Supporting the Design and Analysis of HCI Experiments

Alexander Eiselmayer
University of Zurich
Zurich, Switzerland
eiselmayer@ifi.uzh.ch

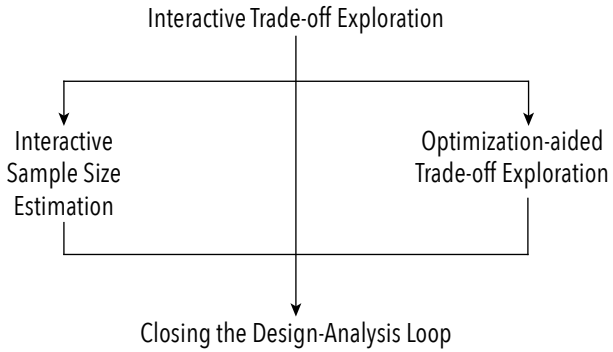

Figure 1: With those four projects, I strive towards improving the design and analysis process for controlled experiments in HCI.

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

ACM ISBN 978-1-4503-6819-3/20/04.
http://dx.doi.org/10.1145/3334480.3375038

Abstract
Researchers in HCI commonly use controlled experiments to evaluate artifacts and interaction techniques. However, experiment design and statistical analysis are complex tasks that are prone to errors, especially for novice researchers. In part, this is because researchers need to make numerous decisions about the design and analysis while it is hard to immediately anticipate their effect. In this dissertation, I aim to study how interactive systems that provide real-time feedback, enable direct manipulation, and facilitate exploration positively influence decision making and reproducibility. My previous work, *Touchstone2*, shows how researchers can benefit from comparing trade-offs among experimental designs. I contribute an empirical understanding of how researchers design and analyze experiments, as well as a set of tools that support researchers during that process.

Author Keywords
Experiment Design; Randomization; Counterbalancing; Power analysis; Reproducibility; Simulation

CCS Concepts
•Human-centered computing → HCI design and evaluation methods; *Laboratory experiments;*

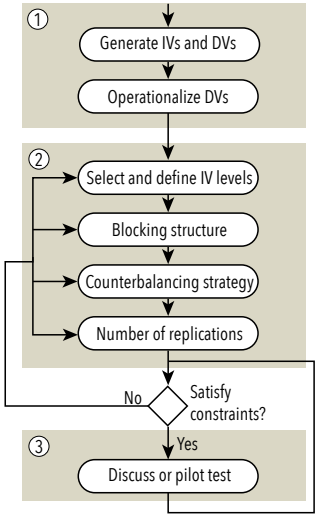

Figure 2: The iterative process of designing experiments:
(1) Conceptualization,
(2) Counterbalancing, and
(3) Testing.

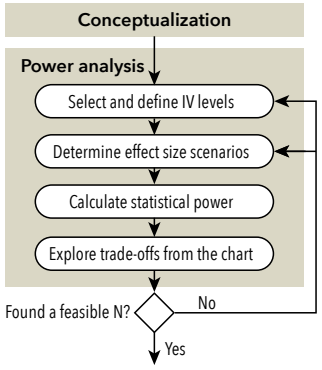

Figure 3: Power analysis can be used to select the number of participants N.

Introduction

In Human-Computer Interaction (HCI), the effectiveness of artifacts and interaction techniques is often evaluated by controlled experiments. Researchers need to carefully design the experiment, collect data from participants, and analyze the data using statistical methods. Designing controlled experiments is an iterative process during which researchers need to weight trade-offs by exploring different design alternatives (Figure 2) including statistical power (Figure 3) [6]. Those decisions, the researcher degrees of freedom, are important if researchers try to replicate and extend experiments from the literature. Natural sciences experienced a replication crisis – a recent survey with over 1500 scientists showed that 70% were not able to replicate experiments from other scientists [2]. To increase the reproducibility of experiments in HCI, Cockburn et al. advocate pre-registering experiments [5]. To prevent "Hypothesising After the Results are Known", it is good practice to summarize the experiment and its analysis plan, and pre-register them on a public repository. However, pre-registering an experiment requires researchers to communicate their experimental design correctly and entirely. At this point, there are tools available that support researchers during parts of the design and analysis process while at the same time increasing the reproducibility of experiments. However, each of the tools only supports a very specific step, e.g. power analysis, thus providing the researcher not with the relevant information while designing an experiment.

