Paths Explored, Paths Omitted, Paths Obscured: Decision Points & Selective Reporting in End-to-End Data Analysis

Yang Liu
University of Washington
yliu0@uw.edu

Tim Althoff
University of Washington
althoff@cs.washington.edu

Jeffrey Heer
University of Washington
jheer@uw.edu

ABSTRACT

Drawing reliable inferences from data involves many, sometimes arbitrary, decisions across phases of data collection, wrangling, and modeling. As different choices can lead to diverging conclusions, understanding how researchers make analytic decisions is important for supporting robust and replicable analysis. In this study, we pore over nine published research studies and conduct semi-structured interviews with their authors. We observe that researchers often base their decisions on methodological or theoretical concerns, but subject to constraints arising from the data, expertise, or perceived interpretability. We confirm that researchers may experiment with choices in search of desirable results, but also identify other reasons why researchers explore alternatives yet omit findings. In concert with our interviews, we also contribute visualizations for communicating decision processes throughout an analysis. Based on our results, we identify design opportunities for strengthening end-to-end analysis, for instance via tracking and meta-analysis of multiple decision paths.

Author Keywords

Data analysis; Analytic decision making; Multiverse analysis; Garden of forking paths; Reproducibility; Interview Study

CCS Concepts

•**Human-centered computing** → **Human computer interaction (HCI);** HCI theory, concepts and models;

INTRODUCTION

A replicability crisis has stirred multiple scientific fields [3], with replication studies failing to validate prior results [5, 6, 43, 46]. In Biology, two laboratories ventured to validate published "landmark" studies, but were successful in replicating the original results in only 11% and 25% of projects, respectively [5, 46]. In Psychology, the Open Science Collaboration replicated 100 published studies using high-powered designs and original materials, but found that on average, "replication effects were half the magnitude of original effects" [43].

Why are these studies, backed by empirical evidence from peer-reviewed data analysis, failing to replicate? Scholars suggest that undisclosed freedom in analytic decisions plays a key role [23, 51]. Researchers routinely make decisions throughout quantitative data analysis, from data collection and wrangling, to statistical modeling and inference. For example, what are the cutoffs for excluding outliers? What variations of model formulae should one choose? Different sequences of analytic decisions might result in different interpretations of empirical data, possibly leading to conflicting conclusions. Failing to constrain this freedom – experimenting with alternative analytic paths and selectively reporting findings – inflates the chance of false discovery [51]. Even well-intentioned experts produce large variations in analysis outcomes [50], suggesting a degree of arbitrariness in analytic decisions.

In response, we investigate decision making within *end-to-end quantitative analysis*: the full lifecycle of quantitative data analysis including phases of data collection, wrangling, modeling, and evaluation. We conduct semi-structured interviews with authors of nine published studies in HCI and other scientific domains. We pore over participants' manuscripts and analysis scripts to assess their decisions, and ask them to recall, brainstorm, and compare alternatives in every analytic step.

In this paper, we contribute the results and analysis of these interviews. We present a visualization design for representing analytical decisions, both to communicate our interview results and as a tool for mapping future studies. We identify recurring rationales for analytic decisions, highlighting conflicts and implicit trade-offs among options. Next we examine the motivations for carrying out alternative analyses, a practice that exercises freedom in analytic decisions. We subsequently discuss how participants choose what to include in research reports if they have explored multiple paths. Finally, based on our observations, we identify design opportunities for strengthening end-to-end analysis, for instance via tracking and meta-analysis of multiple decision paths. Given the HCI community's demonstrated interest in quantitative empirical research, we hope our findings will help inspire the design of both improved analysis tools and community standards.

RELATED WORK

Our work is motivated by the replicability crisis and issues of "researcher degrees of freedom." Our visualizations draw on the scientific workflow literature, and our interview results relate to both provenance tracking and multiverse analysis.

Practices for Improving Replicability

Replicability concerns have prompted scientists to re-examine how data analysis practices might lead to spurious findings.

CHI'20, April 25–30, 2020, Honolulu, HI, USA

ISBN 978-1-4503-6708-0/20/04…$15.00

DOI: https://doi.org/10.1145/3313831.3376533

Simmons *et al.* [51] describe how *researcher degrees of freedom* – the flexibility in making analytic decisions – might inflate false-positive rates (*i.e., p-hacking* [41]). Machine learning researchers note similar issues, for example tuning random seeds can drastically alter results [28]. Gelman & Loken [23, 24] argue that p-hacking need not be intentional, as *implicit* decisions present similar threats. They use a metaphor of a *garden of forking paths*, with each path potentially leading to different outcomes. Failing to address this flexibility gives rise to issues such as multiple comparison problem (MCP) [19, 20, 61], hypothesizing after the results are known (*HARKing*) [33], and overfitting [47]. As indicated by a survey of 2,000 psychologists [29], p-hacking is unfortunately prevalent.

In response, scholars have endorsed a number practices, including pre-registration [10, 56, 58], using estimation instead of dichotomous testing [2, 13, 14], adopting Bayesian statistics [22, 32], and increasing transparency in reporting [17, 39, 41, 53]. Wicherts *et al.* [59] develop a comprehensive decision checklist for study design, data collection, analysis, and reporting. The HCI community has also contributed empirical studies [17, 31], tools [16, 18, 38, 57] and design spaces [47] for improving reproducibility. Closest to our work are the interview studies by Kale *et al.* [31] and Liu *et al.* [36]. We corroborate Kale's findings on analytic decision-making strategies and Liu's observations on motivations for pursuing alternatives. By richly diagramming our participants' analyses, we further observe recurring patterns in analysis processes, such as feedback loops and fixations. In addition, by closely examining specific, published analyses, we identify conflicts between decision rationales and opportunism.

One perspective is that flexibility is unavoidable [50, 52, 54], as well-intentioned experts may produce divergent outcomes. In a crowdsourced study [50], 29 teams analyzed the same dataset to answer the same question, yet the analyses and conclusions differ considerably. This variation is not explained by prior beliefs, expertise, or peer-reviewed quality of the analysis [50]. Fixating on a single analytic path may be less conclusive, with results dependent on arbitrary choices [52].

For more comprehensive assessments, researchers have proposed *multiverse analysis* [52, 54]: evaluating all "reasonable" analysis specifications and interpreting results collectively (where "reasonable" decisions are those with firm theoretical or statistical support [52]). Others have adopted multiverse analysis in practice [48], cited it as an important area for future tool work [31], designed interactive media for multiverse results [15], and proposed ways to quantify multiple analysis findings [44, 60]. Existing multiverse visualizations typically use animation [15], juxtaposition [15, 52], or aggregation [54] to convey analysis outcomes; some visualize the decision space in a matrix view [52, 54]. Our analytic decision graph visualizations convey the multiverse decision space by depicting decisions in relation to the overall analysis process.

Workflow and Provenance

Prior work on computational reproducibility concerns scientific workflows [37] – process networks of analytic steps, that model a data analysis pipeline. Workflow management systems (*e.g.,* [7, 37, 42]) provide languages to specify workflows and record information for automation, reproducibility, and sharing [21]. These workflows are often represented as directed graphs, where nodes are computational steps and edges convey data flow. We similarly design visualizations to communicate data analysis processes, but focus not on a singular dataflow but on the space of potential decisions.

Many workflow management systems also record provenance information, namely the history of execution steps and the environment, such as input data and parameters [21]. To visualize provenance relationships, prior work predominantly uses network diagrams [7, 9, 40, 45, 62]. For example, VisTrails [7] visualizes provenance as a tree where a node denotes a separate dataflow that differs from its parent and an edge records the changes. Recent work also explores human-centered interactions with history [26, 35], for example supporting annotations on automatically collected provenance [26] and providing lightweight interactions within Jupyter Notebooks [35]. However, understanding how analytic decisions affect outcomes is still difficult in existing provenance tools, partly because history interactions are disconnected from the analysis pipeline. Our analytic decision graphs capture all paths taken and might serve as a navigation overview to explore history data.

METHODS

To better understand decision-making in end-to-end quantitative data analysis, we conducted semi-structured interviews with authors of nine published studies. We first inspected the papers and analysis scripts, then engaged researchers in discussion about their decision rationales and possible alternatives.

