Effective Use of Likert Scales in Visualization Evaluations: A Systematic Review

Laura South[iD], David Saffo[iD], Olga Vitek[iD], Cody Dunne[iD], and Michelle A. Borkin[iD]

Northeastern University

Abstract

Likert scales are often used in visualization evaluations to produce quantitative estimates of subjective attributes, such as ease of use or aesthetic appeal. However, the methods used to collect, analyze, and visualize data collected with Likert scales are inconsistent among evaluations in visualization papers. In this paper, we examine the use of Likert scales as a tool for measuring subjective response in a systematic review of 134 visualization evaluations published between 2009 and 2019. We find that papers with both objective and subjective measures do not hold the same reporting and analysis standards for both aspects of their evaluation, producing less rigorous work for the subjective qualities measured by Likert scales. Additionally, we demonstrate that many papers are inconsistent in their interpretations of Likert data as discrete or continuous and may even sacrifice statistical power by applying nonparametric tests unnecessarily. Finally, we identify instances where key details about Likert item construction with the potential to bias participant responses are omitted from evaluation methodology reporting, inhibiting the feasibility and reliability of future replication studies. We summarize recommendations from other fields for best practices with Likert data in visualization evaluations, based on the results of our survey. A full copy of this paper and all supplementary material are available at `https://osf.io/exbz8/`.

CCS Concepts

• **Human-centered computing** → **Visualization design and evaluation methods;** *Empirical studies in visualization;*

1. Introduction

Evaluations are critical for assessing the validity of visualization techniques and systems [IIC*13, Kos16]. The metrics by which visualizations and systems are considered successful depend on the application and the researchers' goals. Many researchers are interested in objective measures, such as time or accuracy, while others are interested in subjective data sources, such as ease of use or user confidence. Subjective response is often measured qualitatively through interviews or free-response questions and consequently analyzed with qualitative techniques such as open coding or rich description. However, qualitative methods are not always appropriate for a subjective evaluation. For example, researchers might be interested in directly comparing the subjective performance of two visualizations, a task that could be more difficult with unstructured qualitative data. **Likert scales** allow researchers to collect quantitative estimates of subjective traits, producing numeric data that can be summarized and visualized in the similar manner to other quantitative data collected in an evaluation.

Guidelines intended to help researchers use Likert scales in a responsible manner are nearly as ubiquitous as Likert scales themselves across a multitude of scientific disciplines, such as agriculture [CD94], pharmaceuticals [Har15], and psychology [LK18]. These guidelines exist for good reason; the process of running a study with a Likert questionnaire is full of potential pitfalls that can jeopardize the scientific validity of a study. Prior to running an experiment, researchers must make decisions about the construction of their Likert scales, such as the number of response options to include and the phrasing of individual Likert item statements. These details of Likert scale construction can affect how participants respond to a questionnaire, so researchers must be careful when constructing their Likert scales and precise when reporting their experimental procedure in their paper. Even after the Likert questionnaire has been constructed, researchers must decide how they will analyze and interpret the data collected from the questionnaire. The measurements produced by Likert scales can be interpreted as ordinal (i.e., discrete) or interval (i.e., continuous) in nature, depending on how the Likert item was constructed. The decision to interpret Likert data as ordinal or interval influences the methods researchers use to summarize and, if applicable, run statistical tests on Likert data. Ordinal data are generally summarized using the median as a summary statistic and analyzed using nonparametric statistical procedures, while arithmetic mean and parametric statistical procedures are appropriate for interval data.

Despite the volume of guidelines that have been written about Likert scales, there is noticeable variation in how Likert data are reported and interpreted in visualization evaluations. For example,

some evaluations handle Likert data according to best practices for quantitative data. These papers report effect sizes alongside measures of uncertainty, include visualizations of the data, and conduct statistical analysis, such as estimation or NHST, if the Likert data is used for confirmatory research. Other evaluations take different approaches to analyzing Likert data, such as reporting means without a measure of uncertainty or describing the responses of individual participants without a discussion about broader patterns in participant responses. Some authors give detailed descriptions of the Likert questionnaires given to participants, while others include only a handful of details about what questions participants were asked. Lack of detail in reporting how Likert questionnaires were constructed can impede valid replications in future studies and mislead readers and reviewers, as we explain in Section 3.1. There are plenty of guidelines for the proper handling of Likert scales, but do visualization researchers follow the best practices set forth in these guidelines? What consequences might we face if we do not adhere to best practices for handling Likert data in our visualization evaluations?

This survey paper contributes the first systematic analysis of Likert scale use in visualization evaluations. Through a literature survey and systematic review of 134 visualization papers, we investigate three aspects of Likert scale usage with the potential to impact the scientific validity of a visualization evaluation. First, we find that authors frequently omit details about how Likert questionnaires were constructed and presented to research participants. These details can affect participant responses, and are necessary to help readers contextualize results and enable other researchers to reliably replicate studies in the future. Second, we identify inconsistencies in the interpretation of Likert data as interval or ordinal within individual papers and discuss the adverse impacts these inconsistencies could have on the validity of results. Finally, we compare the handling of objective and subjective measures in visualization evaluations and find that Likert data are often handled differently than objective measures in studies that collect both types of data. As a resource for the visualization community, we additionally provide a concise summary of best practices and guidelines drawn from other fields including statistics, psychology, and HCI to assist visualization researchers in the effective use of Likert scales (Section 6.2). All paper and survey materials are available at https://osf.io/exbz8/.

2. Related work

2.1. Improving visualization evaluations

Literature surveys of published visualization evaluation papers have historically been used to better understand current evaluation practices within the visualization community and to illustrate the need for improvements. Lam et al. and Isenberg et al. studied the characteristics and goals of visualization evaluations in general [IIC*13, LBI*11], while Hullman et al. conducted a survey specific to evaluations of uncertainty visualizations [HQC*18]. Several other notable critiques of visualization evaluation practices are relevant to our work, such as Correll's discussion of the "heroic age" of visualization [Cor20] and Kosara's examination of what we truly know about visualization after decades of perceptual evaluations [Kos16]. Guidelines specific to qualitative and quantitative
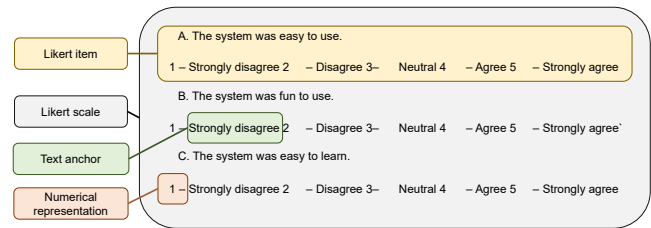

Figure 1: *A Likert scale consists of a series of Likert items, each accompanied by numbered response options and text anchors.*

data in visualization evaluations were also useful to us in compiling our recommendations for the handling of quantitative subjective data. Meyer and Dykes provide guidelines for improving the rigor of qualitative research in visualization [MD19]. Several papers document the current state of quantitative evaluations in HCI and visualization and propose recommendations for improving rigor and external validity [Cai16, KR12, Kos16]. Statistical practices in HCI and visualization have been brought into question in recent years. A particular concern is the field's overreliance on dichotomous inference as a result of NHST [BD19] and the potential for a replication crisis in empirical computer science research [CDBG20]. Several solutions have been proposed to address these issues, including requiring the preregistration of analysis plans prior to running studies [CGD18]. In this paper, we focus on documenting how subjective evaluations with Likert scales are handled in visualization papers, with an emphasis on two aspects of statistical practices: reporting of methodological details for replication (Section 3.1) and methods used to summarize, analyze, and report Likert data (Sections 3.2 and 3.3).

2.2. Understanding the Likert scale

Proper methods for handling Likert data are rarely mentioned in evaluation methodology work within HCI or visualization, aside from Kaptein, Nass, & Markopoulos' recommendations for nonparametric analysis methods for Likert data in HCI studies [KNM10]. For additional background on Likert scales, in Section 3 we pull from the wide variety of fields that rely on these scales to collect quantitative subjective data. Harpe provides a comprehensive overview of the history and appropriate use of Likert scales [Har15]. Liddell and Kruschke demonstrated through a simulation study that applying metric models to ordinal data can lead to erroneous conclusions, after a survey of psychology papers found that most researchers used parametric tests when analyzing Likert data [LK18]. Other studies have found the opposite effect when comparing the power of parametric and nonparametric tests on simulated datasets, demonstrating that both types of tests perform equally well on Likert data [CP08, Nor10]. In this paper, we contribute a survey of how Likert scales are used in-the-wild in visualization evaluations.

