Increasing the Transparency of Research Papers with Explorable Multiverse Analyses

Pierre Dragicevic, Yvonne Jansen, Abhraneel Sarma, Matthew Kay, Fanny Chevalier

Increasing the Transparency of Research Papers with Explorable Multiverse Analyses

Pierre Dragicevic
Inria
Orsay, France
pierre.dragicevic@inria.fr

Yvonne Jansen
CNRS – Sorbonne Université
Paris, France
yvonne.jansen@sorbonne-universite.fr

Abhraneel Sarma
University of Michigan
Ann Arbor, MI, USA
abhsarma@umich.edu

Matthew Kay
University of Michigan
Ann Arbor, MI, USA
mjskay@umich.edu

Fanny Chevalier
University of Toronto
Toronto, Canada
fanny@cs.toronto.edu

ABSTRACT

We present *explorable multiverse analysis reports*, a new approach to statistical reporting where readers of research papers can explore alternative analysis options by interacting with the paper itself. This approach draws from two recent ideas: *i) multiverse analysis*, a philosophy of statistical reporting where paper authors report the outcomes of many different statistical analyses in order to show how fragile or robust their findings are; and *ii) explorable explanations*, narratives that can be read as normal explanations but where the reader can also become active by dynamically changing some elements of the explanation. Based on five examples and a design space analysis, we show how combining those two ideas can complement existing reporting approaches and constitute a step towards more transparent research papers.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**.

KEYWORDS

Multiverse analysis, explorable explanation, statistics, transparent reporting, interactive documents.

CHI 2019, May 4–9, 2019, Glasgow, Scotland UK

ACM ISBN 978-1-4503-5970-2/19/05…$15.00
https://doi.org/10.1145/3290605.3300295

ACM Reference Format:
Pierre Dragicevic, Yvonne Jansen, Abhraneel Sarma, Matthew Kay, and Fanny Chevalier. 2019. Increasing the Transparency of Research Papers with Explorable Multiverse Analyses. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland UK*. ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3290605.3300295

1 INTRODUCTION

The recent replication crisis in psychology and other disciplines has dealt a blow to the credibility of human-subject research and prompted a movement of methodological reform- [70]. Much of this movement calls for more transparency in the way statistics are reported, so that findings become more trustworthy, more likely to be interpreted correctly, and easier to verify and replicate [29, 69, 72]. Concern for transparency in statistical reporting has spread to the HCI community, which has published several articles [24, 31, 58] and hosted several workshops [56, 57, 96] on the topic.

While much of the current discussions around transparent statistics in HCI focus on how the community can improve its practice, it has been suggested that HCI can do more than endorse and promote the transparent statistics movement—it can actively contribute to it by proposing novel user interfaces for better doing and better communicating statistics [31, 97]. In this article, we consider the research paper as a user interface, and seek to understand how we can enrich this user interface to better support and promote transparent statistics reporting.

While there are many ways a statistical report can lack transparency, a common and damaging form of opacity is *undisclosed flexibility* (see Figure 1a), i.e., not reporting the different options that have been tried during the analysis [85, 98], or the options that would have been chosen had the data been different [40]. Undisclosed flexibility is damaging because it substantially increases the chances of reporting erroneous findings, while being invisible to the reader.

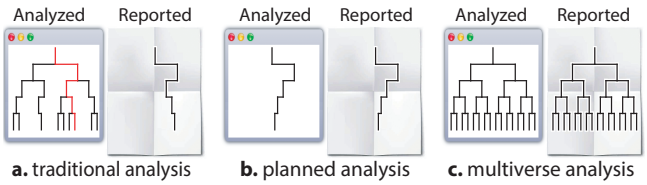

Figure 1: Three reporting strategies, from the least transparent to the most transparent: a) traditional analysis with undisclosed flexibility; b) planned analysis; c) multiverse analysis. Each branching represents a choice between different analysis options [71].

One response to the issue of undisclosed flexibility has been to encourage researchers to commit to a single statistical analysis that has been planned [17, 31] and ideally registered [24] ahead of time (Figure 1b). Although planning eliminates the problem of undisclosed flexibility, some statisticians and methodologists are starting to argue that more transparency can be achieved by letting researchers try many analyses and report all of them in their paper [86, 87] (Figure 1c). This is partly motivated by evidence that different researchers who analyze the same data will make different choices and will thus get slightly—and sometimes widely—different results [84].

In a *multiverse analysis*, researchers identify sets of defensible analysis choices (e.g., different ways of excluding outliers, different data transformations), implement them all, and then report the outcomes of all analyses resulting from all possible choice combinations. This approach increases transparency because readers can appreciate the "fragility or robustness of a claimed effect" [86] by checking whether the findings are dependent on arbitrary analysis choices.

Multiverse analysis promises an unprecedented level of transparency for research papers, but the idea is in its infancy. Writing papers with a multiverse analysis are difficult and there is currently very little guidance. An important part of the difficulty lies in reporting the outcomes of potentially hundreds or thousands of analyses in a single research paper (past examples contain between 120 and 1728 analyses [86, 87]), causing challenges for authors, reviewers, and readers. There are currently two ways to approach this problem.

A first option consists of sharing multiverse analyses as supplementary material, letting readers look under the hood of a default analysis and try alternative analysis options in a different environment. This approach has long been promoted by the *reproducible research* movement, and has been the subject of a vast body of work [21, 41, 73]. Although supplementary material is crucial for reproducibility and reuse, casual readers are very unlikely to engage with it. Thus proponents of multiverse analysis argue for acknowledging, reporting and discussing the multiverse in the research paper itself [86, 87].

A second option consists of summarizing the multiverse in the research paper itself using tables and custom plots. This is the approach put forward by the two papers introducing multiverse analysis [86, 87], which use as examples histograms of p-values [87], grids of p-values [87], and sorted dot-plots of point estimates where color encodes statistical significance [86]. Although such visualizations are very useful as overviews, they force the paper author to collapse the results of each analysis into a single point estimate, p-value or statistical significance verdict. In doing so, a lot of the richness and nuance of well-crafted statistical reports is lost.

A third approach, which we propose here, consists of making the results section of the research paper interactive. More specifically, we introduce and explore the concept of *explorable multiverse analysis report* (**EMAR**). An EMAR is a multiverse analysis report that is presented in the form of an *explorable explanation* in Bret Victor's sense [94]. Explorable explanations are explanations that "enable and encourage truly active reading" and allow "the reader to play with the author's assumptions and analyses, and see the consequences" [94]. At the same time, they can be read like normal prose: "the reader is not forced to interact in order to learn" [94]. Consistent with this design philosophy, an EMAR looks like a regular results section, but readers are given the possibility to change some of the analysis options and immediately see the results within the research paper itself.

We posit that EMARs can be a useful and compelling complement to existing multiverse reporting approaches. Our goal with the present paper is to provide a better understanding of the design space of possible approaches. We will explore this design space through five examples of short interactive papers[1] we have written to demonstrate the concept of explorable multiverse analysis report. We will discuss the trade-offs between different EMAR reporting strategies, as well as the benefits and challenges raised by EMARs compared to alternative reporting strategies such as the use of supplemental material or static multiverse analysis reports.

2 RELATED WORK

We regard the research paper as an interactive medium and focus on how this medium can support the communication of multiverse analyses. We review prior literature on interactive documents and interactive statistical reports, and discuss the state of academic publication practices.

Interactive Documents

We use "document" to refer to any information artifact that is constructed around a textual narrative. Since the invention of hypertext [16], the HCI community has never stopped

[1] available at https://explorablemultiverse.github.io/.

to explore how interactivity can be used to enhance documents, e.g., for supporting annotation tasks [81, 101] or non-linear navigation within document content. Fluid documents [102, 103], for example, allow for supplemental content such as definitions and details to be revealed in-place and on demand. Document Cards [89] operate the opposite way by summarizing content into a set of curated figures in order to produce concise views that facilitate browsing of document collections. Elastic documents explore linking of text and tables to generated contextual visualizations [10]. Finally, explorable explanations are highly-interactive documents for which there exists a proof-of-concept toolkit, Tangle [94], and a comprehensive toolkit, Idyll [25], that was just recently released.

In parallel with this stream of research, rich interactive documents have become prevalent on the web. News outlets now publish stories rich in animated figures and interactive graphs, providing the reader with a more engaging reading experience [44]. Greatly facilitated by the development of specialized web editorial tools (see [25] for a comprehensive review), this trend is spreading to the scientific sphere. For example, the Distill [1] platform specializes in the publication of machine learning articles with interactive figures.

These prior efforts illustrate the many ways interactivity can be used to enhance the reading experience. Inspired by this movement, we propose to add interactivity to the results section of research articles in order to support more complete and more transparent statistical reports.

Academic Publishing: A Long-Awaited Transition

Even though academic journals and conferences still heavily rely on static PDF documents, there has been efforts to move beyond them. The past decade has seen the introduction of many interactive publication concepts such as *semantic publication* [83], *rich interactive publication* [20] or *narrative interactive publication* [90]. Interactive enrichments include, for example, semantic markup of textual terms and structured document summaries [83], two-way linking between the article's narrative and underlying research data enabling generation of interactive tables and visualizations [74], citations in context where quotes from the original text are presented in a tooltip along with the full reference [83], and multimedia enhancements such as animated and interactive figures [11, 47, 95], interactive maps and timelines [20].