Participants

We interviewed 9 academic scientists (3 females, 6 males, age 24–72), including 6 Ph.D. students, 2 research scientists and 1 tenured professor. Our interviewee's research fields include Human-Computer Interaction (5), Proteomics (2), Marine Biology (1), and Geography (1). Participants' analyses cover a spectrum from directed question-answering to open-ended exploration. P1-5 conducted confirmatory analyses: they designed controlled experiments to answer predefined research questions. P6 explored their data to develop a biological assay. P7 and P8 performed exploratory data analyses (EDA). P9 gathered insights from EDA to form a hypothesis for a subsequent confirmatory experiment.

We recruited interviewees by advertising in multiple HCI and data science mailing lists. We also identified 15 local authors from the CHI 2018 proceedings and emailed them directly, netting three participants. Regardless of recruitment method, all interested participants filled out a survey to provide a publication and the accompanying analysis scripts. We recruited every respondent whose publication involved quantitative data analysis and had been published in a peer-reviewed venue.

Interview Procedure

We interviewed one researcher at a time for 60–90 minutes. We began each interview with an introduction describing the purpose of the interview: to understand decision making during data analysis and to collect use case examples for developing prototype tools for robust data analysis. We then proceeded

Theme	Category	Description	Representative Quote	%
Decision rationales	Methodology	Participants defend the decision with methodological concerns, including statistical validity, study design and research scope.	I mainly used t-test for hypothesis testing because my data was parametric.	25
	Prior work	Participants support the analytic decision using previous studies, "standard practice" and/or internalized knowledge.	We adapt the method from a previous paper and we follow the same process to do the analysis.	33
	Data	Participants mention data constraints, including data availability, data size and data quality.	The reason I combined them together is because more data has less variation.	21
	Expertise	Participants feel limited by expertise.	I don't know how to do this really.	12
	Communication	Participants prefer an alternative that is easier to communicate.	Because they were actually very hard to write up.	7
	Sensitivity	Participants believe that the decision has little impact on the results and provide no further rationales.	In my quick mental calculation, it seemed like it wouldn't actually make a big difference.	3
Executing alternatives	Opportunism	Participants willingly explore new alternatives to look for desired results.	I tried three different settings for those parameters and the chosen ones looked slightly better.	45
	Systematicity	Participants outline all reasonable alternatives, implement them, and choose the winning alternative based on an objective metric.	We performed a sensitivity analysis to identify the best combination.	9
	Robustness	Participants implement additional alternatives after making a decision, in order to gauge the robustness of their conclusions.	That is just for robustness, to say, "even if you look at [another option], you see the same thing."	16
	Contingency	Participants have to deviate from their original plans because the planned analysis turned out to be erroneous and/or infeasible.	This [filter] produced anomalous results and we went back [to apply] a more stringent filtering.	30
Selective reporting	Desired results	Participants only report the desired results and omit findings that are non-significant, uninteresting, or incoherent to their theory.	It felt stronger to say five out of seven, rather than four out of six, was one reason to keep it.	29
	Similar results	Participants claim that the results are similar and thus omit interchangeable alternative analyses.	But it didn't make a huge difference so we just kind of went with [the current option].	10
	Correctness	Participants apply rationales, primarily methodology and prior work, to remove analytic approaches they consider incorrect.	I was concerned about whether I had a strong hypothesis to see those interaction effects or not.	31
	Social constraints	Social constraints and communication concerns prevent participants from reporting some findings.	I'm a second author and many decisions made in the manuscript writing were against my wishes.	31

Table 1: Themes and high-level codes that emerged from open coding of the interview data.

with our discussion protocol, which consisted of three phases. The discussion focused specifically on the analysis project provided to us by the participant in the signup survey. Afterwards, all participants were compensated with a $20.00 gift card.

Phase 1: Recall. We first asked participants to freely propose different, yet justifiable analytic decisions. We encouraged participants to recall alternative paths they had considered and executed, and those raised by reviewers. We did this prior to other phases to elicit responses without biasing participants.

Phase 2: Brainstorm. We asked participants to brainstorm alternatives using a checklist (Figure supp. 1) based loosely on the work of Wicherts *et al.* [59]. The checklist contains common analytic decisions across stages of a typical data analysis pipeline, from data collection and wrangling to modeling and inference. We used the checklist to help participants systematically examine all steps in the end-to-end pipeline.

Phase 3: Compare. To raise options overlooked by participants in the previous phases, we discussed additional decisions we had prepared before the interview. We generated alternative analytic proposals by perusing the paper, appendix, and analysis scripts, while consulting the checklist to ensure a comprehensive coverage of different phases of the analysis.

Analysis of Interview Data

All interviews were audio recorded and transcribed verbatim. The first author analyzed the data, with iterative feedback from other authors throughout the analysis process. As our findings might put participants in a vulnerable position, we have replaced identifiable information in figures and quotes. For example, we might replace an identifiable variable name (*autophagy substrate*) with a generic name (*IV*).

We first sought to understand the overall analysis process. From the interview data, we extracted analytic steps and their relationships to re-construct both decision points and data flow. We drew graphs to aid interpretation, and soon realized that the graphs had greater utility beyond summarizing interview results. We thus conducted a dedicated design exercise by outlining design goals, iterating over visual encodings, and producing visualizations, as detailed in the next section.

Next we investigated how participants made analytic decisions. We began by using open coding [11] as a preliminary step to identify recurring themes. Three themes emerged: participants provided *rationales* for decisions, described their experiences in *executing alternative* analyses and subsequently *selective reporting* of the results. We integrated raw codes within each theme to extract common concepts and patterns. Table 1 summarizes the themes and categories, along with example quotes (the quotes were edited for brevity and clarity; full quotes and relevant contexts are in later sections). The table also lists the prevalence of each category, computed as the ratio of unique instances within each theme. We discuss our empirical findings in the section *Interview Results*.

Limitations

One limitation is our convenience sampling approach, which introduces potential bias. For example, our sample is mostly composed of HCI and junior researchers. To be clear, our research goal is to characterize the space of analytic processes and decisions, *not* to quantify the prevalence of any specific activity. Also, while our study reaches saturation in some regards [25] as the last two participants did not surface new categories, our convenience sample might miss known phenomena. Some practices currently gaining adherents, such

as pre-registration followed by exploratory analysis on collected data and planned analysis based on simulated data, are not observed. A future taxonomy might better delineate the distinction between a-priori and a-posteriori decisions.

We note violations of methodological validity when perusing participants' analyses to flag potentially problematic practices, but our judgments are subjective. Some methods we endorse are not universally accepted, such as multiple comparison correction [12]. Other than methodological validity, we interpret from the perspectives of the participants as much as we could.

Participants might withhold information on potentially problematic practices. Where possible, we complement the transcripts with what we found from participants' analysis scripts (*e.g.,* evidence of implementing multiple model formulae in R code), but not all scripts retain the full history. Thus, there are likely additional explorations of alternatives that we are unable to observe. In addition, all accounts of analytic decisions were given post-hoc. Future studies are needed to inspect researchers' decision making process during the analysis event.

ANALYTIC DECISION GRAPHS

To represent participants' process and decisions, we created visualizations that we call *Analytic Decision Graphs* (ADGs). We developed ADGs in conjunction with our analysis of the transcribed interviews. We present the design of ADGs here first, so that we can refer to them in our later discussions.

Design Goals

ADGs aim to visualize analytic decisions in the context of end-to-end analysis pipelines. We expect ADGs to afford two utilities. First, with ADGs as visual illustrations, authors should be able to communicate their decisions and processes more easily. Second, ADGs might prompt reflection on decisions, potentially encouraging consideration of further alternatives.

To review an analysis decision process, users will need to perform at least the following tasks:

- Gain an *overview* of the high-level analytic components.
- Understand the analytic *steps* and their relationships.
- Examine and evaluate the *decisions* made in each step.

From these tasks we can distill some design requirements:

- *Represent the input and the outcomes.* To provide context, ADGs should include inputs such as data sources and outcomes such as deliverables supporting the conclusion.
- *Display granularity of analysis components.* ADGs should visualize both high-level modules and individual decisions.
- *Represent relationships between the steps.* ADGs should capture various types of relationships, such as order and dependency, to organize steps into a coherent process.
- *Visualize the rationales and the ramifications of a decision.* Visualizing rationales might help authors identify weak spots and help readers gauge the validity.

