Tasks and Telephones: Threats to Experimental Validity due to Misunderstandings of Visualisation Tasks and Strategies

Position Paper

Abhraneel Sarma* Northwestern University
Sheng Long† Northwestern University
Michael Correll‡ Northeastern University
Matthew Kay§ Northwestern University

ABSTRACT

Empirical studies in visualisation often compare visual representations to identify the most *effective* visualisation for a particular visual judgement or decision making task. However, the effectiveness of a visualisation may be intrinsically related to, and difficult to distinguish from, individual-level factors such as visualisation literacy. Complicating matters further, visualisation literacy itself is not a singular intrinsic quality, but can be a result of several distinct challenges that a viewer encounters when performing a task with a visualisation. In this paper, we describe how such challenges apply to experiments that we use to evaluate visualisations, and discuss a set of considerations for designing studies in the future. Finally, we argue that aspects of the study design that are often neglected or overlooked (such as the onboarding of participants, tutorials, training, etc.) can have a big role in the results of a study and can potentially impact the conclusions that the researchers can draw from the study.

Index Terms: Experiment design, visualisation effectiveness, visualisation misinterpretation, visualisation literacy

1 INTRODUCTION

The goal of many empirical studies in visualisation is to determine the effectiveness of various visual representations—whether a particular visualisation type is "better" than others in helping a viewer perform a particular task. This empirical groundwork is established through rigorous quantitative experiments. There have been many such empirical studies that have demonstrated that certain visualisation types can improve performance in a range of visualisation tasks (e.g., accuracy and precision in estimation [21, 23, 29, 35, 40] or quality of decision-making [12, 22, 41, 45]).

However, recent work also outlines a number of challenges, such as misinterpretation of the task, misunderstanding of the visual encoding etc., that viewers face when performing tasks using a visualisation [34]. These challenges likely extend to and persist in empirical scenarios where we test the effectiveness of novel visualisation idioms over existing ones. In this paper, we ponder whether these barriers in correctly interpreting data visualisations may constitute a threat to the commonly adopted experimental framework in visualisation research.

For instance, more complex visualisations are considered more likely to be misinterpreted. Yet, many complex visual representations such as the probability density and cumulative distribution plots used by Fernandes et al. [12], which the average viewer is likely not very familiar with, have been found to result in high-quality decisions. From a literacy standpoint, visualisations that are more challenging and are more frequently misinterpreted can be considered as requiring more "literacy", "competency" [17], or ability—if teaching a viewer to interpret a bar chart or a pie chart is considered the equivalent of teaching concepts of arithmetic and geometry in mathematics, one could argue that teaching interpretations of probability density plots or cumulative distribution function plots is equivalent to teaching probability or calculus. Thus, evaluating the effectiveness of a visualisation (such as a probability density plot) without giving participants the requisite training on how to interpret the chart would be akin to testing a student proficient in arithmetic with questions on calculus. This highlights the need to provide participants with the requisite training in visualisation studies that purport to assess the effectiveness of visualisations.

The previous example serves to highlight the gap between what we, as researchers, may intend to assess and what we actually assess. Training is but one tool in our arsenal to bridge this gap. We outline the need to more carefully consider how researchers can leverage aspects of experimental design—tutorials, incentives, training and feedback, pilot studies—to bridge this gap and improve the ecological validity of our empirical research.

2 EFFECTIVENESS, LITERACY AND WHY ANY OF THIS MATTERS FOR EXPERIMENTS

2.1 Visualisation Effectiveness and Literacy

As previously mentioned, much empirical research in visualisation focuses on identifying the most effective visualisation for a particular task. However, visualisation effectiveness[1] [28, 31] is a somewhat ambiguous concept. Prior work has offered many definitions of effectiveness, which whilst appealing, can be challenging to operationalise and suffers from a "nuance trap" [16]. Instead, we adopt a more crude but easily operationalisable interpretation based on prior work that ranks the effectiveness of different chart types [7, 10, 18, 29]—whether a particular visual representation leads to better performance on a task, as measured by some reasonable metric. Examples of such reasonable metrics include accuracy (smaller bias and/or greater precision), decision-making quality, and Just Noticeable Differences (JNDs) [11].

Visualisation literacy describes "a person's ability to understand data presented in graphical elements." However, the term visualisation literacy is perhaps overloaded. Broadly construed, a person's ability to make sense of the information presented in a chart depends on a large range of skills such as critical thinking, spatial awareness, working memory capacity, statistical training, etc. [4, 17]. Yet, for the sake of simplicity, we will persist with the term *visualisation literacy* as an umbrella term for all of these skills and use it in its broadest possible meaning.

The principles of visualisation effectiveness and visualisation literacy are seemingly intrinsically related. Cabouat et al. [4] argue that individual factors, which include but are not limited to visualisation literacy, affect any assessment of the effectiveness of a visualisation. As such, assessments of visualisation effectiveness

*e-mail: abhraneel@u.northwestern.edu
†e-mail: shenglong@u.northwestern.edu
‡e-mail: m.correll@northeastern.edu
§e-mail: mjskay@northwestern.edu

[1] while Munzner [31] primarily discusses effectiveness in terms of visual channels, we adopt a wider notion of the effectiveness of the entire visual representation.