Statement of Thesis

I claim that decision processes and collaboration practices in experiment design can be supported by interactive systems that provide feedback, enable direct manipulation, and facilitate thinking with and about data through simulations and visualizations.

Interactive Trade-off Exploration

My first project, *Touchstone2*, uses a direct-manipulation interface to examine trade-offs between different experimental designs [6]. *Touchstone2* improves upon previous work that uses a step-by-step approach [12] and a question-answer approach [14] to elicit the input parameters for experimental design. *Touchstone2* also adds the trade-off comparison between multiple experimental designs. To enable direction manipulation, experimenters can specify experiment design parameters with a block-baed language, and inspect the results immediately (Figure 4).

Trade-offs of multiple experimental designs can be compared side-by-side in the same environment. For example, changing the counterbalancing strategy might change the number of participants needed for the experiments To compare the trial order across different designs, we use BRUSHING [21], and a TABLE LENS [16] visualization to compare the distribution of conditions within the designs.

The design of *Touchstone2* was informed by an interview study with ten researchers from HCI (6), Psychology (2), Biology (1), and Economics (1). We evaluated *Touchstone2* using a workshop and an observational study. During the workshop, 17 participants in nine teams were asked to reproduce their own experiments in *Touchstone2*, and explore two alternatives. Most teams were able to successfully reproduce their designs, and mentioned that the visual interface helped in communicating the design to the other team-member. During the observational study, ten participants used the power analysis tool to elicit an appropriate number of participants. Participants appreciated the interactive power chart showing the trade-off between number of participants and statistical power but had difficulties understanding the standardized effect size Cohen's *f*. To

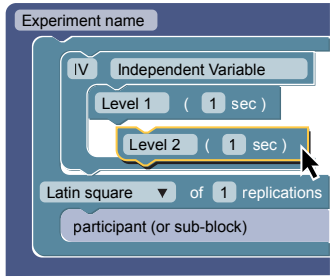

Figure 4: *Touchstone2's* blocked-based interface to specify experimental designs.

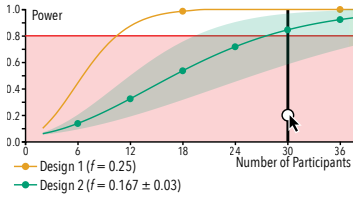

Figure 5: Users can compare the statistical power of two experiments to choose the number of participants.

CHI	Exp	Power
'17	391	6 (1.5%)
'18	464	10 (2.2%)
'19	474	13 (2.7%)

Sidebar 1: Number of papers at CHI that use the term "experiment" compared to the ones using "power analys".

address these difficulties, we created an application for interactive sample size estimation based on statistical power.

Contribution to the Thesis
Touchstone2 and the touchstone language (TSL) form the first building block for my thesis. During the process of the project, we noticed that a more precise domain-specific language (DSL) than [19] is needed to represent common experiments in HCI. I will use *Touchstone2* and TSL in my follow-up projects.

Limitations & Opportunities
Touchstone2 opened opportunities for expanding the interaction capabilities, improving the way to compare trade-offs among many design alternatives, and make power analysis more accessible to researchers (Figure 5). *Touchstone2*'s interface is currently limited to within-participants designs only, while TSL is already able to represent a greater variety of experiments (e.g. between-, mixed-participants, and multi-session designs). However, TSL only includes information about the experimental design itself, but no meta-information which could be useful for data analysis, e.g. dependent variables or effect of interest. The search space of all possible experimental designs is large and complex. Each design has different trade-offs, and the researcher needs to judge if an alternative makes sense or not. While the interface seemed to be easy to use, it is still difficult to efficiently search through all possible experimental designs to compare their trade-offs.