Visual Encodings

To meet these requirements we iterated over several designs. We discuss the tradeoffs made and present the final design. As ADGs should visualize both steps and their relationships, a graph is a natural representation. We use a *node* (○) to encode a *decision point* and an *edge* to encode the *relationship* between two decision points. We further include auxiliary nodes with distinct shapes: rectangles (■) represent analysis outcomes, whereas solid dots (●) are "dummy" nodes. In an earlier design, we visualized all potential alternative choices one could make in addition to the decision point, but the graph soon grew cluttered. We thus omit individual alternatives.

Various types of relationships exist between two decision points. The first type is a *dataflow dependency* (——), where the output of one node is the input to another. The second type is a *procedural dependency* (⟶), where the downstream decision would not exist if some alternative in the upstream decision were chosen. For example, if a researcher had chosen a frequentist model instead of a Bayesian model, she would not need to decide among different priors. The third type is an *information dependency* (⟶), where one decision informs another. For example, insights from exploratory analysis might inform the hypothesis of a subsequent confirmatory experiment. We also have *feedback loops* (⇠⋯), as researchers revise an upstream decision based on the results from a downstream step. All of these relationships appear as edges of different textures. We further arrange the nodes vertically according to their order in the dataflow, with the top being the start. Yet another type of relationship exists – *temporal order* – as some decisions are made earlier than others. We overload the vertical axis to represent temporal order when it does not conflict with dataflow dependency.

We use a categorical color palette to represent type of decision rationale. To reduce visual complexity, we simplify the categories of Table 1 to three groups. We use a red color for *desired results* (●) to call out potentially problematic practice; this is when researchers made the decision by weighing end results, for example discarding options that produced non-significant results. We assign blue to *data, methodology, and prior work* (●), which are relatively primary concerns. The rest of the rationales, denoted *other rationales* (●), receive a desaturated gray color. Finally, as we (the interviewers) might propose alternatives that the participant had not thought of, we use white (○) to indicate additional decisions not considered by the participants at the time of analysis. Further information on color assignment is in the supplemental material.

The size of a node corresponds to the number of enumerated alternatives for the decision point. The thickness of a dataflow edge conveys the number of accumulated alternatives, namely all possible combinations of alternatives of previous decisions leading to that point. Since the accumulated total grows exponentially, we use a logarithmic scale for edge thickness.

ANALYSIS OF ANALYTIC DECISION GRAPHS

We created an ADG for each participant, as shown in Figure 2 (full-size diagrams are available in supplemental material). We first describe P1's ADG (Figure 1) in detail, then summarize recurring patterns drawn from the ADGs for all participants. As comparing unrelated studies is not a design goal of ADGs, care should be taken when interpreting apparent differences between graphs. Some visual properties (*e.g.,* those described below) are meaningful to compare, but other visual differences (*e.g.,* horizontal position of nodes, edge curvature) are not.

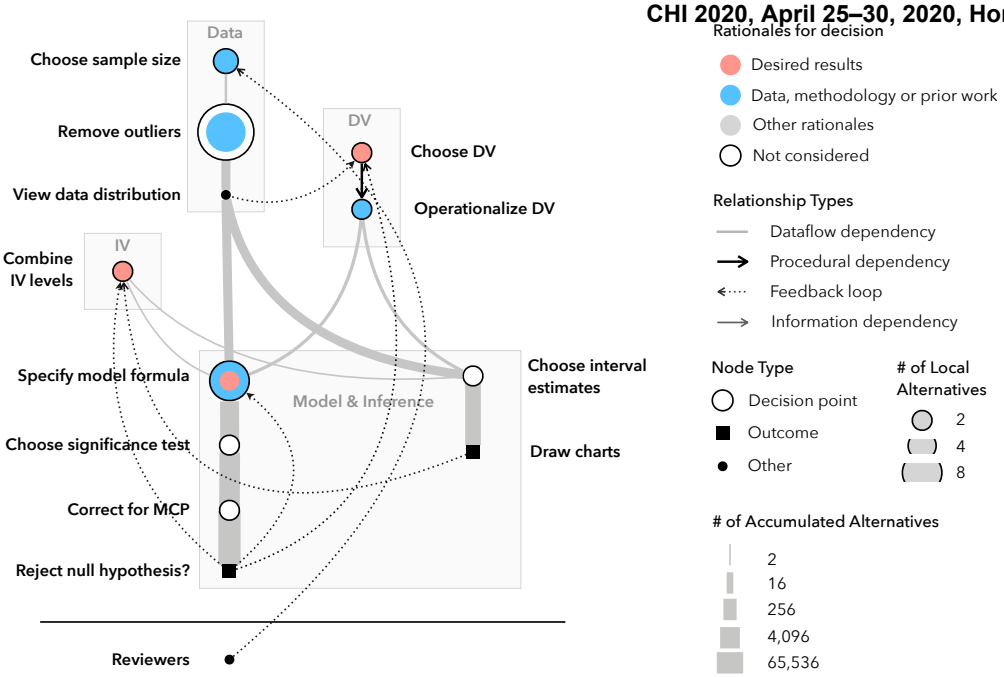


Figure 1: Analytic Decision Graph for P1, representing a controlled experiment to investigate the impact of web design on reading performance. At several steps, P1 revised her analytic decisions based on end results and reviewer feedback, for instance merging two levels of an IV because effect sizes were similar. While she examined model specification options thoroughly, she appeared to place less emphasis on inference decisions such as choosing which significance test to use.

ADG Walkthrough for P1

P1 designed a controlled experiment to investigate the impact of web design on reading performance. She followed a typical confirmatory pipeline: she operationalized (*i.e.,* defined the measurements of) the variables germane to her research questions, collected and processed the data, built a statistical model, and interpreted the results, ultimately producing a bar chart of effect sizes with uncertainty intervals and several p-values.

The dataflow edges funnel into two linear paths leading to the end results, as opposed to a typical exploratory analysis (*e.g.,* Figure 2f) where the dataflow forks into multiple branches. Still, P1's analysis has many feedback loops: P1 revised her analytic decisions at several steps, based on observed data, end results, and reviewer feedback. Despite being a relatively simple pipeline with 9 decision points, P1's analysis gives rise to over 5,000 possible ways to compute the final p-values, as indicated by the width of the dataflow edge into the final node *reject null hypothesis*. Judging by the size and color of decision nodes, P1 examined model specification options thoroughly (indicated by the size of the *specify model formula* node), but she appeared to place less emphasis on *inference* decisions (indicated by empty nodes in the inference section).

Summary of ADG Patterns

Using the interpretation approach above, we analyzed ADGs for all participants. Here are a few recurring observations.

Feedback loops are present in all analysis processes of our participants, regardless of whether the analysis is confirmatory or exploratory (Figure 2, dotted edges). We further examine these iterative fine-tuning behaviors in the next section.

Participants often fixate on a few prominent steps while ignoring decisions in the end-to-end pipeline. Among our participants, we observe that data and inference decisions are often neglected (Figure 2, empty nodes). When prompted by the interview checklist or the interviewer, participants revealed that they did not recognize these steps as decision points and implicitly chose a single viable option. On the other hand, *choosing variables*, *choosing models* and *specifying model formula* are often considered thoroughly (Figure 2, large nodes).

Procedural branches are rare among our participants. P1's process includes one procedural edge and no procedural branches (Figure 1, thick black edges); she could have considered ways to operationalize other candidate dependent variables. The lack of such branches implies a relatively linear process where decisions were made in order, one step at a time.

Across participants, the "multiverse" size ranges from 16 to over 25,000,000 (median 1,632; see Figure 2, thickness of dataflow edges immediately before rectangular nodes). We revisit issues related to scale in the discussion section.