Some of these new forms of publication have been implemented by academic publishers. In particular, interactive exploration of 2D/3D scientific imagery and virtual volume rendering has been around for some time [6, 7, 68]. In order to better support research reproducibility, academic publishers have also been experimenting with various variations of Knuth's vision of literate programming [60], such as the concept of *executable paper* where the reader can re-execute snippets of code to re-compute figures of a research article in a side panel [61]. To date, Distill [1], mentioned previously, might be the scholarly journal that most fully embraced interactivity. Distill papers allow readers to delve into how computational models work by interactively manipulating their parameters.

This stream of work reveals a clear intent from the research community and publishers to push for more interactive publication media. However, the vast majority of the literature on the topic remain at a technical level, with very few concrete examples of interactive papers and virtually no discussion on how interactive papers should be designed.

Interactive Statistical Reports

It has long been suggested that people can better learn statistics if they are allowed to interact with parameters of statistical analyses and observe the results in real time [19, 27, 67]. Such interactive applications can be used for, e.g., demonstrating the central limit theorem [71] or showing how statistical analysis outcomes vary across replications [26]. Today, there is a proliferation of websites such as *Seeing theory* [63], *R Psychologist* [64] or *Setosa* [75], which employ interactive analyses to explain various statistical concepts. These applications differ from our work in two important respects. First, they are not documents as we define them, but user interfaces that combine plots with controls. Second, their purpose is to educate people about general statistical concepts through simulations, not to communicate the findings from a particular study.

Early on, Sawitski [80] advocated for statistical environments that let users mix textual narratives with interactive plots. In his example paper, all figures are linked to the same underlying statistical model and changes in one figure are reflected in all other figures. Aschwanden's web essay [9] is a compelling example of how interactivity can help readers appreciate the influence of analysis choices on outcomes. Embedded in the text is an interactive figure where the reader can dynamically exclude data and manipulate analysis parameters, and observe whether different choices result in a statistically significant outcome or not. However, this work still focuses on general statistical education rather than on how to communicate findings from real empirical data.

There has recently been a surge of work on extending statistical computing and graphing languages to support the authoring of interactive documents. R Shiny [5] allows authors to create statistical reports with dynamic plots and interactive controls. The R Markdown Gallery [3] features many compelling examples of such interactive statistical reports. Jupyter Notebook [2] is another popular environment, with a special focus on literate statistical reports [60, 77] that interleave code snippets with plots, tables and narratives.

All such environments could in principle be used to implement EMARs, with the proper API. In our work we chose to use generic web technologies instead (HTML, CSS and JavaScript), as our focus is on prototyping and exploring non-standard designs rather than developing a toolkit, which we see as the next step.

All these prior efforts demonstrate a real enthusiasm for approaches that offer readers the possibility of interacting with statistical reports. However, we do not know of any example that focuses on how to report a multiverse study analysis. Furthermore, interaction is typically limited to figures, while the text itself is static and non-interactive. We show how by building on Bret Victor's explorable explanations [94] we can explore a richer design space of interactive statistical reports.

3 EXAMPLES

We wrote five mini-papers (2–4 pages each) to explore the design space of EMARs. We refer to each of them by a code name (e.g., FREQUENTIST, LIKERT). All mini-papers are reanalyses of previously published studies for which data and R scripts were publicly available. They are self-contained and can be read online at https://explorablemultiverse.github.io or by clicking directly on a code name below. In addition to the interactions described below, all mini-papers include a toolbar allowing readers to: *i)* switch between a single-column HTML layout and a two-column ACM layout; *ii)* animate the paper by randomly drawing analyses; and *iii)* switch back to the analysis that was shown by default.

Example 1 – FREQUENTIST

The FREQUENTIST example [36] is a reanalysis of a CHI study evaluating physical visualizations [51]. It is meant to illustrate a few basic multiverse analysis ideas for a typical frequentist analysis with confidence intervals (CIs). The results of the analysis are initially identical to the original paper, including the two figures reporting mean task completion time per technique and pairwise comparisons, with 95% CIs. Four aspects of the analysis can be changed by the reader, which has the effect of immediately updating the two plots and some text elements such as explanations and figure captions. Changes are made by clicking or dragging the elements of the text in blue as in Bret Victor's explorable explanations [94] (see Figure 2).

First, horizontally dragging the "95%" text has the effect of changing the confidence level (7 levels are provided from 50% to 99.9%) and updating the length of error bars in the two figures. This allows the reader to appreciate that the 95% level is arbitrary [66] and thus that CIs should not be interpreted in a strictly dichotomous manner [29]. Meanwhile, readers who insist on interpreting effects as significant or non-significant have the option of changing the customary cutoff of α=.05

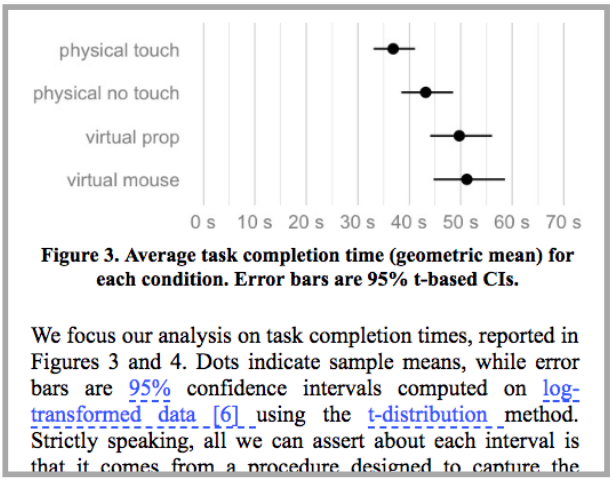

Figure 3. Average task completion time (geometric mean) for each condition. Error bars are 95% t-based CIs.

We focus our analysis on task completion times, reported in Figures 3 and 4. Dots indicate sample means, while error bars are 95% confidence intervals computed on log-transformed data [6] using the t-distribution method. Strictly speaking, all we can assert about each interval is that it comes from a procedure designed to capture the

Figure 2: Excerpt from the mini-paper FREQUENTIST, showing widgets embedded in the text in Bret Victor's [94] style. Operating a widget changes one aspect of the analysis and immediately updates the figure.

(95% CIs), for example to the α=.005 (99.5% CIs) criterion now advocated by some methodologists [15].

Clicking the "transformed data" text toggles the text to "untransformed data" and updates the two figures with results from the corresponding analysis. Although some researchers recommend that completion times be log-transformed [79], other researchers may be suspicious of, or unfamiliar with data transformations—this option reassures them that the results hold for untransformed data. Similarly, clicking on "t-distribution" switches the text to "BCa bootstrap" and shows the results of the analysis using non-parametric bootstrap CIs, which tend to be liberal (i.e., too narrow) with small samples but do not require distributional assumptions [59].

Finally, the plot with the three planned pairwise comparisons (not shown in Figure 2) shows uncorrected CIs, but the reader can apply a Bonferroni correction by clicking on the text "not corrected for multiplicity". Correction for multiplicity is strongly recommended by many but it is not without drawbacks: there is a controversial and complex literature on the topic [31]. To help the reader interpret the CIs correctly, the mini-paper contains a paragraph that gives the individual and the family-wise CI coverage and false positive rates, which are updated whenever Bonferroni correction is turned on or off, or whenever the confidence level is changed. More details can be found in the mini-paper itself [36].

The FREQUENTIST mini-paper covers a total of 7×2×2×2 = 56 unique analyses. The paper concludes that the findings from the original study (i.e., good evidence of a difference for the first two comparisons, inconclusive results for the third one) are reasonably robust, as they hold across the sub-multiverse where the confidence level is at 95% or less.

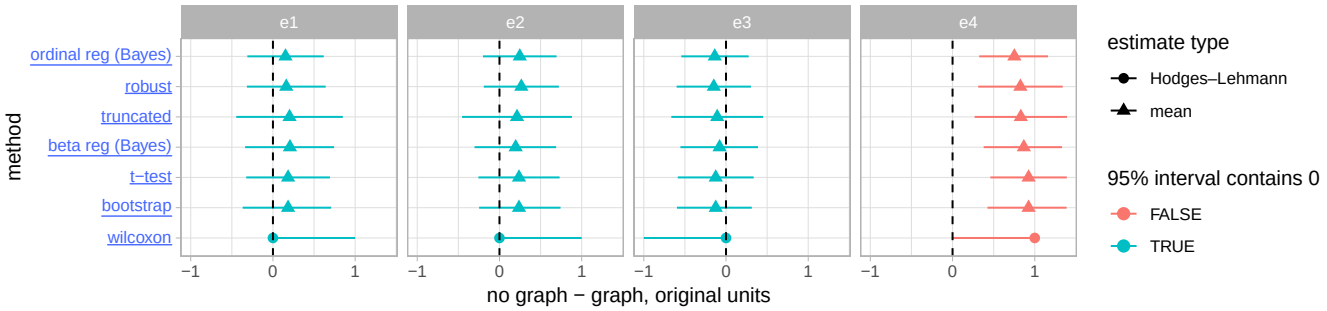

Figure 3: Plot from the mini-paper LIKERT, summarizing point estimates and 95% CIs for an effect measured across 4 different experiments (columns) and analyzed using 9 different methods (rows). Clicking on a row label updates the method section. Here no matter how the data are analyzed, no conclusive effect is found for the first three experiments (blue intervals), while there is convincing evidence for an effect in the fourth (red intervals).