should consider the viewers' literacy and other factors contributing to individual differences. Similarly, any assessment of visualisation literacy should consider visual representations of varying effectiveness. In the case of visualisation literacy assessment tests (e.g., [3, 9, 26]), one possible explanation for the current status quo is likely that these tests are implicitly controlling for visualisation effectiveness by choosing specific combinations of visual representation and task type. More concretely, item response theory models which are often used to evaluate these assessment tests [9, 13] have an item (question) difficulty parameter which, in this case, would depend on both visualisation effectiveness and task difficulty. In empirical studies evaluating the effectiveness of two or more visual representations for a particular task, the status quo has three possible explanations: (1) in randomised controlled experiments, by virtue of randomisation, we assume a baseline balance of visualisation literacy across experimental conditions; (2) we tend to consider visualisation literacy as a singular construct (even though it is likely not [4, 17]; also see §2.2); and (3) we tend to hand wave these concerns away because we tend to get uncomfortable and do not like to answer difficult questions which threaten the validity of our research.[2,3]

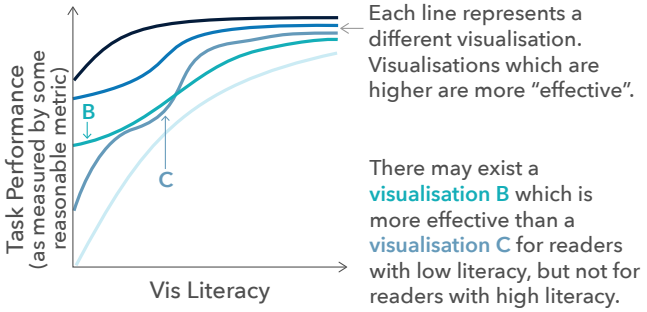

Figure 1: Presumed inter-relationship between literacy and effectiveness. Note that we are merely speculating that this is what these curves might look like. Our best guess is that these are likely to be monotonic (with respect to literacy). "More effective" visualisations should usually result in better average task performance than a "less effective" visualisation, but there are nuances and caveats (as shown by curves B and C).

Figure 1 describes what we presume to be a plausible effect of visualisation effectiveness on the relationship between literacy and task performance. Given a particular visualisation, an individual with greater literacy will, on average, perform better than individuals with lower literacy at tasks designed to be performed using this visualisation as they might have some knowledge about which patterns to look for. On the other hand, there may also exist two visualisations (for e.g., B and C in Figure 1) such that one visualisation is more effective for readers with low literacy, while the other is more effective for readers with high literacy (see [6, 10]). Thus, the conventional notion of visualisation effectiveness—where effectiveness is an intrinsic property of a visualisation—need not always hold (and, as evidenced by recent work, it does not always hold [10]). Finally, it is not unreasonable to consider a diminishing effect of visualisation effectiveness as the literacy of the individual performing the task increases.

2.2 Barriers

Recent work by Nobre et al. [34] outline a set of nine distinct barriers which may prevent readers of a chart from accurately performing tasks using a visualisation (see Figure 2 for an overview). These

[2] we are joking

[3] okay, maybe only slightly joking

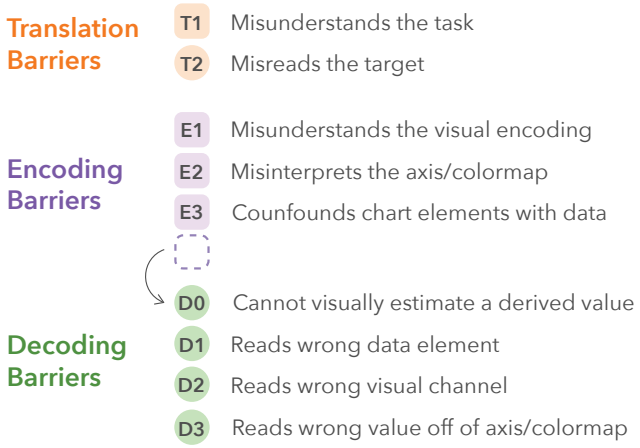

Figure 2: Figure adapted from Nobre et al. [34] providing an overview of the nine visualisation literacy barriers identified in their work. While in their work *D0 (Cannot visually estimate a derived value)* was classified as an encoding barrier (E4), we believe that this might be better classified as a decoding barrier.

barriers highlight the interplay between the design of the visualisation and the literacy (and other individual factors) of the reader of the visualisation. For instance, consider the barrier E1 (misunderstanding the task)—if a reader fails to perform a task due to this barrier, does that mean that they have lower literacy or that the designer of the visualisation could have improved the design of the visualisation to safeguard the reader from this barrier, perhaps for instance by adding annotations? Or perhaps, both? When evaluating the effectiveness of a visualisation in an experiment, the presence of these barriers, in some contexts, can have a potential confounding effect on any measure of effectiveness. To better understand the impact of these barriers in visualisation experiments, we first lay out a framework describing the idealised process of how researchers expect to measure effectiveness through experiments, and recast these barriers as breakdowns in the idealised process.

2.3 The Telephone Framework

As researchers, when we design experiments, we provide participants (a user) with a task T to perform using a visualisation V, which represents some data using some encoding; we implicitly assume that users will be performing the task T using an optimal strategy S ("we," the authors, acknowledge that this is not always the case and there have been a growing number of studies in recent years which have elicited qualitative descriptions from participants about the strategies they have used in performing tasks [e.g., 22, 41]). Based on how well users perform T, we assess the effectiveness of a visualisation for T. This process is represented using the top panel in Figure 3. However, it is very likely that this idealised process is not realised in practice. As participants come to an experiment with varying knowledge and goals—individual factors—which can impact how they perform in an experiment [5]. Thus, one possibility is that the user develops a distinct mental model of what task (T') they are supposed to be performing, and a strategy S' which they think best supports task T'. Thus, what the researchers end up assessing is how well a visualisation V supports a user using the strategy (S'), which they think best supports the task (T')—which may or may not be the task they are supposed to be doing (T)—assuming they are even able to successfully carry out their strategy S'. This alternate pathway now appears to resemble the popular children's game telephone.

