How to Evaluate an Evaluation Study? Comparing and Contrasting Practices in Vis with Those of Other Disciplines

Position Paper

Anamaria Crisan*
University of British Columbia,
Department of Computer Science

Madison Elliott†
University of British Columbia,
Department of Psychology

ABSTRACT

Evaluative practices within vis research are not routinely compared to those of psychology, sociology, or other areas of empirical study, leaving vis vulnerable to the replicability crisis that has embroiled scientific research more generally. In this position paper, we compare contemporary vis evaluative practices against those in those other disciplines, and make concrete recommendations as to how vis evaluative practice can be improved through the use of quantitative, qualitative, and mixed research methods. We summarize our discussion and recommendations as a checklist, that we intend to be used a resource for vis researchers conducting evaluative studies, and for reviewers evaluating the merits of such studies.

Index Terms: Human-centered computing—Visualization—Visualization design and evaluation methods

1 INTRODUCTION

Evaluation practices within information visualization have historically been drawn from those within human computer interaction, and by extension, behavioral psychology and sociology. Over the past several years, however, psychology and sociology have experienced a transformation in the rigor of their evaluative practices that has yet to fully permeate vis itself. This does not mean the vis literature has not examined its own practices, in fact there have been active panel discussions, summative assessments (such as Munzner [31], Carpendale [8], Lam [25], Isenberg [20], and Kay [23]), and spirited collegial dialogues continually occur at vis conferences and beyond. But we do believe that vis literature is investigating its practices without a full awareness of modern, and continually evolving, practices in other empirical disciplines, and that this lack of awareness impacts the community's ability to deeply reflect on contemporary vis evaluations. One of these modern practices is the use of reviewer checklists, which are becoming an increasingly common convention in other disciplines, to not only standardize what is reported, but also verify that a study meets some minimal level of rigor that is necessary for reproducibility and that both researchers and reviewers have carefully considered.

In this position paper, we primarily comment on evaluative studies that help researchers understand the impact of a visual idiom, interaction, or system design choices upon the intended user. In keeping with evolving review practices in other disciplines, we argue for a set of checklist criteria for evaluative studies that both a researcher and reviewer must consult and comment upon in their manuscript and subsequent reviews. The criteria within our proposed checklist are drawn from our investigation of the contemporary practices within behavioral psychology, sociology, and statistics, and as such cover quantitative, qualitative, and mixed methods approaches. We also draw upon our own knowledge and experiences of submitting manuscripts to journals in other disciplines that require clear exposition of the study. We do not believe this checklist to be definitive and immutable in its current form, and welcome active dialogue within the community about the appropriateness of such an initiative, as well as the relevance of the checklist content.

The rest of the paper is as follows. In the first section we describe the scope and motivation for evaluation in vis that we address in this position paper. Ahead of the presenting the checklist, in section 2 and 3 we comment on study methods and techniques from quantitative, qualitative, and mixed methods research approaches, and how we believe vis practice tends to align or diverge from our experiences with these types of studies in other disciplines. In the fourth and final section, we present our checklist, which is informed by the methodological and technical overview presented in the prior sections. We also propose different levels of standards for reporting these evaluative studies, from insufficient, to bare minimum, good, and finally gold standard.

2 EVALUATIVE STUDIES IN VIS

2.1 Is an evaluation necessary?

Whether or not an evaluative study is necessary, and what level of rigor is required, is largely dependent upon what the investigators claim. While is is natural to extrapolate research findings to other potential impacts, researchers often tend to claim far more than their data support. For example, novel visualization idioms, techniques, or systems engineering contributions may not require evaluative studies so long as they limit their claims to those more technical contributions and do not broadly speculate on the utility of those technical contributions or their effects upon the user. Expecting a single paper to contribute a novel idiom, technique, or system, as well as a rigorous perceptual or user study, may in fact hamper progress, and can lead to one or both of the contributions to be rushed and incomplete; in our experience with vis and other human computer interaction research, it's the evaluative component that suffers most. Still, populating a design space replete with visualization idioms, interactions, and techniques, but devoid of substantive context about how the user is impacted by those design decisions ultimately has limited utility. When a user study is undertaken, it is vital that researchers accurately and consistently report their study goals, methods, and findings. We find that, despite prior calls to improve reporting practices (for example [26]), there appears to still be little consensus or standard practices of what should be reported and how. This lack of consensus puts vis behind the current expectations of rigor in other empirical disciplines, which have more clearly and strongly articulated reporting standards.

2.2 Evaluation practices beyond vis

Prior work by Lam et. al. [25] has characterized seven evaluative scenarios within vis research that occur at different stages of the design process. While this work is accurate in its representation

*e-mail: acrisan@cs.ubc.ca
†e-mail: mellio10@psych.ubc.ca

of vis evaluative practice, it does not situate these practices within those of psychology, sociology, or statistics. It may be that vis itself is sufficiently nuanced such as to require its own distinct evaluative practices, but we caution that such a belief does not benefit the community. After all, qualitative, quantitative, and mixed method approaches are proposed explicitly in [25], and these research methods do not exist in the microcosm of vis research alone. Relative to our experience with evaluative studies in healthcare and psychology, we find that the portrayal of empiricism and the evaluative strategies discussed in [25] are broad to a point of failing to distinguish between practical due diligence on a project and a scientifically rigorous result. Failure to distinguish between relatively standard project work and a scientific result confuses the importance of what information should be reported. By discussing the more traditional scientific aspects of quantitative, qualitative, and mixed method research, as we do in Section 3, we hope to clarify what aspects of a study must be reported in order to support reproducibility efforts in the future.