As noted earlier, this and all other mini-papers below offer the option of animating the multiverse by repeatedly drawing an analysis at random. This feature gives an overview of the multiverse and permits the reader to observe which parts of the article change substantially across the multiverse.

Example 2 – LIKERT

The LIKERT example [34] is a reanalysis of a recent InfoVis study on the effect of charts on comprehension and persuasion [35]. It provides an example of multiple alternative analyses that differ substantially in their methodology. The dependent variable of interest is the response to a single Likert-type question, a type of data commonly collected in HCI and for which different analysis approaches have been proposed in the field [54, 99]. As there is currently no consensus on how to best analyze such data, any single analysis method is unlikely to convince all readers and reviewers. This issue is easily addressed with a multiverse analysis.

The LIKERT mini-paper reanalyzes the four experiments in the original InfoVis study [35] using nine different methods covering a broad range of approaches, including parametric vs. non-parametric and frequentist vs. Bayesian. In contrast with the previous mini-paper, all analysis outcomes are summarized in a static overview figure to facilitate comparison. Seven of the nine methods yield simple effect sizes (e.g., mean differences) which are summarized in the plot shown in Figure 3, while the remaining two methods yield log-odds ratios, reported in a different plot (not shown here). By default, the method section in the mini-paper only details the bootstrap method, which was used in the original study. However, clicking on a row label in the figure changes the method section to provide a description and justification of the selected method, an interpretation of its results, and the *p*-value for the fourth experiment (when available).

The LIKERT mini-paper covers a total of 9 unique analyses. It concludes that the results are consistent across analyses:

Fertility

The classification of women into a high or low fertility group based on cycle day can be done in several ways:

☒ Participants with cycle days ranging from 7 to 14 are assigned to the high fertility group, whereas participants with cycle days ranging from 17 to 25 are assigned to the low fertility group [2],

☐ days 6–14 are used for high fertility, whereas days 17–27 are used for low fertility [4],

☐ days 9–17 for high fertility and 18–25 for low fertility [5],

☐ days 8–14 for high fertility and 1–7 and 15–28 for low fertility [6], and

☐ days 9–17 for high fertility and 1–8 and 18–28 for low fertility [7].

Figure 4: Excerpt from the mini-paper DATAVERSE, listing five different ways of dichotomizing a dependent variable. Elsewhere in the mini-paper, an interaction plot gets updated each time an option is chosen.

no matter how the Likert data are analyzed, no conclusive effect is found for the first three experiments (blue intervals in Figure 3), while there is convincing evidence for an effect in the fourth (red intervals). The results differ slightly nevertheless, and the reader can observe which types of analysis are more conservative and which ones are more liberal.

Example 3 – DATAVERSE

The DATAVERSE example [55] reproduces part of the multiverse analysis reported in Steegen et al. [87], which is itself a re-analysis of a famous and controversial study on the effect of ovulatory cycles on voting behavior [38]. The DATAVERSE example is meant to illustrate alternative ways of processing experimental data (e.g., dichotomizing responses, excluding participants), and the use of interactive choice lists.

The "*Constructing the data multiverse*" section in Steegen et al. [87] goes through each data processing choice made in the original study [38] and describes alternative choices that could have been reasonably made. The Dataverse mini-paper essentially reproduces this section with the difference that the reader can select particular choices. The mini-paper first lists five ways of dichotomizing a particular dependent variable, and lets the reader choose one of them (Figure 4). Four other data processing operations are described afterwards, each with two to three options to choose from. The mini-paper ends with a figure showing the result of the selected analysis in the form of an interaction plot, which is updated each time a different option is chosen in the text.

The Dataverse mini-paper covers 5×2×3×3×2 = 180 unique analyses. Steegen et al. [87] summarizes the multiverse by plotting the 180 corresponding *p*-values. While this summary provides an extremely useful overview clearly showing that the original findings are not robust, it does not allow the reader to examine detailed outcomes of specific analyses of interest. By making it possible to select any particular analysis and see the resulting effect sizes, the Dataverse mini-paper conveys more complete results than a simple summary of *p*-values. As in the Frequentist mini-paper the multiverse can be animated, giving a striking demonstration of the variability of effect sizes across the multiverse that can usefully complement the *p*-value summary.

Example 4 – Prior

The Prior example [78] is a reanalysis of a CHI study [52] that used Bayesian analysis to examine the effect of incidental power poses on risk taking behavior, measured using the *number of pumps* in a balloon pumping task. In our mini-paper, we reanalyze the data for experiment 2 using a similar Bayesian model to show the effect of using different priors on the difference in *number of pumps* on the results.

A Bayesian analysis allows researchers to set priors on parameters, enabling them to incorporate domain knowledge into the analysis. However, several different priors might be reasonable for the same analysis. Researchers can face the problem of choosing between different priors, each of which may appear defensible. For example, a researcher might choose a 0-centered, *skeptical*, *regularized* [66] prior to down-weight unreasonably large effect sizes. Alternatively, a researcher may opt for an *optimistic* prior centered on a previously observed effect size in the same domain. A researcher may also be more or less confident in their prior knowledge, which may manifest in narrower (more confident) or wider (less confident) priors. These are just some of the myriad possibilities.

Performing an analysis with just one prior might lead to a biased result; it also imposes the author's prior beliefs on the reader. By contrast, our Prior mini-paper shows how

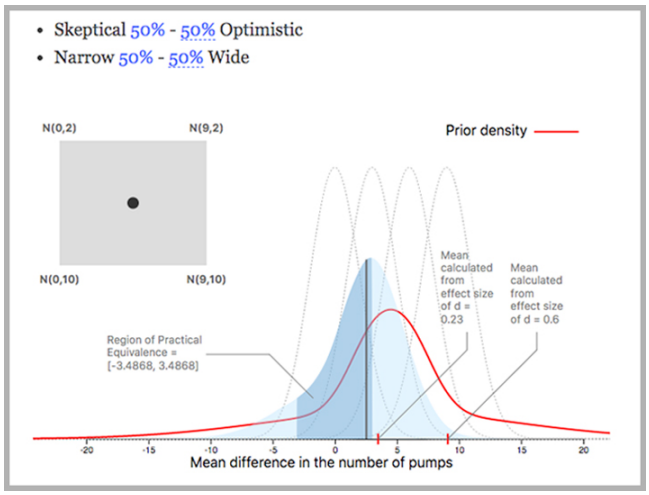

Figure 5: Excerpt from the mini-paper Prior depicting the prior and posterior densities. Readers can use the 2D selection widget (left inset gray box) or drag the highlighted percentages to change the prior.

an emar can let authors present the same analysis with a range of reasonable priors. The reader can change the default prior on the mean difference in *number of pumps*, and the posterior density plot is updated dynamically. The reader can manipulate the prior by changing its location (from 0-centered up to a large effect) and its width (from narrow to wide). Unlike other examples, these two axes are continuous. The reader can change their prior either by clicking and dragging on a point in a 2-dimensional space (see Figure 5), or by clicking and dragging on text sliders (like how confidence level can be adjusted in the Frequentist mini-paper).

Readers of the Prior mini-paper can see how the prior affects the posterior probability of the mean difference of interest. Animating the multiverse depicts the results from a randomly chosen prior along these dimensions and highlights the extent to which the choice of prior can affect the results of the study. In this example, it highlights that large effect sizes are likely only if one has a confident prior centered on the large effect size of the original power pose study [22], ignoring any intervening studies (a meta-analysis of which found an effect less than half the size [46]). This shows how being able to shift the prior allows readers to answer questions like "what would I have to have believed before this study in order to believe there is a large effect here?".

Example 5 – Dance

The Dance example [33] is a reanalysis of a previous InfoVis study on the perception of correlations [48]. It is meant to illustrate the use of simulated datasets to convey inferential information that can be missing from plots.