INTERVIEW RESULTS

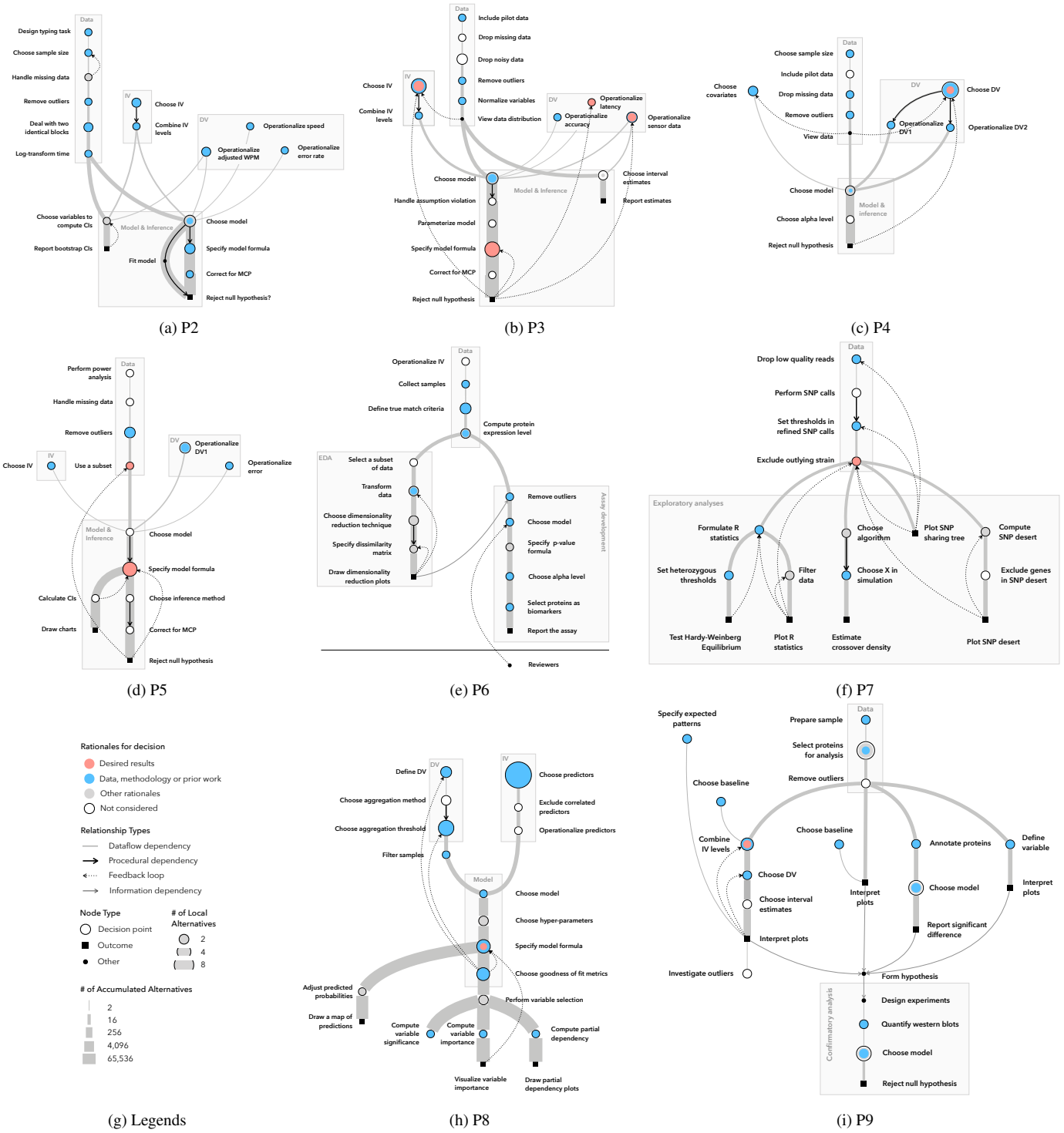
We now describe the patterns that emerged from the qualitative analysis of our interview data, following the organization of themes and categories in Table 1.

Rationales for Analytic Decisions

When participants recognized an analytic step as a decision point, they might *reason* about it, identifying and evaluating options before selecting a path along which to proceed. From 190 such instances, we identified six categories of rationales for analytic decision making.

Methodology

Methodological concerns comprised a major set of rationales (48 instances, 9/9 participants). These arguments typically involved statistical validity, study design, and research scope.



(a) P2 (b) P3 (c) P4

(d) P5 (e) P6 (f) P7

(g) Legends (h) P8 (i) P9

Figure 2: Analytic Decision Graphs for P2–P9.

Many methodological concerns (19 instances, 9/9 participants) were rooted in statistical validity. Meeting model assumptions was a concern for seven participants, as they chose the statistical model best suited for the data distribution, or wrangled the inputs to satisfy model assumptions. Participants used various strategies for the later approach: they might balance the datasets, normalize the inputs, log-transform a variable, or remove collinear variables. Besides model assumptions, five participants supported their decision with logical arguments, pointing out mathematical properties or explaining the intuitions behind customized methods. As a simpler example, P6 explained why she used a less common log-transformation, $log(x+1)$, to process the data: *"because I have a lot of zeros."* (Figure 2e, transform data).

Validity concerns also stem from study design (21 instances, 9/9 participants). Five participants argued that confounders were controlled for and thus were excluded in model specifications. Four participants stressed that variables in their models strictly followed the factors, levels, and measures in their experimental design. Participants followed a preselected plan akin to pre-registration [10, 56, 58], though none of the studies was officially pre-registered.

Other than validity, a few rationales (8 instances, 2/9 participants) are rooted in scope, as researchers discarded alternatives outside the scope of their current research questions. As P2 argued (Figure 2a, design typing task): *"the intention of this research is to evaluate text entry, the real-life text entry. So we'll not type random text."*

Prior Work

Another group of rationales were anchored in prior work (62 instances, 9/9 participants). Here, we use *prior work* to refer to prior studies, standard practices, and internalized knowledge.

All participants cited prior studies to support their decisions. Besides utilizing knowledge from prior studies to inform decisions, three participants mimicked configurations from a prior work. While this enables direct comparison with previous findings, participants might admit that alternatives warranting further considerations might exist. For example, P8 stated (Figure 2h, choose goodness of fit metrics):

> *"... [the chosen method] is what multiple other papers have used. But there would be alternatives and we have a whole host of other model performance metrics."*

Without citing specific sources, seven participants drew on knowledge that likely resulted from a combination of prior studies, consensus, and training. A participant referred to field consensus in outlier removal: *"We did not remove any outliers. Because in the autism field, why it's called Autism Spectrum Disorder, because other Autism are considered outliers."*

Six participants in 22 instances honored *"standard practice"*, *"tradition"*, and *"convention"*, sometimes without questioning its validity. For instance, a participant followed a *"rule-of-thumb"* of recruiting $\sim$20 participants for an experiment, though the study might be under-powered and so fail to resolve effects of smaller size (Figure 2a, choose sample size). A participant chose to *"start with a t-test, because it's standard"* (Figure 2b, choose model), though the data violated normality assumptions. Two participants admitted that standard practices might not be best practices, but they were concerned about social aspects. They believed that readers would accept standard practice more readily and *"reviewers would have asked for it."*

Five participants expressed how the lack of theory prevented them from choosing statistically valid alternatives. Two participants avoided interaction patterns that they *"didn't have a strong hypothesis to include"* (Figure 2d, specify model formula). P2 explained how tweaking alpha, a parameter in a metric to operationalize a variable, might allow one to obtain desirable outcomes, and argued against such practices because *"there's no reasonable theory or rationale underlying that alpha."*(Figure 2a, operationalize adjusted WPM).

Data

Data constraints represented another major group of rationales (39 instances, 8/9 participants). Researchers were constrained by data availability, quality, and size.

Some data constraints were hard constraints. Unavailable data might prevent participants from investigating additional variables. P8 originally identified 23 relevant predictors from prior work, but later dropped 7 of them for which he was unable to obtain sufficient data (Figure 2h, choose predictors). Three participants stated that collecting more data was too costly or infeasible, as P4 complained: *"we set the target beforehand, but we couldn't achieve the target group. We just tried to recruit as many groups as possible."* (Figure 2c, choose sample size). Sticking with a small sample size, two participants noted that certain modeling approaches, for instance time series analysis, were infeasible.

On the other hand, some data issues allowed more room for flexibility. What constituted clean data might be subjective, but three participants excluded noisy data at the expense of study design, for instance dropping an entire variable. Similarly, three participants altered study designs, such as pooling variable levels, in order to achieve a larger sample size.

Expertise

Researchers also felt limited by expertise (23 instances, 8/9 participants). They might not know what alternatives were possible, as P9 commented (Figure 2i, choose model):

> *"And there is almost certainly some other way to do that, but I'm not sure that I would know what it is."*

When researchers had a rough notion of viable alternatives, they opted not to pursue an unfamiliar method. P4 echoed sentiments of three participants about Bayesian analysis: *"I heard something about Bayesian statistics, but I don't have any background to try more than that."* (Figure 2c, choose model). Two researchers deferred a decision to a statistician, who they believed had better authority over the subject.