2.2.1 Subjective and qualitative vs objective and quantitative

One brief aside that is important to note: when reviewing vis evaluative studies we have also noted that there exists the tendency to treat the terms subjective and qualitative as interchangeable synonyms, as well as objective and quantitative. This is an incorrect practice that confuses the design and reporting of evaluative studies. To clarify this issue, we propose that researchers think of *research methods* (qualitative, quantitative, and mixed methods) separately from *data sources* (objective, subjective) when designing their evaluative studies. Objective data sources are those measured by reasonably calibrated devices, or that derive from the frank and precise and procedural reporting of some actions that were either observed by the researcher or recorded in logs (i.e. noting that the user clicked on a button is an objective fact). Subjective data sources are self-reported by a subject in the study (i.e. the participant did not like the color of the button), and can also include a researcher's interpretation of some situation (i.e. noting the participant looked confused when clicking on the button). Small nuances and rare exceptions to these definitions exist, but for our purposes, these are useful operational definitions.

We note that vis researchers have the tendency to automatically assume that subjective data is qualitative, even when the subjective information is numerically encoded via Likert scale and analyzed using standard statistical techniques - a practice that many qualitative researchers would object to. There also exists the tendency to believe that quantitative data and subsequent application of statistical techniques, are nearly always objective, but this is also false since numerical data can be biased in a number of ways.

3 Methodological approaches to evaluative studies and what to report

In this section, we frame our discussion around what we believe are common evaluative practices in psychology and sociology, and comment upon how these practices are used or are absent in vis evaluative studies. By our appraisal, vis research does not tend to meet the reporting standards that are contemporarily used in these other disciplines. This may be because researchers in vis come from many diverse backgrounds, and have likely been exposed to either different research methodologies or none at all (the computer science discipline, for example, tends to provide little instruction on the empirical research methodology necessary to conduct user studies). For this reason, we provide a summary overview of different techniques, as this baseline understanding is critical for the reporting checklist in Section 4.

There are many excellent references within vis and beyond on how to conduct quantitative, qualitative, and mixed method studies. Rather than repeat them here, we have instead summarized those we consider to be important components of empirical studies, as well as aspects of these studies that are absent, or conducted differently, in vis studies. The points we raise in this section serve as justification for our checklist criteria, which is presented in the subsequent section. Throughout our discussion here, we are conscious of the fact that there is an art to the practice of evaluative research— specifically, that it is not concisely written or expressed in one single location, but is learned by the process of conducting research itself and sharing the results. We present our current work as a position paper precisely because we structure our arguments along the learned conventions from our respective disciplines, healthcare/bioinformatics and psychology, and their research practices.

Throughout the discussion in this section, research ethics are assumed to apply across all study designs and are not commented upon individually. Furthermore, we argue that artifacts resulting from the evaluative studies should be made available, with the caveat of participant agreement to disclose these data, as they are critical for evaluating the rigor and validity.

3.1 Quantitative Methods

Quantitative methods are useful for evaluating patterns of behavior and interactions with vis displays or systems. Researchers can adapt methodological paradigms from the fields of psychology and human computer interaction for use in vis studies. It is imperative that vis researchers bear in mind that quantitative behavioral research methods were developed for inspecting, measuring, and understanding behavior, and not systems or displays. For summative assessments, Tinbergen [40] proposed an excellent but oft-forgotten way of classifying evaluative research motivations into four categories: 1) *proximate causation or control*, i.e., investigating the underlying psychological or physiological mechanisms or underpinnings of behavior, 2) *development or ontogeny*, i.e., investigating the factors that influence behavioral development or processes, 3) *function*, i.e., investigating why a behavior is used or observed to occur, and 4) *evolution or phylogeny*, i.e., investigating the evolution, adaptation, or construction of a behavior. It is notable that none of these motivations concern the impact of behavior on a separate entity. This means that a hypothesis which is explored with a psychological or sociological research paradigm must be aimed at understanding behavior itself, and that the relationship between subsequent experimental results and the goodness of a display must be carefully justified and accounted for. A common assertion from the vis community is that evaluation of complex systems or interfaces might be too challenging to operationalize or break down into rigorously observable components. We agree that this is an especially challenging aspect of vis research, but we also assert that this challenge highlights the importance of understanding how to apply quantitative methods correctly. Although a vis experiment may not be able to characterize or measure all variables/factors, it should consider and document them, as well as provide a rationale if any are left unaddressed. This way, readers can understand and make clear decisions about the evaluation study. We propose that vis research follow Tinbergen's suggestions, by first identifying and justifying the primary goal of a user evaluation. If this goal does not fit into one of Tinbergen's classifications, the study may not be measuring *user* behavior. Importantly, to our knowledge, there are no user evaluation methods that examine only a display itself— the interaction with human behavior is typically essential to its study. Therefore, another critical suggestion for implementing rigorous quantitative evaluations is for researchers to conduct thorough literature reviews. This will help researchers successfully connect chosen behavioral measures, predictions, and outcomes, with secondary hypotheses about display or system function.

3.1.1 Description of the Study Design

Quantitative methods have frequently been thought of as synonymous with perceptual or psychological evaluation methods in vis

research [25]. Though this is not always accurate, we first address when it is the case. The evaluation scenarios for user performance in Lam [25] both refer to measuring visual or cognitive performance. Cognitive psychology studies typically use randomized within subjects designs, which use subjects as their own controls [28] by exposing individuals to both experimental and control conditions. In these designs, the behavioral research assumption of independence of observations can be violated in a controlled manner, by capitalizing on running a relatively small number of participants on a large number of trails. Part of the rationale for this design is the fact that perceptual studies typically report at least medium-sized effects (i.e., $\eta^2 = 0.06$) [10]. Pre-determined effect size thresholds are crucial for planning power analyses, which allow researchers to determine and justify a given number of participants needed for their study. This also helps ensure that study results can be used in future meta-analyses. We advise that vis researchers exercise caution if they plan to use other types of study designs or procedures for perceptual evaluation, and re-iterate our preliminary suggestion and completing a rigorous literature review of similar work.