In the context of experimental design, this framework allows us to recast the barriers identified by Nobre et al. [34] as systematic breakdowns or deviations from the ideal pathway. For instance,

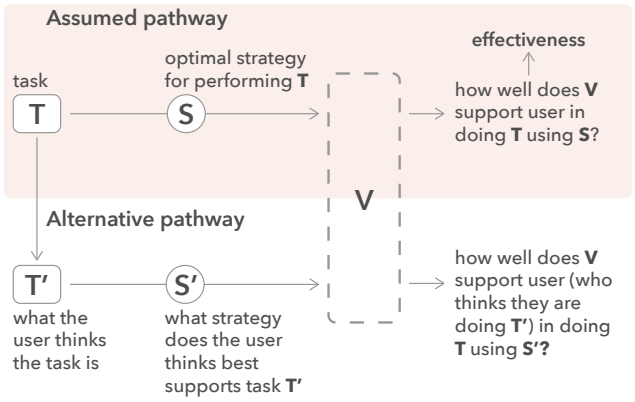

Figure 3: The telephone framework for experimental design to determine effectiveness of a visual representation

the barriers T1 and T2 would tend to arise where there is a gap between T and T'; the barriers E1–E3 result from the user identifying a strategy S' which is far from the ideal strategy S; the barriers D1–D3 are all a result of the user not being able to execute their strategy S' properly (even if $S' = S$). Only the barrier D0, which may manifest in both the idealised and deviated pathways, can likely be considered an indicator of the effectiveness of a visualisation.

This raises two pertinent questions: first, can these breakdowns potentially manifest in ways that can impact the conclusions about the effectiveness of the visualisations being tested? (answer: yes); second, do aspects of the current experimental design framework help in addressing these barriers? (answer: perhaps?) To answer these questions, we first need to review the typical components of experiments in visualisation (§3). We then provide examples of studies which have had to contend with breakdowns from the desired experimental process (§4). Finally, we outline a set of best practices for the design of experiments and make a case to study the effects of some of these factors which are still not well understood (§5).

3 The Components of Experiments in Visualisation

Broadly, visualisation experiments can be broken down into the following components:

1. **Informed consent**, which often consists of descriptions of the experiment, what it entails, duration, etc.
2. **Introduction and tutorial**, which typically describes to participants the exact task that they have to perform, the information that will be provided to them to perform the task, and often, how to interpret that information.
3. **Training trials**, which can help provide the participant familiarity with the task (in the case of incentivised studies, without the expectation that it will impact their final pay). Additionally, if feedback is also provided, it can potentially help clear up misunderstandings related to the task.
4. **Test trials** (often repeated multiple times) are the primary focus of data collection and comprise the bulk of most experiments.
5. **Attention check questions**, which are often interspersed between the test trials, especially when the study involves repeated trials. Attention check questions are meant to eliminate poor actors who are not actually responding to the stimulus and randomly answering the task questions.[4]
6. **Qualitative or Likert-style Questions** eliciting participants experience interacting with the stimuli in the trials. In typical visualisation experiments, these include preference or confidence ratings, cognitive load questionnaires etc.
7. **Demographic questions** are often asked to understand the background of the participants and are often useful to understand the type of participants.

Although we attempt to list these in the order in which they typically appear, and in the order that we would place them if we were designing an experiment in the future, this order is not definitive (e.g., demographics questions can be collected either at the beginning or end of the experiment).

In addition, researchers employ two other tricks to ensure the robustness of experiments: **pilot studies**, and **randomisation of participants between conditions**. Pilot studies help the researcher identify if there is a possibility for potential misinterpretation of either the task or the visual representation. Randomisation, if implemented properly, helps ensure baseline balance—all baseline covariates such as participants' ability are randomly distributed across treatment groups [1, 47].

4 Deviations from the Assumed Pathway, in Practice

We describe how experiments for assessing visualisation effectiveness can suffer from deviations from the desired pathway. Through this discussion, we take the opportunity to highlight how these deviations often occur due to insufficient use of "safeguards"—components of experimental design such as tutorials, training, and pilot studies.

4.1 Task Misunderstandings

How do issues related to misunderstanding the task occur in experiments, and do these issues limit the conclusions that can be drawn from the results of a study? Broadly construed, this breakdown tends to manifest as participants performing a slightly different task than what is intended by the experimenters (i.e. $T' \not\approx T$).

Assessing participants based on a different task than what they are performing. In a study by Zgraggen et al. [46], participants were presented with datasets and were asked to perform exploratory data analysis tasks. The datasets had some genuine correlations as well as some Gaussian noise. Specifically, participants were instructed to "find and report any reliable observations" and "write down textual descriptions [...] about any observations they wanted to report." Zgraggen et al. found that over 60% of the generated insights were false. Examining the study design using our proposed framework, we believe that the task the experimenters wanted the participants to perform was T: *report any reliable insights from the dataset such that the false discovery rate was at most 5%*, while the tasks the participants likely believed that they were performing was T': *report any reliable insights or interesting patterns from the dataset*, without a specific interpretation of term "reliable."

We conducted a conceptual replication [41] of this study using crowdsourced participants on Prolific. This was a particularly tricky study to design, and one of the primary challenges was to ensure that participants were actually performing the intended task—identifying as many reliable insights as possible, while accounting for potential false discoveries in the data at a specific $\alpha = 0.25$ level. A retrospective reflection reveals two aspects of our study design which we believe likely helped prevent participants from misinterpreting T: (1) providing participants with explicit incentives tied to their monetary compensation for participation; and (2) a training and feedback module.