The interactive power chart gave a first intuition about the relationship between power and the number of participants, estimating the effect size Cohen's *f* and possible confounds (e.g. fatigue or practice effect) remains challenging.

We are continuing the work on *Touchstone2* to extend the block-based language to include a wider variety of designs, add a feature to browse through designs in the past, and create a proof-of-concept recommender system for design choices.

Interactive Sample Size Estimation

The number of participants a researcher chooses to recruit for an experiment depends on various factors. For example,

- the availability of the target group (e.g. visually impaired people or children),
- the experiment itself (e.g. longitudinal study over several weeks vs. 30 minutes),
- the counterbalancing strategy, and
- the statistical power.

The researcher needs to take relevant factors into account and weigh them depending on the research question.

Statistical power is the probability of detecting an effect when it exists in the population. *A priori* power analysis can be used to plan the number of participants for an experiment. At CHI, only a handful of publications use statistical power to plan the number of participants (see Sidebar 1). I used CERMINE [20] to extract the content of the papers and text-filtered for "experiment" and "power analys"[1].

There are packages available that support *a priori* power analysis in R, e.g. `pwr` [4] and `skpr` [15], and Python, e.g. `statsmodels` [18]. JMP DOE [17] and G*Power [7] are two applications that allow users to specify parameters in a graphical user interface. However, none of these tools support the decision-making process by enabling comparison between different scenarios in the iterative process of experimental design. *Touchstone2* supports power analysis in an iterative way along with designing experiments. However, *Touchstone2* calculates the power for the overall experiment design, and the user is not able to select

[1] To include both singular and plural form.

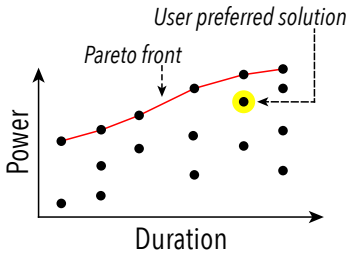

Figure 6: The Pareto front might not include the preferred solution of the researcher. Here, the researcher trades some power for easier execution of the study. For example, reducing the number of block switches by testing one level back-to-back instead of alternating between them. The latter is not captured by the objective function.

conditions of interest. One of the inputs for these tools is the expected effect size. In G*Power for F tests, Cohen's *f* can be specified, or it can be calculated from partial eta-squared (η^2) or the variance. In *Touchstone2*, Cohen's *f* can be specified directly and users are able to input measurements to calculate the effect size. Field [8] describes that the understanding of standardized effect size was a barrier for the researchers which we also found in our second evaluation study [6, p. 9].

Project

We developed an application in which users can specify multiple parameters instead of a standardized effect size to explore their effect on statistical power. Researchers can add expected measures, and are able to explore the effect of possible confounds such as e.g. learning or fatigue effect. Different experimental designs can be selected, and each of the design or parameter changes can be compared. We are revising this work for future publication.

Contribution to my Thesis

With this project, we hope to lower the knowledge barrier for power analysis making it more usable for a wider audience of researchers. In combination with *Touchstone2*, this tool can facilitate the experimental design process and could contribute to more rigorous science.

Optimization-aided Trade-off Exploration

In *Touchstone2*, researchers have to change parameters to explore trade-offs between experimental designs. This trade-off exploration is cumbersome and complex as the search space of possible experimental designs is large. In order to search this space efficiently, researchers need to have a certain level of experience and expertise. Leveraging computational power to generate the search space, I would like to investigate how optimization could support researchers in designing experiments.

First, the algorithm needs to generate all possible experimental designs to populate the search space i.e. the design space, and reduce its size using Pareto optimality. Selecting an appropriate experimental design from the design space can be viewed as a multi-criteria design task. Researchers need to find an experimental design that best fits their research questions and their constraints. However, there might not be a single "best" solution because the objective function, i.e. trade-offs between the alternatives, is difficult to specify and the user preference might be fuzzy [3, 9]. The number of elements in the design space can be reduced by selecting all Pareto optimal experimental designs. An experimental design is Pareto optimal when there exists no alternative that is better in one criterion when all others are equal.