There are of course other quantitative study designs that are usable for vis evaluations. For instance, researchers could perform correlational studies with log files, so long as the system output is carefully designed to support this type of procedure. Descriptive correlational research can be conducted with user surveys and questionnaires, which can also be analyzed with quantitative methods. In the case of any analysis technique, the study must be designed to support it. For analysis of surveys and questionnaires, this typically means recruiting a large number of participants ($n > 100$). Popular choices in psychology research include item analysis [33], multiple regression [11], factor analysis [39], and structural equation modeling [24]. It is not useful to interpret or generalize human behavior or sentiment from quantitative analyses of small-sample survey data. To accomplish this goal, we instead encourage vis researchers to familiarize themselves with and consider study designs for multivariate analysis of human behavior [38].

3.1.2 Data Collection Procedures

Data for quantitative evaluative studies can be collected for both experimental or correlational study designs. Experiments can produce information about behavior and cognition, or physiological response. Correlational studies often yield descriptive information about patterns of user behavior, emotional response, and personality. Vis researchers should reference power analyses from past work with the same study design, or use available tools such as G*Power [17] to determine the number of participants needed to meet power thresholds (85% is a good heuristic) to detect the effect (we recommend at least $\eta^2 = 0.06$, or equivalent effect size for a chosen statistical procedure) in their chosen study design. In all cases, the type of data collected must be reported, along with the following items:

- The choice of participant population and why it is relevant. In experimental studies, this will typically involve random selection. If random selection is not used, a clear justification must be provided.
- The number of participants recruited. Additionally, any preliminary power analyses or past work used to determine this number should be reported/cited.
- The strategy to recruit and motivate participants. This might be student research participation credit, or monetary compensation.
- Clear data collection procedures, including information about software tools or other instruments that were used in data collection. If the data collection was observation-based, observer strategy, knowledge of the hypothesis, and motivation should be reported.
- The duration of time over which data was collected.

Steve Haroz provides a useful template for reporting experimental methods [19].

3.1.3 Data Analysis

There are many available procedures for data analysis in quantitative evaluations. Critical discussions of inferential approaches to hypothesis testing in vis are popular, including Kaptein & Robertson [22], Dragicevic [16], Kay, et al. [23]. There is general agreement that vis researchers should avoid cookbook approaches, or engineered statistical analyses. The integrity of a scientific claim is dependent upon its supporting data analysis, which must fit the assumptions of the raw data collected. One of the biggest revelations from psychology's replication crisis has been the misuse of statistical techniques [2]. We encourage vis researchers to become more engaged in discussion of statistical evaluative practices beyond their own community, in order to keep abreast with evolving analysis recommendations. Here, we will make high-level recommendations that vis researchers should consider when conducting evaluative studies.

Behavioral operationalizations and data collection for evaluation studies should be designed with a planned statistical analysis procedure in mind. It is a good idea to consult other behavioral researchers about this plan before data collection begins. This can increase the likelihood of replicating the study findings and decrease the likelihood that experimental data is unsuitable or unusable for analysis [28]. Additionally, authors should distinguish between planned confirmatory vs. exploratory analyses. Confirmatory analyses should be documented before experimentation begins. They are used to understand how much confidence can be placed on a behavioral effect or observation. Post-hoc summaries, concatenations, aggregations, and searches through raw data are essential components of behavioral research, but they must be reported as exploratory. These descriptive analyses facilitate learning by generating new hypotheses and questions, but critically, they cannot be used to confirm an existing hypothesis about empirical data.

It is important that researchers clearly report their data analysis procedure, specifically:

- How variables were defined, operationalized, and measured in the study.
- Which technique they chose for the planned confirmatory analysis (with a rationale and supporting citations).
- What data they analyzed (including what they might have excluded, and why).
- Whether they transformed or manipulated raw data before applying a statistical technique (including manipulation details and rationale).
- Software used to support the analysis.
- If they performed mulitple comparisons with a frequentist technique (p-values or confidence intervals), significance or coverage probability threshold adjustments must be reported.
- Whether exploratory analyses were conducted, with replicable details.

It is also important to publish questionnaires, analysis scripts, and even raw data (so long as it is ethical, and the authors receive permission) as supplemental materials, which are referred to in the manuscript and easily accessible, so that reviewers and future readers can assess the quality of the data analysis.

3.1.4 Validity

For quantitative evaluations, their internal validity can be defined as the extent to which the chosen behavioral measurement(s) actually fulfill their purpose of measuring or predicting what the researchers

intended. Only valid measures can be used to inform scientific hypotheses and questions. A valid vis evaluation will address and minimize both researcher and participant bias that may affect experimentation and analysis.

Here, we present two major considerations for internal validity in vis evaluations:

- The chosen measures should describe or predict only what was intended by the researchers, and nothing else.
- The measurement procedure, or experimental manipulation, should be statistically unbiased and should attempt to minimize systematic errors. This will ensure that measured values of a construct correspond to its true values.

Validity procedures should be decided upon and documented at the beginning of the study. If there is ambiguity in what is being measured or predicted, or systematic bias is present in the experimental procedure, these issues must be explicitly discussed in the report so readers can judge validity themselves.