Due to the incentives, participants are not only aware of when they are going to be rewarded (if they are able to successfully identify correct insights) and when they will be penalised (if they incorrectly identified insights that are false positives), but also the

[4] e.g., see Prolific's attention check question guidelines: https://researcher-help.prolific.com/hc/en-gb/articles/360009223553-Prolific-s-Attention-and-Comprehension-Check-Policy

acceptable false discovery rate in order to obtain a positive payout. The training and feedback module, which comprised of five trials prior to the actual test trials, helped them understand how they were going to be evaluated. Throughout the study, we provided participants with a cumulative points total to inform them of how they were performing in the test trials.

Qualitative descriptions of the strategies that participants used reveal that several participants were able to learn good strategies for accomplishing the task over the course of the trials. In other words, we believe that training and consistent feedback likely helped many participants identify $S' \approx S$. Further, some participants also mentioned that while they were aware of what they needed to do for the task, they still performed poorly as they could not identify a "good" strategy for doing so—suggesting that the breakdown was in correctly identifying S, and not in understanding T. While we do not claim that these steps are sufficient to completely eliminate breakdowns in identifying T when designing somewhat complex decision-making tasks, we do believe that our study's results would not have been viable without these steps.

Finally, our research question ("whether participants in a multiple comparisons scenario adjust for multiple comparisons") itself was to see if participants are able to identify a good strategy for performing the task. Thus, our study results should reflect that what we are measuring is a composite of *how well does V support identification of* $S \approx S'$ *in performing* T and *are participants able to execute* S' *to perform* T. We leave it to the reader to decide if the reporting of the results holds up to this standard.

Using Phrasing Variation as a Task Manipulation A recent study by Oral et al. [36] claims that studies in visualisation often conflate the terms "judgement" and "decision making", and investigates whether participants are more accurate when performing a judgement task compared to a decision-making task. The distinction between how the two tasks in this study were operationalised was entirely based on how they were phrased: participants were either asked "what is the best option?" (an observation framing for the judgment task) or "what do you choose?" (an action-oriented framing for the decision-making task).

Examining the study design using the telephone framework, we believe that the authors expected the phrasing manipulation to result in the T' identified by the participants in the decision and judgements conditions to be different. The authors likely expected that this would subsequently lead to slightly different S'. Our primary concern with this study design is that it rests on the assumption that this slight phrasing manipulation results in (or should result in) different cognitive processes (i.e. identification of different T' and S') in a participant which would manifest in their responses. While it is not obvious that participants in this study misunderstood the task, it is possible that participants may not be performing the task as the experimenters want them to perform. Their quantitative analysis found a small, but statistically non-significant difference between performance in the two conditions. Moreover, participants' qualitative descriptions also did not reveal any differences in how/whether the participants are performing the two tasks differently [36]. As such, we believe that there may be a disconnect between the intended task on behalf of the experimenters, and the actual task that participants are performing.

Does misunderstanding the task always impact the results? The phrasing of task questions is an important aspect of experiment design. In a recent study investigating when and what type of truncation in bar charts is appropriate [27], we experienced challenges in eliciting the desired response on a *compare ratios* task[5]. We went through multiple iterations of question phrasing as there was a potential for some participants to misunderstand this task. Despite these precautions, in our results, participants were on average less accurate in the *compare ratios* task ($\sim$ 52–74%) than in a *compare gap* task[6] ($\sim$ 93–94%). This could be either because the *compare ratios* task is perceptually and cognitively more difficult than the *compare gap* task, or it could also be because participants were more prone to misunderstand the *compare ratios* task, and the experiment design does not allow us to distinguish between the two possibilities. However, this does not have to mean that the results of this study are invalid. In this study, task was a between-subjects condition, and the primary comparisons of interest were manipulated within-subjects. Thus, participants across all conditions which were being compared were presented with the same task and the same phrasing of the task, implying that task misinterpretation would not manifest differently between the conditions being compared.

The distinction we raise here is admittedly subtle but important—it is important to ensure that factors such as task misunderstanding (and even strategy mis-identification, as we will see later) are balanced across all the conditions being compared, or acknowledged as a possible factor driving the results. Consider the following hypothetical counter-example: you are conducting an experiment, where no training is provided to participants, to test two visualisations for effectiveness. One of the visualisations being compared *makes the correct task more obvious*. The conclusions that can be drawn from this study depend on the communication context. If you (as a designer) intend to deploy the visualisation in situations where training is not possible, then the experiment will indicate how users will behave with the visualisation. However, if you intended to deploy the visualisation in situations where users may receive training, the results of your study would not reflect the effectiveness of the visualisation in these scenarios.

4.2 Strategy Mismatches

Visualisations are frequently misinterpreted. When a bar chart is used to visualise the means of several categories, some viewers are prone to the *within-the-bar bias*—a point that falls within the bar is judged as being more likely than a point that falls outside the bar, even if the two points are equidistant from the mean [8, 25, 32]—a phenomenon which occurs because viewers misunderstand the visual encoding. Similarly, the cone of uncertainty which is commonly used in hurricane forecasts is also frequently misinterpreted—regions immediately outside the cone are judged to be at significantly lower levels of risk of hurricane damage when compared to areas just inside the cone [37, 39]. Broadly, we can consider these misinterpretations of the visual encoding (E1) [34] of a chart to result in users performing the task with visualisation by adopting strategies that do not match the designer's intended strategy (i.e. $S' \not\Leftrightarrow S$).