3. Likert scale background

3.1. Constructing Likert scales

A **Likert scale** consists of one or more statements or questions accompanied by a range of response options. Each individual item in the questionnaire is referred to as a **Likert item**; the term Likert scale refers to all the Likert items in the questionnaire as whole. See Figure 1 for an illustrated overview of terms. Several elements of Likert item construction can influence how participants perceive and respond to a subjective questionnaire, such as the number of response options [CRSH17], the text anchors corresponding to each response option [HB10], the phrasing of each Likert item [FA99], and the numerical representation attached to each response option [SKH*91]. There is no single recipe that is perfect for all scenarios; in most cases, the correct construction for an given Likert item depends on the context of the study.

Validated subjective questionnaires such as the System Usability Scale (SUS) [B*96] or NASA Task Load Index (NASA-TLX) [HS88] can provide a quick and easy way for researchers to determine the usability or cognitive workload of a visualization system using a standardized questionnaire format. The majority of visualization papers in our survey (Section 4) chose to construct custom Likert questionnaires (123 papers, 92%), rather than using standardized questionnaires to elicit subjective response. Of the eleven papers that used a standardized questionnaire, seven chose to add a custom Likert questionnaire to their subjective evaluation. Only three papers relied solely on a standardized questionnaire. Visualization researchers' tendency to create new Likert questionnaires for their evaluations makes it critical for our field to understand how differences in Likert item construction can affect participant response. Here we detail **four elements of Likert item construction** with known response effects and provide examples of common implementations:

1. **Number of choices**: The response options that the participant has to choose from. Most Likert items present 5 or 7 options, although even-numbered items without a "Neutral" option are appropriate in certain scenarios, such as when participants are familiar with the subject matter or operating under social desirability bias (i.e., if respondents feel pressure to select an option that is perceived as more socially accepted) [CRSH17].
2. **Text anchors**: The written descriptors that accompany each numeric response option on the Likert scale. Examples include "Strongly disagree" to "Strongly agree", "Not at all confident" to "Very confident", and "Unsatisfied" to "Satisfied". Providing text anchors for all response options yields more reliable responses than labeling only endpoints [Wen04], except when Likert items are used to estimate linear relationships [WCS10].
3. **Question phrasing**: The statement that participants are asked to respond to. The phrasing most often takes the form of a statement that participants can agree or disagree with, such as "I enjoyed using the visualization tool."
4. **Numerical representation**: The characteristics of the numbers used to distinguish response options. Ascending scales (e.g., 1 to 5, 0 to 6) are most common, but symbolic (e.g., – to ++) and diverging (e.g., -2, to 2) representations are also used.

A conceptual replication (i.e., a replication focused on detecting the same effect in a new experiment) requires well-documented methods and analyses in the original paper in order to be reliable [Kos16]. Replication studies, while currently uncommon in visualization literature, are important for testing the validity of the conclusions that form our scientific understanding. An effect that can be detected by more than one experimenter in separate studies (i.e., replications) is likely to be robust and not merely the result of an experimental or statistical error [Kos16, KH18]. Replications can also help to identify and weed out incorrect conclusions when effects detected in earlier studies *fail* to replicate. Proper reporting of Likert scale design is critical to enabling future replication studies, as *each of the four elements described above can affect how participants respond to a Likert scale*. Without detailed descriptions of Likert scale implementations, it would be impossible to know whether contradictory results from a replication study truly repudiate the effect found in the original study or if participants simply responded differently to the construction of a given Likert item. Our first research question addresses this issue:

RQ1: Do visualization evaluations include sufficient methodological details regarding the collection of Likert data to enable reliable future replication studies?

3.2. Interpreting Likert scales

The original formulation of the Likert scale, as proposed by Likert himself, required that responses to the individual Likert items within the Likert scale be accumulated or aggregated prior to analysis [Lik32]. For example, a participant's response to a Likert scale with three individual Likert items, within the original definition, would be the sum or the average of their response to the three items. Analyzing accumulated Likert scale responses in this manner is very uncommon in visualization literature; only one paper in our literature survey (Section 4) explicitly analyzed aggregate responses to multiple Likert items— [RHR15]. In an effort to understand and catalog how Likert scales are used in visualization evaluations today, our discussion in this paper focuses on analyzing Likert responses on an item-by-item basis rather than accumulated responses. For the purposes of accurately describing current practices in visualization research, the phrase "Likert data" throughout the rest of the paper refers to individual Likert items, rather than accumulated scores from multiple items.

When Likert data are summarized with descriptive statistics, visualized in a chart, or analyzed with a statistical procedure, an assumption is made to interpret the Likert response as **interval** or **ordinal** in nature. If the Likert data are assumed to be from an ordinal scale, they should be summarized with descriptive statistics that require no knowledge beyond the relative ranked ordering of response options (e.g., median, interquartile range, mode) [Ste46]. Within an ordinalist interpretation, Likert responses should be visualized in a way that emphasizes the discreteness of the data, such as a histogram or a stacked bar chart. Nonparametric statistical tests are appropriate for an ordinalist interpretation of Likert data because no assumptions are made about the normality or continuity of the data [Har15]. On the other hand, an intervalist interpretation of Likert responses allows for a wider range of descriptive statistics. In particular, when equal distances between each response option

are implied, the arithmetic mean and standard deviation are permissible for data summarization [Ste46]. The mean has been shown to be a better measure of central tendency than the median when summarizing Likert scale data [Lew93]. Visualization styles that treat Likert data as continuous are appropriate under an intervalist interpretation, such as bar charts of means or violin plots of densities. Parametric tests may be appropriate under the intervalist interpretation if no other violations of test assumptions are present [Har15].

The appropriate interpretation for Likert data depends largely on how the items are constructed and whether the Likert items within the scale are measured individually or accumulated into an average or sum. An accumulated Likert scale (i.e., the sum or average of responses to multiple Likert items) can generally be considered a continuous interval measurement [CP08, Har15, Nor10], although Jamieson provides an argument against interval interpretations of accumulated Likert scales [Jam04]. Interpretation of individual Likert items requires more nuance. Some researchers argue that individual Likert items should always be interpreted as ordinal, regardless of how the Likert item is designed [CP08, Nor10]. Conversely, Harpe argues that responses to individual Likert items with five or more response options can be interpreted as interval, while those with four or fewer response options should be restricted to ordinal interpretations only [Har15, HF69].

Intervalist and ordinalist interpretations are most often considered in the context of applying parametric or nonparametric analysis to Likert data, but they are also implicitly stated by the descriptive statistics and visualization styles researchers use to summarize Likert data. *Inconsistencies between ordinalist and intervalist interpretations can lead to results that are methodologically and theoretically unsound and could also threaten the statistical validity of results.* Nonparametric tests have lower statistical power than their parametric counterparts [Sie57], meaning a nonparametric procedure will more frequently fail to detect an existing effect. If the assumptions of a parametric statistical model are in fact appropriate, a researcher who starts out with an intervalist interpretation (e.g., reporting means and visualizing responses with a violin plot) may sacrifice statistical power if they switch to an ordinalist interpretation for no reason other than the fact that their data came from a Likert scale. At the same time, if a researcher is using nonparametric tests because they have correctly identified that their experimental design is not an appropriate match for an intervalist interpretation, they may risk misleading readers or misstating their statistical conclusions if they use a visualization that emphasizes intervalist statistics, such as a mean bar chart. To better understand how intervalist and ordinalist interpretations are used in visualization evaluations, we introduce our second research question:

RQ2: Are visualization researchers consistent in their usage of intervalist and ordinalist interpretations to summarize, visualize, and analyze Likert data in evaluations?

3.3. Handling Likert data

The conflation of the terms "subjective" and "qualitative" [CE18] when describing data collected with Likert questionnaires may
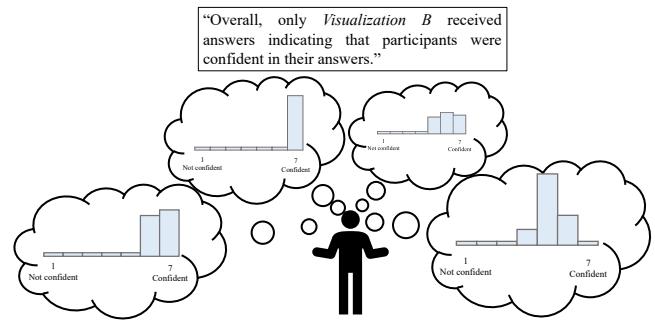

Figure 2: *Describing Likert results verbally without giving the reader quantitative estimates of centrality and uncertainty can be ambiguous and may mislead the reader.*

have contributed to inconsistent practices for handling Likert data in evaluations. However, this hypothesis has never been empirically validated and the causal effect of mistakenly viewing Likert data as qualitative in nature is difficult to determine. Some visualization papers explicitly refer to evaluations containing Likert data as "qualitative" in the body of the paper (e.g., [KOCC13, TRL*18]), while in other papers the assumption is implied when researchers exclusively use qualitative techniques, such as rich description, to analyze Likert data (e.g., [IBDF11]).