3.1.5 Generalizability of findings

A well-designed quantitative evaluation should have external validity, or the ability to generalize to other evaluations and related studies in psychology, human computer interaction, and vis. Generalizable vis evaluations will use precise, consistent measurements and predictions that minimize random error. It is important to distinguish between the reliability of measurement and prediction in a controlled laboratory experiment vs. true ecological validity, or the ability for a study result to generalize to the natural world. Many popular experimental methods and results from psychology and human computer interaction lack *veridicaility*, or the ability to predict a broad spectrum of behavior outside of the testing environment, as well as *verisimilitude*, or the degree to which variable manipulations and observations resemble real-world contexts for the behavior being measured [15]. An established trade-off must be acknowledged, where researchers may choose to prioritize objective quantitative interpretations, which are more readily achieved in controlled laboratory environments. The main advantage of this reductionist approach is that it can clearly establish cause and effect relationships between specific variables. Its disadvantage is that applicability to full-fledged, complicated natural scenarios may be limited.

We provide guidelines for reporting on the generalizability of quantitative vis evaluations to both controlled experiments and the natural world here:

- Authors should report how consistent their measurements and predictions are. This can be investigated with repeated measures designs or self-replication. Measuring the same unmanipulated variable multiple times should produce the same results.
- Authors should identify and possibly report the smallest change in the true value of a variable that can be detected.
- Authors should consider and report how small changes in the true value of a variable are reflected by changes in its measured value.
- Authors should consider and report the ubiquity and level of the natural process they are measuring. For instance, lab experiments quantifying early visual processes such as low-level feature detection in visualizations will inherently be more ecologically valid than those measuring high-level cognitive decision-making about elements in a complex system.

The ability to generalize quantitative results to real behavioral characteristics and differences can be justified based on how reliably they have been measured in the study.

3.2 Qualitative Methods

Qualitative methods are especially valuable when trying to study processes that are not easily quantifiable, but still of relevance to the vis research community. A good example is work done by Kandel [21] that conducts a comprehensive evaluation of data visualization use in enterprise settings. For evaluative studies that address the impact of design choices on the user, qualitative methods can also be very pragmatic, since they do not require the rigid definition of a control and can be conducted "in the wild". These methods can also afford researchers the flexibility to highlight and contextualize unusual or surprising results (outliers in quantitative speak) in greater depth.

Although the legitimacy of qualitative research is debated, (less so lately), a properly conducted qualitative study applies many of the same considerations as quantitative methods toward the relevance of the research questions, the appropriateness of study design and participant population, exposition and choice of analytic techniques, and validity of the results [12]. It is important to realize that interviews conducted with a small convenience sample within contrived laboratory settings lack many of the aforementioned considerations and thus do not constitute a qualitative study. Perhaps what is most discomforting about qualitative methods is the use of subjective data sources, coupled with a perceived lack of generalizability. However, with careful reporting, we will show that rigor and validity is still possible and yields valuable results.

Ahead of the discussion on qualitative studies, we believe it is important to indicate that practice of qualitative vis studies appears to be different from ethnography and sociology (the fields from which these techniques emerged). As we will discuss further in the 'data analysis' portion of this section, vis researchers do not use theory in their qualitative analyses in quite the same way that qualitative researchers do. Instead, vis researchers primarily appear to use elicitation and data analysis techniques from qualitative research methods. Since not all vis researchers may agree with this characterization, we encourage greater discussion on the points raised in this paragraph, and in the remainder of this section.

3.2.1 Description of the Study Design

Qualitative research methods are largely drawn from ethnographic observation of an individual or group of individuals in their "natural environment". Studies can collect data passively through observation, or more actively through focus groups, interviews, or case study investigations. Vis and HCI researchers also have developed unique methods of elicitation that are used in a laboratory setting and that are better tailored toward software systems, such as wizard-of-oz designs [18], chauffeured demos [27], or multi-dimensional in depth longitudinal case studies [36]. A key difference between qualitative and quantitative methods is the role of the researcher in the study process. In qualitative methods, the researcher is herself viewed as a research instrument [12], much like a survey or measurement device is a research instrument in quantitative methods.

It is important for an evaluative study to clearly report a study design and its relevance to the research problem. It can be helpful to explicitly use the words "the design of our evaluative study is". The role of researcher as interviewer, observer, or active participant should also be clearly defined.

3.2.2 Data Collection Procedures

Data collected from qualitative studies can be derived from field notes, interviews (conducted with individuals or in focus groups), interactive sessions (for example using affinity diagramming), images, audio materials, or other documents [12, 25]. In vis research, rapid and iterative prototyping can also produce data in the form of sketches or wireframes that may be marked up with informative notes. In addition to the types of data collected, the researchers need to clearly report:

- The choice of participant population and why it is relevant. Although the participant population is less critical than in quantitative research, it is still useful to pick participants that have relevant and comparable backgrounds, analysis objectives, or level technical skill (for example).
- The strategy used to recruit participants.
- Clear data collection procedures, including information about software tools or other instruments that were used in data collection.
- The environmental context, for example, within the laboratory or in-situ within a users environment, and duration of time over which data was collected.

3.2.3 Data Analysis

There exist a number of techniques for analyzing qualitative data, but the most common objective is to categorize data into themes that are used for further analysis. Codifying elements within data (hence forth, "coding") is the core component of qualitative analysis, and more specifically, grounded theory approaches. Within sociology, there is a more explicit connection between coding and the development of theory [9]. Whereas it is our observation that within vis, qualitative coding is used more as a technique to concisely describe data, which happens to align with how qualitative coding is used in information systems user research more generally [9, 41]. The connections between theory, evaluative studies, and vis are complex and something the community may wish to further discuss. Here, we mainly discuss coding as a descriptive technique. Another noteworthy difference is that while qualitative researchers in sociology primarily analyze textual data, vis researchers also apply coding to the analysis of images [3, 13]. In the context of textual data analysis, researchers in sociology tend to make more extensive use of quotations in their analysis, whereas we've observed that vis researchers have the tendency to quantify the results of their coding and provide more descriptive statistics in their analysis. We caution against the practice of quantizing qualitative data, as this invites many of the critiques that can be directed toward quantitative evaluative studies.