Discrepancy between the optimal strategy and the actual strategy identified by participants. In a study by Hofman et al. [20], participants were asked to estimate the probability of superiority between two probability distributions. The distributions were presented to participants using either 95% confidence intervals (i.e., the uncertainty in the estimate of the mean) or 95% prediction intervals (i.e., uncertainty in the location of future observations). The authors provide participants with a textual description of what the visual idioms—the error bars—meant (see Figure 4). The study found that participants in the 95% prediction interval (PI) condition made judgments of *probability of superiority* (and a *willingness to pay* value) which were closer to the ground truth value.

It is important to note a few things here. First, confidence intervals are notoriously difficult to interpret and numerous studies have shown that even experts with statistical training frequently fail

[5] the compare ratios task entailed comparing the ratios between two pairs of bars in a bar chart

[6] the compare gap task entailed comparing the difference in heights between two pairs of bars in a bar chart

Description for (text matched) 95% CI condition

Your average sliding distance with the special boulder: If you were to slide the special boulder 1,000 times, you would attain an average distance of 104 meters.

Variation in your sliding distances with the special boulder: Roughly speaking 95% of your next 1,000 slides would be between 74 and 134 meters, with approximately equal percentages of slides falling short of 74 meters or long of 134 meters. The graph shows your average distance and this interval without the special boulder (left) and with the special boulder (right), as indicated by the black points and vertical bars

Description for (text matched) 95% PI condition

Your average sliding distance with the special boulder: If you were to slide the special boulder 1,000 times, you would attain an average distance of 104 meters.

Uncertainty in estimating your true average sliding distance with the special boulder: A 95% confidence interval on your aerage sliding distance with a special boulder is 103 to 105 meters. A 95% confidence interval is constructed such that if we watched many such sessions of 1,000 slides and repeated this process, 95% of the constructed intervals would contain your true average.

Figure 4: Description of confidence interval (CI) and prediction interval (PI) in the study by Hofman et al. [20] (adapted from Table 2 in https://osf.io/za2d9). In this study, participants were presented with the mean point estimate and uncertainty distribution for two boulders—regular and special. The regular boulder has an average sliding distance of 100 metres (95% CI: [99, 101]; 95% PI: [70, 130]). They were then asked to provide estimates for *probability of superiority* and *willingness to pay*.

to correctly interpret the information presented with a confidence interval [19, 30] or fail to distinguish between error bars depicting standard errors and confidence intervals [2]. Second, a subsequent study by Kale et al. [22] found that participants largely relied on heuristics for making decisions which involved comparing two distributions represented using confidence intervals, although they did not test prediction intervals; some of the more commonly used strategies included comparing the two distributions based on the distance between the means, distance between the means relative to the uncertainty or distribution overlap. Finally, the information presented in the PI condition would lead to a "more correct" answer (i.e. closer to the ground truth) than the CI condition, if a participant were to use some of the heuristics identified by Kale et al. [22].

As the participants in the study by Hofman et al. [20] were lay people, it raises the question of whether the tutorial provided to them (Figure 4) was sufficient and allowed them to truly understand the *semantic* difference between the presented confidence and prediction intervals, or if they simply relied on heuristics to make their judgments. In other words, the study cannot distinguish between two possible explanations for the result: were participants in both conditions able to correctly identify the optimal strategy S but fail to execute the strategy properly? Or, did they merely identify an S' which coincidentally allowed them to perform better with one representation (95% PI), but not with the other (95% CI)?

As is, we believe that the current study, due to lack of explicit training, best supports claims about the holistic understanding of the visual representation, task, and strategy; thus the most reasonable conclusion here is that a 95% PI would lead to more accurate responses of the probability of superiority measure and the willingness to pay measure, but without much insight into why they might be better. However, if this improvement is indeed due to the use of a heuristic, we might conceivably be able to find a different task (for e.g., inferential tasks as opposed to predictive tasks) where PIs might perform poorly when compared to CIs.[7]

On the other hand, if the authors want to conclude that the 95% PI is better because it allows participants to identify and execute the optimal strategy for performing the task, it is likely necessary to explicitly train participants on how to correctly interpret the visual representation, and be supported through additional qualitative data about strategies that people are using. This would be useful to allay potential concerns such as whether participants are able to meaningfully distinguish between error bars used to depict PIs and CIs. Finally, if the authors are interested in making claims about the perceptual effectiveness of the visual representation (assuming participants adopt the respective optimal strategies), then participants may need training in both how to interpret the visualisation and what the optimal strategy is.

5 Takeaways for Experimental Design

What does all of this mean for the design of experiments? We believe that thinking about experimental design using such a framework can enable us as researchers to strengthen the ecological validity of our results. Moreover, certain components of experiments (§3) can serve as useful tools to ensure that there is minimal deviation from the ideal experiment pathway—disparity between T and T' or between S and S'.

5.1 Task

If the objective of our study is to make claims about the task and the visualisation—visualisation V is more suited for Tasks T_i and T_j—we will need to discuss the steps taken to ensure minimal task misunderstanding as well as demonstrate that the participants understand, and are performing, the task we want them to perform. On the other hand, if we are not making claims about the task (e.g., truncation in the bar charts study [27]), then we can design our experiment to allow our analysis to marginalise over the effect of any potential task misunderstandings.

The various tools of experimental design that are available to the researcher, can be used to make sure there is minimal discrepancy between the task that we researchers want a participant to perform and the task they perceive we want them to perform. A first step would be to conduct pilot studies of the experiment, where the (pilot) participants are asked to think aloud [33] their thoughts as they are performing the task. This is likely the fastest and simplest way to detect potential task misunderstandings. In an ideal world, we would advocate for multiple rounds of pilot studies, with at least one round consisting of participants who are sampled from the same population as the actual study. Specifically, if the intent is to ultimately run the study on participants recruited from a crowdsourcing platform, we should avoid resorting to the convenience of running very small pilot studies with other graduate students in visualisation or computer science—it is very likely that the challenges faced by a graduate student in the same field is not going to be the same as the challenges that the actual study participants are going to encounter. As an additional step, we also recommend researchers collect qualitative data to help verify whether participants understand the task.