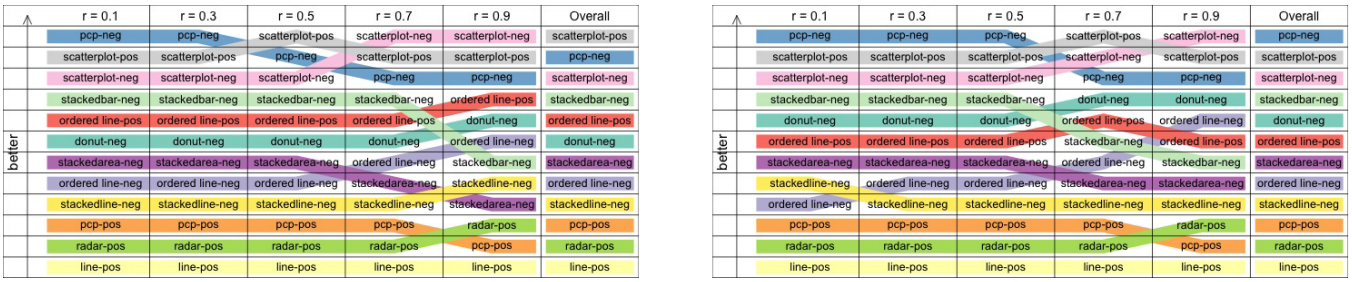

Figure 6: ***Left:*** **plot showing a ranking of visualizations in their ability to convey correlation [48].** ***Right:*** **an alternative plot that could have reasonably come up in an exact replication, created by bootstrapping the experimental dataset. Some results hold (e.g., the bottom of the ranking) while some do not (e.g., the top and middle of the ranking). The mini-paper Dance allows to animate between 100 of those plots.**

Here, the analysis procedure is the same across the multiverse but the raw dataset varies. More specifically, 100 alternative datasets were created from the original study's dataset using bootstrapping [100]. A bootstrap dataset is created by sampling the original dataset with replacement. A remarkable property of bootstrapping is that the distribution of bootstrap samples tends to resemble the true sampling distribution, and thus bootstrap samples can be used to compute CIs [39]. In the Dance example, we use bootstrapping to derive a set of datasets that could have reasonably come up if the study was replicated with different participants. We then subject all datasets to the same statistical analysis and plotting procedure.

The mini-paper reproduces the analysis from the original study, with its four plots. It also lets readers replace the original dataset with any of the 100 bootstrap datasets. When the dataset changes, each of the 4 plots changes slightly. More interestingly, animating the multiverse yields a "dance of plots" similar to Cumming's dance of *p*-values [28] and other statistical dances [32], with the difference that the sampling distribution is estimated from data rather than simulated.

Animating the multiverse of bootstrap datasets allows the reader to appreciate the reliability of the different quantities, trends and patterns depicted by each plot and to carry out "inference by eye" [30]: a pattern that is stable across the multiverse is a good indication that it is reliable. This is an example of the use of hypothetical outcome plots (HOPs) for conveying uncertainty [50, 53]. Compared to static representations of inferential information such as error bars, this technique has the advantage of being applicable to any plot. It is especially useful for revealing statistical uncertainty that is hidden in some plots, such as the ranking plot reproduced in Figure 6. More examples can be found in the mini-paper.

Implementation

We combined a set of existing web frameworks to render academic papers in the browser. We use distill.js[2] as a base framework to provide support for transforming BibTeX references into hyperlinks and adding a list of numbered references at the end of the document. The CSS styling for the ACM SIGCHI format is taken from the Pubcss project[3] which enables writing of ACM style compliant articles using HTML and CSS. Pubcss requires a compilation step to transform the HTML sources into a PDF which is not desirable for our purposes. We thus developed our own conversion scripts in JavaScript which takes an HTML document written to be compliant with the distill format as well as the CSS file from Pubcss and transforms the layout of the distill format such that it appears in the browser window like a paginated PDF in the SIGCHI format. This custom script also takes care of handling differences between the distill.js format and the ACM format to enable restyling for the two template styles.

For handling interaction, we use a customized version of the Tangle[4] JavaScript library developed by Bret Victor for his essay on explorable explanations [94]. This library enables in-text widgets, as illustrated in our examples. In addition to updating text, we use Tangle callbacks that update figures by changing their source URL. We extended the Tangle library to support additional widgets, such as the choice lists from the example Dataverse. In the Likert example, interaction with figures is enabled using HTML image-maps.

For Likert, which uses static figures, the plots were simply generated in R and exported as bitmaps. In Frequentist, Dataverse and Dance, all possible figures were generated in R and exported as bitmaps with a naming scheme allowing JavaScript to retrieve the figure corresponding to a specific combination of analysis options. In all three examples little modification to the existing R analysis was necessary, and

[2]https://github.com/distillpub/template
[3]https://github.com/thomaspark/pubcss
[4]http://worrydream.com/Tangle/

mostly involved wrapping the code inside a function that was called multiple times with different parameters.

For our Prior example, the naive approach might be to re-fit the model interactively or to pre-fit several hundred models to approximate continuous selection, both of which are infeasible due to the time needed to fit Bayesian models. Instead, we pre-fit a small number of versions of the same model M with $J = 8$ slightly different priors. We define a set of equally-spaced priors along two dimensions: location and scale. This defines J different priors; we fit J variants of model M, one with each prior, using Markov chain Monte Carlo (mcmc) in *Stan* [91].

Using a weighted mixture of the J priors, we can derive the posterior distribution for any variant of model M. Given the desired weights on each prior ($w_j^{(0)}$) and the marginal likelihoods of the models fit with each prior (C_j), the posterior distribution can be expressed as a weighted mixture of the posteriors of the models, with posterior weights $w_j^{(1)}$:

$$w_j^{(1)} = \frac{w_j^{(0)} C_j}{\Sigma w_j^{(0)} C_j} \tag{1}$$

We use a kernel density estimator to obtain posterior densities from each mcmc posterior, and we use bridge sampling [45] to calculate marginal likelihoods. This approach admits a range of possible prior shapes and model types—essentially any model and prior that can be fit using *Stan*. These steps are pre-computed in R and written to a csv file.

In the browser, as users interact with Tangle widgets or our 2D widget (Figure 5) to move along the two dimensions (location and scale), we calculate the weights for the prior distributions and the corresponding weights for the posteriors using the above formula. We then calculate the mixture posterior density and visualize it using D3.js in real time.

4 THE DESIGN SPACE OF EMARS

Based on the previous examples, we lay out a terminology and a design space for explorable multiverse analysis reports (emars). These are meant to structure the discussions in this paper, as well as provide a common framework and language for future research. Throughout this section, we also discuss the major trade-offs between different points in the design space and offer general tips for writing emars.

Designing an emar involves two steps: *i)* defining the multiverse, i.e., choosing *what* analyses to conduct and report, and *ii)* designing the report, i.e., choosing *how* to report these analyses. Although step *i)* has been previously discussed by Simonsohn et al. [86] and Steegen et al. [87], here we extend their discussion to cover a broader spectrum of analyses relevant to emars. We start with basic terminology.

Basic Multiverse Terminology

Consider the "tree of analysis" metaphor of Figure 1, inspired by Nolan and Temple Lang [71]: an analysis proceeds from top to bottom, and each branching represents a choice between different analysis options. We refer to an **analysis parameter** as a node in the tree that has more than one child, and to an **analysis option** as one of those children. An **analysis**, meanwhile, is a complete path from the root to a leaf. We are interested in what is reported. For example, in Figures 1a-b, there is no analysis parameter exposed in the report, thus there is only one analysis reported. In Figure 1c, several analysis parameters are exposed, each having 2 analysis options to choose from, for a total of 16 possible analyses, i.e., as many as leaves at the bottom of the tree.

Types of Analysis Parameters by Level of Analysis

Analysis parameters can be classified in different ways. One useful way to classify them is according to their position in the statistical analysis pipeline:

- **Data substitution parameters** offer to switch between different raw datasets, either collected or simulated as in our Dance example.
- **Data processing parameters** offer to process the same raw data in different ways before it is analyzed. Our Dataverse example has five data processing parameters. Another example is the decision to use or not a log transformation in our Frequentist example.
- **Modeling parameters** offer different ways of analyzing the same processed data. Our Likert example has a single modeling parameter with nine options. Another example of modeling parameter is the choice between t-distribution and bootstrap CIs, or between no multiplicity correction and Bonferroni correction in our Frequentist example. The Prior example uses two parameters, each having a theoretically infinite number of options.
- **Presentation parameters** offer different ways of presenting analysis outcomes. The choice of confidence level in our Frequentist example can be considered a presentation parameter [66]. We discuss other examples below.

Of the two published multiverse articles, the one from Steegen et al. [87] essentially focuses on data processing parameters, while Simonsohn et al. [86] focus on both data processing and modeling parameters. Although these are central aspects of statistical analysis, when designing a multiverse for an emar it helps to consider all levels at which a parameter can be exposed. We provided concrete examples for each of these levels, but many more examples can be thought of. For example, presentation parameters can involve choosing between different types of graphical representations, different plotting options (e.g., histogram bin

size, smoothing kernels), or different levels of numerical precision (e.g., one appropriate for communication [8] and one appropriate for verification [49]). Data processing parameters can involve selecting population subgroups of interest (e.g., female participants). As for modeling parameters, we gave an example of choice of prior but many more Bayesian modeling parameters can be exposed, such as which predictors to include or which link function to use in a multiple regression model [66].