It is important that researchers clearly articulate a data analysis procedure, specifically:

- What data were analyzed, and if applicable what was excluded.
- The coding technique used (open and axial coding are common in vis).
- Software used to support the analysis.

It is also important to provide artifacts of the coding process as supplemental materials, which are referred to in the manuscript and easily accessible, so that reviewers and future readers can assess the validity of the data analysis. A good example of this in practice is from Qu [34].

In qualitative research, data analysis and data collection can sometimes be a simultaneous process [12], for example, earlier interviews may modify the interview questions for later participants. If this happens to be the case, then such a synergistc relationship between data collection and analysis must be clearly reported.

3.2.4 Validity of findings

A valuable aspect of a qualitative study is its ecological validity, that is the "realism" of the results, since it is conducted in a users environment in lieu of a contrived laboratory setting [31]. However, it is also necessary to ensure that a qualitative study is internally valid, by applying checks against a researcher's biases and assumptions that can permeate into the analysis process. Assessing the validity of qualitative studies is an active area of research [12], but here we highlight some strategies that vis researchers may benefit from.

- Using member checks, where study participants are presented with a researcher's findings, and have an opportunity to comment on whether the findings are accurate.
- Using different sources of data to triangulate upon some observation and its impact. For example, using both interviews and usage logs to explore how a users interaction with a data visualization systems impacts the development of insights.
- When applicable, for example when using open or axial coding techniques, use intercoder reliability metrics to report the extent of agreement and disagreement in the coding process.
- Peer evaluations ahead of manuscript submission can also be a useful check to assess biases or assumptions a researcher is not aware of.

Validity procedures should be decided upon at the beginning of the study and should also be clearly reported in the manuscript.

3.2.5 Generalizability of findings

Whether results from qualitative studies generalize, in other words have external validity, remains a matter of debate [9, 12, 29]. The value of qualitative research is its ability to describe and analyze a specific contextual setting and individual experience [12]. For example, in instances where a data visualization is intended for a single individual or small group of users, generalizable results have little value relative to in-depth portrayals afforded by a qualitative analysis of how the individual or small group derives some insight from the visualization. Some researchers argue [9] that qualitative research results are inherently generalizable, especially when broad number of individual cases are analyzed, but there remains controversy. Claims of generalizability from qualitative studies must be carefully considered, and should be not be an expected outcome of a qualitative evaluative study.

3.3 Mixed Methods

Mixed methods research involves integrating both quantitative and qualitative research methods, building on the strengths of each to yield a more comprehensive study [12]. These include relatively new methods that have been primarily used in sociological research, and have yet to fully to be integrated into vis research evaluative practice. While powerful, mixed methods approaches can also be very resource-intensive, as they involve conducting **both** a qualitative and quantitative study and *integrating* their results. A concrete use of mixed methods research in conjunction with the design study methodology is presented in Crisan et al. [14]. There is also an interesting paper by Muller et al. [30] written for HCI applications that essentially proposes a mixture of grounded theory methods and machine learning, which is very similar to existing mixed methods designs.

Mixed methods research may be a natural fit to vis evaluative practice, which does often need both quantitative and qualitative analyses to understand the impact of design decisions on a user's ability to generate insights. However, the scale of such evaluative studies may turn many researchers away, since a properly conducted mixed methods evaluative study must adhere to all of the reporting and rigor requirements indicated both the qualitative **and** quantitative methods section of this paper; mixed methods studies are not just lesser qualitative and quantitative studies analyzed together. Still, a well conducted mixed methods study could provide an invaluable resource of information for future projects the broader vis community.

As many considerations and reporting requirements for mixed methods are inherited from qualitative and quantitative methods, we will state here only the reporting requirements that are specific to mixed methods evaluation studies. Note that in this section we refrain from using the terms "quantitative data" and "qualitative data", as many reference texts do, in order to keep with the convention

we described in section 2.2.1 and instead use "data collected for quantitative/qualitative analysis"

Finally, but most importantly, since there are few examples of mixed methods evaluative studies in vis, we can only speculate upon their utility here and **strongly recommend that readers consult other sources beyond this paper** so the community can think critically about this emergent type of evaluative study.

3.3.1 Description of the Study Design

Study designs in mixed methods research continue to evolve [12], but some concrete designs have emerged, and we list those that have been more widely adopted. The key idea is that each study design communicates the order and timing in which data are collected for qualitative and quantitative analysis.

- *Exploratory sequential designs* initially collect data for qualitative analysis that is then used to inform a quantitative analysis. This design is often used in the development of survey instruments. For example, an vis researcher may first conduct a small focus group to identify relevant analysis tasks, and develop a survey around those tasks. That survey is then later used to quantitatively rate the efficacy of a new visualization idiom or system against some existing standard.
- *Explanatory sequential designs* initially collects data for a quantitative analysis and is then followed up with data collection for a qualitative analysis that dives deeper into initial findings. This design could be very pragmatic for vis researchers, because an experimental, laboratory based, quantitative analysis can be enriched with an "in the wild" follow-on qualitative analysis. For example, a researcher may conduct an experiment to quantitatively assess visualization design choice preferences and once the results have been analyzed a qualitative analysis is initiated to try explain some of the findings.
- *Embedded convergent designs* collect data for quantitative and qualitative analysis at the same time and compares and contrasts the findings. In [14] this design was used to help researchers understand the extent to which study participants preferred some design over another, and also why.

3.3.2 Data Collection Procedures

It is critical to report the order of which data were collected per the study design.

3.3.3 Data Analysis

Data analysis proceeds as it would for quantitative and qualitative data analysis, but depending upon the study designs data can be analyzed separately or together. In sequential designs, there is an initial data analysis that informs a subsequent data collection and analysis. In convergent designs, data can be collected simultaneously and analyzed separately at first, but and then compared and contrasted together.