Communication

Sometimes researchers preferred an alternative that was easier to communicate (13 instances, 6/9 participants), quoting a variety of values. Two participants preferred an *"interpretable"* method over *"methods that merely produce black-box predictions"* (Figure 2h, choose model). A participant attempted to be *"consistent"* with the methods he used, because *"otherwise*

the readers will be confused" (Figure 2b, choose model). Another participant aimed for higher *generalizability* by targeting for practical use cases (Figure 2a, design typing task). Finally, a participant just wanted to keep things *simple*, avoiding *"more complex"* options (Figure 2a, choose IV). Communication concerns can come at the expense of validity. P3 chose a statistical model suboptimal for their data distribution because *"to make the analysis consistent across the whole study, we just stick with one statistical test."* (Figure 2b, choose model).

Sensitivity

Finally, researchers sometimes claimed that choosing another alternative would have little impact on the results (5 instances, 4/9 participants). Two researchers supported the claim with logic. P8 said: *"in my quick mental calculation, it seemed like it wouldn't actually make a big difference."* (Figure 2h, adjust predicted probabilities). Others recalled from past experience that two methods tended to produce similar results. As they did not evaluate their current situation, perceived sensitivity might differ from actual sensitivity.

Interactions of Rationales

We observed an interplay between decision rationales, particularly in terms of which rationales tended to dominate others. Both our own interviews and previous studies [52, 56] identify *methodology* and *prior work* as dominant rationales that researchers primarily rely upon. A bottom-up, exploratory approach might include *data* as a dominant rationale category, as researchers develop tentative theories to account for observed phenomena. However, in practical situations, the analysis plan supported by the dominant rationales nevertheless accommodates various constraints concerning *data*, *expertise* and *communication*. *Sensitivity* ignores other rationales by focusing instead on the impacts of the decision.

These decision rationales interact with each other, often creating conflicts. The previous section described two ways in which the dominant categories, *methodology* and *prior work*, are contradictory. First, standard practices are not always best practices; by adhering to conventions, participants might adopt a statistically faulty method. Second, a statistically valid approach might lack theoretical support, as five participants described how they avoided such situations. The previous section also contains ample evidence of how secondary rationales constrain and override dominant concerns. *Data*, *expertise* and *communication* all limit the viable methods researchers choose from, as researchers prefer a method that is familiar, easy to communicate, and feasible for the current data size. *Data*-related issues also impact study design, for instance researchers might drop a noisy variable or combine multiple levels within a variable to increase sample size.

Motivations for Executing Alternative Analyses

While some researchers *reasoned* about alternatives, ruled out options, and implemented a single final decision, others *executed* alternative analyses. What spurred researchers to actualize possibilities and travel multiple analytic paths? We found 44 instances in which participants explicitly described, or we could reasonably infer, their motivations to pursue alternatives. We then identified four categories of motivations.

Opportunism

When being opportunistic, researchers willingly explored new alternatives, searching for desired results in the garden of forking paths (20 instances, 7/9 participants). Such exploratory behavior comes in two forms: one might search for patterns without a hypothesis to defend, or one might actively search for a confirmation of existing hypothesis. The first form is sensible as long as the exploratory nature is clearly acknowledged in the publications [56, 58]. In fact, exploratory data analysis (EDA) literature often advocates an open mindset and a comprehensive exploration before focusing on pre-defined questions [1, 4]. Participants doing EDA all demonstrated an opportunistic attitude, as P8 described:

> *"It was like a little experiment . . . It wasn't to test any hypothesis, but it was to explore the data in a more complete way where we could actually investigate the effects that we were interested in."*

However, we also observed opportunism among participants who reported strictly confirmatory findings (Figure 1 & 2a-d, feedback loops into red nodes). Participants tried multiple analytic options and selected a path leading to desired results. Such endeavors might happen in the data wrangling phase, when participants qualitatively explored data distributions and avoided analytic options unlikely to produce desired outcomes. P1 discarded a dependent variable because it failed to yield differential results across conditions (Figure 1, choose DV):

> *"The distributions of accuracy are similar across questions. So, instead of looking at how different conditions affect it, we use [accuracy] as another exclusion criteria."*

Others adopted a deliberate and structured search. P3 tried *"all the different combinations"* of independent variables in a model specification (Figure 2b, specify model formula):

> *"You can think of it as a cross product, we did all of them, right? . . . we have ANOVA to test the difference of accuracy with and without considering age, and with and without considering gender, and with considering both gender and age. We did all of them."*

After an exhaustive search for patterns, he selectively reported *"interesting findings."* These examples of opportunism in confirmatory analysis might increase the chance of false discovery and lead to non-replicable conclusions [23, 41, 51].

Systematicity

Voluntary exploration was not always driven by a desire to find interesting results. Researchers could *systematically* enumerate reasonable alternatives, implement them, and evaluate the outcomes based on an objective metric (4 instances, 3/9 participants). The key evidence to help us distinguish *systematicity* from *opportunism* was that the evaluation metric did not hinge on anticipated conclusions; the metric was not the end result. Two participants enumerated model specifications and chose the best one based on the goodness of fit. P8 also ran a local multiverse analysis and used the goodness of fit to choose the best combination of two decisions (Figure 2h).

Robustness

In another type of voluntary exploration, researchers tested alternatives *after* making a decision to gauge the *robustness* of

the outcomes (7 instances, 4/9 participants). After the model yielded expected results, P2 implemented two redundant tests *"to gain an inner confidence of the metric"* (Figure 2a, choose model). P9 applied two protein annotation methods to corroborate the same conclusion (Figure 2i, annotate proteins):

> *"That is just for robustness, to say, 'Hey, even if you look at orthologs of proteins that in mammals and so on are EV proteins, you see the same thing."'*

Contingency

In the case of contingency, researchers had no choice. They had to deviate from their original plans because the planned path proved to be erroneous or infeasible (13 instances, 5/9 participants). Contingency might arise internally, as five participants ran into a dead end and retracted to an upstream analytic step. At a filtering step, P7 initially set loose thresholds because *"having more data was probably better"*, but the decision backfired (Figure 2f, drop low quality reads):

> *"But two years into the project, it was realized that this [filter] produced some very anomalous results, and we went back, and for some of the subsequent analysis we went through a more stringent filtering of the data which removed some of these anomalies."*

External contingency came from reviewers, who urged researchers to revise the analysis. P6 switched to a Fisher's exact test from a t-test: *"well, the reviewer made me do it, but I'm not sure it's the best choice."* (Figure 2e, choose model).

Motivations for Selective Reporting

After researchers executed alternative analytic paths and observed multiple outcomes, they must choose which analyses to include in publications. We observed 52 instances in which researchers did not report all analytic paths taken. Why did researchers report some findings but omit others? We identified four categories of motivations underlying selective reporting.

Desired Results

Evaluating multiple options allowed researchers to view and weigh the outcomes. Unsurprisingly, the quality of the outcome was a major criterion in selecting which alternative to report. In opportunistic exploration, researchers searched the garden of forking paths for desired results; consequently, they typically only reported the desired results and omitted findings that were non-significant, uninteresting, or incoherent to the theory they intended to support (15 instances, 7/9 participants).

A majority of participants conducting confirmatory analysis (4/5) omitted statistically non-significant results. When multiple results proved significant, participants selected the option with stronger implications for their intended theory. P5 tested two ways to filter the data and both produced significant results, so she chose the larger subset such that she could argue for a greater impact of the proposed mechanism (Figure 2d, use a subset). Two participants included non-significant results and devised further criteria for "interesting" findings worthy of reporting. To P3, interesting findings meant all significant results plus unexpected null results *"which we thought it might be significant but it turns out not."* He truthfully documented initially plausible hypotheses that failed an empirical test, yet his reporting strategy also includes any hypothesis that seemed plausible post-hoc – which is a form of HARKing [33].

Two participants conducting EDA also omitted explored analysis paths that did not corroborate the conclusions. Only one participant comprehensively documented alternative analyses they had performed during exploration.

Similar Results

In a few cases (5 instances, 3/9 participants), researchers relied on analytic outcomes, but argued that the outcomes were similar in terms of both the actual results and their implications. Thus, reporting one of the alternatives was deemed sufficient. Participants did not elaborate any criteria for selecting among similar options, implying that sensitivity alone was the reason for suppressing interchangeable analysis alternatives.

