visualization. *IEEE transactions on visualization and computer graphics*, 23(1):471–480, 2017.

[11] P. Dragicevic and Y. Jansen. Blinded with Science or Informed by Charts? A Replication Study. *IEEE transactions on visualization and computer graphics*, 24(1):781–790, 2018.

[12] G. Ellis and A. Dix. An explorative analysis of user evaluation studies in information visualisation. In *Proceedings of the 2006 AVI workshop*

on BEyond time and errors: novel evaluation methods for information visualization, pp. 1–7. ACM, 2006.

[13] M. Feng, C. Deng, E. M. Peck, and L. Harrison. HindSight: encouraging exploration through direct encoding of personal interaction history. *IEEE transactions on visualization and computer graphics*, 23(1):351–360, 2017.

[14] C. Forsell. A guide to scientific evaluation in information visualization. In *Information Visualisation (IV), 2010 14th International Conference*, pp. 162–169. IEEE, 2010.

[15] B. R. Gaines. Modeling and forecasting the information sciences. *Information Sciences*, 57:3–22, 1991.

[16] G. Gigerenzer and U. Hoffrage. How to improve Bayesian reasoning without instruction: frequency formats. *Psychological review*, 102(4):684, 1995.

[17] T. M. Green and B. Fisher. Towards the personal equation of interaction: The impact of personality factors on visual analytics interface interaction. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*, pp. 203–210. IEEE, 2010.

[18] T. M. Green, D. H. Jeong, and B. Fisher. Using personality factors to predict interface learning performance. In *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*, pp. 1–10. IEEE, 2010.

[19] S. Greenberg and B. Buxton. Usability evaluation considered harmful (some of the time). In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 111–120. ACM, 2008.

[20] C. Gutwin, A. Cockburn, and N. Gough. A Field Experiment of Spatially-Stable Overviews for Document Navigation. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 5905–5916. ACM, 2017.

[21] L. Harrison, D. Skau, S. Franconeri, A. Lu, and R. Chang. Influencing visual judgment through affective priming. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2949–2958. ACM, 2013.

[22] L. Harrison, F. Yang, S. Franconeri, and R. Chang. Ranking visualizations of correlation using weber's law. *IEEE transactions on visualization and computer graphics*, 20(12):1943–1952, 2014.

[23] J. Heer and M. Bostock. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 203–212. ACM, 2010.

[24] K. Hornbæk et al. Some whys and hows of experiments in human–computer interaction. *Foundations and Trends® in Human–Computer Interaction*, 5(4):299–373, 2013.

[25] K. Hornbæk, S. S. Sander, J. A. Bargas-Avila, and J. Grue Simonsen. Is once enough?: on the extent and content of replications in human-computer interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 3523–3532. ACM, 2014.

[26] J. Hullman, E. Adar, and P. Shah. The impact of social information on visual judgments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1461–1470. ACM, 2011.

[27] I. Hur, S.-H. Kim, A. Samak, and J. S. Yi. A comparative study of three sorting techniques in performing cognitive tasks on a tabular representation. *International Journal of Human-Computer Interaction*, 29(6):379–390, 2013.

[28] M. R. Jakobsen and K. Hornbæk. Interactive visualizations on large and small displays: The interrelation of display size, information space, and scale. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2336–2345, 2013.

[29] K. S. Jones, P. L. Derby, and E. A. Schmidlin. An investigation of the prevalence of replication research in human factors. *Human factors*, 52(5):586–595, 2010.

[30] M. Kay and J. Heer. Beyond weber's law: A second look at ranking visualizations of correlation. *IEEE transactions on visualization and computer graphics*, 22(1):469–478, 2016.

[31] S.-H. Kim, Z. Dong, H. Xian, B. Upatising, and J. S. Yi. Does an eye tracker tell the truth about visualizations?: findings while investigating visualizations for decision making. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2421–2430, 2012.

[32] N. Kong, J. Heer, and M. Agrawala. Perceptual guidelines for creating rectangular treemaps. *IEEE transactions on visualization and computer graphics*, 16(6):990–998, 2010.

[33] R. Kosara. An Empire Built On Sand: Reexamining What We Think We Know About Visualization. In *Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization*, pp. 162–168. ACM, 2016.

[34] K. Kurzhals, C. F. Bopp, J. Bässler, F. Ebinger, and D. Weiskopf. Benchmark data for evaluating visualization and analysis techniques for eye tracking for video stimuli. In *Proceedings of the fifth workshop on beyond time and errors: novel evaluation methods for visualization*, pp. 54–60. ACM, 2014.

[35] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale. Empirical studies in information visualization: Seven scenarios. *IEEE transactions on visualization and computer graphics*, 18(9):1520–1536, 2012.

[36] H. Lam and T. Munzner. Increasing the utility of quantitative empirical studies for meta-analysis. In *Proceedings of the 2008 Workshop on BEyond time and errors: novel evaLuation methods for Information Visualization*, p. 2. ACM, 2008.

[37] J. Lazar, J. H. Feng, and H. Hochheiser. *Research methods in human-computer interaction*. Morgan Kaufmann, 2017.

[38] J. W. Lucas. Theory-testing, generalization, and the problem of external validity. *Sociological Theory*, 21(3):236–253, 2003.

[39] S. A. Malkoc, W. Hedgcock, and S. Hoeffler. Between a rock and a hard place: The failure of the attraction effect among unattractive alternatives. *Journal of Consumer Psychology*, 23(3):317–329, 2013.

[40] S. E. Maxwell, M. Y. Lau, and G. S. Howard. Is psychology suffering from a replication crisis? What does failure to replicate really mean? *American Psychologist*, 70(6):487, 2015.