5.2 Strategy

The goal of a study could be to either determine whether a user is able to identify the "optimal" strategy for performing a task with a visualisation, or it could be to identify how well a user is able to execute this "optimal" strategy. Empirical work in visualisation often conflates the two, and this is understandable—it may be reasonable to consider a visualisation not good for a particular task if it does not make readily obvious the optimal strategy for performing the task. Yet, strictly speaking, only the latter (how well a user is able to execute the "optimal" strategy with the visualisation to perform

[7] see also: https://statmodeling.stat.columbia.edu/2023/08/16/confusions-about-inference-prediction-and-probability-of-superiority/

the task) is providing a measure of effectiveness, in the traditional sense of the term. Moreover, both of these are reasonable research questions, but we should be explicit in choosing which we intend to study as there are different implications for the design of the experiment.

If the goal of a study is to determine whether a user is able to identify the "optimal" strategy for performing a task with a visualisation, researchers would need to ensure that all participants at least have a comparable understanding of the visual representation (and what it encodes). One way of achieving this can be through proper introduction and tutorial of the visual representations, as "good instruction-writing" can ensure that all participants have the necessary knowledge to perform the experiment [5]. If the introduction and tutorial sections for one of the conditions in a study is poorly worded, and the participants in that condition have lower scores on an effectiveness metric, researchers will not be able to distinguish between whether the visualisation is actually ineffective or whether it was misunderstood. At the risk of over-correction, it is also possible to provide an excessive amount of information in this section, beyond what a user in most contexts in which the visualisation is going to be used has access to, threatening the ecological validity of the study.

On the other hand, if the goal is to specifically assess how well a visualisation supports the optimal strategy, we would need to ensure that most participants in our studies adopt the optimal strategy (or close-to-optimal strategy). In this case, training and feedback can help dispel concerns of mis-identification of the optimal strategy by, at the very least, informing the participants if the strategies that they had initially identified were sub-optimal and providing them with information to refine their strategy. As before, asking participants to describe the strategies that they used to perform the task through qualitative questionnaires can often reveal valuable insights. The use of explicit incentive structures, along with feedback can also indirectly help reinforce notions of the optimal strategy. This is because deviations from the optimal strategy can become quickly evident through sub-optimal payoffs.

5.3 Effectiveness and External Validity

So far our discussion has overlooked the potential impact of motivation and engagement. In the context of experiment design in behavioral economics, Camerer and Hogarth [5] claims that certain people might be intrinsically motivated to participate in an experiment; recent work in visualisation has found that viewers may have varying levels of intrinsic motivation to engage with the information presented in a visualisation [15, 24, 38]. Beyond making sure that there is a baseline balance across experimental conditions, such that participants in one condition are not more motivated or engaged than those in other conditions, we consider here other ways in which these factors can impact experimental results—we identify external validity of the visualisation as one such instance.

Consider a visualisation deployed in a public space or in a journalism article—in such scenarios, users engaging with the visualisation may have varying levels of skills and motivation. Thus, as briefly discussed at the end of §4.1, the most ecologically valid evaluation of the effectiveness of such a visualisation would be to not provide participants extensive tutorials and training on how to use the graph. Conversely, if a visualisation is going to be used in a highly specialised domain, it would not be unrealistic to assume that the users will receive extensive training on how to use the visualisation. Hence, ecologically valid evaluations of the effectiveness of such visualisations should reflect the training that the target users are likely to receive (in addition to ensuring that users have the appropriate amount of knowledge to perform the intended tasks with the visualisation, for instance, by recruiting from the intended target population).

5.4 The Multiverse of the Oft Overlooked Experimental Design Components

The preceding discussion highlights how the often overlooked components of a study—onboarding, introduction, tutorial, and training to name a few—can have a non-trivial impact on its results. Yet, there is no standard process (nor could there be one) for creating these components. As a result, researchers are required to decide and make implicit choices on how to create them, or even whether to include them in a study at all. These decisions, if mentioned at all, often tend to be relegated to the supplementary materials of a paper. Much like the recent discussion on how various analytical choices, for a particular research question, can result in substantial variation in the results [14, 42, 43, 44], so too can the decisions involved in creating the experimental design components.

As such we feel it is important to ask—what is the extent to which such decisions impact the results? For instance, in an experiment assessing the effectiveness of visualisations, how much of an impact does including or not including a training module have on the results? Suppose we assess a visualisation using an experiment where no training was provided. Could we extrapolate these results to an "alternative universe" where the target users received some form of training? Or, could the results be so drastically different that they imply qualitatively opposite conclusions?[8] We believe that it might be valuable to systematically study the impact of the various decisions that go into creating the experimental design components.

6 Conclusion

In this paper, we posit that there exist many threats to the experimental validity of studies that claim to assess the effectiveness of visualisations. Much of these potential threats stem from potential misunderstandings of the task that experimenters want participants to perform or mis-identification of the strategies that participants should "ideally" adopt in performing the task with a visual representation. These misunderstandings are intrinsically related to visualisation literacy, which in itself is a complex construct. We highlight how these threats can often manifest in practice, and how they can impact the conclusions that researchers can draw from a study. Finally, we discuss how various, often overlooked, components of experiments can be used to strengthen the design of studies and ensure greater ecological validity.