Types of Analysis Options by Function

Each analysis parameter (e.g., type of data transformation) involves a set of analysis options (e.g., log, inverse, untransformed). It helps to distinguish options by their *function*. We can distinguish between:

- **Author-consensual options**, which are analysis options considered reasonable and worth reporting by all authors of the paper. The primary goal behind including these options is to cover a range of analyses in order to assess and convey the robustness of the study findings [86, 87].
- **Author-specific options**, which are analysis options endorsed by only a subset of authors. For example, one author may insist on using method *A* while others may insist on using method *B*. An EMAR allows reporting both.
- **Anticipatory options** are analysis options authors consider invalid or irrelevant but think others may want to see included. They can be provided for the sole purpose of shielding the paper from criticism, e.g., in anticipation of (or as a response to) reviewer requests. Anticipatory options may also be provided as a courtesy, e.g., an author who dislikes reporting *p*-values [31] may add an option to display them for readers who are more comfortable with them.
- **Educational options** are analysis options that no one would normally consider reasonable (in the sense that no one would choose them in a single-universe report), but that are included for pedagogical purposes. For example, a paper may include unusual confidence levels to reinforce the idea that the limits of interval estimates are arbitrary (as in our FREQUENTIST example), or may include bootstrapped datasets to emphasize the numerical uncertainty of the analysis outcomes (as in our DANCE example). These options too increase transparency, because a transparent statistical report should be "an exercise of pedagogy as much as an exercise of rhetoric" and should "anticipate misinterpretations" [31].

The two previous articles on multiverse analysis [86, 87] focus largely on reporting author-consensual options in order to assess and communicate the robustness of the authors' findings. Simonsohn et al. [86] additionally discuss what we call anticipatory options and propose to construct multiverses as unions of author-consensual and anticipatory options. Though these are key considerations, EMARs also provide opportunities for offering the other types of options discussed previously, all of which can contribute to increasing research transparency.

So far we have covered the "what" and the "why", i.e., what type of analysis may be included in an EMAR and why. From now on we cover the "how", i.e., the different ways of presenting these analyses in an EMAR, and discuss their trade-offs.

EMAR Content Terminology

Like most statistical reports, an EMAR interleaves text with figures. Text can be prose (e.g., explanations, discussions) or non-prose (e.g., numerical results, formulas, code). In addition, some portions of an EMAR consist of non-statistical content unrelated to the multiverse analysis (e.g., introduction, study methods), while other portions are **analysis reports** referring to a particular analysis (or several analyses) in the multiverse. An analysis report is a combination of: **analysis explanations**, which consist of content (text and figures) whose purpose is to explain and justify the analysis; and **analysis outcomes**, which consist of content (text and figures) whose function is to communicate the results of the analysis.

This terminology is applicable to any statistical report (EMARs, traditional reports and static multiverse reports), but what distinguishes EMARs from the rest is that not all analysis explanations and outcomes are simultaneously visible.

Default Analyses and Reporting Style

A **default analysis** is an analysis whose report is fully visible (i.e., both explanations and outcomes) when the article is opened for the first time. An EMAR can have a single default analysis, which is typically the analysis that the authors would favor over all others if given only one analysis "slot". As in our FREQUENTIST, DANCE and PRIOR mini-papers, such an article can look like a regular paper and can be read as such. LIKERT and DATAVERSE also have a single default analysis (the one whose report is initially fully visible), but in addition they simultaneously show multiple outcomes (in LIKERT) and multiple explanations (in DATAVERSE). In a non-explorable multiverse report [86, 87], all analyses are default in the sense that all their reports are fully visible.

EMAR authors are free to choose where they want their paper to sit in this continuum between classical single-universe reports and full multiverse reports. Making the multiverse explicit by providing, e.g., outcome overviews as in LIKERT makes it much easier for the reader to get a good sense of the multiverse. Meanwhile, adopting the style of classical reports

and conveying the multiverse more subtly (as in Frequentist and Dance) can make the paper more accessible and perhaps less daunting to an audience who is unfamiliar with multiverse analyses. Furthermore, since non-default analyses are de-emphasized, this reporting style may encourage authors to include author-specific, anticipatory and educational options they may not include otherwise (e.g, *p*-values). On the other hand, it hides the complexity of the multiverse from all but the most engaged readers.

Next we examine important usability trade-offs by comparing different multiverse reporting strategies in more detail. Although much of the following discussion applies to explorable explanations more generally, the trade-offs within their design space have never been examined in detail. Thus, we review them here. In addition, explorable explanations have never been used in the context of academic statistical reporting, where explanations can span several pages, potentially introducing additional design challenges.

Multiplexing and Aggregation

There are three main approaches for conveying multiple analysis reports. **Space multiplexing** consists of showing multiple analysis explanations or outcomes simultaneously, by juxtaposing them. **Time multiplexing** means showing multiple explanations or outcomes at the same location, but at different times (thus requiring interaction or animation). For example, Dataverse uses space multiplexing for explanations and time multiplexing for outcomes, while Likert does the opposite. A third category is **aggregation**, which consists of combining multiple explanations or outcomes in a static representation within the same space. All our examples use aggregation in the sense that they contain discussions summarizing results from the multiverse. Outcome aggregation can also be done graphically, as in Steegen et al.'s [87] histograms of *p*-values.

Space multiplexing can be space-demanding: it would require ≈10 pages to fit all the 112 figures of the Frequentist example. Thus it is best used when the outcomes or explanations have a visual representation that fits a tight space, and it is especially useful when the representation supports easy side-by-side visual comparison. For example, in the Likert mini-paper (Figure 3) each outcome takes up a row, while outcomes are one-pixel columns in Simonsohn et al.'s specification curves [86]. Space multiplexing however comes at the expense of the level of detail, just like aggregation.

Whenever detailed analyses are worth reporting, time multiplexing is a useful alternative. Dynamic (interactive or animated) views also facilitate the detection of subtle differences and can be effective at conveying concepts such as uncertainty [50, 53]. However, they are not printable and less easy to navigate [93]. For example, it can be difficult to search for a particular view, such as an extreme outcome. Dynamic views can also render a report unstable, e.g., if different analyses yield paragraphs of different sizes. This can be addressed by making sure that all paragraphs in the multiverse have equal height, or by using figures of fixed size. Further potential costs of time multiplexing will be discussed in the next section.

Controls

So far we have discussed how to show multiple analyses in an emar, and mentioned time multiplexing as a strategy that depends heavily on interaction. Here, we discuss the different ways a reader can interact with time-multiplexed reports.

We refer to **controls** as elements in an emar that let readers change analysis parameters and thus change the visibility of analysis outcomes and/or descriptions in the paper. A first decision is whether to place a control in the text (as in all our examples) or in a figure (as in Likert and Prior).

In-text controls support narrative-guided exploration of the multiverse and allow authors to introduce parameters one by one, at the right time. Because analysis descriptions are typically textual, readers are likely to learn about details of the analysis from the text. Thus, if there are default options the authors know are controversial or unfamiliar to some readers, it is sensible to offer immediate access to alternative (anticipatory) options by placing controls in the text itself. In contrast, controls in figures are harder to connect with the main narrative and are easier to miss. One drawback of text controls is that they can be quite spread out, so it can be difficult to gain an overview of (and access to) all parameters from the multiverse. This can be addressed by adding "control panels" with multiple parameters. In particular, sets of controls can be placed within figures in order to support free exploration.

The proper choice and placement of controls can be further informed by the instrumental interaction framework [14]. In this framework, controls are *instruments* and their targets are *domain objects*. We define a **target** as a portion of text or a figure that shows analysis explanations or outcomes and that is modified by a control. The framework recommends to minimize **spatial indirection**, which is the distance between instruments and domain objects [14]. A reader may indeed experience difficulties if a control and its targets are situated far apart. In some of our mini-papers, some control-target pairs do not simultaneously fit the browser window. In these cases it can be difficult to follow changes, or even know what has changed in the paper. This issue can be mitigated by a clever placement of targets (typically figures), but if multiple interdependent controls and targets are distributed across several pages (as in Frequentist), a perfect paper layout might be unattainable. This is undoubtedly a key limitation of the time multiplexing approach, although

possible solutions involving changes to the article reading UI will be considered in the discussion section.

Following the instrumental interaction framework, controls should also be ideally designed to maximize the **degree of compatibility**, which is "the similarity between the physical actions of the users on the instrument and the response of the object" [14]. For example, in Prior, the skeptical/optimistic mixing is controlled in the text by setting the weight of the optimistic prior, in such a way that dragging the control moves the plotted prior in the same direction. Alternatively, allowing the prior to be directly manipulated in the figure would have simultaneously minimized spatial indirection and maximized the degree of compatibility.

Finally, controls too can follow a **space multiplexing** or a **time multiplexing** approach. In the former, all options are simultaneously visible (e.g., choice lists of Dataverse), while in the latter, only the currently selected option is shown (e.g., draggable values of Frequentist and Prior). The trade-offs are similar as before: space multiplexing supports overview but tends to occupy space. One exception is the 2D widget in Prior, a space-multiplexed control where each option takes only a single pixel. For in-text controls, space multiplexing ensures text stability (see Dataverse) but can break the flow of the narrative. Time multiplexing as in Frequentist can be useful if one wants to preserve the traditional reporting style and emphasize a preferred analysis.