Describing participant responses qualitatively may be helpful in some instances, for example if the sample size is very small, but can also lead to statements that are ambiguous or misleading to the reader. For example, Figure 2 shows a quote from a visualization paper in our survey (Section 4) in which responses to a Likert item are described textually, without visualizations or quantitative summary statistics ("Overall, only Visualization B received answers indicating that participants were confident in their answers"). The quote is ambiguous because it could plausibly describe all four of the Likert response distributions shown in Figure 2, and likely many others. The reader has no way of knowing which of the distributions conjured up by this sentence is accurate, assuming the authors did not release their study data (which seems likely, given the findings of our survey in Section 5.3).

Comparing the treatment of quantitative objective (e.g., speed, accuracy) and quantitative subjective data (e.g., confidence, aesthetic value) in visualization evaluations can give us more information about the hypothesized subjective-qualitative confusion regarding Likert data. There are valid reasons why objective and subjective data could be handled differently within the same study, even if both data sources are quantitative. For example, Likert data might be more commonly used for exploratory research (i.e., "research that mainly seeks to explore patterns with no *a priori* articulated hypotheses" [NBL20]) rather than confirmatory research (i.e., "research that explicitly tests *a priori* formulated hypotheses" [NBL20]). In this situation, Likert data will not be included in the quantitative analysis run during confirmatory research.

If researchers *are* interested in using Likert data for confirmatory research, *a priori* formulated hypotheses about the subjective attributes of interest are evaluated using statistical testing procedures. NHST produces a *p*-value representing the probability of

obtaining results as extreme (or more extreme) than those actually observed, given the null hypothesis is true [Dra15]. NHST is commonly used in visualization and HCI communities despite strong criticism about how NHST encourages researchers to put undue emphasis on "statistical significance" and direct insufficient attention on the practical significance of their results. Estimation is an alternative to NHST that emphasizes reporting effect sizes and confidence intervals with nuanced interpretations intended to avoid the pitfalls of dichotomous inference that often accompany traditional NHST [Dra15]. If Likert data are being used in confirmatory research and researchers are making inferences based on Likert data, those inferences should be backed up by some form of quantitative analysis, whether that is estimation or NHST. While we do not advocate the use of NHST by default and acknowledge that its drawbacks are well-established [Dra15], the decision to not apply statistical analysis to Likert data during confirmatory research could be an indication that the standards for analyzing Likert data are not sufficiently rigorous, particularly if statistical analysis is conducted with other quantitative data in the same study.

Running a power analysis prior to conducting confirmatory hypothesis testing can help researchers determine the smallest sample size necessary to detect an effect of a given size at the desired level of significance [PGC18]. While power analyses can be beneficial when planning a study, researchers should be careful to view statistical power as a continuous measure influenced by multiple variables, rather than a binary indicator of "adequate" and "inadequate" sample sizes [Bac10]. Power analysis can be more complicated when working with Likert data than other data types. For example, researchers can use G*Power [FELB07] to conduct power analysis for parametric procedures, but such tools often do not support the nonparametric techniques that are commonly used with Likert data. Simulations can be helpful for conducting nonparametric power analysis (e.g., [Mum02]). Furthermore, researchers might not have reliable estimates of effect size when working with Likert data due to poor reporting and data openness practices in the visualization community, or they might be working with a small sample of experts. In these instances, researchers can conduct power analysis with a range of possible effect sizes or the minimum effect size they would be practically interested in. Sensitivity analysis allows researchers to determine the smallest effect they will be able to practically detect given their restricted sample size [PGC18]. While running *a priori* power analysis with Likert data might require more effort than with other types of quantitative data, it remains a feasible task and a more responsible way of conducting NHST.

Even if only objective measures are considered *critical parameters* [Hor13] and relevant for confirmatory research, we can compare other aspects of data handling to look for instances where Likert data are held to a lower standard than other data types. For example, do authors give estimates of effect size for both objective data and Likert data? Do they include measures of uncertainty (e.g., standard deviation, confidence intervals, interquartile range) for both data types? Are results summarized in a visualization or with descriptive statistics? Do the authors publicly release experimental datasets for both objective and Likert data? Guidelines for handling evaluation data agree that these practices are important for producing reliable and unambiguous quantitative results [CE18, Dra15], regardless of the subjective or objective nature of data sources. If

these practices are used when handling non-Likert quantitative data in a study but not when handling Likert data, this could be evidence that Likert data are being held to a lower standard of rigor. Our third research question addresses this issue:

RQ3: Do visualization researchers handle *subjective* Likert data differently than *objective* quantitative data?

---

4. Literature Survey

To answer the research questions introduced in Section 3 (RQ1–3), we conducted a literature survey of how Likert scales are constructed, reported, and analyzed in visualization evaluations.

4.1. Survey methodology

Our survey includes 134 papers from the IEEE VIS conference proceedings published in special issues of the IEEE Transactions of Visualization and Computer Graphics journal from 2009 to 2019. This includes all papers from the three primary tracks of IEEE VIS: Information Visualization (InfoVis), Scientific Visualization (SciVis), and Visual Analytics Science and Technology (VAST). With an impact factor of 3.78, IEEE TVCG is one of the largest and most diverse premier publication venues in the field of visualization [cla18]. We did not include proceedings from ACM CHI, another popular venue for visualization publications, as there was not a dedicated track for visualization papers until 2019, making it difficult to objectively categorize papers published at CHI as "visualization" or not. We included all papers with either "Likert" or "subjective questionnaire" in the abstract or main body of the text, producing a subset of 134 papers out of the total 1,119 papers published in the conference proceedings from 2009 to 2019. By including the "subjective questionnaire" keyword, we intended to account for papers that use Likert-style responses (e.g., a response indicating levels of agreement corresponding to an ordinal scale) without explicitly mentioning the conventional name. However, the two papers in our dataset that included the "subjective questionnaire" keyword also included the phrase "Likert scale", indicating that although we successfully identified a large sample of papers with Likert data, we were not successful in identifying unlabelled Likert scales. We describe this limitation in greater detail in Section 6.1. Our review was not registered and we do not include a review protocol. All papers were accessed using the IEEEXplore database in March 2020 and manually screened to ensure they met our inclusion criteria by two independent coauthors.

4.2. Survey construction

We collected the following information for each paper to answer the research questions defined in Section 3. Papers were manually coded by two authors independently.

RQ1: Likert scale construction reporting

To understand how visualization researchers describe their Likert scales in evaluations, we recorded the following information for each paper in our survey:

- *Number of response options*: How many response options were participants presented with (e.g., 5, 7, 10)?
- *Numerical representation*: What did the numbers attached to each response option look like (e.g., ascending, diverging)?
- *Text anchors*: What text was used to describe each response option (e.g., "Strongly disagree", "Neutral", "Very useful")?
- *Question phrasing*: How was the question phrased for each Likert item (e.g., "The system was easy to use")?

RQ2: Ordinalist and intervalist interpretations

We recorded the following information for each paper in order to identify inconsistencies between ordinalist and intervalist interpretations of Likert scales in visualization evaluations:

- *Summary statistics*: What type of statistics were used to summarize Likert responses (nominal, ordinal, or interval)? In cases where several summary statistics associated with different scales were reported (e.g., median and standard deviation or mean and IQR), we recorded the type associated with the measure of centrality. We included all summary statistics presented in the body of the paper, including tables and figures.
- *Visualization style*: If a visualization of Likert responses was provided, did the visualization portray Likert responses as discrete (e.g., histogram) or continuous (e.g., mean bar chart)?
- *Quantitative analysis*: If statistical analysis was conducted using Likert data (NHST or estimation), did the researchers use parametric or nonparametric methods?

RQ3: Comparing treatment of subjective Likert data and objective data

To compare the treatment of subjective quantitative data collected with Likert scales and quantitative objective data (e.g., time or accuracy) in visualization evaluations, we recorded the following information for each paper:

- *Objective and subjective measures*: Were quantitative objective data collected in addition to subjective Likert data?
- *Quantitative analysis*: What type of quantitative analysis (NHST, estimation) was done on the Likert data? If applicable, what kind of quantitative analysis was done on objective measures?
- *Visualization of results*: Did the researchers include a visualization of the subjective Likert data collected during the study? If the study incorporated objective data in addition to subjective Likert data, did the researchers include a visualization of objective data collected during the study?
- *Uncertainty measures*: Were any measures of uncertainty (e.g., standard deviation, interquartile range, confidence intervals) provided for subjective Likert data? If applicable, were measures of uncertainty provided for quantitative objective measures?
- *Data availability*: Were Likert data made publicly available upon publication of the paper? If applicable, were quantitative objective data also made publicly available? We considered datasets to be publicly available if they are included in the paper's Supplementary Material on IEEEXplore or in an open repository that is linked to in the text of the paper.
- *Power analysis*: If quantitative analysis (NHST or estimation) was run with both objective data and subjective Likert data, were both forms of measurement explicitly included in a power or sensitivity analysis?