3.3.4 Validity of findings

When integrating the results of qualitative and quantitative analysis, researchers should also report:

- The number of participants in each component, and whether or not they used the same group of participants (for sequential designs, the same group should not be involved in both phases).
- Whether data collected in each component is truly comparable and able to be integrated.
- The extent to which each type of data (qualitative and quantitative) is used in the final analysis.

3.3.5 Generalizability of findings

Mixed methods evaluative studies collect far richer data than either quantitative or qualitative methods alone. So long as the individual analysis components are valid, and in the case of the quantitative component, are designed to be generalizable, then the results from mixed methods studies may generalize.

4 A CHECKLIST FOR EVALUATIVE STUDIES IN VIS

Having described the different methods vis researchers can use in their evaluative studies, and how they compared to our understand of current modern practices in other disciplines, we now summarize our proposed reporting criteria as a checklist of questions researchers and reviewers should ask themselves. Subsequent to the checklist, we also present different levels of reporting standards for reviewers to consider. For researchers, this checklist is intended to be used as a guide when preparing and presenting the results of an evaluation study. For reviewers, this checklist is intended to help verify the merits of the study, and we would like to see reviewers explicitly comment on the checklist items in their reviews of evaluative studies. As we indicated in the introduction, this check list is not immutable, and we encourage active discussion as to its appropriateness and utility. Recommendations for specific research methods are indicated using square brackets ([]) and are listed *after* general considerations that apply across all studies.

- **Description of Study Designs**
  - Is there a clearly articulated research motivation and question?
  - Is a specific research methodology (quantitative, qualitative, mixed) stated?
  - Is a specific study design clearly articulated?
  - Is the study design appropriate for the research motivation?
- **Data Collection Procedures**
  - Is the data collected appropriate for the stated study design?
  - Is the data collected appropriate for the research motivation?
  - If they can be reasonably made available, are data collection instruments (surveys, interview questions, etc.) included as supplemental materials?
  - Is there a clear description of the study participant population and its size?
  - Is the study population appropriate for the study design and research motivation?
  - Is there a clearly articulated strategy to recruit study participants? Is this strategy appropriate?
  - Are data collection procedures clearly articulated?
  - Is it clear over what duration, and in what conditions, data were collected?
  - Are data collection instruments (devices, questionnaires, interview questions etc) and software described?
  - Was data collected in an ethical manner, and with appropriate consideration for a study participant's privacy? Specifically for mechanical turk studies, were the participants fairly compensated (i.e. paid minimum wage or higher)?
  - [Quantitative Methods] Is the number of participants recruited justified by a power analysis or relevant prior work?
  - [Quantitative Methods] Is the data collected appropriate for the variable operationalizations and definitions?
  - [Qualitative Methods] Is the researcher's role in data collection (observer, interviewer, active participant) indicated?
  - [Mixed Methods] Is the order of data collection and analysis specified?

- **Data Analysis**
  - Is there a clearly articulated data analysis plan? Is it appropriate?
  - Is it clear what data were included or excluded in the data analysis?
  - Is software used in analysis listed?
  - Within reason, and without violating the study participant's privacy, do researchers make analysis artifacts available so that others can independently verify the results?
  - Is it clear whether the researchers are conducting a confirmatory or exploratory study analysis?
  - Within reason and without violating ethics and study participant privacy, are the study data made available? Is there an explicit justification for why data was not made available?
  - If the necessary supporting documents are not made available and appropriate justification for their absence is not provided should this paper still be considered for publication?
  - [Quantitative Methods] Was the data analysis plan determined ahead of the study? Was the procedure pre-registered?
  - [Quantitative Methods] Is it clear what the researcher is measuring?
  - [Quantitative Methods] Is analysis code made available?
  - [Quantitative Methods] Is the choice of statistical analysis techniques appropriate? Do the researchers articulate how the choice of methods is appropriate for the type of data collected?
  - [Mixed Methods] Is is clear when results were analyzed? In sequential study designs, separately and different points in times, in convergent study designs simultaneously.
  - [Qualitative Methods] Are artifacts of the data coding processing made available?
- **Validity of Findings**
  - Is the study internally valid? That is, have the researchers taken care to avoid introducing systematic biases or errors and have the researchers chosen the appropriate analysis methods for the intended research motivation? At a bare minimum, all vis evaluative studies **must** be internally valid.
  - Is the study externally valid? That is, do researchers demonstrate that the results apply to other settings? It may not be necessary for a study to be externally valid so long a researchers do not make such claims.
  - Is the study ecologically valid? That is, do the study conditions reflect real-world, or "in the wild" settings? It may not be necessary for a study to be ecologically valid so long a researchers do not make such claims.
- **Generalizability**
  - From the description of the study design, data collection, data analysis, and validity statements, is it legitimate to claim generalizability of the study findings?

4.1 Levels of Reporting Standards

We propose different levels of reporting standards to achieve a level of reproducibility that aligns with contemporaneous efforts in psychology, sociology, and statistics, to improve study rigor and, by extension, reproducibility. Most importantly, achieving high levels of reporting standards does not guarantee that a study's findings are accurate or reproducible. Instead, accurate reporting provides researchers with sufficient information to assess a study and reproduce its findings. We also encourage critical discussion within the vis community about which checklist items should be prioritized during review. Vis researchers should decide which reporting standard they consider necessary in order to achieve an acceptable level of quality in user evaluations.

The first level is ***Insufficient Reporting***. An evaluative study that provides little to no details on study design, data collection, and data analysis, provides no supplemental materials, and makes broad and speculative claims. Such studies should not be published.