Narrative Design

We finish with some general recommendations about the design of textual narratives. Narratives are key in any academic statistical report, but building the narrative of an emar differs in several respects. First, the paper needs to inform the reader of the multiverse exploration possibilities. Although the most important is to ensure the presence of effective affordance cues on the controls (e.g., the link styling and tooltips used by Bret Victor and in our mini-papers), the textual narrative can also explicitly invite the reader to interact.

Perhaps the most important common sense rule for writing emars is that *the textual narrative should always be consistent with the reported outcomes* (e.g., plots, numbers, tables). In other words, a reader should be able to freeze the paper at any point, and the paper should make sense [92]. One way to ensure consistency is to have the narrative update itself with the displayed analysis. This can be fairly easy for short pieces of narratives that only require numbers or statistical terms to be updated, such as figure captions or technical explanations (e.g., the explanation of family-wise error rates in Frequentist). However, in many cases writing multiple narratives can quickly become overwhelming, both for the author and later for the reader. A much simpler approach is to write narratives that are consistent with the entire multiverse.

In line with our main objective to increase transparency, *we recommend against the use of multiple narratives when interpreting results and drawing conclusions.* Again, a better alternative is to write statements that are true for the entire multiverse—this has the benefit of forcing authors to focus on reliable effects, and refrain from commenting on fragile effects. Another alternative is to write narratives that acknowledge the entire multiverse, by summarizing it or by contrasting different analyses (e.g., see final discussions in Frequentist or Likert). In particular, we recommend emar authors to incorporate a short discussion summarizing how robust or fragile their results are in the context of the multiverse, as well as explain "the key choices in data processing that are most consequential in the fluctuation of statistical results" [87], should conclusions vary across universes.

5 DISCUSSION

We first discuss limitations of our work, and then the potential challenges involved in the adoption of emars.

First of all, our design space is meant to capture elementary emar techniques, but many other more sophisticated techniques are possible to further enhance emars. Coordinated views and linking [76] could be used to facilitate navigation in the multiverse, for example by highlighting the currently visible analysis in a multiverse summary. Mechanisms could be introduced for automatically creating multiverse overviews and summaries, such as blending all plots or laying them out as thumbnails on a grid. A reverse direct manipulation mechanism [37] could be added to let readers manipulate plots and observe which analysis options lead to certain plots (e.g., "which options yield the largest effect?"). Finally, techniques could be implemented to support comparisons of two or more analyses of interest, or to record and visualize the reader's navigation history in the multiverse. Other techniques and ideas could be borrowed from the domain of *visual parameter space analysis*, which shares many conceptual similarities with emars [82].

Our examples cover relatively simple statistical analyses. Although their level of complexity is typical of HCI papers, analyses can in principle get much more complex. Complex analyses would probably need to expose less parameters to preserve usability: as pointed out by Steegen et al. [87], multiverses do not need to be large to be useful. When multiple models are used, one issue is finding meaningful statistics that can be compared across models. Simonsohn et al. [86]'s analysis does not suffer from this issue since every output is a p-value, but our Likert example was complicated by the fact that different models yielded different types of outputs. When this arises, emar authors may want to identify a set of end-goal outcomes that are the same for all models, in order to facilitate cross-universe comparisons.

Importantly, our mini-papers are only proofs of concepts. Although we provide a template for other researchers to write their own mini-paper, our tool is experimental and its workflow involves much manual work. For EMARs to be adopted, it is crucial that usable toolkits are developed. A future toolkit could build on modern interactive document authoring tools such as Idyll [25] and capitalize on initiatives to integrate academic writing with statistics environments such as the recently released Radix [4], which combines the Distill framework with R Markdown. Ideally, such a toolkit will combine explorable explanation features with reproducible research functionality [77]. One pending question is whether EMARs will need to be standalone documents that can be viewed with minimal infrastructure (such as our mini-papers) or live articles that require a statistical environment. If EMARs were to become a norm in the foreseeable future other challenges would need to be addressed, including support for digital preservation and archival longevity [62, 73, 88], support for citations (e.g., being able to cite a specific analysis in an EMAR), and accessibility: EMARs will need to be made compatible with screen-readers and support accessible user navigation in the multiverse. Although these issues are beyond the scope of the present work, by laying out an elementary design space for EMARs, we hope this article will facilitate the design of future toolkits and infrastructures.

Even if tools are developed to support the authoring of EMARs, however, some objections to their widespread use will likely remain. We examine four potential objections.

1. Writing EMARs will remain hard. With the proper tools, it is unclear whether writing EMARs will be harder than writing static multiverse analysis reports as in [86, 87]. It is clear however that writing EMARs will always be harder than writing single-universe analyses, and as with providing supplemental material, the extra effort may not come with tangible rewards [20]. The statistical analysis itself is more work-intensive and even if new libraries can be developed to facilitate it, it is unlikely that the job can be automated [87]. Similarly, it will never be possible to ensure that a multiverse analysis is complete or even well-chosen. However, as we already pointed out, even a small multiverse analysis is superior to a single-universe analysis in terms of transparency [87]. Thus for researchers who want to signal or promote transparency the option of writing EMARs can still be attractive, and may become more and more attractive as signals of research integrity and transparency get more and more rewarded [42, 43].

2. Reading EMARs will remain hard. We already pointed out potential usability issues with EMARs, including difficulties with following changes that occur outside the viewport and predicting where changes will occur. It would be interesting in the future to investigate whether these issues can be mitigated by improvements to the paper reading UI. For example, the interface could present two views of the same article that can be scrolled independently, allowing readers to monitor targets that are located far from the operated controls. Alternatively, changes could be highlighted with different colors both within and outside the viewport [12, 23], or replayed back when hidden portions of the article are brought into view [13, 18]. Regardless, engaging with EMARs will always require extra effort from the reader. Because of this, it is again crucial that EMARs can be read at two levels and understood without interacting [94] by readers who quickly want to learn about the authors' conclusions and how they arrived at them.

3. Reviewing EMARs will remain hard. At first sight, EMARs seem to pose a challenge to reviewing policies. Publishers typically impose strict cut-offs in terms of page length, and enabling interactive content implies that authors can provide a theoretically unbounded amount of content. EMARs also break the linear structure reviewers are used to. The problem is however not new, and already arose with the introduction of supplemental materials [65]. Publishers and reviewers can simply consider all non-default analyses as supplemental material: in case reviewers are not required to review supplemental material (e.g., as in CHI), only the default analysis would need to be reviewed. Reviewers are still free to scrutinize other analyses and demand analysis options to be added [86].

4. Preregistration will remain preferable. Multiverse analyses are fully compatible with planning and preregistration [86, 87]. One approach is to preregister the entire multiverse analysis [86]. This would require to specify which is the default analysis in the EMAR (if any), as the choice of which analysis to emphasize constitutes a important researcher's degree of freedom. A weaker but easier form of preregistration could include the default analysis only.

6 CONCLUSION

We presented *explorable multiverse analysis reports* (EMARs), a new approach to statistical reporting where readers of research papers can explore alternative analysis options by interacting with the paper. Through examples and a design space analysis, we illustrated the many opportunities offered by EMARs, as well as the pending challenges. We hope our work will inspire more HCI research where the academic paper is treated as a user interface whose purpose is to convey scientific knowledge in an accurate and transparent manner.

7 ACKNOWLEDGEMENTS

Many thanks to: Chat Wacharamanotham and Stéphane Huot for early discussions; Steve Haroz and dgp members for discussions and feedback; Catherine Plaisant, Theophanis Tsandilas, Tobias Isenberg and Gilles Bailly for detailed feedback on our paper draft; our reviewers for their suggestions.

REFERENCES

[1] [n. d.]. Distill. https://distill.pub/. Accessed: 2018-09-15.

[2] [n. d.]. Jupyter. http://jupyter.org/. Accessed: 2018-09-15.

[3] [n. d.]. R Markdown Gallery. https://rmarkdown.rstudio.com/gallery.html. Accessed: 2018-09-15.

[4] [n. d.]. Radix for R Markdown. https://rstudio.github.io/radix/. Accessed: 2018-09-16.

[5] [n. d.]. Shiny. https://shiny.rstudio.com/. Accessed: 2018-09-15.

[6] IJsbrand Jan Aalbersberg, Pilar Cos Alvarez, Julien Jomier, Charles Marion, and Elena Zudilova-Seinstra. 2014. Bringing 3D visualization into the online research article. *Information Services & Use* 34, 1-2 (2014), 27–37.

[7] Michael J Ackerman. 2015. The Educational Value of Truly Interactive Scientific Publishing. *Journal of Electronic Publishing* 18, 2 (2015).

[8] APA. 2010. *The Publication manual of the APA (6th ed.).* Washington, DC.

[9] Christie Aschwanden. 2015. Science is not broken. https://fivethirtyeight.com/features/science-isnt-broken/. Accessed: 2018-09-15.