5. Results

In this section, we summarize the findings of our systematic review and answer the research questions (RQ1-3) defined in Section 3. As described in Section 4, 134 papers were selected via keyword search ("Likert scale" or "subjective questionnaire") out of 1,119 papers published in IEEE VIS conference proceedings from 2009 to 2019. A full copy of all metadata collected for each paper in the survey is available in our Supplementary Materials and at https://osf.io/exbz8/ or at https://airtable.com/shrrCOcWlzS7UQvSG.

Construction element	Replicable?	# papers	% papers
Number of choices			
5 or 7 (i.e., including neutral option)	✓	116	87%
4 or 6 (i.e., no neutral option)	✓	9	7%
Other	✓	5	4%
Not stated	✗	**4**	**3%**
Numerical representation			
Ascending (starting at 1)	✓	88	64%
Symbolic (e.g., –, -, 0, +, ++)	✓	7	5%
Ascending (starting at 0)	✓	6	4%
Diverging (e.g., -2 to 2)	✓	3	2%
Ascending and symbolic	✓	1	0.7%
Not stated	✗	**29**	**21%**
Text anchors			
All anchors specified	✓	14	10%
Only end anchors specified	✗	**78**	**57%**
One or more anchors specified	✗	**3**	**2%**
No anchors specified	✗	**39**	**28%**
Question phrasing			
Full phrasing specified for all items	✓	45	33%
Item topics specified, not full phrasing	✗	**76**	**55%**
Full phrasing specified for some items	✗	**3**	**2%**
No phrasing specified	✗	**10**	**7%**

Table 1: *We recorded which elements of Likert item construction methodology were reported in the text or Supplementary Materials (if the authors indicated additional details could be found there) of 134 visualization papers. Categories with insufficient detail for reliable replication are shown in bold.*

5.1. Reporting Likert construction details (RQ1)

Detailed descriptions of how Likert items were constructed are critical for enabling reliable future replication studies and helping readers contextualize results. To determine how well the visualization community adheres to reporting standards for Likert questionnaires (RQ1, Section 3.1), we recorded how each paper in our survey documented the implementation details of their Likert scale questions. We looked for these details in the body of the paper and in Supplementary Materials (if the authors indicated additional details could be found there). The distribution of reporting behaviors encountered in our survey is shown in Table 1.

The number of response options provided to study participants was the most common detail of Likert scale construction provided by papers in the survey, appearing in 130 papers (97%). We observed little variation in the number of response options included in Likert scales: 116 papers (87%) provided 5 or 7 options (e.g., [HTL13]). The numerical representation used within the Likert scale was reported in 105 papers (78%). Ascending scales starting at 1 were most popular (88 papers, 64%; e.g., [DBD16]), while only a handful of papers used symbolic (7 papers, 5%; e.g., [GLH*14])

or diverging (3 papers, 2%; e.g., [RHY14]) scales. One paper included Likert scales with two different numeric representations (ascending and symbolic) [MOJB*18]. Numerical representation was not stated in 29 papers (21%). This means that in a hypothetical replication study of one in five papers included in our survey, it would be impossible to know whether the participant responses were biased by differences in Likert scale construction (e.g., replicators might use an ascending scale instead of the symbolic scale used by the original experimenters). Ten papers did not specify their numerical representation but summarized Likert scale responses using mean or median as a summary statistic (e.g., [SFP*18]). Without knowing the possible range of values, it is impossible to effectively interpret a mean or median. For example, an average response of 3.8 might be more convincing if the response range is 0 to 4 than if the range is 1 to 7.

We also recorded how much information each paper provided about the text anchors used within their Likert scales. Ideally, all text anchors provided to participants should be fully specified in the description of Likert scale construction, limiting the room for potential bias in future replications as much as possible. Only fifteen papers in our survey (10%) fully specified all text anchors (e.g., [STM16]). The majority of papers provided only the endpoints of their scales (78 papers, 57%; e.g., [BIAI16]). While this is likely sufficient detail for text anchors with a standard structure (e.g., "Strongly disagree" to "Strongly agree"), this practice could be problematic for less common text anchors. Forty papers (30%) reported only endpoint text anchors for scales that deviated from the standard structure, such as "The worst" to "The best" and "Not at all" to "Extremely" (e.g., [DBD16, SFMB12]). It is unlikely that a replicator would select precisely the same phrasing for intermediate text anchors, indicating a threat to validity for future replication studies. While it is possible that some of the 40 papers genuinely did not present intermediate text anchors, none explicitly stated this design choice. 39 papers (28%) provided no information at all about the text anchors used in their Likert scale questionnaires (e.g., [BS15]). As described in Section 3.1, text anchor phrasing affects participant response to Likert scales [FA99]. This means in a hypothetical replication study of a quarter of the papers included in our survey, replicators would have no way of knowing if their experiment is detecting the same underlying effect found in the original paper or if participant response has been biased by the replicators' choice of text anchors.

Finally, we examined how each paper described the question phrasing attached to individual Likert items. In the same way that text anchor phrasing can affect perception of Likert scales, the phrasing of Likert scale questions can also affect participant response [FA99]. For optimal transparency and replicability, full phrasings should be provided for all Likert items given to participants. Approximately one third of the papers in our survey met this standard for replicability by including full phrasings for all Likert items (45 papers, 33%; e.g., [GS14]). The majority of papers (76, 55%; e.g. [LTPH16]) gave a short summary of the Likert item topic but refrained from providing the full phrasing (i.e., "system usability" in place of "I thought the system was easy to use"). Shortened versions of question phrasing might be sufficient in some cases, such as "Easy to use" instead of "The system was easy to use", but this practice can leave room for error if the shortened

version is generic and the full question phrasing cannot be intuited from the shorthand version (e.g., "How important" [RAL*16]). Three papers gave full question phrasing for some Likert items but not all (e.g., [CBY10]), while ten papers failed to provide statement phrasings or even shortened item topics for all Likert items (e.g., [WS09]). Without knowing what questions participants were asked to respond to, it is difficult for readers and other researchers to fully contextualize and build on the results presented in each paper. Five papers gave no information about the question phrasing or text anchor phrasing used in their Likert scale questionnaires, all but guaranteeing the unreliability of future replication studies based on their results (e.g., [WLMB*14]).

Summary: While 97% of the papers in our survey report the number of response options included in their Likert scale questionnaires, we find that other essential details such as numerical representation, text anchor phrasing, and question phrasing, are often omitted or described in insufficient detail for contextualizing results and enabling reliable replication studies.
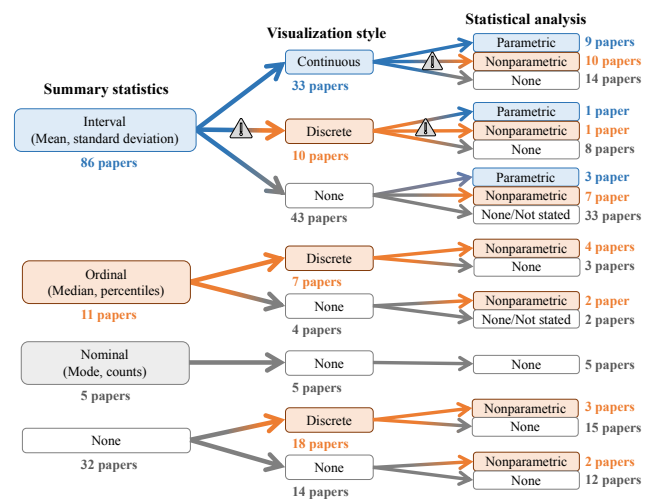

Figure 3: *We recorded each paper's decision to use intervalist or ordinalist interpretations for Likert scales at three points in reporting and analysis phases. Papers that used ordinal summary statistics maintained a consistent ordinalist interpretation, while several papers that began with interval summary statistics shifted to ordinalist interpretations later on in their analysis (indicated by ⚠ icon).*

5.2. Intervalist and ordinalist interpretations (RQ2)

To understand how visualization evaluations use intervalist and ordinalist interpretations of Likert scales (Section 3.2), we recorded which interpretation was used at three steps in the research process: reporting summary statistics, visualizing Likert responses, and applying statistical tests. Figure 3 shows the distribution of intervalist and ordinalist interpretations across these three decision points.