Second, the optimal experimental designs for the researcher might not be one of the Pareto optimal elements i.e. it might not be on the Pareto front (Figure 6). As not all constraints can be operationalized by the researcher because they might be too complex or vaguely defined, e.g. the blocking, or the counterbalancing strategies (Figure 7). This means that researchers need to be able to efficiently search through the space behind the Pareto front; Khire et al. named this area "Pareto band" [1].

Third, an additional challenge is that in experimental design some parameters can be either design (i.e. input) parameters, performance (i.e. output) parameters, or both. This depends on the research context, and might differ from research question to research question. For example, the researcher needs to decide how many participants should be recruited for the experiment – number of participants is a design parameter. However, if the researcher uses a

Design 1 *computationally better*

Blocked by Device with Latin square. Width and Distance are blocked together with *Latin square*.

Design 2

Blocked by Device with Latin square. Width and Distance are blocked together with *Random* counterbalancing.

D1	D2	N	Power
●	●	18	79%
	●	20	83%
	●	22	88%
		⋮	⋮
●	●	36	99%

user preferred solution

Figure 7: Two Fitts's law style experiments. Design 1 would be preferable as it uses Latin square counterbalancing in both blocks. However, Design 1 needs a multiple of 18 participants and the counterbalancing for $W \times D$ can be random, hence, the researcher chooses Design 2 with 22 participants.

counterbalancing strategy such as Latin square, or tries to exceed a certain power threshold (e.g. $1 - \beta \geq 80\%$)[2], the number of participants to recruit is a performance parameter.

Project

I plan to extend *Touchstone2* so that researchers have access to the Pareto optimization and can interactively explore the Pareto band. To start, a preliminary study in the form of a semi-structured interview will help revealing possible interaction capabilities of the new prototype. After the said prototype is implemented, I plan to conduct a design workshop with senior HCI researchers to identify possible breakdowns and opportunities for design. To evaluate the application, I will recruit researchers from HCI and other fields, and observe them while they are designing their real experiments. I plan to observe the researchers while doing their data analysis to obtain some data for the next project.

Contribution to my Thesis

With this project, I hope to make experimental design more usable and efficient to a wider range of researchers by harnessing computational power. By using Pareto optimization, I strive to create a Human-Computer Partnership to support researchers in the experimental design process.

Closing the Design-Analysis Loop

Researchers oftentimes design and conduct a study without taking the statistical analysis into account. They operationalize their research question into dependent variables, i.e. measurements, without further consideration of how they might be analyzed. However, without generating a detailed analysis plan before conducting the study, the researchers might expose themselves to non-intentional HARKing[3] or *p-hacking* to obtain statistically significant results [11]. To combat HARKing, *p-hacking*, and the *file-drawer* effect[4], Cockburn et al. advocate the use of preregistration of experiments which should include an analysis plan. Generating a detailed analysis plan is difficult for researchers who only have some statistical training and analyze experiments infrequently throughout the year. Researchers need to make assumptions about their data to choose an analysis from the jungle of statistical tests. Furthermore, the result of the experiment might be heavily influenced by the design of the experiment. For example, if an experiment is too long, participants might be tired at the end resulting in a performance decrease. If an experiment is too short, the data might be too noisy to draw clear implications. The statistical result can also be used to inform and shape the design of the experiment when using simulated data to create a better experiment.