[10] Sriram Karthik Badam, Zhicheng Liu, and Niklas Elmqvist. 2019. Elastic Documents: Coupling Text and Tables through Contextual Visualizations for Enhanced Document Reading. *IEEE Transactions on Visualization & Computer Graphics* (2019). http://www.umiacs.umd.edu/~elm/projects/elastic-documents/elastic-documents.pdf

[11] David G Barnes and Christopher J Fluke. 2008. Incorporating interactive three-dimensional graphics in astronomy research papers. *New Astronomy* 13, 8 (2008), 599–605.

[12] Patrick Baudisch and Ruth Rosenholtz. 2003. Halo: a technique for visualizing off-screen objects. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 481–488.

[13] Patrick Baudisch, Desney Tan, Maxime Collomb, Dan Robbins, Ken Hinckley, Maneesh Agrawala, Shengdong Zhao, and Gonzalo Ramos. 2006. Phosphor: explaining transitions in the user interface using afterglow effects. In *Proceedings of the 19th annual ACM symposium on User interface software and technology*. ACM, 169–178.

[14] Michel Beaudouin-Lafon. 2000. Instrumental interaction: an interaction model for designing post-WIMP user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, 446–453.

[15] Daniel J Benjamin, James O Berger, Magnus Johannesson, Brian A Nosek, E-J Wagenmakers, Richard Berk, Kenneth A Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, et al. 2018. Redefine statistical significance. *Nature Human Behaviour* 2, 1 (2018), 6.

[16] Timothy J Berners-Lee. 1989. *Information management: A proposal.* Technical Report.

[17] Donald Berry. 2012. Multiplicities in cancer research: ubiquitous and necessary evils. *Journal of the National Cancer Institute* 104, 15 (2012), 1125–1133.

[18] Anastasia Bezerianos, Pierre Dragicevic, and Ravin Balakrishnan. 2006. Mnemonic rendering: an image-based approach for exposing hidden changes in dynamic displays. In *Proceedings of the 19th annual ACM symposium on User interface software and technology*. ACM, 159–168.

[19] Rolf Biehler. 1997. Software for learning and for doing statistics. *International Statistical Review* 65, 2 (1997), 167–189.

[20] Leen Breure, Hans Voorbij, and Maarten Hoogerwerf. [n. d.]. Rich Internet Publications:" Show What You Tell". *Journal of Digital Information* 12, 1 ([n. d.]).

[21] Jonathan B Buckheit and David L Donoho. 1995. Wavelab and reproducible research. In *Wavelets and statistics*. Springer, 55–81.

[22] Dana R Carney, Amy JC Cuddy, and Andy J Yap. 2010. Power posing: Brief nonverbal displays affect neuroendocrine levels and risk tolerance. *Psychological science* 21, 10 (2010), 1363–1368.

[23] Fanny Chevalier, Pierre Dragicevic, Anastasia Bezerianos, and Jean-Daniel Fekete. 2010. Using text animated transitions to support navigation in document histories. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 683–692.

[24] Andy Cockburn, Carl Gutwin, and Alan Dix. 2018. Hark no more: on the preregistration of CHI experiments. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 141.

[25] Matt Conlen and Jeffrey Heer. 2018. Idyll: A Markup Language for Authoring and Publishing Interactive Articles on the Web. In *ACM User Interface Software & Technology (UIST)*. http://idl.cs.washington.edu/papers/idyll

[26] Geoff Cumming. [n. d.]. Exploratory Software for Confidence Intervals. https://thenewstatistics.com/itns/esci/. Accessed: 2018-09-15.

[27] Geoff Cumming. 2002. Live figures: Interactive diagrams for statistical understanding. In *Proceedings of the Sixth International Conference on Teaching of Statistics, Cape Town. Voorburg, The Netherlands: International Statistical Institute*.

[28] Geoff Cumming. 2009. The dance of p-values (video). https://www.youtube.com/watch?v=ez4DgdurRPg

[29] Geoff Cumming. 2014. The new statistics: Why and how. *Psychological science* 25, 1 (2014), 7–29.

[30] Geoff Cumming and Sue Finch. 2005. Inference by eye: confidence intervals and how to read pictures of data. *American Psychologist* 60, 2 (2005), 170.

[31] Pierre Dragicevic. 2016. Fair statistical communication in HCI. In *Modern Statistical Methods for HCI*. Springer, 291–330.

[32] Pierre Dragicevic. 2017. Statistical Dances: Why no Statistical Analysis is Reliable and What to do About it (video). https://www.youtube.com/watch?v=UKX9iN0p5_A

[33] Pierre Dragicevic. 2018. Adding Inferential Information to Plots using Resampling and Animations. (2018). https://explorablemultiverse.github.io/examples/dance/

[34] Pierre Dragicevic. 2018. An Explorable Multiverse Analysis of Durante et al. (2013). (2018). https://explorablemultiverse.github.io/examples/dataverse/

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

[36] Pierre Dragicevic and Yvonne Jansen. 2018. Re-Evaluating the Efficiency of Physical Visualizations: A Simple Multiverse Analysis. (2018). https://explorablemultiverse.github.io/examples/frequentist/

[37] Pierre Dragicevic, Gonzalo Ramos, Jacobo Bibliowitcz, Derek Nowrouzezahrai, Ravin Balakrishnan, and Karan Singh. 2008. Video browsing by direct manipulation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 237–246.

[38] Kristina M Durante, Ashley Rae, and Vladas Griskevicius. 2013. The fluctuating female vote: Politics, religion, and the ovulatory cycle. *Psychological Science* 24, 6 (2013), 1007–1016.

[39] Bradley Efron. 1992. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*. Springer, 569–593.

[40] 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).

[41] Robert Gentleman and Duncan Temple Lang. 2007. Statistical analyses and reproducible research. *Journal of Computational and Graphical Statistics* 16, 1 (2007), 1–23.

[42] Roger Giner-Sorolla. 2012. Science or art? How aesthetic standards grease the way through the publication bottleneck but undermine science. *Perspectives on Psychological Science* 7, 6 (2012), 562–571.

[43] Roger Giner-Sorolla. 2012. Will we march to utopia, or be dragged there? Past failures and future hopes for publishing our science. *Psychological Inquiry* 23, 3 (2012), 263–266.

[44] Jonathan Gray, Lucy Chambers, and Liliana Bounegru. 2012. *The data journalism handbook: How journalists can use data to improve the news.* " O'Reilly Media, Inc.".

[45] Quentin F Gronau, Henrik Singmann, and Eric-Jan Wagenmakers. 2017. Bridgesampling: an r package for estimating normalizing constants. *arXiv preprint arXiv:1710.08162* (2017).

[46] Quentin F Gronau, Sara Van Erp, Daniel W Heck, Joseph Cesario, Kai J Jonas, and Eric-Jan Wagenmakers. 2017. A Bayesian model-averaged meta-analysis of the power pose effect with informed and default priors: The case of felt power. *Comprehensive Results in Social Psychology* 2, 1 (2017), 123–138.

[47] Tovi Grossman, Fanny Chevalier, and Rubaiat Habib Kazi. 2016. Bringing research articles to life with animated figures. *interactions* 23, 4 (2016), 52–57.

[48] Lane Harrison, Fumeng Yang, Steven Franconeri, and Remco Chang. 2014. Ranking Visualizations of Correlation Using Weber's Law. *IEEE Trans. Vis. Comput. Graph.* 20, 12 (2014), 1943–1952.

[49] James Heathers. 2017. Life In The Tinderbox. Online. https://medium.com/@jamesheathers/life-in-the-tinderbox-6b2e9760f3aa.

[50] Jessica Hullman, Paul Resnick, and Eytan Adar. 2015. Hypothetical outcome plots outperform error bars and violin plots for inferences about reliability of variable ordering. *PloS one* 10, 11 (2015), e0142444.

[51] Yvonne Jansen, Pierre Dragicevic, and Jean-Daniel Fekete. 2013. Evaluating the efficiency of physical visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2593–2602.

[52] Yvonne Jansen and Kasper Hornbæk. 2018. How Relevant are Incidental Power Poses for HCI?. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 14.

[53] Alex Kale, Francis Nguyen, Matthew Kay, and Jessica Hullman. 2018. Hypothetical Outcome Plots Help Untrained Observers Judge Trends in Ambiguous Data. *IEEE transactions on visualization and computer graphics* (2018).

[54] Maurits Clemens Kaptein, Clifford Nass, and Panos Markopoulos. 2010. Powerful and Consistent Analysis of Likert-type Ratingscales. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, New York, NY, USA, 2391–2394. https://doi.org/10.1145/1753326.1753686

[55] Matthew Kay and Pierre Dragicevic. 2018. A Multiverse Reanalysis of Likert-Type Responses. (2018). https://explorablemultiverse.github.io/examples/likert/

[56] Matthew Kay, Steve Haroz, Shion Guha, and Pierre Dragicevic. 2016. Special interest group on transparent statistics in HCI. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 1081–1084.