Interval summary statistics (e.g., mean, standard deviation) were most commonly used to describe Likert responses in our survey (86 papers, 63%; e.g., [LBW18]). Papers that reported interval summary statistics chose to include both continuous (33 papers, 38%;

Visualization type	# papers	% papers
None	**65**	**49%**
Discrete	**35**	**26%**
Histogram (e.g., [LBW18])	10	7%
Percentage stacked bar chart (e.g., [STM16])	8	6%
Boxplot of medians (e.g., [GLH*14])	6	4%
Diverging stacked bar chart (e.g., [DBD16])	5	3%
Heatmap with discrete color map (e.g., [SLK*16])	4	3%
Bar chart of medians (e.g., [CLB*15])	2	2%
Continuous	**34**	**25%**
Bar chart of means (e.g., [AL19])	18	13%
Scatterplot of means (e.g., [GS14])	7	5%
Boxplot of means (e.g., [WG12])	3	2%
Violin plot (e.g., [LBB*19])	2	1%
Line chart of means (e.g., [LPCRH18])	2	1%
Heatmap of means (e.g., [ZLC*18])	2	1%

Table 2: *Papers in our survey were equally likely to choose discrete or continuous visualization styles.*

e.g., [WCA*16]) and discrete (10 papers, 12%; e.g., [DBD16]) visualization styles. Table 2 summarizes the specific visualization styles used to display Likert scale responses in our survey. We observed inconsistencies between intervalist and ordinalist interpretations in researchers' selection of parametric and nonparametric statistical tests among papers that used interval-level summary statistics. Of the 31 papers (36%) that ran a statistical test on Likert data after reporting an interval summary statistic, 18 chose to use a nonparametric test inconsistent with an intervalist interpretation. There are valid reasons why a researcher might choose nonparametric tests despite reporting interval summary statistics, such as if the data fail to pass the Shapiro-Wilk test for normality [SW65]. Only 4 out of 18 papers using interval summary statistics and nonparametric tests provided a justification for their decision (e.g., [WCA*16]). It is unclear how the remaining 14 papers decided to use nonparametric analysis because no justification is provided. Depending on the context, it is possible that these papers might have had greater statistical power if they had used parametric tests consistent with their intervalist interpretation.

Only 11 papers (8%; e.g., [WMZ*19]) used an ordinalist interpretation when reporting summary statistics (e.g., median, quartiles). While we observed deviations from the intervalist interpretation within papers that used interval summary statistics, all eleven papers that used ordinal statistics and chose to visualize their data used a discrete visualization style consistent with an ordinalist interpretation. Additionally, all of the five papers that reported ordinal statistics and ran statistical tests chose to use nonparametric tests consistent with an ordinalist interpretation. Of the five papers (4%; e.g., [EEL*19]) that reported nominal summary statistics (e.g., mode, counts), none chose to visualize Likert responses or run statistical analysis.

Visualizations of Likert data appeared in approximately half of the papers in our survey (69 papers, 51%). Among papers that chose to include a visualization, discrete (35 papers, 26%) and continuous (34 papers, 25%) styles were equally popular. Table 2 summarizes the visualization types used for Likert responses in our survey. Histograms and stacked bar charts of responses to individual Lik-

ert items were the most popular discrete visualization types, while most continuous visualizations displayed the mean response to individual Likert items.

Summary: We observed that several papers in our survey reported interval summary statistics (e.g., mean, standard deviation) but deviated into ordinalist interpretations later in their analysis by using discrete visualization styles and nonparametric tests, rather than continuous visualizations and parametric tests. While this may be appropriate in some cases, we identified 13 papers where nonparametric tests were not explicitly justified and where parametric tests could have been used to obtain greater statistical power. All papers in our study that reported ordinal summary statistics used discrete visualization styles and nonparametric tests consistent with an ordinalist interpretation of their Likert scales.

	Consistent?	# papers	% papers
Summary statistics			
Not included	⊘	1	1%
Both objective and subjective measures	✓	65	71%
Objective measures only	✕	**25**	**27%**
Visualization of results			
Neither objective or subjective data visualized	⊘	20	22%
Objective and subjective data both visualized	✓	46	51%
Objective measures only	✕	**25**	**27%**
Quantitative analysis			
Not included	⊘	19	21%
Both objective and subjective measures	✓	39	43%
Objective measures only	✕	**33**	**36%**
Uncertainty measures			
Not included	⊘	27	30%
Both objective and subjective measures	✓	38	42%
Objective measures only	✕	**26**	**29%**
Power analysis			
Not applicable	⊘	52	57%
Not included	✕	38	42%
Both objective and subjective measures	✓	0	0%
Objective measures only	✕	**1**	**1%**
Data availability			
No data released	✕	82	90%
Both objective and subjective data released	✓	7	8%
Only subjective data released	✕	**1**	**1%**
Only objective data released	✕	**1**	**1%**

Table 3: *For the 91 papers in our survey that collected both objective and subjective quantitative measures, we looked for discrepancies in data handling between the two evaluation components. Instances where quantitative best practices were applied to objective measures and not to subjective measures are highlighted in bold.*

5.3. Comparing treatment of subjective Likert data and objective data (RQ3)

Likert data are quantitative but they are not always held to the same standards for reporting and analysis as other quantitative forms of data. To answer RQ3 (Section 3.3), we recorded whether each paper in our survey collected both objective and subjective quantitative measures or only subjective measures. Among papers that collected both types of data, we compared the treatment of subjective and objective data sources. Within our survey, 43 papers (32%) collected only subjective quantitative measures while 91 papers (68%) collected both subjective and objective quantitative measures. Table 3 summarizes the discrepancies we observed between the handling of subjective and objective quantitative data in 91 papers that collected both types of data.

We recorded whether a visualization was included for objective

and subjective data and found that more than a quarter of papers provided a visualization only for objective measures in their study (25 papers, 27%; e.g., [YEII15]). Not visualizing Likert data is not necessarily bad; many papers are restricted by page limits and use summary statistics to report results instead. Nevertheless, the gap we observed could indicate that Likert responses are viewed as less essential than objective counterparts in visualization evaluations.

We also recorded whether researchers chose to run quantitative statistical analyses (i.e., NHST or estimation) on subjective and objective data. The limits of NHST are well-known [Dra15, KR12] and we do not suggest that researchers should automatically run statistical tests on Likert data without considering potential drawbacks. However, it is worth examining more closely why some papers choose to run NHST, despite the risks, on objective data but not on Likert responses. 39 papers (42%) ran quantitative analysis on both objective and subjective measures (e.g., [WCA*16]). Over a third of papers with both types of data ran quantitative analysis on objective data but not Likert data (33 papers, 36%; e.g., [BCC*19, LBB*19]). Choosing not to apply NHST or estimation to Likert data when it is deemed appropriate for other quantitative measures used for confirmatory research could indicate that Likert data are not held to the same standards as objective data sources. Not applying quantitative analysis could lead to overstated or misleading results if, for example, Likert item means are presented as confirmatory evidence for a system's superiority without confidence intervals, p-values, or measures of uncertainty to provide necessary context.

Measures of uncertainty provide valuable context about the practical significance of experimental effects identified in a study [Dra15]. We recorded what measures of uncertainty were included for objective and subjective data in all 91 papers that collected both types of data. 27 papers (30%) included a measure of uncertainty, such as standard deviation or interquartile range (IQR), for objective data sources but not for Likert responses (e.g., [BKH*11]), while 38 papers (42%) included an uncertainty measure for both types of data (e.g., [BRH*16]). Many of the papers that did not provide a measure of uncertainty used descriptive statistics such as mean or median alone to summarize the data. Although it is possible to have a valid quantitative analysis without including uncertainty measures, they are helpful for contextualizing and interpreting the significance of results and should be considered for Likert data if they are deemed appropriate for objective measures in a study [Dra15].

Power calculations allow a researcher to estimate the probability of correctly rejecting the null hypothesis when it is false with a given sample size. Although there are risks associated with overemphasizing the cutoff point between sample sizes that are hypothetically "adequate" and "inadequate" [Bac10], *a priori* power analyses are generally recommended when using NHST [CE18, PGC18]. Of the 45 papers (34%) in which a power analysis would have been appropriate (i.e., confirmatory research was conducted), only two papers mentioned running a power analysis. One paper included only objective measures in their power analysis [KOCC13], while the other included only Likert data [LPCRH18].

Finally, we recorded whether objective and subjective datasets were made publicly available upon publication (i.e., included in Supplementary Materials on IEEEXplore or uploaded to a public repository that is linked to in the body of the paper). Less than ten percent of the papers in our survey publicly released study datasets in Supplementary Materials (12 papers, 9%; e.g., [LBW18]). We did not observe a disparity in data availability between Likert and non-Likert data: seven papers released both objective and Likert data (e.g., [YDJ*18]), while one paper released only objective data [BBB*18] and another released only Likert data [LBB*19].