So far, there are tools that support the learning process of statistical methods and tools which facilitate the analysis itself. Wacharamanotham et al. created *Statsplorer*, an application designed for novice users to learn and understand statistical methods [22]. Subramanian et al. created a web-application, *StatPlaygound*, that supports exploratory learning of statistics by providing the possibility to directly manipulate visualizations of data characteristics, e.g. the distribution in the data. Both tools focus on novice users and are not designed for integrating into the analysis workflow of researchers in practice. Martens created *Illmo*, an application with which users can do data analysis based on Thurstone modeling [13]. The analysis is displayed, and the users can choose between different models based on their research question. Participants liked the visual display of information, however, it still requires the user to know

[2] β is the probability of making a type II error i.e. rejecting a false null hypothesis.

[3] Hypothesizing After the Results are Known.

[4] Also know as *publication bias*. Only significant results are published.

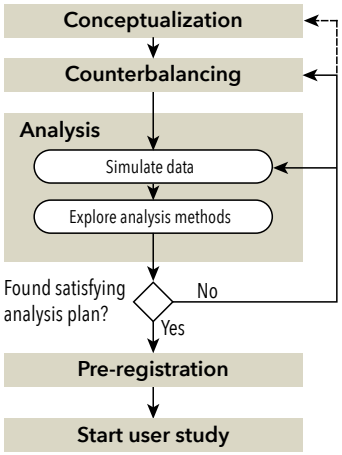

Figure 8: Researchers see analysis results based on simulated data while designing an experiment. The analysis plan can be tweaked and pre-registered effortlessly before conducting the user study.

about the modeling procedure before being able to make a decision. Jun et al. created *Tea*, a Python package where users can specify their hypothesis in a domain-specific language (DSL) and receive valid statistical tests based on their data [10].

None of these tools support the user in designing and analyzing an experiment at the same time to generate an optimal design for the research question. If an experiment is designed thoroughly and pre-registered, its result should yield valuable implications disregarding the statistical significance.

Project

I plan to create an application in which researchers are able to design experiments and see the statistical results based on simulated data in order to study how researchers design experiments when they can better anticipate the results (Figure 8). To start, I plan to conduct semi-structured interviews with expert and novice HCI researchers to learn more about their practices for doing statistical analysis. To study the decision-making process of experimenters, I plan to recruit HCI researchers who design, conducted, and analyzed experiments in the recent past. Researchers will receive an executable template in R or Python, which they can use and tweak further, as well as a summary report that can be directly pre-registered.

Contribution to my Thesis

With this project, I hope to close the loop between experimental design and data analysis by enabling researchers to think about and explore the statistical results. Researchers will be able to make better decisions in favor of their research question and reproducibility.

Research Situation

I am in my second year of my four-year Ph.D. program under the supervision of Prof. Chat Wacharamanotham at the University of Zurich. During my first year, I completed all my coursework requirements and will be able to focus on research going forth. I am currently working on my Ph.D. proposal which I will defend in spring 2019.

Expected Contributions

The result of my thesis is a set of applications which supports researchers in designing their controlled experiments by giving them in place information about statistical power, possible alternatives and their trade-offs, and their statistical result. My work will add an empirical understanding of how researchers currently design experiments and how existing systems support these tasks. Ultimately, I hope to contribute to a more rigorous design and analysis of controlled experiments.

Acknowledgments

I want to thank Chat Wacharamanotham, Wendy E. Mackay, and Michel Beaudouin-Lafon for their ongoing support during my thesis. I want to thank Jan Gugler, Wanyu Liu, and Xiaoyi Wang for their insightful discussions and collaborations. This work was partially supported by European Research Council (ERC) grants № 321135 “CREATIV: Creating Co-Adaptive Human-Computer Partnerships” and № 695464 “ONE: Unified Principles of Interaction”.

REFERENCES

[1] 2008. *Product Family Commonality Selection Through Interactive Visualization*. International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Vol. Volume 1: 34th Design Automation Conference, Parts

A and B. DOI: http://dx.doi.org/10.1115/DETC2008-49335

[2] Monya Baker. 2016. 1,500 scientists lift the lid on reproducibility. *Nature News* 533, 7604 (2016), 452.

[3] Richard Balling. 1999. Design by shopping: A new paradigm?. In *Proceedings of the Third World Congress of structural and multidisciplinary optimization (WCSMO-3)*, Vol. 1. International Soc. for Structural and Multidisciplinary Optimization Berlin, 295–297.