[41] J. E. McGrath. Methodology matters: Doing research in the behavioral and social sciences. In *Readings in Human–Computer Interaction*, pp. 152–169. Elsevier, 1995.

[42] L. Micallef, P. Dragicevic, and J.-D. Fekete. Assessing the effect of visualizations on bayesian reasoning through crowdsourcing. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2536–2545, 2012.

[43] T. Munzner. Process and pitfalls in writing information visualization research papers. In *Information visualization*, pp. 134–153. Springer, 2008.

[44] J. W. Neuliep. Editorial bias against replication research. *Journal of Social Behavior and Personality*, 5(4):85, 1990.

[45] W. Newman. A preliminary analysis of the products of HCI research, using pro forma abstracts. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 278–284. ACM, 1994.

[46] A. L. Nichols and J. K. Maner. The good-subject effect: Investigating participant demand characteristics. *The Journal of general psychology*, 135(2):151–166, 2008.

[47] A. Ottley, E. M. Peck, L. T. Harrison, D. Afergan, C. Ziemkiewicz, H. A. Taylor, P. K. Han, and R. Chang. Improving Bayesian reasoning: The effects of phrasing, visualization, and spatial ability. *IEEE transactions on visualization and computer graphics*, 22(1):529–538, 2016.

[48] C. L. Park. What is the value of replicating other studies? *Research Evaluation*, 13(3):189–195, 2004.

[49] R. D. Peng. Reproducible research and biostatistics. *Biostatistics*, 10(3):405–408, 2009.

[50] C. Plaisant. The challenge of information visualization evaluation. In *Proceedings of the working conference on Advanced visual interfaces*, pp. 109–116. ACM, 2004.

[51] D. Reed and E. H. Chi. Online privacy; replicating research results. *Communications of the ACM*, 55(10):8–9, 2012.

[52] K. Reinecke, D. R. Flatla, and C. Brooks. Enabling designers to foresee which colors users cannot see. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 2693–2704. ACM, 2016.

[53] R. A. Rensink and G. Baldridge. The perception of correlation in scatterplots. In *Computer Graphics Forum*, vol. 29, pp. 1203–1210. Wiley Online Library, 2010.

[54] H. Romat, C. Appert, B. Bach, N. Henry-Riche, and E. Pietriga. Animated Edge Textures in Node-Link Diagrams: a Design Space and Initial Evaluation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, p. 187. ACM, 2018.

[55] M. Rønne Jakobsen and K. Hornbæk. Sizing up visualizations: effects of display size in focus+ context, overview+ detail, and zooming interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1451–1460. ACM, 2011.
[56] R. Rosenthal. Experimenter effects in behavioral research. 1966.
[57] R. Rosenthal. Replication in behavioral research. *Journal of Social Behavior and Personality*, 5(4):1, 1990.
[58] B. Saket, A. Endert, and J. Stasko. Beyond usability and performance: A review of user experience-focused evaluations in visualization. In *Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization*, pp. 133–142. ACM, 2016.
[59] M. Sedlmair, M. Meyer, and T. Munzner. Design study methodology: Reflections from the trenches and the stacks. *IEEE transactions on visualization and computer graphics*, 18(12):2431–2440, 2012.
[60] E. Segel and J. Heer. Narrative visualization: Telling stories with data. *IEEE transactions on visualization and computer graphics*, 16(6):1139–1148, 2010.
[61] B. Shneiderman and C. Plaisant. Strategies for evaluating information visualization tools: multi-dimensional in-depth long-term case studies. In *Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization*, pp. 1–7. ACM, 2006.
[62] M. Stone and L. Bartram. Alpha, contrast and the perception of visual metadata. In *Color and Imaging Conference*, vol. 2008, pp. 355–359. Society for Imaging Science and Technology, 2008.
[63] D. A. Szafir. Modeling Color Difference for Visualization Design. *IEEE transactions on visualization and computer graphics*, 24(1):392–401, 2018.
[64] A. Tal and B. Wansink. Blinded with science: Trivial graphs and formulas increase ad persuasiveness and belief in product efficacy. *Public Understanding of Science*, 25(1):117–125, 2016.
[65] J. Talbot, V. Setlur, and A. Anand. Four experiments on the perception of bar charts. *IEEE transactions on visualization and computer graphics*, 20(12):2152–2160, 2014.
[66] M. Tohidi, W. Buxton, R. Baecker, and A. Sellen. Getting the right design and the design right. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pp. 1243–1252. ACM, 2006.
[67] M. Tory and T. Moller. Human factors in visualization research. *IEEE transactions on visualization and computer graphics*, 10(1):72–84, 2004.
[68] J. Walny, S. Huron, C. Perin, T. Wun, R. Pusch, and S. Carpendale. Active Reading of Visualizations. *IEEE transactions on visualization and computer graphics*, 24(1):770–780, 2018.
[69] D. H. Wedell. Distinguishing among models of contextually induced preference reversals. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 17(4):767, 1991.
[70] M. L. Wilson, E. H. Chi, S. Reeves, and D. Coyle. RepliCHI: the workshop II. In *CHI'14 Extended Abstracts on Human Factors in Computing Systems*, pp. 33–36. ACM, 2014.
[71] F. Yang, L. Harrison, R. A. Rensink, S. Franconeri, and R. Chang. Correlation Judgment and Visualization Features: A Comparative Study. *IEEE Transactions on Visualization and Computer Graphics*, 2018.
[72] C. Ziemkiewicz, A. Ottley, R. J. Crouser, A. R. Yauilla, S. L. Su, W. Ribarsky, and R. Chang. How visualization layout relates to locus of control and other personality factors. *IEEE transactions on visualization and computer graphics*, 19(7):1109–1121, 2013.