Summary: We found that studies with both objective and subjective data are often apply different standards when handling the two types of quantitative data. Papers in our survey were less likely to use visualizations, summary statistics, quantitative analysis, and uncertainty measures to report and analyze subjective data from Likert questionnaires than from objective data sources. On the other hand, we observed that power analyses and publicly available datasets are equally rare for both objective and subjective studies.

6. Discussion

6.1. Limitations & future work

Although our survey included all relevant papers published in the IEEE VIS conference proceedings, it did not include visualization papers published in ACM CHI, EuroVis, or other publication venues and may not be representative of all visualization evaluations. Additionally, our keyword search strategy may have missed some papers that used Likert-style questionnaires without explicitly using the terms "Likert" or "subjective questionnaire". While we believe our results are valuable for the visualization community, analysis of a wider variety of venues with different inclusion criteria could reveal additional trends in the handling of Likert data in our field. Finally, there may be rater errors in our survey because all papers were coded by a single author. Cataloging how other forms of measurement, such as adjectival scales [SNC15] or slider scales [RLA15] are used in visualization research is a promising area of future research. Empirical studies assessing the effectiveness of the various techniques for visualizing Likert data observed in our survey (Table 2) could provide better guidance than currently available heuristics (e.g., [RH*11]).

6.2. Recommendations

As demonstrated in Section 5, there are persistent issues present in how Likert data are collected, reported, and analyzed in visualization evaluations. In this section, we provide concrete recommendations for researchers and reviewers to improve the validity of Likert-based evaluations.

1. Is a Likert scale right for you? Before conducting your study, think about whether a Likert scale is appropriate for your goals. Is your research question sufficiently specific that it can be reduced to quantitative estimates without losing valuable nuance and context? If not, consider incorporating a qualitative component to your study with a semi-structured interview or free-response question. Several papers in our survey analyzed Likert responses qualitatively (Section 3.3), suggesting that qualitative methods might have been a better fit for their research (e.g., [IBDF11, PYHZ14]).

2. Use best practices for objective data as a model for handling

quantitative subjective data. When developing a research plan that includes a Likert questionnaire, decide whether your analysis will be exploratory or confirmatory (Section 3.3). If you are running a confirmatory analysis with Likert data, consider whether your procedure meets accepted best practices (e.g., [Dra15, KR12]) for other forms of quantitative data, such as accuracy or speed.

3. Provide detailed descriptions of your Likert methodology. Construct individual Likert items in accordance with established guidelines and describe your methodology with as much detail as possible (see Section 3.1). An ideal Likert construction statement would include the number of response options, all text anchors, all question statements, and the numerical representation used for each response option (e.g., [ALBR15, LBW18]).

4. Match your axes to your scales. When visualizing Likert data, ensure that the numerical representation and text anchors built into your Likert items are accurately represented on your visualization axes (e.g., [VZS17]). Readers may not remember how many options were presented to participants or the text anchors attached to each number, so visualizations with inconsistent or imprecise axes can be misleading (e.g., [GBFM15]).

5. Analyze all Likert items given to participants. Once you have settled on an analysis plan, be sure to run it on all Likert data you have collected and report results for each question (e.g., [AL19]). Providing results for only a handful of the questions weakens overall credibility and can be misleading.

6. Do not automatically rule out parametric analysis with Likert data. While nonparametric tests are most commonly used with Likert data [Sie57], parametric analysis may allow researchers to gain statistical power without violating test assumptions in certain situations, such as when multiple Likert items are accumulated or if participants have five or more response options to choose from. Researchers can also use parametric models that are built for handling ordinal data, such as proportional odds or continuation ratio ordinal logistic models [Agr03, H*15].

7. Be aware of the pitfalls of quantitative analysis as they apply to Likert data. Use statistical best practices (e.g., [Dra15]) when analyzing Likert data, such as reporting effect sizes, estimates of uncertainty, and, if applicable, *p*-values (e.g., [KCWK19]).

6.3. Improving replication studies

While some of the issues identified in Section 5 are stylistic choices that could lead to misinterpretations (e.g., not including visualizations of Likert data), others have more serious consequences for visualization research as a whole, such as the lack of methodological details described in Section 5.1. Replication studies are rare in the visualization community [KH18], making it difficult to know the validity and soundness of our field's conclusions. A natural solution to this problem is to conduct and publish more replications, but this only addresses half of the issue: We need to not only publish more replication studies but also ensure that the replication studies we *do* publish are valid and reliable. We demonstrated in Section 5.1 that while many papers provide basic information about their Likert scales, such as the number of response options, far fewer provide additional details required for a reliable replication study,

such as numerical representation or text anchors. No matter how many times a study with insufficient details about Likert item construction is replicated, it will be impossible to know whether participants have been biased by differences between the Likert questionnaires used by the original researchers and the replicators. This means that the burden of enabling replication must fall not only on future researchers who may one day produce replications but also on current researchers producing original studies today.

6.4. Subjective-qualitative confusion

Throughout our survey of visualization papers with Likert scales to measure subjective qualities, we noticed a recurring confusion between the terms "subjective" and "qualitative". Some papers explicitly described their Likert questionnaires as "qualitative" studies (e.g., [BKH*11]), while others implicitly applied qualitative analysis methods to Likert responses (e.g., [IFM*10]). This confusion could explain the discrepancies in data handling identified in our literature survey (Section 5.3). Summary statistics cannot be meaningfully applied to interview transcripts or other qualitative records, so it follows that a researcher who has mistakenly assumed their Likert responses are qualitative might not include summary statistics for Likert items. In a semi-structured interview or free-response question, research participants can explain their thoughts and provide reasoning for their subjective response. Collecting subjective data in a qualitative manner allows participants and researchers to share context and nuance that is difficult to convey in a quantitative response. We recommend that researchers tread carefully when attempting to provide context through qualitative analysis when reporting Likert responses alone. To avoid confusing the terms "subjective" and "qualitative", we encourage researchers to think of research methods (quantitative, qualitative, or mixed methods) separately from data sources (objective or subjective) when planning evaluations.

7. Conclusion

We have presented a literature survey of 134 visualization evaluations that used Likert scales to measure subjective responses. Our results demonstrate that many evaluations do not describe implementation details of Likert scales in sufficient detail for reliable future replication studies and are inconsistent in applying intervalist and ordinalist interpretations when presenting and analyzing Likert data. We also observe that subjective Likert data are not held to the same standards as objective data in confirmatory studies involving both data types. As a service to the visualization community, we summarize guidelines for best practices with Likert data from other fields. We hope our survey and summarized guidelines will help researchers and reviewers in the visualization community more effectively use Likert scales in future subjective evaluations.

8. Acknowledgments

References

[Agr03]   AGRESTI A.: Ordinal responses: Cumulative logit models. In *Categorical data analysis*. John Wiley & Sons, 2003, ch. 8.2. 10

[AL19]   AHN Y., LIN Y.-R.: Fairsight: Visual analytics for fairness in decision making. *IEEE Transactions on Visualization and Computer Graphics 26*, 1 (2019), 1086–1095. 8, 10

[ALBR15]   ALBO Y., LANIR J., BAK P., RAFAELI S.: Off the radar: Comparative evaluation of radial visualization solutions for composite indicators. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (2015), 569–578. 10

[B*96]   BROOKE J., ET AL.: Sus-A quick and dirty usability scale. *Usability Evaluation in Industry 189*, 194 (1996), 4–7. 3

[Bac10]   BACCHETTI P.: Current sample size conventions: flaws, harms, and alternatives. *BMC medicine 8*, 1 (2010), 1–7. 5, 9

[BBB*18]   BLASCHECK T., BESANÇON L., BEZERIANOS A., LEE B., ISENBERG P.: Glanceable visualization: Studies of data comparison performance on smartwatches. *IEEE Transactions on Visualization and Computer Graphics 25*, 1 (2018), 630–640. 9

[BCC*19]   BATCH A., CUNNINGHAM A., CORDEIL M., ELMQVIST N., DWYER T., THOMAS B. H., MARRIOTT K.: There is no spoon: Evaluating performance, space use, and presence with expert domain users in immersive analytics. *IEEE Transactions on Visualization and Computer Graphics 26*, 1 (2019), 536–546. 9

[BD19]   BESANÇON L., DRAGICEVIC P.: The continued prevalence of dichotomous inferences at CHI. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (2019), pp. 1–11. 2

[BIAI16]   BESANÇON L., ISSARTEL P., AMMI M., ISENBERG T.: Hybrid tactile/tangible interaction for 3d data exploration. *IEEE Transactions on Visualization and Computer Graphics 23*, 1 (2016), 881–890. 7