References

[1] A. D. Althouse, K. Z. Abebe, G. S. Collins, and F. E. H. Jr. Response to "Why all randomized controlled trials produce biased results". *Annals of Medicine*, 50(7):545–548, 2018. PMID: 30122065. doi: 10.1080/07853890.2018.1514529 3

[2] S. Belia, F. Fidler, J. Williams, and G. Cumming. Researchers misunderstand confidence intervals and standard error bars. *Psychological methods*, 10:389–96, Dec. 2005. doi: 10.1037/1082-989X.10.4.389 5

[3] J. Boy, R. A. Rensink, E. Bertini, and J.-D. Fekete. A principled way of assessing visualization literacy. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1963–1972, 2014. doi: 10.1109/TVCG.2014.2346984 2

[4] A.-F. Cabouat, T. He, F. Cabric, T. Isenberg, and P. Isenberg. Position paper: A case to study the relationship between data visualization readability and visualization literacy. In *CHI 2024 - Workshop Toward a More Comprehensive Understanding of Visualization Literacy*. O'ahu (Honolulu), United States, May 2024. 1, 2

[8] As is the case with most things, the answer likely lies somewhere between these two extremely optimistic and pessimistic scenarios.

[5] C. F. Camerer and R. M. Hogarth. *The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework*, p. 7–48. Springer Netherlands, Dordrecht, 1999. doi: 10.1007/978-94-017-1406-8_2 2, 6

[6] S. C. Castro, P. S. Quinan, H. Hosseinpour, and L. Padilla. Examining effort in 1d uncertainty communication using individual differences in working memory and nasa-tlx. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):411–421, 2022. doi: 10.1109/TVCG.2021.3114803 2

[7] 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. doi: 10.1080/01621459.1984.10478080 1

[8] M. Correll and M. Gleicher. Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2142–2151, 2014. doi: 10.1109/TVCG.2014.2346298 4

[9] Y. Cui, L. W. Ge, Y. Ding, F. Yang, L. Harrison, and M. Kay. Adaptive assessment of visualization literacy. *IEEE Transactions on Visualization and Computer Graphics*, 30(1):628–637, 2024. doi: 10.1109/TVCG.2023.3327165 2

[10] R. Davis, X. Pu, Y. Ding, B. D. Hall, K. Bonilla, M. Feng, M. Kay, and L. Harrison. The risks of ranking: Revisiting graphical perception to model individual differences in visualization performance. *IEEE Transactions on Visualization and Computer Graphics*, 30(3):1756–1771, 2024. doi: 10.1109/TVCG.2022.3226463 1, 2

[11] M. A. Elliott, C. Nothelfer, C. Xiong, and D. A. Szafir. A design space of vision science methods for visualization research. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1117–1127, 2021. doi: 10.1109/TVCG.2020.3029413 1

[12] M. Fernandes, L. Walls, S. Munson, J. Hullman, and M. Kay. Uncertainty displays using quantile dotplots or cdfs improve transit decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, p. 1–12. Association for Computing Machinery, New York, NY, USA, 2018. doi: 10.1145/3173574.3173718 1

[13] L. W. Ge, Y. Cui, and M. Kay. Calvi: Critical thinking assessment for literacy in visualizations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23. Association for Computing Machinery, New York, NY, USA, 2023. doi: 10.1145/3544548.3581406 2

[14] A. Gelman and E. Loken. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*, 348:1–17, 2013. 6

[15] H. A. He, J. Walny, S. Thoma, S. Carpendale, and W. Willett. Enthusiastic and grounded, avoidant and cautious: Understanding public receptivity to data and visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 30(1):1435–1445, 2024. doi: 10.1109/TVCG.2023.3326917 6

[16] K. Healy. Fuck nuance. *Sociological Theory*, 35(2):118–127, 2017. doi: 10.1177/0735275117709046 1

[17] M. Hedayati, A. Hunt, and M. Kay. From pixels to practices: Reconceptualizing visualization literacy. In *CHI 2024 - Workshop Toward a More Comprehensive Understanding of Visualization Literacy*. O'ahu (Honolulu), United States, May 2024. 1, 2

[18] 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*, CHI '10, p. 203–212. Association for Computing Machinery, New York, NY, USA, 2010. doi: 10.1145/1753326.1753357 1

[19] R. Hoekstra, R. D. Morey, J. N. Rouder, and E.-J. Wagenmakers. Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, 21(5):1157–1164, Oct. 2014. doi: 10.3758/s13423-013-0572-3 5

[20] J. M. Hofman, D. G. Goldstein, and J. Hullman. How visualizing inferential uncertainty can mislead readers about treatment effects in scientific results. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, p. 1–12. Association for Computing Machinery, New York, NY, USA, 2020. doi: 10.1145/3313831.3376454 4, 5

[21] J. Hullman, P. Resnick, and E. Adar. Hypothetical outcome plots outperform error bars and violin plots for inferences about reliability of variable ordering. *PLOS ONE*, 10(11):1–25, 11 2015. doi: 10.1371/journal.pone.0142444 1

[22] A. Kale, M. Kay, and J. Hullman. Visual reasoning strategies for effect size judgments and decisions. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):272–282, 2021. doi: 10.1109/TVCG.2020.3030335 1, 2, 5

[23] M. Kay, T. Kola, J. R. Hullman, and S. A. Munson. When (ish) is my bus? user-centered visualizations of uncertainty in everyday, mobile predictive systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, p. 5092–5103. Association for Computing Machinery, New York, NY, USA, 2016. doi: 10.1145/2858036.2858558 1