The second level is ***Bare Minimum Reporting***. An evaluative study that achieves the bare minimum level of reporting should comment upon the study design, data collection, and data analysis procedures somewhere in the manuscript and comment upon some, but not all, check list items. The reader should be able to assess whether the study is at least internally valid and whether authors speculated too broadly on their results. Bare minimum reporting does not include supplemental material. While this is quite a low standard, there are in fact evaluative user studies to date that are published and fail to achieve bare minimum reporting standards.

The third level is ***Good Reporting***. An evaluative study that achieves a good reporting standard would have a dedicated section of the manuscript to describe the study design, data collection procedures, and data analysis steps and would comment upon some, but not all, check list items. It should be clear that the study is at least internally valid. To evaluate their claims, the authors would make explicit comments on how their findings tie back to their methods and clearly indicate the limitations of their methods to speculate upon specific outcomes. The authors provide some supplemental materials to support their analysis findings, in particular analysis dataset and scripts, but do not make all study available. Following the replication crisis, many fields are pushing for a Good Level of reporting with explicit reviewer checklists, some even push for gold standard reporting.

The final level is ***Gold Standard Reporting***. An evaluative study that achieves a gold standard of reporting would have a dedicated section of the manuscript to describe the study design, data collect procedures, and data analysis steps reporting all, or very-nearly all of the checklist items. If the authors pre-register their quantitative studies, it should be easy for the reader to find a dated version of the pre-registered study. It should be clear that the study is at least internally valid. To evaluate their claims, the authors should make explicit comments on how their findings tie back to their methods, and clearly indicate the limitations of their methods to speculate upon specific outcomes. The authors provide all non-confidential study data and analysis scripts in a manner that allows the reader to link to these supplemental materials in perpetuity.

5 Discussion

5.1 The evolution of science and where vis currently stands

Information visualization is a young discipline with evolving methodological practices. In this position paper, we have commented upon the interplay between contemporary evaluative practices in vis relative to developments in psychology, sociology, and statistical practice, and have suggested ways to improve the rigor of vis evaluative studies. The replication crisis in psychology, sociology, and beyond, has forced researchers to introspectively reassess many long-held assumptions, and we have attempted to integrate this much broader and evolving discussion into vis evaluative practices. For example, in perceptual science, researchers have investigated the validity, reliability, and replicability of small-n, many-trial lab studies using massive, representative samples recruited through Mechanical Turk. Berinsky, et al. [1] and Buhrmester et al. [7] show high agreement on data and subsequent findings from controlled laboratory experiments such as Brady & Alvarez [4, 5], and Brady & Tenenbaum [6]. In social and quantitative psychology, there is active discussion and research into the virtue of methodological

rigor and transparency over alternative hypothesis testing frameworks [32, 35, 37], an introspection which we echo in this position paper.

Within sociology, and even human computer interaction, there continues to be active discussion about the role of qualitative research methods and ways to improve rigor and validity. Exciting developments in mixed methods research can support new and interesting research questions that vis researchers have yet to fully explore.

5.2 General recommendations

In addition to our in-depth discussion of available methods for vis evaluations, we hope both researchers and reviewers will use our checklist in future work. It is not necessary to comment on every single line item in the checklist, but instead to comment on the general aspects of the study exposition (Study Designs, Data Collection, Data Analysis, Validity, Generalizability, and Supporting Documentation). The detailed checklist items serve as prompts for what should be reported or considered when reviewing a manuscript. Use of the checklist is a guard against passive and convenient data collection and the inappropriate application of analysis techniques. Finally, we actively encourage formal dialogue and consultation with colleagues in human subjects, psychology, statistics, and sociology research disciplines, so we can improve evaluative practices within vis.

Vis evaluation exists at the crossroads of several research disciplines with evolving agendas and rapidly improving research methodologies. Increasing education and establishing dialogue with related academic disciplines will not only improve our ability to implement existing methods more rigorously, but even more importantly, it will offer vis researchers a seat at the pan-disciplinary decision-making table to influence the ongoing growth and development of modern evaluative research methods.

Acknowledgments

We would like to thank the following members of the infovis research group at the University of British Columbia: Dr. Tamara Munzner, Zipeng Liu, Kimberly Dextras-Romagnino, Michael Oppermann, Emily Hindalong, and Shannah Fisher, as well as Dr. Ron Rensink, for many long discussions on evaluation practices between vis and our own disciplines of epidemiology/biostatistics and cognitive science.

References

[1] A. J. Berinsky, G. A. Huber, and G. S. Lenz. Evaluating online labor markets for experimental research: Amazon.com's mechanical turk. *Political Analysis*, 20(3):351–368, 2012.

[2] J. Bohannon. Many psychology papers fail replication test. *Science*, 2015.

[3] M. A. Borkin, Z. Bylinskii, N. W. Kim, C. M. Bainbridge, C. S. Yeh, D. Borkin, H. Pfister, and A. Oliva. Beyond Memorability: Visualization Recognition and Recall. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):519–528, 2016. doi: 10.1109/TVCG.2015.2467732

[4] T. F. Brady and G. A. Alvarez. Hierarchical encoding in visual working memory: Ensemble statistics bias memory for individual items. *Psychological Science*, 22(3):384–392, 2011.

[5] T. F. Brady and G. A. Alvarez. No evidence for a fixed object limit in working memory: Spatial ensemble representations inflate estimates of working memory capacity for complex objects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 41(3):921, 2015.

[6] T. F. Brady and J. B. Tenenbaum. A probabilistic model of visual working memory: Incorporating higher order regularities into working memory capacity estimates. *Psychological review*, 120(1):85, 2013.

[7] M. Buhrmester, T. Kwang, and S. D. Gosling. Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on psychological science*, 6(1):3–5, 2011.