Correctness

Despite having access to the analytic outcomes, sometimes participants did not utilize this information. Instead, they fell back to using rationales described in the *decision rationales* theme, most frequently *methodology* and *prior work*, to remove analytic approaches they considered incorrect (16 instances, 7/9 participants). Such practices might ensue from an exploration out of contingency or robustness. For example, researchers switched to an alternative method requested by reviewers, omitting the original, presumably flawed, method. However, sometimes the motivation for exploring alternatives was unclear and we do not know whether the correctness argument was formed before or after seeing the results. The latter scenario, namely coming up with post-hoc explanations for *desired results*, is precisely HARKing [33].

Social Constraints

Finally, social constraints could prevent participants from reporting certain findings (16 instances, 7/9 participants). Colleagues and reviewers might disapprove of particular analysis methods. P2 did not report his experimental code on Bayesian analysis because his *"colleagues don't seem to favor that"* (Figure 2a, choose model). P8 similarly complained that he did not have full control over reporting:

> *"I'm a second author and many decisions made in the publication, in the manuscript writing, and figure making were decisions against my wishes."*

Two participants mentioned that reporting every detail would exceed the page limit. In response, P3 deleted the alternative taking up more space and P2 removed a finding perceived by the authors to be *"not of interest."*

Researchers might voluntarily cater to communicative concerns to make figures and manuscripts easier to understand. Two participants applied additional filtering to a visualization to reduce over-plotting; they omitted the original plot and parameters. Another two participants removed analysis methods unfamiliar to the audience. P2 stated that describing Bayesian analysis in an accessible way would be too much work, and P9 simply claimed that a method would confuse readers.

DISCUSSION

In this work, we pored over nine published studies and interviewed the authors to discuss analytic decisions in the end-

to-end quantitative data analysis. We presented common rationales for analytic decisions and discussed how researchers trade off between options. We observed various reasons for exploring alternatives and selectively reporting results. We also introduced Analytic Decision Graphs and discussed recurring patterns along analysis processes. Together, these results help us better understand current practices in the midst of the replicability crisis and how we might start to revise them. Below, we discuss design opportunities for supporting users in making and communicating analytic decisions.

Analysis Diagramming & Provenance Tracking

In many instances our respondents were limited in coming up with alternatives: they might fail to recognize an analytic step as a decision (*e.g.,* following default settings), adopt a single option without considering alternatives (*e.g.,* making the same decision as a previous study), or overlook possible alternatives due to *expertise*. A corresponding avenue for future research concerns analysis *linters* or *recommenders*, in which tools flag potentially problematic practices (such as the feedback loops observed in our interviews), recommend alternative methods, or even automatically suggest a preferred method based on statistical validity [8, 30, 57]. One strategy for such tools is to enable higher-level specifications of analysis goals (*e.g.,* specifying annotated model inputs and outputs rather than explicit test types or formulae), from which appropriate analysis methods might be synthesized in conjunction with the data [30]. Another strategy is to leverage the abundance of online analysis code [49] to mine patterns of decisions and alternatives, which might be useful for building automatic recommenders.

In some cases, our respondents evaluated multiple alternatives and then engaged in selective reporting. Integrating diagramming methods with provenance tracking could provide some level of automated documentation, for example by analyzing executed code paths to model and visualize the various alternatives that were explored (*c.f.* [34]). Similar elicitation and tracking strategies have also been suggested for reducing false discovery during exploratory visualization [61].

Even with complete documentation of analysis history, hindsight bias might lead researchers to unintentionally misremember *post hoc* explanations developed after conducting analysis as motivating *a priori* hypotheses [33]. Tools for mapping analysis decisions might promote more comprehensive assessment *a priori*. By instantiating decision points and providing analytic checklists [59], analysis tools might do more to promote *planning*, not just *implementation*. For example, an analysis team might manually author, annotate, and debate an analytic decision graph and corresponding rationales *a priori*. The results could then document and aid communication of decision points and rationales. Overviews of the end-to-end analysis process could also guide implementation work, for example with decision graph nodes linked to corresponding analysis code snippets (*i.e.,* cells in a computational notebook).

Multiverse Specification & Analysis

While the above methods focus on documenting decisions and selecting a preferred path, many "reasonable" alternatives may exist. Proponents of *multiverse analysis* [52, 54] have argued for preserving such decisions and evaluating them collectively. However, the design and evaluation of tools for both specifying and evaluating multiverse analyses remains an open challenge.

Authoring a multiverse analysis may be tedious, as analysts have to write scripts to manually execute all possible combinations of reasonable alternatives. Future tools could provide better scaffolding for defining decision points and procedural branches without devolving into a morass of multiple, largely redundant analysis scripts [26]. Inspiration might be taken from design tools for parallel prototyping [27, 55].

Second, interpreting the outcomes of a vast number of analyses is difficult. Visualizations that juxtapose or animate individual outcomes [15] may not scale, and may fail to accurately convey the relative sensitivity of decision points. In addition, some of our participants bypassed decision making if they perceived the *sensitivity* to be low; they did not always verify if the decision indeed had limited influence on the results. Future tools might aggregate subsets of outcomes, and quantify the end-to-end statistical variance via a meta-analysis of multiverse results [44, 60]. Multiverse analysis tools might assess sensitivity across decision points and identify high-impact decisions for further consideration.

Finally, multiverse analysis also poses a number of underlying systems challenges. How might one optimize multiverse evaluation, for example by efficiently reusing shared computation across "universes," or by using adaptive sampling methods to more efficiently explore a parameter space?

Sociotechnical Concerns

While new analysis tools might help improve systematic consideration and communication of analysis alternatives, they must operate within an accepting social environment. We are hardly the first to note that the urges to "tell a good story," sidestep unfamiliar methods, and appease reviewers can undermine a full and accurate accounting of one's research [33], and our interviews confirm their persistence. If publication incentives and reviewer criteria remain unchanged, a provenance tracking tool that reveals problematic choices, or multiverse tools that produce more comprehensive yet more complex and unfamiliar outputs, may be abandoned in favor of the status quo. Accordingly, improving the reliability of end-to-end analysis must also be a community priority, ranging from the standards and practices of peer review to how we educate researchers, new and old. We hope that the decision making and selective reporting rationales identified in our interview analysis provide useful insights for the design of both improved analysis tools *and* community processes.

ACKNOWLEDGEMENTS

We thank our participants, the anonymous reviewers (especially the shepherd), Alex Kale, Eunice Jun, Rene Just, Tongshuang Wu, and IDL members for their help. This work was supported by a Moore Foundation Data-Driven Discovery Investigator Award and NSF Award 1901386.

SUPPLEMENTAL MATERIAL

Additional supporting information on methods and graphs may be found at https://osf.io/m5cph/.

REFERENCES

[1] Sara Alspaugh, Nava Zokaei, Andrea Liu, Cindy Jin, and Marti A. Hearst. 2019. Futzing and moseying: Interviews with professional data analysts on exploration practices. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2019), 22–31. DOI: http://dx.doi.org/10.1109/TVCG.2018.2865040

[2] David R. Anderson, William A. Link, Douglas H. Johnson, and Kenneth P. Burnham. 2001. Suggestions for presenting the results of data analyses. *The Journal of Wildlife Management* 65, 3 (2001), 373–378. DOI: http://dx.doi.org/10.2307/3803088

[3] Monya Baker. 2016. 1,500 scientists lift the lid on reproducibility. *Nature* 533, 7604 (2016), 452–454. DOI: http://dx.doi.org/10.1038/533452a

[4] Leilani Battle and Jeffrey Heer. 2019. Characterizing exploratory visual analysis: A literature review and evaluation of analytic provenance in tableau. *Computer Graphics Forum (Proc. EuroVis)* (2019). DOI: http://dx.doi.org/10.1111/cgf.13678

[5] C. Glenn Begley and Lee M. Ellis. 2012. Raise standards for preclinical cancer research. *Nature* 483, 7391 (2012), 531–533. DOI:http://dx.doi.org/10.1038/483531a

[6] Richard Border, Emma C. Johnson, Luke M. Evans, Andrew Smolen, Noah Berley, Patrick F. Sullivan, and Matthew C. Keller. 2019. No support for historical candidate gene or candidate gene-by-interaction hypotheses for major depression across multiple large samples. *American Journal of Psychiatry* 176, 5 (2019), 376–387. DOI: http://dx.doi.org/10.1176/appi.ajp.2018.18070881 PMID: 30845820.