[4] Stephane Champely, Claus Ekstrom, Peter Dalgaard, Jeffrey Gill, Stephan Weibelzahl, Aditya Anandkumar, Clay Ford, Robert Volcic, and Helios De Rosario. 2018. Basic functions for power analysis. *R Package Version https://cran. r-project. org/web/packages/pwr/pwr. pdf* (2018).

[5] 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 (CHI '18)*. ACM, New York, NY, USA, Article 141, 12 pages. DOI: http://dx.doi.org/10.1145/3173574.3173715

[6] 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 *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 217, 11 pages. DOI: http://dx.doi.org/10.1145/3290605.3300447

[7] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods* 41, 4 (01 Nov 2009), 1149–1160. DOI: http://dx.doi.org/10.3758/BRM.41.4.1149

[8] A. Field. 2009. *Discovering Statistics Using SPSS*. SAGE Publications. https://books.google.ch/books?id=a6FLF1YOqtsC

[9] Salvatore Greco, J Figueira, and M Ehrgott. 2016. *Multiple criteria decision analysis*. Springer.

[10] 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

[11] 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 PMID: 15647155.

[12] 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 *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*. ACM, New York, NY, USA, 1425–1434. DOI: http://dx.doi.org/10.1145/1240624.1240840

[13] Jean-Bernard Martens. 2014. Interactive Statistics with Illmo. *ACM Trans. Interact. Intell. Syst.* 4, 1, Article 4 (April 2014), 28 pages. DOI: http://dx.doi.org/10.1145/2509108

[14] Xiaojun Meng, Pin Sym Foong, Simon Perrault, and Shengdong Zhao. 2017. NexP: A Beginner Friendly Toolkit for Designing and Conducting Controlled Experiments. In *Human-Computer Interaction – INTERACT 2017*, Regina Bernhaupt, Girish Dalvi, Anirudha Joshi, Devanuj K. Balkrishan, Jacki O'Neill, and Marco Winckler (Eds.). Springer International Publishing, Cham, 132–141.

[15] Tyler Morgan-Wall and George Khoury. 2018. *skpr: Design of Experiments Suite: Generate and Evaluate Optimal Designs*. `https://CRAN.R-project.org/package=skpr` R package version 0.54.3.

[16] Ramana Rao and Stuart K. Card. 1994. The Table Lens: Merging Graphical and Symbolic Representations in an Interactive Focus + Context Visualization for Tabular Information. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '94)*. ACM, New York, NY, USA, 318–322. `DOI: http://dx.doi.org/10.1145/191666.191776`

[17] SAS Institute Inc. 2016. *JMP ®13 Design of experiments guide*. SAS Institute Inc., SAS Institute Inc., Cary, NC, USA.

[18] Skipper Seabold and Josef Perktold. 2010. Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference*, Vol. 57. Scipy, 61.

[19] Larisa N Soldatova and Ross D King. 2006. An ontology of scientific experiments. *Journal of The Royal Society Interface* 3, 11 (2006), 795–803. `DOI: http://dx.doi.org/10.1098/rsif.2006.0134`

[20] Dominika Tkaczyk, Paweł Szostek, Mateusz Fedoryszak, Piotr Jan Dendek, and Łukasz Bolikowski. 2015. CERMINE: automatic extraction of structured metadata from scientific literature. *International Journal on Document Analysis and Recognition (IJDAR)* 18, 4 (2015), 317–335.

[21] Lisa Tweedie, Robert Spence, Huw Dawkes, and Hus Su. 1996. Externalising Abstract Mathematical Models. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '96)*. ACM, New York, NY, USA, 406–ff. `DOI: http://dx.doi.org/10.1145/238386.238587`

[22] 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 (CHI '15)*. ACM, New York, NY, USA, 2693–2702. `DOI: http://dx.doi.org/10.1145/2702123.2702347`