[57] Matthew Kay, Steve Haroz, Shion Guha, Pierre Dragicevic, and Chat Wacharamanotham. 2017. Moving transparent statistics forward at CHI. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 534–541.

[58] 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 *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 4521–4532.

[59] Kris N Kirby and Daniel Gerlanc. 2013. BootES: An R package for bootstrap confidence intervals on effect sizes. *Behavior research methods* 45, 4 (2013), 905–927.

[60] Donald Ervin Knuth. 1984. Literate programming. *Comput. J.* 27, 2 (1984), 97–111.

[61] Hylke Koers, Ann Gabriel, and Rebecca Capone. 2013. Executable papers in computer science go live on ScienceDirect. *Available at)(Accessed August 6, 2018) ElsevierConnect* (2013).

[62] David Koop, Emanuele Santos, Phillip Mates, Huy T Vo, Philippe Bonnet, Bela Bauer, Brigitte Surer, Matthias Troyer, Dean N Williams, Joel E Tohline, et al. 2011. A provenance-based infrastructure to support the life cycle of executable papers. *Procedia Computer Science* 4 (2011), 648–657.

[63] Daniel Kunin, Jingru Guo, Tyler Dae Devlin, and Daniel Xiang. [n. d.]. Seeing theory. https://students.brown.edu/seeing-theory/index.html. Accessed: 2018-09-15.

[64] Kristoffer Magnusson. [n. d.]. R Psychologist. http://rpsychologist.com/d3/CI/. Accessed: 2018-09-15.

[65] Emilie Marcus. 2009. Taming supplemental material. *Immunity* 31, 5 (2009), 691.

[66] Richard McElreath. 2018. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. CRC Press.

[67] Amelia Ahlers McNamara. 2015. *Bridging the gap between tools for learning and for doing statistics*. Ph.D. Dissertation. UCLA.

[68] Tara A Morgan, Adam E Flanders, William W Olmsted, Scott D Steenburg, and Eliot L Siegel. 2012. Challenges encountered and lessons learned in initial experience with the next generation of interactive radiology literature in RadioGraphics. *Radiographics* 32, 3 (2012), 929–934.

[69] Marcus R Munafò, Brian A Nosek, Dorothy VM Bishop, Katherine S Button, Christopher D Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J Ware, and John PA Ioannidis. 2017. A manifesto for reproducible science. *Nature Human Behaviour* 1, 1 (2017), 0021.

[70] Leif D Nelson, Joseph Simmons, and Uri Simonsohn. 2018. Psychology's renaissance. *Annual review of psychology* 69 (2018).

[71] Deborah Nolan and Duncan Temple Lang. 2007. Dynamic, interactive documents for teaching statistical practice. *International Statistical Review* 75, 3 (2007), 295–321.

[72] Brian A Nosek, George Alter, George C Banks, Denny Borsboom, Sara D Bowman, Steven J Breckler, Stuart Buck, Christopher D Chambers, Gilbert Chin, Garret Christensen, et al. 2015. Promoting an open research culture. *Science* 348, 6242 (2015), 1422–1425.

[73] Daniel Nüst, Markus Konkol, Marc Schutzeichel, Edzer Pebesma, Christian Kray, Holger Przibytzin, and Jörg Lorenz. 2017. Opening the publication process with executable research compendia. *D-Lib Magazine* 23, 1/2 (2017).

[74] Steve Pettifer, Philip McDermott, James Marsh, David Thorne, Alice Villéger, and Terri K Attwood. 2011. Ceci n'est pas un hamburger: modelling and representing the scholarly article. *Learned Publishing* 24, 3 (2011), 207–220.

[75] Victor Powell and Lewis Lehe. [n. d.]. Setosa. http://setosa.io/. Accessed: 2018-09-15.

[76] Jonathan C Roberts. 2007. State of the art: Coordinated & multiple views in exploratory visualization. In *Coordinated and Multiple Views in Exploratory Visualization, 2007. CMV'07. Fifth International Conference on*. IEEE, 61–71.

[77] Anthony Rossini and Friedrich Leisch. 2003. Literate statistical practice. (2003).

[78] Abhraneel Sarma, Yvonne Jansen, and Matthew Kay. 2018. A Multiverse Analysis Considering Different Priors for Incidental Power Poses in HCI. (2018). https://explorablemultiverse.github.io/examples/prior/

[79] Jeff Sauro and James R Lewis. 2010. Average task times in usability tests: what to report?. In *Proceedings of the SIGCHI Conference on*

Human Factors in Computing Systems. ACM, 2347–2350.

[80] Gnther Sawitzki. 2002. Keeping Statistics Alive in Documents. *Computational Statistics* 1, 17 (2002), 65–88.

[81] Bill N Schilit, Gene Golovchinsky, and Morgan N Price. 1998. Beyond paper: supporting active reading with free form digital ink annotations. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM Press/Addison-Wesley Publishing Co., 249–256.

[82] Michael Sedlmair, Christoph Heinzl, Stefan Bruckner, Harald Piringer, and Torsten Möller. 2014. Visual parameter space analysis: A conceptual framework. *Visualization and Computer Graphics, IEEE Transactions on* 99 (2014).

[83] David Shotton, Katie Portwin, Graham Klyne, and Alistair Miles. 2009. Adventures in semantic publishing: exemplar semantic enhancements of a research article. *PLoS computational biology* 5, 4 (2009), e1000361.

[84] Raphael Silberzahn, Eric Luis Uhlmann, Dan Martin, Pasquale Anselmi, Frederik Aust, Eli C Awtrey, Štěpán Bahník, Feng Bai, Colin Bannard, Evelina Bonnier, et al. 2017. Many analysts, one dataset: Making transparent how variations in analytical choices affect results. (2017).

[85] 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.

[86] Uri Simonsohn, Joseph P Simmons, and Leif D Nelson. 2015. Specification Curve: Descriptive and Inferential Statistics on All Reasonable Specifications. (2015).

[87] 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.

[88] Rudolf J Strijkers, Reginald Cushing, Dmitry Vasyunin, Cees de Laat, Adam Belloum, Robert J Meijer, et al. 2011. Toward Executable Scientific Publications.. In *ICCS*. 707–715.

[89] Hendrik Strobelt, Daniela Oelke, Christian Rohrdantz, Andreas Stoffel, Daniel A Keim, and Oliver Deussen. 2009. Document cards: A top trumps visualization for documents. *IEEE Transactions on Visualization and Computer Graphics* 15, 6 (2009), 1145–1152.

[90] Kenji Takeda, Graeme Earl, Jeremy Frey, Simon Keay, and Alex Wade. 2013. Enhancing research publications using rich interactive narratives. *Phil. Trans. R. Soc. A* 371, 1983 (2013), 20120090.

[91] Stan Development Team. 2016. RStan: the R interface to Stan. *R package version* 2, 1 (2016).

[92] George R Thoma, Glenn Ford, Sameer Antani, Dina Demner-Fushman, Michael Chung, and Matthew Simpson. 2010. Interactive publication: the document as a research tool. *Web Semantics: Science, Services and Agents on the World Wide Web* 8, 2-3 (2010), 145–150.

[93] Barbara Tversky, Julie Bauer Morrison, and Mireille Betrancourt. 2002. Animation: can it facilitate? *International journal of human-computer studies* 57, 4 (2002), 247–262.

[94] Bret Victor. 2011. Explorable Explanations. Online. http://worrydream.com/ExplorableExplanations/.

[95] Bret Victor. 2011. Scientific Communication As Sequential Art. Online. http://worrydream.com/ScientificCommunicationAsSequentialArt/.

[96] Chat Wacharamanotham, Matthew Kay, Steve Haroz, Shion Guha, and Pierre Dragicevic. 2018. Special Interest Group on Transparent Statistics Guidelines. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, SIG08.

[97] Chat Wacharamanotham, Krishna Subramanian, Sarah Theres Völkel, and Jan Borchers. 2015. Statsplorer: Guiding novices in statistical analysis. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2693–2702.

[98] Jelte M Wicherts, Coosje LS Veldkamp, Hilde EM Augusteijn, Marjan Bakker, Robbie Van Aert, and Marcel ALM 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.

[99] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only Anova Procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 143–146. https://doi.org/10.1145/1978942.1978963

[100] Michael Wood. 2005. Bootstrapped confidence intervals as an approach to statistical inference. *Organizational Research Methods* 8, 4 (2005), 454–470.

[101] Dongwook Yoon, Nicholas Chen, and François Guimbretière. 2013. TextTearing: opening white space for digital ink annotation. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*. ACM, 107–112.

[102] Polle T Zellweger, Bay-Wei Chang, and Jock D Mackinlay. 1998. Fluid links for informed and incremental link transitions. In *Proceedings of the ninth ACM conference on Hypertext and hypermedia: links, objects, time and space—structure in hypermedia systems: links, objects, time and space—structure in hypermedia systems*. ACM, 50–57.

[103] Polle T Zellweger, Anne Mangen, and Paula Newman. 2002. Reading and writing fluid hypertext narratives. In *Proceedings of the thirteenth ACM conference on Hypertext and hypermedia*. ACM, 45–54.