[7] Steven P. Callahan, Juliana Freire, Emanuele Santos, Carlos E. Scheidegger, Cláudio T. Silva, and Huy T. Vo. 2006. VisTrails: Visualization meets data management. In *Proc. ACM SIGMOD International Conference on Management of Data*. 745–747. DOI: http://dx.doi.org/10.1145/1142473.1142574

[8] Dylan Cashman, Shah Rukh Humayoun, Florian Heimerl, Kendall Park, Subhajit Das, John R. Thompson, Bahador Saket, Abigail Mosca, John Stasko, Alex Endert, Michael Gleicher, and Remco Chang. 2019. A user-based visual analytics workflow for exploratory model analysis. *Computer Graphics Forum (Proc. EuroVis)* (2019). DOI: http://dx.doi.org/10.1111/cgf.13681

[9] Kwok Cheung and Jane Hunter. 2006. Provenance explorer – customized provenance views using semantic inferencing. In *International Semantic Web Conference*. 215–227. DOI:http://dx.doi.org/10.1007/11926078_16

[10] Andy Cockburn, Carl Gutwin, and Alan Dix. 2018. HARK no more: On the preregistration of CHI Experiments. In *Proc. ACM Human Factors in Computing Systems*. 141:1–141:12. DOI: http://dx.doi.org/10.1145/3173574.3173715

[11] John W. Creswell and Cheryl N. Poth. 2018. *Qualitative inquiry and research design: Choosing among five approaches*. SAGE publications.

[12] Robert A. Cribbie. 2017. Multiplicity control, school uniforms, and other perplexing debates. *Canadian Journal of Behavioural Science* 49, 3 (2017), 159. DOI: http://dx.doi.org/10.1037/cbs0000075

[13] Geoff Cumming and Robert Calin-Jageman. 2016. *Introduction to the new statistics: Estimation, open science, and beyond* (1 ed.). Routledge. DOI: http://dx.doi.org/10.4324/9781315708607

[14] Pierre Dragicevic. 2016. Fair statistical communication in HCI. In *Modern Statistical Methods for HCI*. Springer, 291–330. DOI: http://dx.doi.org/10.1007/978-3-319-26633-6_13

[15] Pierre Dragicevic, Yvonne Jansen, Abhraneel Sarma, Matthew Kay, and Fanny Chevalier. 2019. Increasing the transparency of research papers with explorable multiverse analyses. In *Proc. ACM Human Factors in Computing Systems*. 65:1–65:15. DOI: http://dx.doi.org/10.1145/3290605.3300295

[16] Alexander Eiselmayer, Chat Wacharamanotham, Michel Beaudouin-Lafon, and Wendy E. Mackay. 2019. Touchstone2: An interactive environment for exploring trade-offs in HCI experiment design. In *Proc. ACM Human Factors in Computing Systems*. 217:1–217:11. DOI:http://dx.doi.org/10.1145/3290605.3300447

[17] Sebastian S. Feger, Sünje Dallmeier-Tiessen, Albrecht Schmidt, and Paweł W. Woźniak. 2019a. Designing for reproducibility: A qualitative study of challenges and opportunities in high energy physics. In *Proc. ACM Human Factors in Computing Systems*. 455:1–455:14. DOI:http://dx.doi.org/10.1145/3290605.3300685

[18] Sebastian S. Feger, Sünje Dallmeier-Tiessen, Paweł W. Woźniak, and Albrecht Schmidt. 2019b. Gamification in science: A study of requirements in the context of reproducible research. In *Proc. ACM Human Factors in Computing Systems*. 460:1–460:14. DOI: http://dx.doi.org/10.1145/3290605.3300690

[19] Wolfgang Forstmeier and Holger Schielzeth. 2011. Cryptic multiple hypotheses testing in linear models: Overestimated effect sizes and the winner's curse. *Behavioral Ecology and Sociobiology* 65, 1 (2011), 47–55. DOI: http://dx.doi.org/10.1007/s00265-010-1038-5

[20] Wolfgang Forstmeier, Eric-Jan Wagenmakers, and Timothy H. Parker. 2017. Detecting and avoiding likely false-positive findings — A practical guide. *Biological Reviews* 92, 4 (2017), 1941–1968. DOI: http://dx.doi.org/10.1111/brv.12315

[21] Juliana Freire, David Koop, Emanuele Santos, and Cláudio T Silva. 2008. Provenance for computational tasks: A survey. *Computing in Science & Engineering* 10, 3 (2008), 11–21. DOI: http://dx.doi.org/10.1109/MCSE.2008.79

[22] Andrew Gelman, Jennifer Hill, and Masanao Yajima. 2012. Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness* 5, 2 (2012), 189–211. DOI: http://dx.doi.org/10.1080/19345747.2011.618213

[23] Andrew Gelman and Eric Loken. 2013. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University* (2013).

[24] Andrew Gelman and Eric Loken. 2014. The statistical crisis in science. *American Scientist* 102, 6 (2014), 460. DOI:http://dx.doi.org/10.1511/2014.111.460

[25] Greg Guest, Arwen Bunce, and Laura Johnson. 2006. How many interviews are enough? An experiment with data saturation and variability. *Field methods* 18, 1 (2006), 59–82. DOI: http://dx.doi.org/10.1177/1525822X05279903

[26] Philip J Guo and Margo I Seltzer. 2012. BURRITO: Wrapping your lab notebook in computational infrastructure. In *USENIX Workshop on the Theory and Practice of Provenance*. https://www.usenix.org/conference/tapp12/workshop-program/presentation/Guo

[27] Björn Hartmann, Loren Yu, Abel Allison, Yeonsoo Yang, and Scott R. Klemmer. 2008. Design as exploration: Creating interface alternatives through parallel authoring and runtime tuning. In *Proc. ACM User Interface Software and Technology*. 91–100. DOI: http://dx.doi.org/10.1145/1449715.1449732

[28] Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. 2018. Deep reinforcement learning that matters. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[29] Leslie K. John, George Loewenstein, and Drazen Prelec. 2012. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science* 23, 5 (2012), 524–532. DOI: http://dx.doi.org/10.1177/0956797611430953

[30] Eunice Jun, Maureen Daum, Jared Roesch, Sarah E. Chasins, Emery D. Berger, René Just, and Katharina Reinecke. 2019. Tea: A high-level language and runtime system for automating statistical analysis. *CoRR* abs/1904.05387 (2019). http://arxiv.org/abs/1904.05387

[31] Alex Kale, Matthew Kay, and Jessica Hullman. 2019. Decision-making under uncertainty in research synthesis: Designing for the garden of forking paths. In *Proc. ACM Human Factors in Computing Systems*. 202:1–202:14. DOI: http://dx.doi.org/10.1145/3290605.3300432

[32] Matthew Kay, Gregory L. Nelson, and Eric B. Hekler. 2016. Researcher-centered design of statistics: Why Bayesian statistics better fit the culture and incentives of HCI. In *Proc. ACM Human Factors in Computing Systems*. 4521–4532. DOI: http://dx.doi.org/10.1145/2858036.2858465

[33] Norbert L. Kerr. 1998. HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review* 2, 3 (1998), 196–217. DOI: http://dx.doi.org/10.1207/s15327957pspr0203_4

[34] Mary B. Kery, Bonnie E. John, Patrick O'Flaherty, Amber Horvath, and Brad A. Myers. 2019. Towards effective foraging by data scientists to find past analysis choices. In *Proc. ACM Human Factors in Computing Systems*. 92:1–92:13. DOI: http://dx.doi.org/10.1145/3290605.3300322

[35] Mary B. Kery and Brad A. Myers. 2018. Interactions for untangling messy history in a computational notebook. In *2018 IEEE Symposium on Visual Languages and Human-Centric Computing*. 147–155. DOI: http://dx.doi.org/10.1109/VLHCC.2018.8506576

[36] Jiali Liu, Nadia Boukhelifa, and James R. Eagan. 2019. Understanding the role of alternatives in data analysis practices. *IEEE Transactions on Visualization and Computer Graphics* (2019), 1–1. DOI: http://dx.doi.org/10.1109/TVCG.2019.2934593

[37] Bertram Ludäscher, Ilkay Altintas, Chad Berkley, Dan Higgins, Efrat Jaeger, Matthew Jones, Edward A. Lee, Jing Tao, and Yang Zhao. 2006. Scientific workflow management and the Kepler system. *Concurrency and Computation: Practice and Experience* 18, 10 (2006), 1039–1065. DOI:http://dx.doi.org/10.1002/cpe.994