[24] H. Kennedy and R. L. Hill. The feeling of numbers: Emotions in everyday engagements with data and their visualisation. *Sociology*, 52(4):830–848, 2018. doi: 10.1177/0038038516674675 6

[25] S. H. Kerns and J. B. Wilmer. Two graphs walk into a bar: Readout-based measurement reveals the Bar-Tip Limit error, a common, categorical misinterpretation of mean bar graphs. *Journal of Vision*, 21(12):17–17, 11 2021. doi: 10.1167/jov.21.12.17 4

[26] S. Lee, S.-H. Kim, and B. C. Kwon. Vlat: Development of a visualization literacy assessment test. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):551–560, 2017. doi: 10.1109/TVCG.2016.2598920 2

[27] S. Long and M. Kay. To cut or not to cut? a systematic exploration of y-axis truncation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24. Association for Computing Machinery, New York, NY, USA, 2024. doi: 10.1145/3613904.3642102 4, 5

[28] J. Mackinlay. Automating the design of graphical presentations of relational information. *ACM Trans. Graph.*, 5(2):110–141, apr 1986. doi: 10.1145/22949.22950 1

[29] C. M. McColeman, F. Yang, T. F. Brady, and S. Franconeri. Rethinking the ranks of visual channels. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):707–717, 2022. doi: 10.1109/TVCG.2021.3114684 1

[30] R. D. Morey, R. Hoekstra, J. N. Rouder, M. D. Lee, and E.-J. Wagenmakers. The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, 23(1):103–123, feb 2016. doi: 10.3758/s13423-015-0947-8 5

[31] T. Munzner. *Visualization analysis and design*. A K Peters visualization series. CRC Press, Taylor & Francis Group, CRC Press is an imprint of the Taylor & Francis Group, an informa business, Boca Raton, 2015. 1

[32] G. E. Newman and B. J. Scholl. Bar graphs depicting averages are perceptually misinterpreted: The within-the-bar bias. *Psychonomic Bulletin & Review*, 19(4):601–607, Aug. 2012. doi: 10.3758/s13423-012-0247-5 4

[33] J. Nielsen. *Usability engineering*. Interactive Technologies. AP Professional, Cambridge, Mass, 1st edition ed., 1993. 5

[34] C. Nobre, K. Zhu, E. Mörth, H. Pfister, and J. Beyer. Reading between the pixels: Investigating the barriers to visualization literacy. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24. Association for Computing Machinery, New York, NY, USA, 2024. doi: 10.1145/3613904.3642760 1, 2, 4

[35] B. D. Ondov, F. Yang, M. Kay, N. Elmqvist, and S. Franconeri. Revealing perceptual proxies with adversarial examples. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1073–1083, 2021. doi: 10.1109/TVCG.2020.3030429 1

[36] B. Oral, P. Dragicevic, A. Telea, and E. Dimara. Decoupling judgment and decision making: A tale of two tails. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–12, 2023. doi: 10.1109/TVCG.2023.3346640 4

[37] L. M. Padilla, I. T. Ruginski, and S. H. Creem-Regehr. Effects of ensemble and summary displays on interpretations of geospatial uncertainty data. *Cognitive Research: Principles and Implications*, 2(1):40, Oct. 2017. doi: 10.1186/s41235-017-0076-1 4

[38] E. M. Peck, S. E. Ayuso, and O. El-Etr. Data is personal: Attitudes and perceptions of data visualization in rural pennsylvania. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, p. 1–12. Association for Computing Machinery, New York, NY, USA, 2019. doi: 10.1145/3290605.3300474 6

[39] I. T. Ruginski, A. P. Boone, L. M. Padilla, L. Liu, N. Heydari, H. S. Kramer, M. Hegarty, W. B. Thompson, D. H. House, and S. H. Creem-Regehr. Non-expert interpretations of hurricane forecast uncertainty visualizations. *Spatial Cognition & Computation*, 16(2):154–172, 2016. doi: 10.1080/13875868.2015.1137577 4

[40] A. Sarma, S. Guo, J. Hoffswell, R. Rossi, F. Du, E. Koh, and M. Kay. Evaluating the use of uncertainty visualisations for imputations of data missing at random in scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):602–612, 2023. doi: 10.1109/TVCG.2022.3209348 1

[41] A. Sarma, X. Pu, Y. Cui, M. Correll, E. T. Brown, and M. Kay. Odds and insights: Decision quality in exploratory data analysis under uncertainty. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24. Association for Computing Machinery, New York, NY, USA, 2024. doi: 10.1145/3613904.3641995 1, 2, 3

[42] J. P. Simmons, L. D. Nelson, and U. Simonsohn. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, 22(11):1359–1366, 2011. 6

[43] U. Simonsohn, J. P. Simmons, and L. D. Nelson. Specification curve analysis. *Nature Human Behaviour*, 4(11):1208–1214, 2020. 6

[44] S. Steegen, F. Tuerlinckx, A. Gelman, and W. Vanpaemel. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5):702–712, 2016. doi: DOI: 10.1177/1745691616658637 6

[45] F. Yang, M. Hedayati, and M. Kay. Subjective probability correction for uncertainty representations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23. Association for Computing Machinery, New York, NY, USA, 2023. doi: 10.1145/3544548.3580998 1

[46] E. Zgraggen, Z. Zhao, R. Zeleznik, and T. Kraska. Investigating the effect of the multiple comparisons problem in visual analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pp. 479:1–479:12. ACM, New York, NY, USA, 2018. doi: 10.1145/3173574.3174053 3

[47] W. Zhao and V. Berger. Imbalance control in clinical trial subject randomization—from philosophy to strategy. *Journal of Clinical Epidemiology*, 101:116, 2018. 3