[BKH*11]   BURCH M., KONEVTSOVA N., HEINRICH J., HÖFERLIN M., WEISKOPF D.: Evaluation of traditional, orthogonal, and radial tree diagrams by an eye tracking study. *IEEE Transactions on Visualization and Computer Graphics 17*, 12 (2011), 2440–2448. 9, 10

[BRH*16]   BACH B., RICHE N. H., HURTER C., MARRIOTT K., DWYER T.: Towards unambiguous edge bundling: Investigating confluent drawings for network visualization. *IEEE Transactions on Visualization and Computer Graphics 23*, 1 (2016), 541–550. 9

[BS15]   BUTKIEWICZ T., STEVENS A. H.: Effectiveness of structured textures on dynamically changing terrain-like surfaces. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (2015), 926–934. 7

[Cai16]   CAINE K.: Local standards for sample size at CHI. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (2016), pp. 981–992. 2

[CBY10]   CHEN Y., BARLOWE S., YANG J.: Click2annotate: Automated insight externalization with rich semantics. In *2010 IEEE Symposium on Visual Analytics Science and Technology* (2010), IEEE, pp. 155–162. 7

[CD94]   CLASON D. L., DORMODY T. J.: Analyzing data measured by individual Likert-type items. *Journal of agricultural education 35*, 4 (1994), 4. 1

[CDBG20]   COCKBURN A., DRAGICEVIC P., BESANÇON L., GUTWIN C.: Threats of a replication crisis in empirical computer science. *Communications of the ACM 63*, 8 (2020), 70–79. 2

[CE18]   CRISAN A., ELLIOTT M.: How to evaluate an evaluation study? Comparing and contrasting practices in Vis with those of other disciplines: Position paper. In *2018 IEEE Evaluation and Beyond-Methodological Approaches for Visualization (BELIV)* (2018), IEEE, pp. 28–36. 4, 5, 9

[CGD18]   COCKBURN A., GUTWIN C., DIX A.: Hark no more: On the preregistration of chi experiments. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (2018), pp. 1–12. 2

[cla18]   Journal Citation Report. *Clarivate Analytics* (2018). 5

[CLB*15]   COHÉ A., LIUTKUS B., BAILLY G., EAGAN J., LECOLINET E.: Schemelens: A content-aware vector-based fisheye technique for navigating large systems diagrams. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (2015), 330–338. 8

[Cor20]   CORRELL M.: What do we actually learn from evaluations in the "Heroic Era" of visualization?: Position paper. In *2020 IEEE Workshop on Evaluation and Beyond-Methodological Approaches to Visualization (BELIV)* (2020), IEEE, pp. 48–54. 2

[CP08]   CARIFIO J., PERLA R.: Resolving the 50-year debate around using and misusing Likert scales. *Medical education 42*, 12 (2008), 1150–1152. 2, 4

[CRSH17]   CHYUNG S. Y., ROBERTS K., SWANSON I., HANKINSON A.: Evidence-based survey design: The use of a midpoint on the likert scale. *Performance Improvement 56*, 10 (2017), 15–23. 3

[DBD16]   DIMARA E., BEZERIANOS A., DRAGICEVIC P.: The attraction effect in information visualization. *IEEE Transactions on Visualization and Computer Graphics 23*, 1 (2016), 471–480. 6, 7, 8

[Dra15]   DRAGICEVIC P.: *HCI Statistics without p-values*. PhD thesis, 2015. 5, 9, 10

[EEL*19]   EULZER P., ENGELHARDT S., LICHTENBERG N., DE SIMONE R., LAWONN K.: Temporal views of flattened mitral valve geometries. *IEEE Transactions on Visualization and Computer Graphics 26*, 1 (2019), 971–980. 8

[FA99]   FRIEDMAN H. H., AMOO T.: Rating the rating scales. *Journal of Marketing Management, Winter* (1999), 114–123. 3, 7

[FELB07]   FAUL F., ERDFELDER E., LANG A.-G., BUCHNER A.: G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods 39*, 2 (2007), 175–191. 5

[GBFM15]   GSCHWANDTNEI T., BÖGL M., FEDERICO P., MIKSCH S.: Visual encodings of temporal uncertainty: A comparative user study. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (2015), 539–548. 10

[GLH*14]   GLASSER S., LAWONN K., HOFFMANN T., SKALEJ M., PREIM B.: Combined visualization of wall thickness and wall shear stress for the evaluation of aneurysms. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (2014), 2506–2515. 6, 8

[GS14]   GOTZ D., STAVROPOULOS H.: Decisionflow: Visual analytics for high-dimensional temporal event sequence data. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (2014), 1783–1792. 7, 8

[H*15]   HARRELL F. E., ET AL.: Ordinal logistic regression. In *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*, vol. 3. Springer, 2015, ch. 13. 10

[Har15]   HARPE S. E.: How to analyze Likert and other rating scale data. *Currents in Pharmacy Teaching and Learning 7*, 6 (2015), 836–850. 1, 2, 3, 4

[HB10]   HARTLEY J., BETTS L. R.: Four layouts and a finding: the effects of changes in the order of the verbal labels and numerical values on Likert-type scales. *International Journal of Social Research Methodology 13*, 1 (2010), 17–27. 3

[HF69]   HSU T.-C., FELDT L. S.: The effect of limitations on the number of criterion score values on the significance level of the f-test. *American Educational Research Journal 6*, 4 (1969), 515–527. 4

[Hor13]   HORNBÆK K.: Some whys and hows of experiments in human–computer interaction. *Foundations and Trends in Human-Computer Interaction 5*, 4 (2013), 299–373. 5

[HQC*18]   HULLMAN J., QIAO X., CORRELL M., KALE A., KAY M.: In pursuit of error: A survey of uncertainty visualization evaluation. *IEEE Transactions on Visualization and Computer Graphics 25*, 1 (2018), 903–913. 2

[HS88] HART S. G., STAVELAND L. E.: Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in Psychology*, vol. 52. Elsevier, 1988, pp. 139–183. 3

[HTL13] HAJIZADEH A. H., TORY M., LEUNG R.: Supporting awareness through collaborative brushing and linking of tabular data. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (2013), 2189–2197. 6

[IBDF11] ISENBERG P., BEZERIANOS A., DRAGICEVIC P., FEKETE J.-D.: A study on dual-scale data charts. *IEEE Transactions on Visualization and Computer Graphics 17*, 12 (2011), 2469–2478. 4, 9

[IFM*10] ISENBERG P., FISHER D., MORRIS M. R., INKPEN K., CZERWINSKI M.: An exploratory study of co-located collaborative visual analytics around a tabletop display. In *2010 IEEE Symposium on Visual Analytics Science and Technology* (2010), IEEE, pp. 179–186. 10

[IIC*13] ISENBERG T., ISENBERG P., CHEN J., SEDLMAIR M., MÖLLER T.: A systematic review on the practice of evaluating visualization. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (2013), 2818–2827. 1, 2

[Jam04] JAMIESON S.: Likert scales: How to (ab) use them? *Medical education 38*, 12 (2004), 1217–1218. 4

[KCWK19] KREKHOV A., CMENTOWSKI S., WASCHK A., KRÜGER J.: Deadeye visualization revisited: Investigation of preattentiveness and applicability in virtual environments. *IEEE Transactions on Visualization and Computer Graphics 26*, 1 (2019), 547–557. 10

[KH18] KOSARA R., HAROZ S.: Skipping the replication crisis in visualization: Threats to study validity and how to address them: Position paper. In *2018 IEEE Evaluation and Beyond-Methodological Approaches for Visualization (BELIV)* (2018), IEEE, pp. 102–107. 3, 10

[KNM10] KAPTEIN M. C., NASS C., MARKOPOULOS P.: Powerful and consistent analysis of Likert-type rating scales. In *Proceedings of the SIGCHI conference on human factors in computing systems* (2010), pp. 2391–2394. 2

[KOCC13] KERSTEN-OERTEL M., CHEN S. J.-S., COLLINS D. L.: An evaluation of depth enhancing perceptual cues for vascular volume visualization in neurosurgery. *IEEE Transactions on Visualization and Computer Graphics 20*, 3 (2013), 391–403. 4, 9

[Kos16] KOSARA R.: 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* (2016), pp. 162–168. 1, 2, 3

[KR12] KAPTEIN M., ROBERTSON J.: Rethinking statistical analysis methods for CHI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2012), pp. 1105–1114. 2, 9, 10

[LBB*19] LEKSCHAS F., BEHRISCH M., BACH B., KERPEDJIEV P., GEHLENBORG N., PFISTER H.: Pattern-driven navigation in 2d multiscale visualizations with scalable insets. *IEEE Transactions on Visualization and Computer Graphics 26*, 1 (2019), 611–621. 8, 9