[38] Wendy E. Mackay, Caroline Appert, Michel Beaudouin-Lafon, Olivier Chapuis, Yangzhou Du, Jean-Daniel Fekete, and Yves Guiard. 2007. Touchstone: Exploratory design of experiments. In *Proc. ACM Human Factors in Computing Systems*. 1425–1434. DOI: http://dx.doi.org/10.1145/1240624.1240840

[39] Marcus R. Munafò, Brian A. Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie Du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware, and John P. A. Ioannidis. 2017. A manifesto for reproducible science. *Nature Human Behaviour* 1 (2017), 0021. DOI: http://dx.doi.org/10.1038/s41562-016-0021

[40] James D. Myers, Carmen M. Pancerella, Carina S. Lansing, Karen L. Schuchardt, Brett T. Didier, and Goble C. Ashish, N. 2003. Multi-scale science: Supporting emerging practice with semantically derived provenance. *International Semantic Web Conference Workshop: Semantic Web Technologies for Searching and Retrieving Scientific Data* (2003). https://www.osti.gov/biblio/15016920

[41] Leif D. Nelson, Joseph Simmons, and Uri Simonsohn. 2018. Psychology's renaissance. *Annual Review of Psychology* 69, 1 (2018), 511–534. DOI:http://dx.doi.org/10.1146/annurev-psych-122216-011836

[42] Tom Oinn, Matthew Addis, Justin Ferris, Darren Marvin, Martin Senger, Mark Greenwood, Tim Carver, Kevin Glover, Matthew R. Pocock, Anil Wipat, and Peter Li. 2004. Taverna: A tool for the composition and enactment of bioinformatics workflows. *Bioinformatics* 20, 17 (2004), 3045–3054. DOI: http://dx.doi.org/10.1093/bioinformatics/bth361

[43] Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349, 6251 (2015). DOI: http://dx.doi.org/10.1126/science.aac4716

[44] Chirag J. Patel, Belinda Burford, and John P. A. Ioannidis. 2015. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *Journal of Clinical Epidemiology* 68, 9 (2015), 1046–1058. DOI: http://dx.doi.org/10.1016/j.jclinepi.2015.05.029

[45] Joao Felipe Nicolaci Pimentel, Vanessa Braganholo, Leonardo Murta, and Juliana Freire. 2015. Collecting and analyzing provenance on interactive notebooks: When IPython meets noWorkflow. In *USENIX Workshop on the Theory and Practice of Provenance*. https://www.usenix.org/conference/tapp15/workshop-program/presentation/pimentel

[46] Florian Prinz, Thomas Schlange, and Khusru Asadullah. 2011. Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery* 10, 9 (2011), 712. DOI: http://dx.doi.org/10.1038/nrd3439-c1

[47] Xiaoying Pu and Matthew Kay. 2018. The garden of forking paths in visualization: A design space for reliable exploratory visual analytics. In *2018 IEEE Evaluation and Beyond - Methodological Approaches for Visualization (BELIV)*. 37–45. DOI: http://dx.doi.org/10.1109/BELIV.2018.8634103

[48] James R. Rae, Selin Gülgöz, Lily Durwood, Madeleine DeMeules, Riley Lowe, Gabrielle Lindquist, and Kristina R. Olson. 2019. Predicting early-childhood gender transitions. *Psychological Science* (2019). DOI: http://dx.doi.org/10.1177/0956797619830649

[49] Adam Rule, Aurélien Tabard, and James D. Hollan. 2018. Exploration and explanation in computational notebooks. In *Proc. ACM Human Factors in Computing Systems*. 32. DOI: http://dx.doi.org/10.1145/3173574.3173606

[50] Raphael Silberzahn, Eric Luis Uhlmann, Dan Martin, Pasquale Anselmi, Frederik Aust, Eli C Awtrey, Štěpán Bahník, Feng Bai, Colin Bannard, Evelina Bonnier, R. Carlsson, F. Cheung, G. Christensen, R. Clay, M. A. Craig, A. Dalla Rosa, L. Dam, M. H. Evans, I. Flores Cervantes, N. Fong, M. Gamez-Djokic, A. Glenz, S. Gordon-McKeon, T. J. Heaton, K. Hederos, M. Heene, A. J. Hofelich Mohr, F. Högden, K. Hui, M. Johannesson, J. Kalodimos, E. Kaszubowski, D. M. Kennedy, R. Lei, T. A. Lindsay, S. Liverani, C. R. Madan, D. Molden, E. Molleman, R. D. Morey, L. B. Mulder, B. R. Nijstad, N. G. Pope, B. Pope, J. M. Prenoveau, F. Rink, E. Robusto, H. Roderique, A. Sandberg, E. Schlüter, F. D. Schönbrodt, M. F. Sherman, S. A. Sommer, K. Sotak, S. Spain, C. Spörlein, T. Stafford, L. Stefanutti, S. Tauber, J. Ullrich, M. Vianello, E.-J. Wagenmakers, M. Witkowiak, S. Yoon, and B. A. Nosek. 2018. Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science* 1, 3 (2018), 337–356. DOI: http://dx.doi.org/10.1177/2515245917747646

[51] Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* 22, 11 (2011), 1359–1366. DOI: http://dx.doi.org/10.1177/0956797611417632

[52] Uri Simonsohn, Joseph P Simmons, and Leif D Nelson. 2015. Specification curve: Descriptive and inferential statistics on all reasonable specifications. (2015). DOI: http://dx.doi.org/10.2139/ssrn.2694998

[53] Transparent statistics in Human–Computer Interaction working group. 2019. Transparent statistics guidelines. (Feb 2019). DOI: http://dx.doi.org/10.5281/zenodo.1186169 (Available at https://transparentstats.github.io/guidelines).

[54] Sara Steegen, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. 2016. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science* 11, 5 (2016), 702–712. DOI: http://dx.doi.org/10.1177/1745691616658637

[55] Michael Terry, Elizabeth D. Mynatt, Kumiyo Nakakoji, and Yasuhiro Yamamoto. 2004. Variation in element and action: Supporting simultaneous development of alternative solutions. In *Proc. ACM Human Factors in Computing Systems*. 711–718. DOI: http://dx.doi.org/10.1145/985692.985782

[56] Anna E. van't Veer and Roger Giner-Sorolla. 2016. Pre-registration in social psychology – A discussion and suggested template. *Journal of Experimental Social Psychology* 67 (2016), 2–12. DOI: http://dx.doi.org/10.1016/j.jesp.2016.03.004

[57] Chat Wacharamanotham, Krishna Subramanian, Sarah Theres Völkel, and Jan Borchers. 2015. Statsplorer: Guiding novices in statistical analysis. In *Proc. ACM Human Factors in Computing Systems*. 2693–2702. DOI: http://dx.doi.org/10.1145/2702123.2702347

[58] Eric-Jan Wagenmakers, Ruud Wetzels, Denny Borsboom, Han L. J. van der Maas, and Rogier A. Kievit. 2012. An agenda for purely confirmatory research. *Perspectives on Psychological Science* 7, 6 (2012), 632–638. DOI: http://dx.doi.org/10.1177/1745691612463078

[59] Jelte M. Wicherts, Coosje L. S. Veldkamp, Hilde E. M. Augusteijn, Marjan Bakker, Robbie C. M. van Aert, and Marcel A. L. M. van Assen. 2016. Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology* 7 (2016), 1832. DOI: http://dx.doi.org/10.3389/fpsyg.2016.01832

[60] Cristobal Young and Katherine Holsteen. 2017. Model uncertainty and robustness: A computational framework for multimodel analysis. *Sociological Methods & Research* 46, 1 (2017), 3–40. DOI: http://dx.doi.org/10.1177/0049124115610347

[61] Emanuel Zgraggen, Zheguang Zhao, Robert Zeleznik, and Tim Kraska. 2018. Investigating the effect of the multiple comparisons problem in visual analysis. In *Proc. ACM Human Factors in Computing Systems*. 479:1–479:12. DOI: http://dx.doi.org/10.1145/3173574.3174053

[62] Jun Zhao, Chris Wroe, Carole Goble, Robert Stevens, Dennis Quan, and Mark Greenwood. 2004. Using semantic web technologies for representing e-science provenance. In *International Semantic Web Conference*. 92–106. DOI: http://dx.doi.org/10.1007/978-3-540-30475-3_8