[8] S. Carpendale. Evaluating Information Visualizations Challenges in Evaluating Information Visualizations. *Information Visualization*, pp. 19–45, 2008. doi: 10.1007/978-3-540-70956-5_2

[9] K. Charmaz. *Constructing grounded theory: a practical guide through qualitative analysis*. Sage, London, 2006.

[10] J. Cohen. Statistical power analysis for the behavioral sciences. 2nd, 1988.

[11] J. Cohen, P. Cohen, S. G. West, and L. S. Aiken. *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge, 2013.

[12] J. Creswell. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. Sage, London, 2014.

[13] A. Crisan, J. L. Gardy, and T. Munzner. A method for systematically surveying data visualizations in infectious disease genomic epidemiology. *BioArxiv*, 2018. doi: 10.1101/325290

[14] A. Crisan, G. McKee, T. Munzner, and J. L. Gardy. Evidence-based design and evaluation of a whole genome sequencing clinical report for the reference microbiology laboratory. *PeerJ*, 6:e4218, jan 2018. doi: 10.7717/peerj.4218

[15] D. R. Dawson and T. D. Marcotte. Special issue on ecological validity and cognitive assessment, 2017.

[16] P. Dragicevic. *HCI Statistics without p-values*. PhD thesis, Inria, 2015.

[17] F. Faul, E. Erdfelder, A.-G. Lang, and A. Buchner. G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, 39(2):175–191, 2007.

[18] B. Hanington and B. Martin. *Universal methods of design: 100 ways to research complex problems, develop innovative ideas, and design effective solutions*. Rockport Publishers, 2012.

[19] S. Haroz. Experiment Methods Template, 2018.

[20] T. Isenberg, P. Isenberg, J. Chen, M. Sedlmair, and T. Möller. A systematic review on the practice of evaluating visualization. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2818–2827, 2013.

[21] S. Kandel, A. Paepcke, J. M. Hellerstein, and J. Heer. Enterprise data analysis and visualization: An interview study. *IEEE Transactions on Visualization and Computer Graphics*, 18(October):2917–2926, 2012. doi: 10.1109/TVCG.2012.219

[22] M. Kaptein and J. Robertson. Rethinking statistical analysis methods for chi. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1105–1114. ACM, 2012.

[23] M. Kay, S. Haroz, S. Guha, and P. Dragicevic. Special Interest Group on Transparent Statistics in HCI. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '16*, pp. 1081–1084, 2016. doi: 10.1145/2851581.2886442

[24] R. A. Klein, K. A. Ratliff, M. Vianello, R. B. Adams Jr, Š. Bahník, M. J. Bernstein, K. Bocian, M. J. Brandt, B. Brooks, C. C. Brumbaugh, et al. Investigating variation in replicability: A many labs replication project. *Social psychology*, 45(3):142, 2014.

[25] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale. Empirical studies in information visualization: Seven scenarios. *IEEE Transactions on Visualization and Computer Graphics*, 18(9):1520–1536, 2012.

[26] H. Lam and T. Munzner. Increasing the utility of quantitative empirical studies for meta-analysis. *Proceedings of the 2008 conference on BEyond time and errors novel evaLuation methods for Information Visualization - BELIV '08*, p. 1, 2008. doi: 10.1145/1377966.1377969

[27] D. Lloyd and J. Dykes. Human-centered approaches in geovisualization design: Investigating multiple methods through a long-term case study. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2498–2507, 2011. doi: 10.1109/TVCG.2011.209

[28] P. Martin, P. P. G. Bateson, and P. Bateson. *Measuring behaviour: an introductory guide*. Cambridge University Press, 1993.

[29] J. A. Maxwell. *Qualitative Research Design: An Interactive Approach*, vol. 41. 2013.

[30] M. Muller, S. Guha, E. P. S. Baumer, D. Mimno, and N. S. Shami. Machine Learning and Grounded Theory Method: Convergence, Di-

vergence, and Combination. *Proc. GROUP*, pp. 0–6, 2016. doi: 10.1145/2957276.2957280

[31] T. Munzner. Process and pitfalls in writing information visualization research papers. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 4950 LNCS, pp. 134–153, 2008. doi: 10.1007/978-3-540-70956-5_6

[32] L. D. Nelson, J. Simmons, and U. Simonsohn. Psychology's renaissance. *Annual review of psychology*, 69:511–534, 2018.

[33] J. C. Nunnally and I. Bernstein. *Psychometric Theory (McGraw-Hill Series in Psychology)*, vol. 3. McGraw-Hill New York, 1994.

[34] Z. Qu and J. Hullman. Keeping Multiple Views Consistent: Constraints, Validations, and Exceptions in Visualization Authoring. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):468–477, 2018. doi: 10.1109/TVCG.2017.2744198

[35] V. Savalei and E. Dunn. Is the call to abandon p-values the red herring of the replicability crisis? *Frontiers in psychology*, 6:245, 2015.

[36] B. Shneiderman and C. Plaisant. Strategies for Evaluating Information Visualization Tools: Multi-dimensional In-depth Long-term Case Studies. *Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization - BELIV '06*, pp. 1–7, 2006. doi: 10.1145/1168149.1168158

[37] U. Simonsohn, L. D. Nelson, and J. P. Simmons. P-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science*, 9(6):666–681, 2014.

[38] B. G. Tabachnick and L. S. Fidell. *Using multivariate statistics*. Allyn & Bacon/Pearson Education, 2007.

[39] B. Thompson. *Exploratory and confirmatory factor analysis: Understanding concepts and applications.* American Psychological Association, 2004.

[40] N. Tinbergen. On aims and methods of ethology. *Ethology*, 20(4):410–433, 1963.

[41] C. Urquhart. Strategies for Conversation and Systems Analysis in Requirements Gathering: A Qualitative View of Analyst-Client Communication. *The Qualitative Report*, 4(41):1–19, 2000.