[LBI*11] LAM H., BERTINI E., ISENBERG P., PLAISANT C., CARPENDALE S.: Empirical studies in information visualization: Seven scenarios. *IEEE Transactions on Visualization and Computer Graphics 18*, 9 (2011), 1520–1536. 2

[LBW18] LAW P.-M., BASOLE R. C., WU Y.: Duet: Helping data analysis novices conduct pairwise comparisons by minimal specification. *IEEE Transactions on Visualization and Computer Graphics 25*, 1 (2018), 427–437. 7, 8, 9, 10

[Lew93] LEWIS J. R.: Multipoint scales: Mean and median differences and observed significance levels. *International Journal of Human-Computer Interaction 5*, 4 (1993), 383–392. 4

[Lik32] LIKERT R.: A technique for the measurement of attitudes. *Archives of psychology* (1932). 3

[LK18] LIDDELL T. M., KRUSCHKE J. K.: Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology 79* (2018), 328–348. 1, 2

[LPCRH18] LIU L., PADILLA L., CREEM-REGEHR S. H., HOUSE D. H.: Visualizing uncertain tropical cyclone predictions using representative samples from ensembles of forecast tracks. *IEEE Transactions on Visualization and Computer Graphics 25*, 1 (2018), 882–891. 8, 9

[LTPH16] LAWONN K., TROSTMANN E., PREIM B., HILDEBRANDT K.: Visualization and extraction of carvings for heritage conservation. *IEEE Transactions on Visualization and Computer Graphics 23*, 1 (2016), 801–810. 7

[MD19] MEYER M., DYKES J.: Criteria for rigor in visualization design study. *IEEE Transactions on Visualization and Computer Graphics 26*, 1 (2019), 87–97. 2

[MOJB*18] MEUSCHKE M., OELTZE-JAFRA S., BEUING O., PREIM B., LAWONN K.: Classification of blood flow patterns in cerebral aneurysms. *IEEE Transactions on Visualization and Computer Graphics 25*, 7 (2018), 2404–2418. 7

[Mum02] MUMBY P. J.: Statistical power of non-parametric tests: A quick guide for designing sampling strategies. *Marine pollution bulletin 44*, 1 (2002), 85–87. 5

[NBL20] NILSEN E. B., BOWLER D. E., LINNELL J. D.: Exploratory and confirmatory research in the open science era. *Journal of Applied Ecology 57*, 4 (2020), 842–847. 4

[Nor10] NORMAN G.: Likert scales, levels of measurement and the "laws" of statistics. *Advances in health sciences education 15*, 5 (2010), 625–632. 2, 4

[PGC18] PERUGINI M., GALLUCCI M., COSTANTINI G.: A practical primer to power analysis for simple experimental designs. *International Review of Social Psychology 31*, 1 (2018), 5, 9

[PYHZ14] POLK T., YANG J., HU Y., ZHAO Y.: TenniVis: Visualization for tennis match analysis. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (2014), 2339–2348. 9

[RAL*16] REN D., AMERSHI S., LEE B., SUH J., WILLIAMS J. D.: Squares: Supporting interactive performance analysis for multiclass classifiers. *IEEE Transactions on Visualization and Computer Graphics 23*, 1 (2016), 61–70. 7

[RH*11] ROBBINS N. B., HEIBERGER R. M., ET AL.: Plotting Likert and other rating scales. In *Proceedings of the 2011 Joint Statistical Meeting* (2011), pp. 1058–1066. 9

[RHR15] ROBERTS J. C., HEADLEAND C., RITSOS P. D.: Sketching designs using the five design-sheet methodology. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (2015), 419–428. 3

[RHY14] REN D., HÖLLERER T., YUAN X.: ivisdesigner: Expressive interactive design of information visualizations. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (2014), 2092–2101. 7

[RLA15] ROSTER C. A., LUCIANETTI L., ALBAUM G.: Exploring slider vs. categorical response formats in web-based surveys. *Journal of Research Practice 11*, 1 (2015), D1–D1. 9

[SFMB12] SEDLMAIR M., FRANK A., MUNZNER T., BUTZ A.: Relex: Visualization for actively changing overlay network specifications. *IEEE Transactions on Visualization and Computer Graphics 18*, 12 (2012), 2729–2738. 7

[SFP*18] SCHMIDT J., FLEISCHMANN D., PREIM B., BRÄNDLE N., MISTELBAUER G.: Popup-plots: Warping temporal data visualization. *IEEE Transactions on Visualization and Computer Graphics 25*, 7 (2018), 2443–2457. 7

[Sie57] SIEGEL S.: Nonparametric statistics. *The American Statistician 11*, 3 (1957), 13–19. 4, 10

[SKH*91] SCHWARZ N., KNÄUPER B., HIPPLER H.-J., NOELLE-NEUMANN E., CLARK L.: Rating scales numeric values may change the meaning of scale labels. *Public Opinion Quarterly 55*, 4 (1991), 570–582. 3

[SLK*16] SMIT N., LAWONN K., KRAIMA A., DERUITER M., SOKOOTI H., BRUCKNER S., EISEMANN E., VILANOVA A.: Pelvis: Atlas-based surgical planning for oncological pelvic surgery. *IEEE*

Transactions on Visualization and Computer Graphics 23, 1 (2016), 741–750. 8

[SNC15] STREINER D. L., NORMAN G. R., CAIRNEY J.: *Health measurement scales: a practical guide to their development and use*. Oxford University Press, USA, 2015. 9

[Ste46] STEVENS S. S.: On the theory of scales of measurement. *Science 103*, 2684 (1946), 677–680. 3, 4

[STM16] SARVGHAD A., TORY M., MAHYAR N.: Visualizing dimension coverage to support exploratory analysis. *IEEE Transactions on Visualization and Computer Graphics 23*, 1 (2016), 21–30. 7, 8

[SW65] SHAPIRO S. S., WILK M. B.: An analysis of variance test for normality (complete samples). *Biometrika 52*, 3/4 (1965), 591–611. 8

[TRL*18] TANG T., RUBAB S., LAI J., CUI W., YU L., WU Y.: istoryline: Effective convergence to hand-drawn storylines. *IEEE Transactions on Visualization and Computer Graphics 25*, 1 (2018), 769–778. 4

[VZS17] VALDEZ A. C., ZIEFLE M., SEDLMAIR M.: Priming and anchoring effects in visualization. *IEEE Transactions on Visualization and Computer Graphics 24*, 1 (2017), 584–594. 10

[WCA*16] WU Y., CAO N., ARCHAMBAULT D., SHEN Q., QU H., CUI W.: Evaluation of graph sampling: A visualization perspective. *IEEE Transactions on Visualization and Computer Graphics 23*, 1 (2016), 401–410. 8, 9

[WCS10] WEIJTERS B., CABOOTER E., SCHILLEWAERT N.: The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing 27*, 3 (2010), 236–247. 3

[Wen04] WENG L.-J.: Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educational and Psychological Measurement 64*, 6 (2004), 956–972. 3

[WG12] WONGSUPHASAWAT K., GOTZ D.: Exploring flow, factors, and outcomes of temporal event sequences with the outflow visualization. *IEEE Transactions on Visualization and Computer Graphics 18*, 12 (2012), 2659–2668. 8

[WLMB*14] WALDNER M., LE MUZIC M., BERNHARD M., PURGATHOFER W., VIOLA I.: Attractive flicker—guiding attention in dynamic narrative visualizations. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (2014), 2456–2465. 7

[WMZ*19] WEI Y., MEI H., ZHAO Y., ZHOU S., LIN B., JIANG H., CHEN W.: Evaluating perceptual bias during geometric scaling of scatterplots. *IEEE Transactions on Visualization and Computer Graphics 26*, 1 (2019), 321–331. 8

[WS09] WONGSUPHASAWAT K., SHNEIDERMAN B.: Finding comparable temporal categorical records: A similarity measure with an interactive visualization. In *2009 IEEE Symposium on Visual Analytics Science and Technology* (2009), IEEE, pp. 27–34. 7

[YDJ*18] YANG Y., DWYER T., JENNY B., MARRIOTT K., CORDEIL M., CHEN H.: Origin-destination flow maps in immersive environments. *IEEE Transactions on Visualization and Computer Graphics 25*, 1 (2018), 693–703. 9

[YEII15] YU L., EFSTATHIOU K., ISENBERG P., ISENBERG T.: Cast: Effective and efficient user interaction for context-aware selection in 3d particle clouds. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (2015), 886–895. 9

[ZLC*18] ZHAO Y., LUO F., CHEN M., WANG Y., XIA J., ZHOU F., WANG Y., CHEN Y., CHEN W.: Evaluating multi-dimensional visualizations for understanding fuzzy clusters. *IEEE Transactions on Visualization and Computer Graphics 25*, 1 (2018), 12–21. 8